EPFL

Thèse n° 10 134

Reliable data-driven decision-making through optimal transport

Présentée le 14 juin 2024

Collège du management de la technologie
Chaire analyse de risque et optimisation
Programme doctoral en management de la technologie

pour l'obtention du grade de Docteur ès Sciences

par

Bahar TAŞKESEN

Acceptée sur proposition du jury

Prof. D. Filipovic, président du jury
Prof. D. Kuhn, directeur de thèse
Prof. J. Lygeros, rapporteur
Prof. G. Peyré, rapporteur
Prof. N. Kiyavash, rapporteuse

École
polytechnique
fédérale
de Lausanne

2024

Abstract

Decision-making permeates every aspect of human and societal development, from individuals' daily choices to the complex decisions made by communities and institutions. Central to effective decision-making is the discipline of optimization, which seeks the best choice from a set of alternatives based on specific criteria. This thesis focuses on optimization problems fueled by the ever-growing abundance of data. In an era where data is ubiquitous, machine learning algorithms offer unprecedented potential to enhance decision-making across diverse sectors such as healthcare, finance, and technology. The enthusiastic adoption of machine learning in various sectors has necessitated a more cautious approach upon realizing that the reliability of these systems in complex real-world situations is not always guaranteed. At the heart of this investigation is the ambition to design algorithms equipped to make reliable data-driven decisions. This entails addressing the challenges of ensuring robust performance outside training environments, incorporating fairness measures when needed, and achieving decision interpretability while maintaining computational efficiency. Attempting to satisfy all these desires simultaneously is a formidable task, given the challenges in the data collection phase and modeling. In its most comprehensive form, our objective in this thesis entails modeling, developing tools for, and auditing data-driven decision-making systems based on data generated by an unknown mechanism. The common theme shared within the lines of works in this thesis is the use of optimal transport. Thus, the first part of this thesis introduces the optimal transport problem, studies its computational complexity, and proposes numerical solutions. The rest of the thesis explores two interrelated learning paradigms: static decision-making, in which decisions have no immediate impact on the data used in training, and dynamic decision-making, in which decisions actively influence the data acquisition process. The third chapter then investigates the development of estimators in scenarios marked by data scarcity in the target domain despite abundant data in a related source domain. Utilizing optimal transport, we propose robust estimators that capitalize on source data while accommodating the sparse target data. In the fourth chapter, we focus on creating fair and robust models. We introduce a distributionally robust logistic regression model with an unfairness penalty, which helps to prevent discrimination based on sensitive attributes such as gender or ethnicity. This model is tractable when an optimal transport-based ambiguity set is utilized. While it is important to train fair models, it is equally crucial to rigorously examine machine learning models before deploying them in practice. In the fifth chapter, we use ideas from the optimal transport theory and propose a statistical test for detecting unfair classifiers. The sixth chapter extends linear quadratic Gaussian control problems to their distributionally robust counterparts using an optimal transport-based ambiguity set, offering structural insights that aid in the efficient design of numerical solutions.

Estratto

Il processo decisionale permea ogni aspetto dello sviluppo umano e sociale, dalle scelte quotidiane degli individui alle decisioni complesse prese da comunità e istituzioni. Al centro di un processo decisionale efficace c'è la disciplina dell'ottimizzazione, che cerca la scelta migliore tra un insieme di alternative in base a criteri specifici. Questa tesi si concentra sui problemi di ottimizzazione alimentati dalla crescente abbondanza di dati. In un'epoca in cui i dati sono onnipresenti, gli algoritmi di apprendimento automatico offrono un potenziale senza precedenti per migliorare il processo decisionale in diversi settori come la sanità, la finanza e la tecnologia. L'adozione entusiastica dell'apprendimento automatico in vari settori ha reso necessario un approccio più cauto, dopo aver capito che l'affidabilità di questi sistemi in situazioni complesse del mondo reale non è sempre garantita. Il cuore di questa ricerca è l'ambizione di progettare algoritmi in grado di prendere decisioni affidabili basate sui dati. Ciò comporta la necessità di garantire prestazioni robuste al di fuori degli ambienti di addestramento, di incorporare misure di equità quando necessario e di ottenere l'interpretabilità delle decisioni mantenendo l'efficienza computazionale. Tentare di soddisfare tutti questi desideri simultaneamente è un compito formidabile, date le sfide nella fase di raccolta dei dati e nella modellazione. Nella sua forma più completa, il nostro obiettivo in questa tesi comporta la modellazione, lo sviluppo di strumenti e la verifica di sistemi decisionali guidati dai dati e basati su dati generati da un meccanismo sconosciuto. Il tema comune alle linee di lavoro di questa tesi è l'uso del trasporto ottimale. La prima parte di questa tesi introduce il problema del trasporto ottimale, studia la sua complessità computazionale e propone soluzioni numeriche. Il resto della tesi esplora due paradigmi di apprendimento interconnessi: il processo decisionale statico, in cui le decisioni non hanno un impatto immediato sui dati utilizzati nell'addestramento, e il processo decisionale dinamico, in cui le decisioni influenzano attivamente il processo di acquisizione dei dati. Il terzo capitolo analizza lo sviluppo di stimatori in scenari caratterizzati dalla scarsità di dati nel dominio di destinazione, nonostante l'abbondanza di dati in un dominio di origine correlato. Utilizzando il trasporto ottimale, proponiamo stimatori robusti che sfruttano i dati di partenza e al tempo stesso si adattano ai dati scarsi dell'obiettivo. Nel quarto capitolo, ci concentriamo sulla creazione di modelli equi e robusti. Introduciamo un modello di regressione logistica distributivamente robusto con una penalità di iniquità, che aiuta a prevenire la discriminazione basata su attributi sensibili come il genere o l'etnia. Questo modello è fattibile quando si utilizza un set di ambiguità ottimale basato sul trasporto. Se è importante addestrare modelli equi, è altrettanto cruciale esaminare rigorosamente i modelli di apprendimento automatico prima di impiegarli nella pratica. Nel quinto capitolo utilizziamo le idee della teoria del trasporto ottimale e proponiamo un test statistico per individuare i classificatori ingiusti. Il sesto capitolo estende i problemi di controllo lineare quadratico gaussiano alle loro controparti robuste dal punto di vista distributivo, utilizzando un insieme di ambiguità basato sul trasporto ottimale, offrendo spunti strutturali che aiutano a progettare in modo efficiente le soluzioni numeriche.

Acknowledgements

I will remain always grateful to my advisor, Prof. Daniel Kuhn for giving me the wonderful opportunity to learn and grow. Right from the start, he has shown unwavering belief in my abilities and provided invaluable guidance, support, feedback, and trust throughout my journey. I am particularly grateful for his encouragement when I want to explore new topics and his wisdom, as well as the kindness he has displayed in all our interactions. *I truly cherished our countless meetings, submission celebrations, and knowledge I gained from you. Thank you for opening the doors to this marvelous path for me.*

I have had the privilege of working with amazing collaborators, Prof. Viet Anh Nguyen, Prof. Jose Blanchet, Prof. Dan A. Iancu, Prof. Çağıl Koçyiğit, Dr. Jiajin Li, Prof. Karthik Natarajan, Yves Rychener and Prof. Man-Chung Yue. *Working with each of you has been a tremendous pleasure.* I would like to express my gratitude to my dissertation committee members: Prof. Damir Filipovic, Prof. Negar Kivayash, Prof. John Lygeros and Prof. Gabriel Peyré. I would like to thank to the members of College of Management and past and present members of RAO.

I was blessed with wonderful friends who have become my chosen family. Together, we have shared countless special moments, and I have learned so much from each one of them. Even though pursuing PhD is a journey one takes alone, thanks to these incredible individuals, I have never felt like I was walking alone. After several attempts at writing this acknowledgment, I have realized that it is simply impossible to create a version that does not risk leaving someone out or failing to mention them adequately, which would deeply sadden me. Ironically, achieving a truly *fair* inclusion of everyone who has deeply cared about me, and whom I care about profoundly, may then constitute an infeasible problem given the margins. Ultimately, to all my friends who may read this at some point, *know that I thought of you while writing this. Thank you for the laughter, the tears, and all the unforgettable moments we have shared. Your support has been a cornerstone of this journey.*

Finally, I would like to thank to my family. To Yara, who made Lausanne feel like home; to Elif, for always showing me new perspectives; to my grandmother who raised me and is the strongest woman I have ever known; to my dad for inspiring me by his passion for learning and work ethics; to my sister Burcu who brought color, love and a lot of laugh to my life since the day she was born; and to my mom whom I inherited my curiosity and, who has the endless support and patience for me. At last, I wish to thank Andrea for inspiring me with his determination, sharing my passion for science, and his extraordinary support. *I love learning from you, with you and I can not wait to explore life with you.*

Bahar Taşkesen
Lausanne

Dedicated to all women who fight to have a room of their own.

Introduction

A good decision is based on knowledge and not on numbers.

Plato

Decisions are ubiquitous. In nature, choosing the best option among others might drastically increase the species' survival rate. The ability to gather information, learn, generate knowledge, and finally make a pondered choice plays a crucial role in defining humanity's rational nature. Since the beginning of known human history, we have been continuously faced with the challenge of decision-making, as individuals or as a group, with one overarching goal: secure the best possible outcome. As individuals, we have decided to live as a group, and as a group, we started to build cities close to a water source. This collective decision-making process was not arbitrary; it was rooted in the understanding that proximity to water not only facilitated daily living but also enabled agriculture, trade, and security. Over time, these settlements evolved into complex societies, each with its systems of governance, culture, and technology, reflecting the cumulative knowledge and decisions of countless generations. As societies became more complex, so did the decisions they faced.

A discipline that seeks to leverage mathematical and computational methods to find the most effective solution is *optimization*. This discipline is dedicated to the best utilization of our degrees of freedom, known as variables, within certain constraints, guided by a quantitative measure known as the cost (loss) function. The field of optimization is rich and varied, encompassing a wide array of problem types. This thesis, in particular, is mostly concerned with optimization problems that use *data* as a primary source of information and knowledge for the problem formulation and its solution. As we talk about data-driven tasks, we inevitably touch upon the realm of *machine learning* – a field of study within artificial intelligence that is dedicated to developing statistical algorithms capable of learning from data to make predictions or decisions in unseen scenarios, autonomously. The potential of these machine learning algorithms has been further amplified recently by the abundant surge in the availability of data across various domains, including medicine, education, policy-making, marketing, civics, and many more. This data deluge has created opportunities for the development of systems capable of implementing highly precise and personalized predictions for individuals at unprecedented scales, which has led to their use in *decision-making* systems. The healthcare industry leverages these algorithms to decide whom to prioritize for disease screening [Agg+22] and allocate scarce resources [Fri+13]. In the banking sector, they play a crucial role in deciding who qualifies for loans [LSM19] and in detecting potential cases of money laundering [ZT19]. Similarly, technology companies rely on algorithms to target advertisements for various goods, services, and housing [Spe+18], and employment [LT19] opportunities. Today, thanks to the availability of precise sensors, including high-definition cameras and high-capacity servers, we are witnessing a surge in machine learning models integrated also into our physical environment. Robotic surgeries are becoming increasingly common, enabling minimally invasive procedures, enhancing precision, and reducing recovery times [CMG21]. There are approximately 30 million self-driving cars around, and in our virtual interactions, intelligent conversational agents like ChatGPT serviced 180.5 million users by August 2023, offering real-time assistance and information. These advances highlight the transition of intelligent systems from theoretical models to active participants in shaping our reality, where their decisions have a significant impact on our lives. This evolution prompts an elemental question: *How can we trust the decisions made on our behalf?*

The enthusiastic adoption of machine learning in various sectors has necessitated a more cautious approach upon realizing that the *reliability* of these systems in complex situations is not always guaranteed. For example, Amazon used a hiring tool to automatically rate job applicants based on their resumes. Due to the male dominance across the tech industry, the tool was trained on a biased dataset and taught itself that male candidates were preferable [Das18]. Another example is the inaccuracy of pulse oximeters in measuring blood oxygen levels in various racial groups during the COVID-19 pandemic in 2020 [Sjo+20] causing delays in their treatment [Ort22]. On the autonomous vehicles side, recent years have witnessed both driver and pedestrian fatalities due to malfunctions in automated vehicle control systems [Nat17; Nat20]. These examples underscore the critical need for vigilance and ethical considerations in the deployment of machine learning systems, highlighting the delicate balance between technological innovation and its implications on society. Consequently, the essence of our endeavor extends well beyond the mere capability of these systems to make decisions, and this thesis is therefore dedicated to contributing to answering the following overarching question:

How do we design algorithms to make reliable data-driven decisions?

We stand at a pivotal moment, seeking to clarify our expectations from decision-making systems, particularly our interpretation of “reliability”—a term that has been somewhat nebulous until now. Ultimately, we aspire for our decision-making models not to disappoint us in terms of their performance when they are deployed in the real world. We will refer to this as *out-of-sample performance*. Hence, our objective is for our models to establish conservative performance benchmarks during training, with the aim of surpassing these benchmarks upon deployment, all while maintaining *computational efficiency*. Ensuring computational efficiency is integral for several reasons. First, models that require less computational resources can be scaled easily for larger instances, allowing them to be deployed across a wider range of platforms and devices, from high-end servers to mobile devices, making them accessible to a broader audience. Second, computational efficiency translates to faster decision-making, which is crucial in time-sensitive applications such as autonomous driving, market-making, and emergency response systems. Third, it contributes to sustainability by reducing energy consumption and allowing companies to benefit from carbon credit markets, aligning with the growing need for eco-friendly technologies. Lastly, efficient computation helps in managing operational costs, making it economically feasible to implement and maintain these systems on a large scale. Thus, computational efficiency is not merely a technical requirement but a cornerstone of making advanced decision-making models viable, responsive, and sustainable in real-world applications. At the same time, we strive to ensure accountability in institutional decision-making processes by aiming for guarantees of *fairness*, depending on the domain of application. The underlying models must be designed to identify, mitigate, and, where possible, eliminate biases that may exist in the available data, which can stem from historical inequalities or systemic issues. We also want these systems to come with *interpretable* decisions, enabling individuals to understand the rationale behind decisions made for themselves or others.

Attempting to satisfy all these desires simultaneously is a formidable task primarily due to the difficulties in articulating these requirements as technical specifications for auditing models before their deployment at scale. Additionally, the phase of data generation and collection introduces its own challenges, including, for example, that the observational data may exhibit biases stemming from historical discrimination or the underrepresentation of certain demographics, leading to predictions that inadvertently sustain or introduce inequalities. Unobserved confounders in data present another critical challenge, especially when proxies, like exam scores, are used to gauge an individual's intellectual capabilities for recruitment purposes. Additionally, serial dependencies in data can arise, such as in interactive decision-making systems where past decisions affect future data collection. This creates a feedback loop, complicating the model's analysis as the collected data is directly influenced by prior decisions. Furthermore, there could be disparities in learning and deployment environments. For example, a predictive model for healthcare applications trained with data from urban hospitals in wealthy countries may face performance issues when applied in rural clinics of developing countries, owing to unrepresented variations in disease prevalence, environmental factors, and healthcare access. Even without such discrepancies between environments, securing reliable performance on data not seen during training presents its own set of challenges. This is notably true in situations involving limited data availability, such as in the development of machine learning models for rare disease diagnosis. In such cases, researchers might only have a small dataset of case studies available, which complicates the model's ability to adapt and perform accurately in new, unseen circumstances. The underlying *data distribution* is the invisible rule that dictates the chances of each possible outcome occurring within a dataset. It is like a recipe that shapes how the data points are spread out or clustered and only visible through the collected data. When the dataset is scarce, it is as if we are trying to understand the full recipe of a dish by only tasting a single spoonful. Just as this spoonful might miss many of the ingredients and nuances that give the dish its full flavor, a small dataset lacks the breadth and depth to capture the properties of the underlying data distribution. A conventional strategy is the sample average approximation, where decision-makers minimize the average of the loss function evaluated at available data points. However, this approach may overlook potential, yet unseen, scenarios that could drastically impact performance; see *ludic fallacy* detailed in [Tal10]. In contrast, Distributionally Robust Optimization (DRO) offers a more resilient solution. DRO aims for decisions that hold up well even under the most adverse conditions within a defined ambiguity set—a range of probable data distributions that likely includes the unknown real-world data distribution. Conceptually, DRO sets up a zero-sum game between the decision-maker and a fictitious adversary, often represented as nature, who selects the most undesirable distribution from the ambiguity set to 'test' the decision. This model strategy prepares for the worst-case scenario, thus ideally *under-promising* during training to *over-deliver* in deployment. However, the complexity of modeling such zero-sum games means (so far) accepting a trade-off: enhanced robustness might come at the expense of computational efficiency. Hence, the modeling phase introduces its own set of challenges, such as the formulation of an optimization problem for the underlying task, the assessment of its computational complexity, and the pursuit of tractable reformulations, if there are any. Subsequent algorithm design mandates scalability to industry-sized applications, sub-optimality guarantees, autonomy, sustainable implementation, and deployment.

In its most comprehensive form, our objective in this thesis entails modeling, developing tools for, and auditing data-driven decision-making systems based on data generated by an unknown mechanism. The common theme shared within the forthcoming lines of works in this thesis is the use of *optimal transport*, a mathematical field that was born by the French mathematician Gaspard Monge [Mon81a] at the end of the eighteenth century. The field emerged to find the most efficient way to transport a given quantity of soil from one location to another for the purpose of construction. Over time, optimal transport has captured the imagination and intellectual curiosity of an array of scholars spanning various disciplines, including mathematics, physics, chemistry, economics, and engineering; see [Vil03, § 3]. Hence, optimal transport has evolved into a critical tool and plays a fundamental role in comparing probability distributions, showcasing its extensive applicability since its inception. With the gradual acknowledgment of optimal transport's practical implications, a pressing question emerges: *How hard is it to solve the optimal transport problem?* Indeed, Part I of this thesis is dedicated to addressing this very question. The rest of the thesis, as shown in Figure 1, explores two interrelated learning paradigms: first, static (or offline) decision-making, where decisions have no immediate impact on data collected discussed in Part II; and second, dynamic (or interactive) decision-making, in which decisions actively influence the data acquisition process detailed in Part III.

This manuscript reflects collaborative works. Throughout, it employs the collective "*we*" to acknowledge various contributions, even as the specific individuals involved may change. The outline below specifies related publications and thus involved individuals for every chapter of each part.

[1] The image featuring the Earth and hands was created with the assistance of DALL·E.

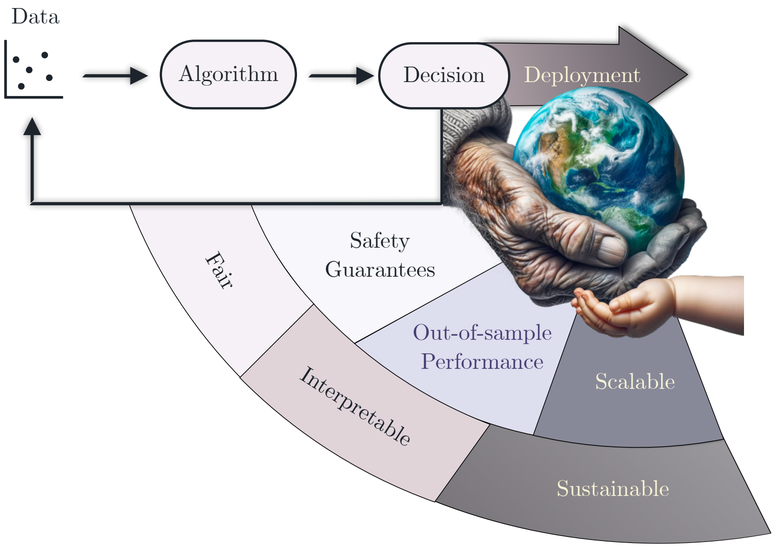


Figure 1: Illustration of interactive decision making and its properties[1]

Outline

In this section, we provide a concise overview of the key contributions of this thesis. For a more comprehensive analysis of the findings and a discussion on related literature, please refer to the respective chapters. Furthermore, each chapter introduces its own notation and is designed to be read independently.

Part-I: Computation of Optimal Transport

Optimal transport (OT) defines minimum cost of transforming a probability measure to some other probability measure with respect to some prescribed transportation cost function and can be viewed as a measure of distance between these two distributions. In the remainder of this thesis, we distinguish discrete, semi-discrete and continuous optimal transport problems in which either both, only one or none of the two probability measures are discrete, respectively. This part is exclusively spared to computational optimal transport, first discrete and then semi-discrete.

Chapter 1: Discrete Optimal Transport [Ta23]

This chapter formally introduces the optimal transport problem and studies its computational complexity when it is evaluated between the distributions of two K-dimensional discrete random vectors. The best known algorithms for this problem run in polynomial time in the maximum of the number of atoms of the two distributions. However, if the components of either random vector are independent, then this number can be exponential in K even though the size of the problem description scales linearly with K. We prove that the described optimal transport problem is #**P**-hard even if all components of the first random vector are independent uniform Bernoulli random variables, while the second random vector has merely two atoms, and even if only approximate solutions are sought. We also develop a dynamic programming-type algorithm that approximates the OT distance in pseudo-polynomial time when the components of the first random vector follow arbitrary independent discrete distributions, and we identify special problem instances that can be solved exactly in strongly polynomial time.

Chapter 2: Semi-discrete Optimal Transport [TSAK23]

Semi-discrete optimal transport problems, which evaluate the Wasserstein distance between a discrete and a generic (possibly non-discrete) probability measure, are believed to be computationally hard. Even though such problems are ubiquitous in statistics, machine learning and computer vision, however, this perception has not yet received a theoretical justification. To fill this gap, we prove that computing the Wasserstein distance between a discrete probability measure supported on two points and the Lebesgue measure on the standard hypercube is already #**P**-hard. This insight prompts us to seek *approximate* solutions for semi-discrete optimal transport problems. We thus perturb the underlying transportation cost with an additive disturbance governed by an ambiguous probability distribution, and we introduce a distributionally robust *dual* optimal transport problem whose objective function is smoothed with the most adverse disturbance distributions from within a given ambiguity set. We further show that smoothing the dual objective function is equivalent to regularizing the primal objective function, and we identify several ambiguity sets that give rise to several known and new regularization schemes. As a byproduct, we discover an intimate relation between semi-discrete optimal transport problems and discrete choice models traditionally studied in psychology and economics. To solve the regularized optimal transport problems efficiently, we use a stochastic gradient descent algorithm with imprecise stochastic gradient oracles. A new convergence analysis reveals that this algorithm improves the best known convergence guarantee for semi-discrete optimal transport problems with entropic regularizers.

Part-II: Static Decision-making

Chapter 3: Distributionally Robust Domain Adaptation [Taş+21b]

Estimators, when trained on a few target domain samples, may predict poorly. Supervised domain adaptation aims to improve predictive accuracy by exploiting additional labeled training samples from a source distribution that is close to the data-generating distribution in the target domain. For example, consider understanding the dynamics of ride-sharing platforms, which requires insights about the demand and supply from both sides of the market. These insights are signaled through the ride fares, which can be explained by characteristics such as the travel distances and the origin-destination pairs of the trips, the time of the day, and the weather conditions. The capability to correctly predict ride fares directly translates into improved profit forecasts, and thus it vitally supports the growth of new-coming platforms. In a competitive market, a follower (e.g., Lyft) needs to target a slightly different market segment than the leader (e.g., Uber) who had entered earlier. Thus, the demand and supply characteristics of the follower may differ from those of the leader. Nevertheless, as both platforms provide on-demand transportation, it is reasonable to assume that their supply and demand dynamics are similar. The follower, who possesses limited data, can query demand on the leader's platform to collect data in order to leap forward in its predictive precision. We investigate novel strategies to synthesize a family of least squares estimator experts that are robust with regard to moment conditions. When these moment conditions are specified using Kullback-Leibler divergence or OT, we can find robust estimators efficiently using convex optimization.

Chapter 4: Learning Fair and Robust Models [Taş+20]

We propose a distributionally robust logistic regression model with an unfairness penalty that prevents discrimination with respect to sensitive attributes such as gender or ethnicity. This model is equivalent to a tractable convex optimization problem if a OT ball centered at the empirical distribution on the training data is used to model distributional uncertainty and if a new convex unfairness measure is used to incentivize equalized opportunities. We demonstrate that the resulting classifier improves fairness at a marginal loss of predictive accuracy on both synthetic and real datasets. We also derive linear programming-based confidence bounds on the level of unfairness of any pre-trained classifier by leveraging techniques from optimal uncertainty quantification over OT balls.

Chapter 5: Auditing for Fairness [Taş+21a]

Before deployment in practice, machine learning models must be rigorously examined to identify any inherent algorithmic biases. We use ideas from the theory of OT to propose a statistical hypothesis test for detecting unfair classifiers. Leveraging the geometry of the feature space, the test statistic quantifies the distance of the empirical distribution supported on the test samples to the manifold of distributions that render a pre-trained classifier fair. We develop a rigorous hypothesis testing mechanism for assessing the probabilistic fairness of any pre-trained logistic classifier, and we show both theoretically and empirically that the proposed test is asymptotically correct. In addition, the proposed framework offers interpretability by identifying the most favorable perturbation of the data so that the given classifier becomes fair.

Part-III: Dynamic Decision-making

Chapter 6: Distributionally Robust Control [Taş+23]

Linear-Quadratic-Gaussian (LQG) control is a fundamental control paradigm that is studied in various fields such as engineering, computer science, economics, and neuroscience. It involves controlling a system with linear dynamics and imperfect observations, subject to additive noise, with the goal of minimizing a quadratic cost function for the state and control variables. We consider a generalization of the discrete-time, finite-horizon LQG problem, where the noise distributions are unknown and belong to OT-based ambiguity sets centered at nominal (Gaussian) distributions. The objective is to minimize a worst-case cost across all distributions in the ambiguity set, including non-Gaussian distributions. Despite the added complexity, we prove that a control policy that is linear in the observations is optimal for this problem, as in the classic LQG problem. We propose a numerical solution method that efficiently characterizes this optimal control policy. Our method uses the Frank-Wolfe algorithm to identify the least-favorable distributions within the OT ambiguity sets and computes the controller's optimal policy using Kalman filter estimation under these distributions.

Contents

Part I

Computation of Optimal Transport

1. Discrete Optimal Transport

> God made the integers, all else is the work of humanity.
>
> *Leopold Kronecker*

1.1. Introduction

Optimal transport (OT) theory is closely intertwined with probability theory and statistics [BLM13; Vil08] as well as with economics and finance [Gal16], and it has spurred fundamental research on partial differential equations [BB00; Bre91]. In addition, OT problems naturally emerge in numerous application areas spanning machine learning [ACB17; CCO17; RCP16], signal processing [Fer+14; KR15; PR17; TPG16], computer vision [RTG00; Sol+14; Sol+15] and distributionally robust optimization [BM19; GK22; EK18]. For a comprehensive survey of modern applications of OT theory we refer to [Kol+17; PC19a]. Historically, the first OT problem was formulated by Gaspard Monge as early as in 1781 [Mon81b]. Monge's formulation aims at finding a measure-preserving map that minimizes some notion of transportation cost between two probability distributions, where all probability mass at a given origin location must be transported to the same target location. Due to this restriction, an optimal transportation map is not guaranteed to exist in general, and Monge's problem could be infeasible. In 1942, Leonid Kantorovich formulated a convex relaxation of Monge's problem by introducing the notion of a transportation plan that allows for mass splitting [Kan42]. Interestingly, an optimal transportation plan always exists. This paradigm shift has served as a catalyst for significant progress in the field.

In this chapter we study Kantrovich's OT problem between two discrete distributions

$$\mu = \sum_{i\in\mathcal{I}} \mu_i \delta_{\boldsymbol{x}_i} \quad \text{and} \quad \nu = \sum_{j\in\mathcal{J}} \nu_j \delta_{\boldsymbol{y}_j},$$

on $\mathbb{R}^K$, where $\boldsymbol{\mu} \in \mathbb{R}^I$ and $\boldsymbol{\nu} \in \mathbb{R}^J$ denote the probability vectors, whereas $\boldsymbol{x}_i \in \mathbb{R}^K$ for $i \in \mathcal{I} = \{1, \dots, I\}$ and $\boldsymbol{y}_j \in \mathbb{R}^K$ for $j \in \mathcal{J} = \{1, \dots, J\}$ represent the discrete support points of μ and ν, respectively. Throughout the chapter we assume that μ and ν denote the probability distributions of two K-dimensional discrete random vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, respectively. Given a transportation cost function $c : \mathbb{R}^K \times \mathbb{R}^K \to [0, +\infty]$, we define the OT distance between the discrete distributions μ and ν as

$$W_c(\mu, \nu) = \min_{\boldsymbol{\pi}\in\Pi(\boldsymbol{\mu},\boldsymbol{\nu})} \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} c(\boldsymbol{x}_i, \boldsymbol{y}_j)\pi_{ij}, \tag{1.1}$$

where $\Pi(\boldsymbol{\mu}, \boldsymbol{\nu}) = \{\boldsymbol{\pi} \in \mathbb{R}_+^{I\times J} : \boldsymbol{\pi}\mathbf{1} = \boldsymbol{\mu},\ \boldsymbol{\pi}^\top\mathbf{1} = \boldsymbol{\nu}\}$ denotes the polytope of probability matrices $\boldsymbol{\pi}$ with marginal probability vectors $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$. Thus, every

$\boldsymbol{\pi} \in \Pi(\boldsymbol{\mu}, \boldsymbol{\nu})$ defines a discrete probability distribution

$$\pi = \sum_{i \in \mathcal{I}} \sum_{j \in \mathcal{J}} \pi_{ij} \delta_{(\boldsymbol{x}_i, \boldsymbol{y}_j)}$$

of $(\boldsymbol{x}, \boldsymbol{y})$ under which $\boldsymbol{x}$ and $\boldsymbol{y}$ have marginal distributions μ and ν, respectively. Distributions with these properties are referred to as transportation plans. If there exists $p \geq 1$ such that $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^p$ for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^K$, then $W_c(\mu, \nu)^{1/p}$ is termed the p-th Wasserstein distance between μ and ν. The OT problem (1.1) constitutes a linear program that admits a strong dual linear program of the form

$$\begin{array}{ll} \max & \boldsymbol{\mu}^\top \boldsymbol{\psi} + \boldsymbol{\nu}^\top \boldsymbol{\phi} \\ \text{s.t.} & \boldsymbol{\psi} \in \mathbb{R}^I, \ \boldsymbol{\phi} \in \mathbb{R}^J \\ & \psi_i + \phi_j \leq c(\boldsymbol{x}_i, \boldsymbol{y}_j) \quad \forall i \in \mathcal{I}, j \in \mathcal{J}. \end{array}$$

Strong duality holds because $\boldsymbol{\pi} = \boldsymbol{\mu}\boldsymbol{\nu}^\top$ is feasible in (1.1) and the optimal value is finite. Both the primal and the dual formulations of the OT problem can be solved exactly using the simplex algorithm [Dan51], the more specialized network simplex algorithm [Orl97] or the Hungarian algorithm [Kuh55]. Both problems can also be addressed with dual ascent methods [BT97], customized auction algorithms [Ber81; Ber92] or interior point methods [Kar84; LS14; NN94]. More recently, the emergence of high-dimensional OT problems in machine learning has spurred the development of efficient approximation algorithms. Many popular approaches for approximating the OT distance between two discrete distributions rely on solving a regularized variant of problem (1.1). For instance, when augmented with an entropic regularizer, problem (1.1) becomes amenable to greedy methods such as the Sinkhorn algorithm [Sin67; Cut13] or the related Greenkhorn algorithm [AG18; AWR17; CK20], which run orders of magnitude faster than the exact methods. Other promising regularizers that have attracted significant interest include the Tikhonov [BSR18; DPR18; ES18; Seg+18], Lasso [LOG16], Tsallis entropy [Muz+17] and group Lasso regularizers [Cou+16]. In addition, Newton-type methods [Bla+18; Qua19], quasi-Newton methods [BSR18], primal-dual gradient methods [DGK18; GHJ20; JST19; LHJ19a; LHJ19b], iterative Bregman projections [Ben+15] and stochastic average gradient descent algorithms [Gen+16] are also used to find approximate solutions for discrete OT problems.

The existing literature mainly addresses OT problems between discrete distributions that are specified by enumerating the locations and the probabilities of the underlying atoms. In this case, the worst-case time-complexity of solving the linear program (1.1) with an interior point algorithm, say, grows polynomially with the problem's input description. In contrast, we focus here on OT problems between discrete distributions supported on a number of points that grows *exponentially* with the dimension K of the sample space even though these problems admit an input description that scales only *polynomially* with K. In this case, the worst-case time-complexity of solving the linear program (1.1) directly with an interior point algorithm grows exponentially with the problem's input description. More precisely, we henceforth assume that μ is the distribution of a random vector $\boldsymbol{x} \in \mathbb{R}^K$ with independent components. Hence, μ is uniquely determined by the specification of its K marginals, which can be encoded using $\mathcal{O}(K)$ bits. Yet, even if each marginal has only two atoms, μ accommodates already 2^K atoms. OT problems involving such distributions are studied by [Çel+21] with the aim to find a discrete distribution with independent marginals that minimizes the Wasserstein distance from a prescribed discrete distribution. While [Çel+21] focus on solving small instances of this nonconvex problem, our results surprisingly imply that even evaluating this problem's objective function is hard. In summary, we are interested in scenarios where the discrete OT problem (1.1) constitutes a linear program with exponentially many variables and constraints. We emphasize that such linear programs are not necessarily hard to solve [GLS12], and therefore a rigorous complexity analysis is needed. We briefly review some useful computational complexity concepts next.

Recall that the complexity class **P** comprises all decision problems (*i.e.*, problems with a Yes/No answer) that can be solved in polynomial time. In contrast, the complexity class **NP** comprises all decision problems with the property that each 'Yes' instance admits a certificate that can be verified in polynomial time. A problem is **NP**-hard if every problem in **NP** is polynomial-time reducible to it, and an **NP**-hard problem is **NP**-complete if it belongs to **NP**. In this chapter we will mainly focus on the complexity class #**P**, which encompasses all counting problems associated with decision problems in **NP** [Val79b; Val79a]. Loosely speaking, an instance of a #**P** problem thus counts the number of distinct polynomial-time verifiable certificates of the corresponding **NP** instance. Consequently, a #**P** problem is at least as hard as its **NP** counterpart, and #**P** problems cannot be solved in polynomial time unless #**P** coincides with the class **FP** of polynomial-time solvable function problems. A Turing reduction from a function problem A to a function problem B is an algorithm for solving problem A that has access to a fictitious oracle for solving problem B in one unit of time. Note that the oracle plays the role of a subroutine and may be called several times. A polynomial-time Turing reduction from A to B runs in time polynomial in the input size of A. We emphasize that, even though each oracle call requires only one unit of time, the time needed for computing all oracle inputs and reading all oracle outputs is attributed to the runtime of the Turing reduction. A problem is #**P**-hard if every problem in #**P** is polynomial-time Turing reducible to it, and a #**P**-hard problem is #**P**-complete if it belongs to #**P** [Val79a; Jer03].

Several hardness results for variants and generalizations of the OT problem have recently been discovered. For example, multi-marginal OT and Wasserstein barycenter problems were shown to be **NP**-hard [ABA21; ABA22], whereas the problem of computing the Wasserstein distance between a continuous and a discrete distribution was shown to be #**P**-hard even in the simplest conceivable scenarios [TSAK23]. In this chapter, we focus on OT problems between two discrete distributions μ and ν. We formally prove that such problems are also #**P**-hard when μ and/or ν may have independent marginals. Specifically, we establish a fundamental limitation of all numerical algorithms for solving

OT problems between discrete distributions μ and ν, where μ has independent marginals. We show that, unless $\mathbf{FP} = \#\mathbf{P}$, it is not possible to design an algorithm that approximates $W_c(\mu, \nu)$ in time polynomial in the bit length of the input size (which scales only polynomially with the dimension K) and the bit length $\log_2(1/\varepsilon)$ of the desired accuracy $\varepsilon > 0$. This result prompts us to look for algorithms that output ε-approximations in *pseudo-polynomial time*, that is, in time polynomial in the input size, the magnitude of the largest number in the input and the inverse accuracy $1/\varepsilon$. It also prompts us to look for special instances of the OT problem with independent marginals that can be solved in *weakly* or *strongly polynomial time*. An algorithm runs in weakly polynomial time if it computes $W_c(\mu, \nu)$ in time polynomial in the bit length of the input. Similarly, an algorithm runs in strongly polynomial time if it computes $W_c(\mu, \nu)$ in time polynomial in the bit length of the input and if, in addition, it requires a number of arithmetic operations that grows at most polynomially with the dimension of the input (*i.e.*, the number of input numbers).

The key contributions of this chapter can be summarized as follows.

- We prove that the discrete OT problem with independent marginals is $\#\mathbf{P}$-hard even if μ represents the uniform distribution on the vertices of the K-dimensional hypercube and ν has only two support points, and even if only approximate solutions of polynomial bit length are sought (see Theorem 1.3.3).

- We demonstrate that the discrete OT problem with independent marginals can be solved in strongly polynomial time by a dynamic programming-type algorithm if both μ and ν are supported on a fixed bounded subset of a scaled integer lattice with a fixed scaling factor and if ν has only two atoms—even if μ represents an arbitrary distribution with independent marginals (see Theorem 1.4.1, Corollary 2 and the subsequent discussion). The design of this algorithm reveals an intimate connection between OT and the conditional value-at-risk arising in risk measurement.

- Using a rounding scheme to approximate μ and ν by distributions $\tilde{\mu}$ and $\tilde{\nu}$ supported on a scaled integer lattice with a sufficiently small grid spacing constant, we show that if ν has only two atoms, then ε-accurate approximations of the OT distance between μ and ν can always be computed in pseudo-polynomial time via dynamic programming—even if μ represents an arbitrary distribution with independent marginals (see Theorem 1.4.4). This result implies that the OT problem with independent marginals is in fact $\#\mathbf{P}$-hard in the weak sense [GJ79, § 4].

Our complexity analysis complements existing hardness results for two-stage stochastic programming problems. Indeed, [DS06; DS15], [HKW16] and [DDN21] show that computing optimal first-stage decisions of linear two-stage stochastic programs and evaluating the corresponding expected costs is hard if the uncertain problem parameters follow independent (discrete or continuous) distributions. This chapter establishes similar hardness results for discrete OT problems. Our work also complements the work of [Gen+16], who describe a stochastic gradient descent method for computing ε-optimal transportation plans in $\mathcal{O}(1/\varepsilon^2)$ iterations. Their method can in principle be applied to the discrete OT problems with independent marginals studied here. However, unlike our pseudo-polynomial time dynamic programming-based algorithm, their method is non-deterministic and does not output an approximation of the OT distance $W_c(\mu, \nu)$.

The remainder of this chapter is structured as follows. In Section 1.2 we review a useful #**P**-hardness result for a counting version of the knapsack problem. By reducing this problem to the OT problem with independent marginals, we prove in Section 1.3 that the latter problem is also #**P**-hard even if only approximate solutions are sought. In Section 1.4 we develop a dynamic programming-type algorithm that computes approximations of the OT distance in pseudo-polynomial time, and we identify special problem instances that can be solved exactly in strongly polynomial time.

Notation. We use boldface letters to denote vectors and matrices. The vectors of all zeros and ones are denoted by $\mathbf{0}$ and $\mathbf{1}$, respectively, and their dimensions are always clear from the context. The calligraphic letters $\mathcal{I}, \mathcal{J}, \mathcal{K}$ and $\mathcal{L}$ are reserved for finite index sets with cardinalities I, J, K and L, that is, $\mathcal{I} = \{1, \ldots, I\}$ etc. We denote by $\|\cdot\|$ the 2-norm, and for any $\boldsymbol{x} \in \mathbb{R}^K$ we use $\delta_{\boldsymbol{x}}$ to denote the Dirac distribution at $\boldsymbol{x}$.

1.2. A Counting Version of the Knapsack Problem

Counting the number of feasible solutions of a 0/1 knapsack problem is a seemingly simple but surprisingly challenging task. Formally, the problem of interest is stated as follows.

#Knapsack

Instance. A list of items with weights $w_k \in \mathbb{Z}_+$, $k \in \mathcal{K}$, and a capacity $b \in \mathbb{Z}_+$.

Goal. Count the number of subsets of the items whose total weight is at most b.

The #Knapsack problem is known to be #**P**-complete [Dye+93], and thus it admits no polynomial-time algorithm unless **FP** = #**P**. [Dye+93] discovered a randomized sub-exponential time algorithm that provides almost correct solutions with high probability by sampling feasible solutions using a random walk. By relying on a rapidly mixing Markov chain, [MS04] then developed the first fully polynomial randomized approximation scheme. Later, [Dye03] interweaved dynamic programming and rejection sampling approaches to obtain a considerably simpler fully polynomial randomized approximation scheme. However, randomization remains essential in this approach. Deterministic dynamic programming-based algorithms were developed more recently by [Gop+11], and [ŠVV12]. In the next section we will demonstrate that a certain class of discrete OT problems with independent marginals is at least as hard as the #Knapsack problem.

1.3. OT with Independent Marginals

Consider now a variant of the OT problem (1.1), where the discrete multivariate distribution $\mu = \otimes_{k\in\mathcal{K}}\mu_k$ is a product of K independent univariate marginal distributions $\mu_k = \sum_{l\in\mathcal{L}} \mu_k^l \delta_{x_k^l}$ with support points $x_k^l \in \mathbb{R}$ and corresponding probabilities μ_k^l for every $l \in \mathcal{L}$. This implies that μ accommodates a total of $I = L^K$ support points. The assumption that each μ_k, $k \in \mathcal{K}$, accommodates the same number L of support points simplifies notation but can be imposed without loss of generality. Indeed, the probability of any unneeded support point can be set to zero. The other discrete multivariate distribution $\nu = \sum_{j\in\mathcal{J}} \nu_j \delta_{\boldsymbol{y}_j}$ has no special structure. Assume for the moment that all components of the support points as well as all probabilities of μ_k, $k \in \mathcal{K}$, and ν are rational numbers and thus representable as ratios of two integers, and denote by U the maximum absolute numerical value among all these integers, which can be encoded using $\mathcal{O}(\log_2 U)$ bits. Thus, the total number of bits needed to represent the discrete distributions μ and ν is bounded above by $\mathcal{O}(\max\{KL, J\}\log_2 U)$. Note that this encoding does *not* require an explicit enumeration of the locations and probabilities of the $I = L^K$ atoms of the distribution μ. It is well known that the linear program (1.1) can be solved in polynomial time by the ellipsoid method, for instance, if μ is encoded by such an inefficient exhaustive enumeration, which requires up to $\mathcal{O}(\max\{I, J\}\log_2 U)$ input bits. Thus, the runtime of the ellipsoid method scales at most polynomially with I, J and $\log_2 U$. As $I = L^K$ grows exponentially with K, however, this does *not* imply tractability of the OT problem at hand, which admits an efficient encoding that scales only linearly with K. In the remainder of this chapter we will prove that the OT problem with independent maringals is #**P**-hard, and we will identify special problem instances that can be solved efficiently.

In order to prove #**P**-hardness, we focus on the following subclass of OT problems with independent marginals, where μ is the uniform distribution on $\{0, 1\}^K$, and ν has only two support points.

#OT (for $p \geq 1$ fixed)

Instance. Two support points $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{R}^K$, $\boldsymbol{y}_1 \neq \boldsymbol{y}_2$, and a probability $t \in [0,1]$.

Goal. For μ denoting the uniform distribution on $\{0,1\}^K$ and $\nu = t\delta_{\boldsymbol{y}_1} + (1-t)\delta_{\boldsymbol{y}_2}$, compute an approximation $\widetilde{W}_c(\mu,\nu)$ of $W_c(\mu,\nu)$ for $c(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|^p$ such that the following hold.

(i) The bit length of $\widetilde{W}_c(\mu,\nu)$ is polynomially bounded in the bit length of the input $(\boldsymbol{y}_1, \boldsymbol{y}_2, t)$.

(ii) We have $|\widetilde{W}_c(\mu,\nu) - W_c(\mu,\nu)| \leq \overline{\varepsilon}$, where

$$\overline{\varepsilon} = \frac{1}{4I} \min \left\{ \left| \|\boldsymbol{x}_i - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p \right| : i \in \mathcal{I},\ \|\boldsymbol{x}_i - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p \neq 0 \right\}$$

with $I = 2^K$ and $\boldsymbol{x}_i$, $i \in \mathcal{I}$, representing the different binary vectors in $\{0,1\}^K$.

We first need to show that the #OT problem is well-posed, that is, we need to ascertain the existence of a sufficiently accurate approximation that can be encoded in a polynomial number of bits. To this end, we first prove that the maximal tolerable approximation error $\overline{\varepsilon}$ is not too small.

Lemma 1.3.1. *There exists $\varepsilon \in (0, \overline{\varepsilon}]$ whose bit length is polynomially bounded in the bit length of $(\boldsymbol{y}_1, \boldsymbol{y}_2, t)$.*

Proof. Note first that encoding an instance of the #OT problem requires at least K bits because the K coordinates of $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ need to be enumerated. Note also that, by the definition of $\overline{\varepsilon}$, there exists an index $i^\star \in \mathcal{I}$ with $\overline{\varepsilon} = \frac{1}{4I} \left| \|\boldsymbol{x}_{i^\star} - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_{i^\star} - \boldsymbol{y}_2\|^p \right|$. As $p \in [1,\infty)$, $\|\boldsymbol{x}_{i^\star} - \boldsymbol{y}_1\|^p$ and $\|\boldsymbol{x}_{i^\star} - \boldsymbol{y}_2\|^p$ may be irrational numbers that cannot be encoded with any finite number of bits even if the vectors $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$ have only rational entries. Thus, $\overline{\varepsilon}$ is generically irrational, in which case we need to construct $\varepsilon \in (0, \overline{\varepsilon})$. To simplify notation, we henceforth use the shorthands $a = \|\boldsymbol{x}_{i^\star} - \boldsymbol{y}_1\|^2$ and $b = \|\boldsymbol{x}_{i^\star} - \boldsymbol{y}_2\|^2$, which can be computed in polynomial time using $\mathcal{O}(K)$ additions and multiplications. Without loss of generality, we may assume throughout the rest of the proof that $a \geq b$. If $a \geq b \geq 1$, then we have

$$\overline{\varepsilon} = \frac{1}{2^{K+2}} \left| a^{p/2} - b^{p/2} \right| = \frac{1}{2^{K+2}} \left| \frac{a^p - b^p}{a^{p/2} + b^{p/2}} \right| \geq \frac{1}{2^{K+2}} \left| \frac{a^{\lfloor p \rfloor} - b^{\lfloor p \rfloor}}{a^{\lceil p/2 \rceil} + b^{\lceil p/2 \rceil}} \right| \triangleq \varepsilon > 0.$$

Here, the first (weak) inequality holds because $a^{p - \lfloor p \rfloor} \geq 1$ and $(b/a)^{p-\lfloor p \rfloor} \leq 1$, which guarantees that

$$|a^p - b^p| = a^{p-\lfloor p \rfloor} \left| a^{\lfloor p \rfloor} - (b/a)^{p - \lfloor p \rfloor} b^{\lfloor p \rfloor} \right| > \left| a^{\lfloor p \rfloor} - b^{\lfloor p \rfloor} \right|,$$

whereas the second (strict) inequality follows from the construction of $\overline{\varepsilon}$ as a strictly positive number, which implies that $a \neq b$. The tolerance ε constructed in this way can be computed via $\mathcal{O}(K)$ additions and multiplications, and as p is not part of the input, its bit length is thus polynomially bounded. If $a \geq 1 \geq b$ or $a, b \leq 1$, then ε can be constructed in a similar manner. Details are omitted for brevity. □

Lemma 1.3.1 readily implies that for any instance of the #OT problem there exists an approximate OT distance $\widetilde{W}_c(\mu, \nu)$ that satisfies both conditions (i) as well as (ii). For example, we could construct $\widetilde{W}_c(\mu, \nu)$ by rounding the exact OT distance $W_c(\mu, \nu)$ to the nearest multiple of ε. By construction, this approximation differs from $W_c(\mu, \nu)$ at most by ε, which is itself not larger than $\overline{\varepsilon}$. In addition, this approximation trivially inherits the polynomial bit length from ε. We emphasize that, in general, $\widetilde{W}_c(\mu, \nu)$ cannot be set to the exact OT distance $W_c(\mu, \nu)$, because $W_c(\mu, \nu)$ may be irrational and thus have infinite bit length. However, Corollary 1 below implies that if p is even, then $\widetilde{W}_c(\mu, \nu) = W_c(\mu, \nu)$ satisfies both conditions (i) as well as (ii).

Note that the #OT problem is parametrized by p. The following example shows that if p was treated as an input parameter, then the problem would have exponential time complexity.

Example 1.3.2. *Consider an instance of the* #OT *problem with* $K = 1$, $y_1 = 1$, $y_2 = 2$ *and* $t = 0$. *In this case we have* $\mu = \frac{1}{2}\delta_0 + \frac{1}{2}\delta_1$, $\nu = \delta_2$ *and* $\overline{\varepsilon} = \frac{1}{8}$. *An elementary analytical calculation reveals that* $W_c(\mu, \nu) = \frac{1}{2}(1 + 2^p)$. *The bit length of any* $\overline{\varepsilon}$*-approximation* $\widetilde{W}_c(\mu, \nu)$ *of* $W_c(\mu, \nu)$ *is therefore bounded below by* $\log_2(\frac{1}{2}(1 + 2^p) - \frac{1}{8}) \geq p - 1$, *which grows exponentially with the bit length* $\log_2(p)$ *of* p. *Note that the time needed for computing* $\widetilde{W}_c(\mu, \nu)$ *is at least as large as its own bit length irrespective of the algorithm that is used. If* p *was an input parameter of the* #OT *problem, the problem's worst-case time complexity would therefore grow at least exponentially with its input size.*

The following main theorem shows that the #OT problem is hard even if $p = 2$.

Theorem 1.3.3 (Hardness of #OT)**.** #OT *is* #***P****-hard for any* $p \geq 1$.

We prove Theorem 1.3.3 by reducing the #KNAPSACK problem to the #OT problem via a polynomial-time Turing reduction. To this end, we fix an instance of the #KNAPSACK problem with input $\boldsymbol{w} \in \mathbb{Z}_+^K$ and $b \in \mathbb{Z}_+$, and we denote by $\nu_t = t\delta_{\boldsymbol{y}_1} + (1-t)\delta_{\boldsymbol{y}_2}$ the two-point distribution with support points $\boldsymbol{y}_1 = \boldsymbol{0}$ and $\boldsymbol{y}_2 = 2b\boldsymbol{w}/\|\boldsymbol{w}\|^2$, whose probabilities are parameterized by $t \in [0, 1]$. Recall also that μ is the uniform distribution on $\{0, 1\}^K$, that is, $\mu = \frac{1}{I}\sum_{i \in \mathcal{I}} \delta_{\boldsymbol{x}_i}$, where $I = 2^K$ and $\{\boldsymbol{x}_i : i \in \mathcal{I}\} = \{0, 1\}^K$. Without loss of generality, we may assume that the support points of μ are ordered so as to satisfy

$$\|\boldsymbol{x}_1 - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_1 - \boldsymbol{y}_2\|^p \leq \|\boldsymbol{x}_2 - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_2 - \boldsymbol{y}_2\|^p \leq \cdots \leq \|\boldsymbol{x}_I - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_I - \boldsymbol{y}_2\|^p.$$

Below we will demonstrate that computing $W_c(\mu,\nu_t)$ approximately is at least as hard as solving the #Knapsack problem, which amounts to evaluating the cardinality of $\mathcal{I}(\boldsymbol{w},b)=\{\boldsymbol{x}\in\{0,1\}^K:\boldsymbol{w}^\top\boldsymbol{x}\le b\}$.

Lemma 1.3.4. *If $c(\boldsymbol{x},\boldsymbol{y})=\|\boldsymbol{x}-\boldsymbol{y}\|^p$ for some $p\ge 1$, then the OT distance $W_c(\mu,\nu_t)$ is continuous, piecewise affine and convex in $t\in[0,1]$. Moreover, it admits the closed-form formula*

$$\begin{aligned}W_c(\mu,\nu_t)=\frac{1}{I}\sum_{i=1}^{\lfloor tI\rfloor}\|\boldsymbol{x}_i-\boldsymbol{y}_1\|^p+\frac{1}{I}\sum_{i=\lfloor tI\rfloor+1}^{I}\|\boldsymbol{x}_i-\boldsymbol{y}_2\|^p\\+\frac{(tI-\lfloor tI\rfloor)}{I}\left(\|\boldsymbol{x}_{\lfloor tI\rfloor+1}-\boldsymbol{y}_1\|^p-\|\boldsymbol{x}_{\lfloor tI\rfloor+1}-\boldsymbol{y}_2\|^p\right).\end{aligned}\tag{1.2}$$

Proof. For any fixed $t\in[0,1]$, the discrete OT problem (1.1) satisfies

$$\begin{aligned}W_c(\mu,\nu_t)&=\min_{\boldsymbol{\pi}\in\Pi(\mu,\nu_t)}\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\|\boldsymbol{x}_i-\boldsymbol{y}_j\|^p\pi_{ij}\\&=\left\{\begin{array}{cl}\displaystyle\min_{\boldsymbol{q}_1,\boldsymbol{q}_2\in\mathbb{R}^I_+}&\displaystyle t\sum_{i\in\mathcal{I}}\|\boldsymbol{x}_i-\boldsymbol{y}_1\|^p q_{1,i}+(1-t)\sum_{i\in\mathcal{I}}\|\boldsymbol{x}_i-\boldsymbol{y}_2\|^p q_{2,i}\\ \text{s.t.}& t\boldsymbol{q}_1+(1-t)\boldsymbol{q}_2=\boldsymbol{1}/I,\ \boldsymbol{1}^\top\boldsymbol{q}_1=1,\ \boldsymbol{1}^\top\boldsymbol{q}_2=1.\end{array}\right.\end{aligned}$$

The second equality holds because the transportation plan can be expressed as

$$\pi=\sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}}\pi_{ij}\delta_{(\boldsymbol{x}_i,\boldsymbol{y}_j)}=t\cdot q_1\otimes\delta_{\boldsymbol{y}_1}+(1-t)\cdot q_2\otimes\delta_{\boldsymbol{y}_2},$$

with $q_j=\sum_{i\in\mathcal{I}}q_{j,i}\delta_{\boldsymbol{x}_i}$ representing the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}=\boldsymbol{y}_j$ under π for every $j=1,2$. This is a direct consequence of the law of total probability. By applying the variable transformations $\boldsymbol{q}_1\leftarrow tI\boldsymbol{q}_1$ and $\boldsymbol{q}_2\leftarrow(1-t)I\boldsymbol{q}_2$ to eliminate all bilinear terms, we then find

$$W_c(\mu,\nu_t)=\left\{\begin{array}{cl}\displaystyle\min_{\boldsymbol{q}_1,\boldsymbol{q}_2\in\mathbb{R}^I_+}&\displaystyle\frac{1}{I}\sum_{i\in\mathcal{I}}\|\boldsymbol{x}_i-\boldsymbol{y}_1\|^p q_{1,i}+\frac{1}{I}\sum_{i\in\mathcal{I}}\|\boldsymbol{x}_i-\boldsymbol{y}_2\|^p\,q_{2,i}\\ \text{s.t.}&\boldsymbol{1}^\top\boldsymbol{q}_1=tI,\ \boldsymbol{1}^\top\boldsymbol{q}_2=(1-t)I,\ \boldsymbol{q}_1+\boldsymbol{q}_2=\boldsymbol{1}.\end{array}\right.\tag{1.3}$$

Observe that (1.3) can be viewed as a parametric linear program. By [DT03, Theorem 6.6], its optimal value $W_c(\mu,\nu_t)$ thus constitutes a continuous, piecewise affine and convex function of t. It remains to be shown that $W_c(\mu,\nu_t)$ admits the analytical expression (1.2). To this end, note that the decision variable $\boldsymbol{q}_2$ and the constraint $\boldsymbol{q}_1+\boldsymbol{q}_2=\boldsymbol{1}$ in problem (1.3) can be eliminated by applying the substitution $\boldsymbol{q}_2\leftarrow\boldsymbol{1}-\boldsymbol{q}_1$. Renaming $\boldsymbol{q}_1$ as $\boldsymbol{q}$ to reduce clutter, problem (1.3) then simplifies to

$$\begin{array}{cl}\displaystyle\min_{\boldsymbol{q}\in\mathbb{R}^I}&\displaystyle\frac{1}{I}\sum_{i\in\mathcal{I}}\left(\|\boldsymbol{x}_i-\boldsymbol{y}_1\|^p-\|\boldsymbol{x}_i-\boldsymbol{y}_2\|^p\right)q_i+\frac{1}{I}\sum_{i\in\mathcal{I}}\|\boldsymbol{x}_i-\boldsymbol{y}_2\|^p\\ \text{s.t.}&\boldsymbol{1}^\top\boldsymbol{q}=tI,\ \boldsymbol{0}\le\boldsymbol{q}\le\boldsymbol{1}.\end{array}\tag{1.4}$$

Recalling that the atoms of μ are ordered such that $\|\boldsymbol{x}_1 - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_1 - \boldsymbol{y}_2\|^p \leq \cdots \leq \|\boldsymbol{x}_I - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_I - \boldsymbol{y}_2\|^p$, one readily verifies that problem (1.4) is solved analytically by

$$q_i^\star = \begin{cases} 1 & \text{if } i \leq \lfloor tI \rfloor \\ tI - \lfloor tI \rfloor & \text{if } i = \lfloor tI \rfloor + 1 \\ 0 & \text{if } i > \lfloor tI \rfloor + 1. \end{cases}$$

Substituting $\boldsymbol{q}^\star$ into (1.4) yields (1.2), and thus the claim follows. □

Lemma 1.3.4 immediately implies that the bit length of $W_c(\mu, \nu_t)$ is polynomially bounded.

Corollary 1. *If $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^p$ and p is even, then the bit length of the OT distance $W_c(\mu, \nu_t)$ grows at most polynomially with the bit length of $(\boldsymbol{y}_1, \boldsymbol{y}_2, t)$.*

Proof. The bit length of $(\boldsymbol{y}_1, \boldsymbol{y}_2, t)$ is finite if and only if all of its components are rational and thus representable as ratios of two integers. We denote by $U \in \mathbb{N}$ the maximum absolute value of these integers.

For ease of exposition, we assume first that $p = 2$ and $t = 1$. In addition, we use $D \in \mathbb{N}$ to denote the least common multiple of the denominators of the K components of $\boldsymbol{y}_1$. It is easy to see that $D \leq U^K$. By Lemma 1.3.4, the OT distance $W_c(\mu, \nu_t)$ can thus be expressed as the average of the I quadratic terms $\|\boldsymbol{x}_i - \boldsymbol{y}_1\|^2 = \boldsymbol{x}_i^\top \boldsymbol{x}_i + 2\boldsymbol{x}_i^\top \boldsymbol{y}_1 + \boldsymbol{y}_1^\top \boldsymbol{y}_1$ for $i \in \mathcal{I}$. Each such term is equivalent to a rational number with denominator D^2 and a numerator that is bounded above by $K(1 + 2U + U^2)D^2$. Indeed, each component of $\boldsymbol{x}_i$ is binary, whereas each component of $\boldsymbol{y}_1$ can be expressed as a rational number with denominator D and a numerator with absolute value at most UD. By Lemma 1.3.4, $W_c(\mu, \nu_t)$ is thus representable as a rational number with denominator ID^2 and a numerator with absolute value at most $IK(1+U)^2 D^2$. Therefore, the number of bits needed to encode $W_c(\mu, \nu_t)$ is at most of the order

$$\mathcal{O}\left(\log_2(IKU^2D^2))\right) \leq \mathcal{O}\left(\log_2(2^K KU^2 U^{2K})\right) = \mathcal{O}\left(K \log_2(U)\right),$$

where the inequality holds because $I = 2^K$ and $D \leq U^K$. As both K and $\log_2(U)$ represent lower bounds on the bit length of $(\boldsymbol{y}_1, \boldsymbol{y}_2, t)$, we have thus shown that the bit length of $W_c(\mu, \nu_t)$ is indeed polynomially bounded in the bit length of $(\boldsymbol{y}_1, \boldsymbol{y}_2, t)$. If p is any even number and t any rational probability, then the claim can be proved using similar—yet more tedious—arguments. Details are omitted for brevity. □

Corollary 1 implies that the OT distance $W_c(\mu, \nu_t)$ is rational whenever p is an even integer and t is rational. Otherwise, $W_c(\mu, \nu_t)$ is generically irrational because the Euclidean norm of a vector $\boldsymbol{v} = (v_1, \ldots, v_K)$ is irrational unless $(v_1, \ldots, v_K, \|\boldsymbol{v}\|)$ is proportional to a Pythagorean $(K+1)$-tuple, where the inverse proportionality factor is itself equal to the square of an integer. We will now show that the cardinality of the set $\mathcal{I}(\boldsymbol{w}, b)$ can be computed by solving the univariate minimization problem

$$\min_{t \in [0,1]} W_c(\mu, \nu_t). \tag{1.5}$$

Lemma 1.3.5. *If $c(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|^p$ for some $p \geq 1$, then $t^\star = |\mathcal{I}(\boldsymbol{w},b)|/I$ is an optimal solution of problem (1.5). If in addition each component of $\boldsymbol{w}$ is even and b is odd, then $t^\star$ is unique.*

Proof. From the proof of Lemma 1.3.4 we know that the OT distance $W_c(\mu,\nu_t)$ coincides with the optimal value of (1.3). Thus, problem (1.5) can be reformulated as

$$\begin{array}{cl} \min\limits_{\substack{t\in[0,1] \\ \boldsymbol{q}_1,\boldsymbol{q}_2\in\mathbb{R}^I_+}} & \dfrac{1}{I}\sum\limits_{i\in\mathcal{I}} \|\boldsymbol{x}_i-\boldsymbol{y}_1\|^p q_{1,i} + \dfrac{1}{I}\sum\limits_{i\in\mathcal{I}} \|\boldsymbol{x}_i-\boldsymbol{y}_2\|^p\, q_{2,i} \\ \text{s.t.} & \boldsymbol{1}^\top \boldsymbol{q}_1 = tI,\ \boldsymbol{1}^\top \boldsymbol{q}_2 = (1-t)I,\ \boldsymbol{q}_1+\boldsymbol{q}_2 = \boldsymbol{1}. \end{array} \tag{1.6}$$

Note that the decision variable t as well as the two normalization constraints for $\boldsymbol{q}_1$ and $\boldsymbol{q}_2$ are redundant and can thus be removed without affecting the optimal value of (1.6). In other words, there always exists $t \in [0,1]$ such that $\boldsymbol{1}^\top\boldsymbol{q}_1 = tI$ and $\boldsymbol{1}^\top\boldsymbol{q}_2 = (1-t)I$. Hence, (1.6) simplifies to

$$\begin{array}{cl} \min\limits_{\boldsymbol{q}_1,\boldsymbol{q}_2\in\mathbb{R}^I_+} & \dfrac{1}{I}\sum\limits_{i\in\mathcal{I}} \|\boldsymbol{x}_i-\boldsymbol{y}_1\|^p q_{1,i} + \dfrac{1}{I}\sum\limits_{i\in\mathcal{I}} \|\boldsymbol{x}_i-\boldsymbol{y}_2\|^p\, q_{2,i} \\ \text{s.t.} & \boldsymbol{q}_1+\boldsymbol{q}_2 = \boldsymbol{1}. \end{array} \tag{1.7}$$

Next, introduce the disjoint index sets

$$\begin{aligned} \mathcal{I}_0 &= \{i\in\mathcal{I} : \|\boldsymbol{x}_i-\boldsymbol{y}_1\| = \|\boldsymbol{x}_i-\boldsymbol{y}_2\|\} \\ \mathcal{I}_1 &= \{i\in\mathcal{I} : \|\boldsymbol{x}_i-\boldsymbol{y}_1\| < \|\boldsymbol{x}_i-\boldsymbol{y}_2\|\} \\ \mathcal{I}_2 &= \{i\in\mathcal{I} : \|\boldsymbol{x}_i-\boldsymbol{y}_1\| > \|\boldsymbol{x}_i-\boldsymbol{y}_2\|\}, \end{aligned}$$

which form a partition of $\mathcal{I}$. Using these sets, optimal solution of problem (1.7) can be expressed as

$$q^\star_{1,i} = \begin{cases} \theta_i & \text{if } i\in\mathcal{I}_0 \\ 1 & \text{if } i\in\mathcal{I}_1 \\ 0 & \text{if } i\in\mathcal{I}_2 \end{cases} \quad\text{and}\quad q^\star_{2,i} = \begin{cases} 1-\theta_i & \text{if } i\in\mathcal{I}_0 \\ 0 & \text{if } i\in\mathcal{I}_1 \\ 1 & \text{if } i\in\mathcal{I}_2 \end{cases} \tag{1.8}$$

Therefore, we have

$$\min_{t\in[0,1]} W_c(\mu,\nu_t) = \frac{1}{I}\sum_{i\in\mathcal{I}} \min\left\{\|\boldsymbol{x}_i-\boldsymbol{y}_1\|^p, \|\boldsymbol{x}_i-\boldsymbol{y}_2\|^p\right\}.$$

Any minimizer $(\boldsymbol{q}^\star_1,\boldsymbol{q}^\star_2)$ of (1.7) gives thus rise to a minimizer $(t^\star,\boldsymbol{q}^\star_1,\boldsymbol{q}^\star_2)$ of (1.6), where $t^\star = (\boldsymbol{1}^\top\boldsymbol{q}^\star_1)/I$. Moreover, the minimizers of (1.5) are exactly all numbers of the form $t^\star = (\boldsymbol{1}^\top\boldsymbol{q}^\star_1)/I$ corresponding to the minimizer $(\boldsymbol{q}^\star_1,\boldsymbol{q}^\star_2)$ of (1.7). In view of (1.8), this observation allows us to conclude that

$$\operatorname*{argmin}_{t\in[0,1]} W_c(\mu,\nu_t) = \left[|\mathcal{I}_1|/I, |\mathcal{I}_0\cup\mathcal{I}_1|/I\right]. \tag{1.9}$$

By the definitions of $\mathcal{I}(\boldsymbol{w}, b)$, $\boldsymbol{y}_1$ and $\boldsymbol{y}_2$, it is further evident that

$$\begin{aligned} |\mathcal{I}(\boldsymbol{w}, b)| &= \left|\left\{i \in \mathcal{I} : \boldsymbol{w}^\top \boldsymbol{x}_i \leq b\right\}\right| \\ &= \left|\left\{i \in \mathcal{I} : \|\boldsymbol{x}_i - \boldsymbol{y}_1\|^2 \leq \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^2\right\}\right| = |\mathcal{I}_1 \cup \mathcal{I}_0|. \end{aligned}$$

Therefore, we may finally conclude that

$$|\mathcal{I}(\boldsymbol{w}, b)|/I \in \underset{t\in[0,1]}{\operatorname{argmin}} W_c(\mu, \nu_t).$$

Assume now that each component of $\boldsymbol{w}$ is even and b is odd. In this case, there exists no $\boldsymbol{x} \in \{0,1\}^K$ that satisfies $\boldsymbol{x}^\top \boldsymbol{w} = b$ and consequentially $\mathcal{I}_0$ is empty. Consequently, the interval of minimizers in (1.9) collapses to the singleton $|\mathcal{I}_1|/I = |\mathcal{I}(\boldsymbol{w}, b)|/I$. This observation completes the proof. □

Armed with Lemmas 1.3.4 and 1.3.5, we are now ready to prove Theorem 1.3.3.

Proof of Theorem 1.3.3. Select an instance of the #Knapsack problem with input $\boldsymbol{w} \in \mathbb{Z}_+^K$ and $b \in \mathbb{Z}_+$. Throughout this proof we will assume without loss of generality that each component of $\boldsymbol{w}$ is even and that b is odd. Indeed, if this was not the case, we could replace $\boldsymbol{w}$ with $\boldsymbol{w}' = 2\boldsymbol{w}$ and b with $b' = 2b + 1$. It is easy to verify that the two instances of the #Knapsack problem with inputs $(\boldsymbol{w}, b)$ and $(\boldsymbol{w}', b')$ have the same solution. In addition, the bit length of $(\boldsymbol{w}', b')$ is polynomially bounded in the bit length of $(\boldsymbol{w}, b)$.

Given $\boldsymbol{w}$ and b, define the distributions μ and ν_t for $t \in [0,1]$ as well as the set $\mathcal{I}(\boldsymbol{w}, b)$ in the usual way. From Lemma 1.3.4 we know that $W_c(\mu, \nu_t)$ is continuous, piecewise affine and convex in t. The analytical formula (1.2) further implies that $W_c(\mu, \nu_t)$ is affine on the interval $[(i-1)/I, i/I]$ with slope $a_i \cdot I$, where

$$a_i = W_c(\mu, \nu_{i/I}) - W_c(\mu, \nu_{(i-1)/I}) \qquad \forall i \in \mathcal{I}. \tag{1.10}$$

Thus, (1.5) constitutes a univariate convex optimization problem with a continuous piecewise affine objective function. As each component of $\boldsymbol{w}$ is even and b is odd, Lemma 1.3.5 implies that $t^\star = |\mathcal{I}(\boldsymbol{w}, b)|/I$ is the unique minimizer of (1.5). Therefore, the given instance of the #Knapsack problem can be solved by solving (1.5) and multiplying its unique minimizer $t^\star$ with I.

In the following we will first show that if we had access to an oracle that computes $W_c(\mu, \nu_t)$ exactly, then we could construct an algorithm that finds $t^\star$ and the solution $t^\star I$ of the #Knapsack problem by calling the oracle $2K$ times (Step 1). Next, we will prove that if we had access to an oracle that solves the #OT problem and thus outputs only approximations of $W_c(\mu, \nu_t)$, then we could extend the algorithm from Step 1 to a polynomial-time Turing reduction from the #Knapsack problem to the #OT problem (Step 2). Step 2 implies that #OT is #**P**-hard.

Step 1. Assume now that we have access to an oracle that computes $W_c(\mu, \nu_t)$ exactly. In addition, introduce an array $\boldsymbol{a} = (a_0, a_1, \ldots, a_I)$ with entries a_i,

$i \in \mathcal{I}$, defined as in (1.10) and with $a_0 = -\infty$. Thus, each element of $\boldsymbol{a}$ can be evaluated with at most two oracle calls. The array $\boldsymbol{a}$ is useful because it contains all the information that is needed to solve the univariate convex optimization problem (1.5). Indeed, as $W_c(\mu, \nu_t)$ is a convex piecewise linear function with slope $a_i \cdot I$ on the interval $[i/I, (i-1)/I]$, the array $\boldsymbol{a}$ is sorted in ascending order, and the unique minimizer $t^\star$ of (1.5) satisfies

$$|\mathcal{I}(\boldsymbol{w}, b)| = t^\star I = \max\left\{i \in \mathcal{I} \cup \{0\} : a_i \leq 0\right\}. \tag{1.11}$$

In other words, counting all elements of the set $\mathcal{I}(\boldsymbol{w}, b)$ and thereby solving the #Knapsack problem is equivalent to finding the maximum index $i \in \mathcal{I} \cup \{0\}$ that meets the condition $a_i \leq 0$. The binary search method detailed in Algorithm 1 efficiently finds this index. Binary search methods are also referred to as half-interval search or bisection algorithms, and they represent iterative methods for finding the largest number within a sorted array that is smaller or equal to a given threshold (0 in our case). Algorithm 1 first checks whether the number in the middle of the array is non-positive. Depending on the outcome, either the part of the array to the left or to the right of the middle element may be discarded because the array is sorted. This procedure is repeated until the array collapses to the single element corresponding to the sought number. As the length of the array is halved in each iteration, the binary search method applied to an array of length I returns the solution in $\log_2 I = K$ iterations [Cor+09, § 12].

Algorithm 1 Binary search method

```
Input: An array $\boldsymbol{a} \in \mathbb{R}^I$ with $I = 2^K$ sorted in ascending order
Initialize $\underline{n} = 0$ and $\overline{n} = I$
for $k = 1, \ldots, K$ do
    Set $n \leftarrow (\overline{n} + \underline{n})/2$
    if $a_n \leq 0$ then $\underline{n} \leftarrow n$ else $\overline{n} \leftarrow n$
end for
if $a_{\underline{n}} \leq 0$ then $n \leftarrow \underline{n}$ else $n \leftarrow \overline{n}$
Output: $n$
```

One can use induction to show that, in any iteration k of Algorithm 1, n is given by a multiple of 2^{K-k} and represents indeed an eligible index. Similarly, in any iteration k we have $\overline{n} - \underline{n} = 2^{K-k+1}$.

Step 2. Assume now that we have only access to an oracle that solves the #OT problem, which merely returns an approximation $\widetilde{W}_c(\mu, \nu_t)$ of $W_c(\mu, \nu_t)$. Setting $\widetilde{a}_0 = -\infty$ and

$$\widetilde{a}_i = \widetilde{W}_c(\mu, \nu_{i/I}) - \widetilde{W}_c(\mu, \nu_{(i-1)/I}) \qquad \forall i \in \mathcal{I}, \tag{1.12}$$

we can then introduce a perturbed array $\widetilde{\boldsymbol{a}} = (\widetilde{a}_0, \widetilde{a}_1, \ldots, \widetilde{a}_I)$ which provides an approximation for $\boldsymbol{a}$. In the following we will prove that, even though $\widetilde{\boldsymbol{a}}$ is no longer necessarily sorted in ascending order, the sign of $\widetilde{a}_i$ coincides with the sign of a_i for every $i \in \mathcal{I}$. Algorithm 1 therefore outputs the exact solution $|\mathcal{I}(\boldsymbol{w}, b)|$ of the #Knapsack problem even if its input $\boldsymbol{a}$ is replaced with $\widetilde{\boldsymbol{a}}$. To see this, we first note that

$$a_i = \frac{1}{I}\left(\|\boldsymbol{x}_i - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p\right) \qquad \forall i \in \mathcal{I}, \tag{1.13}$$

which is an immediate consequence of the analytical formula (1.2) for $W_c(\mu, \nu_t)$. We emphasize that (1.13) has only theoretical relevance but cannot be used to evaluate a_i in practice because it relies on our assumption that the support points $\boldsymbol{x}_i$, $i \in \mathcal{I}$, are ordered such that $\|\boldsymbol{x}_i - \boldsymbol{y}_1\|^p - \|\boldsymbol{x}_i - \boldsymbol{y}_2\|^p$ is non-decreasing in i. Indeed, there is no efficient algorithm for ordering these 2^K points in practice. Using (1.13), we then find

$$\overline{\varepsilon} = \frac{1}{4}\min_{i\in\mathcal{I}}\{|a_i| : a_i \neq 0\} = \frac{1}{4}\min_{i\in\mathcal{I}}|a_i|,$$

where the first equality follows from the definition of $\overline{\varepsilon}$, and the second equality holds because each component of $\boldsymbol{w}$ is even and b is odd, which implies that $\|\boldsymbol{x}_i - \boldsymbol{y}_1\| \neq \|\boldsymbol{x}_i - \boldsymbol{y}_2\|$ and thus $a_i \neq 0$ for all $i \in \mathcal{I}$. The last formula for $\overline{\varepsilon}$ immediately implies that $|a_i| \geq 4\overline{\varepsilon}$ for all $i \in \mathcal{I}$. Together with the estimate

$$|\widetilde{a}_i - a_i| \leq \left|\widetilde{W}_c(\mu, \nu_{i/I}) - W_c(\mu, \nu_{i/I})\right| + \left|\widetilde{W}_c(\mu, \nu_{(i-1)/I}) - W_c(\mu, \nu_{(i-1)/I})\right| \leq 2\overline{\varepsilon},$$

this implies that $\widetilde{a}_i$ has indeed the same sign as a_i for every $i \in \mathcal{I}$. As the execution of Algorithm 1 depends on the input array only through the signs of its components, Algorithm 1 with input $\widetilde{\boldsymbol{a}}$ computes indeed the exact solution $|\mathcal{I}(\boldsymbol{w}, b)|$ of the #Knapsack problem. If the perturbed slope $\widetilde{a}_n$ in line 4 of Algorithm 1 is evaluated via (1.12) by calling the #OT oracle twice, then Algorithm 1 constitutes a Turing reduction from the #**P**-hard #Knapsack problem to the #OT problem.

To prove that the #OT problem is #**P**-hard, it remains to be shown that if any oracle call requires unit time, then the Turing reduction constructed above runs in polynomial time in the bit length of $(\boldsymbol{w}, b)$. This is indeed the case because Algorithm 1 calls the #OT oracle only $2K$ times in total and because all other operations can be carried out efficiently. In particular, the time needed for reading the oracle outputs is polynomially bounded in the size of $(\boldsymbol{w}, b)$. Indeed, the bit length of $\widetilde{W}_c(\mu, \nu_{i/I})$ is polynomially bounded in the bit length of $(\boldsymbol{y}_1, \boldsymbol{y}_2, i/I)$ thanks to the definition of the #OT problem, and the time needed for computing $(\boldsymbol{y}_1, \boldsymbol{y}_2, i/I)$ is trivially bounded by a polynomial in the bit length of $(\boldsymbol{w}, b)$ for any $i \in \mathcal{I}$. These observations complete the proof. □

We emphasize that the Turing reduction derived in the proof of Theorem 1.3.3 can be implemented without knowing the accuracy level $\overline{\varepsilon}$ of the #OT oracle. This is essential because $\overline{\varepsilon}$ is defined as the minimum of exponentially many terms, and we are not aware of any method to compute it efficiently. Without such a method, a Turing reduction relying on $\overline{\varepsilon}$ could not run in polynomial time.

Remark 1 (Polynomial-Time Turing Reductions)**.** *Recall that a polynomial-time Turing reduction from problem A to problem B is a Turing reduction that runs in polynomial time in the input size of A under the hypothetical assumption that there is an oracle for solving B in unit time. The time needed for computing oracle inputs and reading oracle outputs is attributed to the Turing reduction and is not absorbed in the oracle. Thus, a Turing reduction can run in polynomial time only if the oracle's output size is guaranteed to be polynomially bounded. The existence of a polynomial-time Turing reduction from A to B implies that if there was an efficient algorithm for solving B, then we could solve A in polynomial time (this operationalizes the assertion that "A is not harder than B"). One could use this implication as an alternative definition, that is, one could define a polynomial-time Turing reduction as a Turing reduction that runs in polyonomial time provided that the oracle runs in polynomial time. In our opinion, this alternative definition would be perfectly reasonable. However, it is not equivalent to the original definition by [Val79a], which compels us to ascertain that the oracle output has polynomial size irrespective of the oracle's actual runtime. Instead, the alternative definition directly refers to the oracle's actual runtime. In that it conditions on oracles that run in polynomial time, it immediately guarantees that their outputs have polynomial size. In short, the original definition requires the bit length of the oracle's output to be polynonmially bounded for* every *oracle that solves B (which requires a proof), whereas the alternative definition requires such a bound only for oracles that solve B in polynomial time (which requires no proof). As Theorem 1.3.3 relies on the original definition of a polynomial-time Turing reduction, we had to introduce condition (ii) in the definition of the* #OT *problem. We consider the differences between the original and alternative definitions of polynomial-time Turing reductions as pure technicalities, but discussing them here seems relevant for motivating our formulation of the* #OT *problem.*

Assume now that p is an even number, and consider any instance of the #OT problem. In this case, all coefficients of the linear program (1.1) are rational, and thus $W_c(\mu, \nu_t)$ is a rational number that can be computed in finite time (*e.g.*, via the simplex algorithm). From Corollary 1 we further know that $W_c(\mu, \nu_t)$ has polynomially bounded bit length. Thus, $\widetilde{W}_c(\mu, \nu_t) = W_c(\mu, \nu_t)$ satisfies both properties (i) and (ii) that are required of an admissible approximation of the OT distance. Nevertheless, Theorem 1.3.3 asserts that computing $W_c(\mu, \nu_t)$ approximately is already #**P**-hard. This trivially implies that computing $W_c(\mu, \nu_t)$ *exactly* is also #**P**-hard.

1.4. Dynamic Programming-Type Solution Methods

We now return to the generic OT problem with independent marginals, where μ is representable as $\otimes_{k\in\mathcal{K}}\mu_k$, the marginals of μ constitute arbitrary univariate distributions supported on L points, and ν constitutes an arbitrary multivariate distribution supported on J points. This problem class covers all instances of the #OT problem, and by Theorem 1.3.3 it is therefore #**P**-hard even if only approximate solutions are sought. In fact, *any* problem class that is rich enough to contain all instances of the #OT problem is #**P**-hard. We will now demonstrate that, for $p = 2$, particular instances of the OT problem with independent marginals can be solved in polynomial or pseudo-polynomial time by a dynamic programming-type algorithm even though the distribution μ involves exponentially many atoms and the linear program (1.1) has exponential size. Throughout this discussion we call $\mathcal{N} \subseteq \mathbb{R}$ a one-dimensional regular grid with cardinality N if there exist $\hat{s}_1, \ldots, \hat{s}_N \in \mathbb{R}$ and a grid spacing constant $d > 0$ such that $\hat{s}_{i+1} = \hat{s}_i + d$ for all $i = 1, \ldots, N-1$ and $\mathcal{N} = \{\hat{s}_1, \ldots, \hat{s}_N\}$. We say that a set $\mathcal{M} \subseteq \mathbb{R}$ spans the one-dimensional regular grid $\mathcal{N}$ if $\mathcal{M} \subseteq \mathcal{N}$, $\min \mathcal{M} = \min \mathcal{N}$ and $\max \mathcal{M} = \max \mathcal{N}$.

Theorem 1.4.1 (Dynamic Programming-Type Algorithm for OT Problems with Independent Marginals)**.** *Suppose that $\mu = \otimes_{k\in\mathcal{K}}\mu_k$ is a product of K independent univariate distributions of the form $\mu_k = \sum_{l\in\mathcal{L}} \mu_k^l \delta_{x_k^l}$ and that $\nu_t = t\delta_{\boldsymbol{y}_1} + (1-t)\delta_{\boldsymbol{y}_2}$ is a two-point distribution. If $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2$ and if $\mathcal{M} = \{x_k^l(y_{1,k} - y_{2,k}) : k \in \mathcal{K},\, l \in \mathcal{L}\}$ spans a regular one-dimensional grid $\mathcal{N}$ with (known) cardinality N, then the OT distance between μ and ν_t can be computed exactly by a dynamic programming-type algorithm using $\mathcal{O}(KL\log_2(KL) + KLN + K^2N^2)$ arithmetic operations. If all problem parameters are rational and representable as ratios of two integers with absolute values at most U, then the bit lengths of all numbers computed by this algorithm are polynomially bounded in K, L, N and $\log_2(U)$.*

Assuming that $\mathcal{M}$ spans some regular one-dimensional grid $\mathcal{N}$, Theorem 1.4.1 establishes an upper bound on the number of arithmetic operations needed to solve the OT problem with independent marginals. We will see that the proof of Theorem 1.4.1 is constructive in that it develops a concrete dynamic programming-type algorithm that attains the indicated upper bound (see Algorithm 2). However, this bound depends on the cardinality N of the grid $\mathcal{N}$, and Theorem 1.4.1 does not relate N to K, L or U. More importantly, it provides no guidelines for constructing $\mathcal{N}$ or even proving its existence.

Remark 2 (Existence of $\mathcal{N}$)**.** *If all support points of μ and ν have rational components, then a regular one-dimensional grid $\mathcal{N}$ satisfying the assumptions of Theorem 1.4.1 is guaranteed to exist. In general, however, its cardinality scales exponentially with K and L, implying that the dynamic programming-type algorithm of Theorem 1.4.1 is inefficient. To see this, assume that for all $k \in \mathcal{K}$, $l \in \mathcal{L}$ and $j \in \{1, 2\}$ there exist integers $a_{k,l}, c_{j,k} \in \mathbb{Z}$ and $b_{k,l}, d_{j,k} \in \mathbb{N}$ such that $x_k^l = a_{k,l}/b_{k,l}$ and $y_{j,k} = c_{j,k}/d_{j,k}$. Thus, we have*

$$x_k^l(y_{1,k} - y_{2,k}) = \frac{a_k^l(c_{1,k}d_{2,k} - c_{2,k}d_{1,k})}{b_{k,l}d_{1,k}d_{2,k}} \quad k \in \mathcal{K},\ \forall l \in \mathcal{L},$$

which implies that all elements of $\mathcal{M}$ can be expressed as rational numbers with common denominator $D = \prod_{k\in\mathcal{K},l\in\mathcal{L}} b_{k,l}d_{1,k}d_{2,k}$. Clearly, $\mathcal{M}$ therefore spans *a regular one-dimensional grid $\mathcal{N}$ with grid spacing constant $d = D^{-1}$ and cardinality $N = D(\max \mathcal{M} - \min \mathcal{M}) + 1$. If U denotes as usual an upper bound on the absolute values of the integers $a_{k,l}$, $b_{k,l}$, $c_{j,k}$ and $d_{j,k}$ for all $k \in \mathcal{K}$, $l \in \mathcal{L}$ and $j \in \{1, 2\}$, then we have $D \leq U^{3KL}$, and all elements of $\mathcal{M}$ have absolute values of at most $2U^3$. The cardinality of $\mathcal{N}$ therefore satisfies $N \leq 4U^{3(KL+1)} + 1$. This reasoning suggests that, in the worst case, the dynamic programming-type algorithm of Theorem 1.4.1 may require up to $\mathcal{O}(K^2 U^{3(KL+1)})$ arithmetic operations.*

Remark 2 guarantees that a regular one-dimensional grid $\mathcal{N}$ satisfying the assumptions of Theorem 1.4.1 exists whenever the input bit length of the OT problem with independent marginals is finite. However, Remark 2 also reveals that the algorithm of Theorem 1.4.1 may be highly inefficient in general. Remark 3 below discusses special conditions under which this algorithm is of practical interest.

Remark 3 (Efficiency of the Dynamic Programming-Type Algorithm)**.** *The algorithm of Theorem 1.4.1 is efficient on problem instances that display the following properties.*

(i) *If $\mathcal{M}$ spans a regular one-dimensional grid whose cardinality N grows only polynomially with K and L but is independent of U, then the number of arithmetic operations required by the algorithm of Theorem 1.4.1 grows polynomially with K and L but is independent of U, and the bit lengths of all numbers computed by this algorithm are polynomially bounded in K, L and $\log_2(U)$. Hence, the algorithm runs in* strongly polynomial time *on a Turing machine.*

(ii) *If $\mathcal{M}$ spans a regular one-dimensional grid whose cardinality N grows polynomially with K, L and $\log_2(U)$, then the number of arithmetic operations required by the algorithm of Theorem 1.4.1 as well as the bit lengths of all numbers computed by this algorithm are polynomially bounded in K, L and $\log_2(U)$. Hence, the algorithm runs in* weakly polynomial time *on a Turing machine.*

(iii) *If $\mathcal{M}$ spans a regular one-dimensional grid whose cardinality grows polynomially with K, L and U (but exponentially with $\log_2(U)$), then the number of arithmetic operations required by the algorithm of Theorem 1.4.1 grows polyonomially with K, L and U, and the bit lengths of all numbers computed by this algorithm are polynomially bounded in K, L and $\log_2(U)$. Hence, the algorithm runs in* pseudo-polynomial time *on a Turing machine.*

Before proving Theorem 1.4.1, we recall the definition of the Conditional Value-at-Risk (CVaR) by [RU02]. Specifically, if the random vector $\boldsymbol{x}$ is governed by the probability distribution μ, then the CVaR at level $t \in (0, 1)$ of any Borel measurable loss function $\ell(\boldsymbol{x})$ is defined as

$$\mathrm{CVaR}_t[\ell(\boldsymbol{x})] = \inf_{\beta \in \mathbb{R}} \ \beta + \frac{1}{t} \, \mathbb{E}_{\boldsymbol{x} \sim \mu} \left[\max\{\ell(\boldsymbol{x}) - \beta, 0\} \right].$$

Here, the minimization problem over β is solved by the Value-at-Risk (VaR) at level t [RU02, Theorem 10], which is defined as the left $(1-t)$-quantile of the loss distribution, that is,

$$\mathrm{VaR}_t[\ell(x)] = \inf\left\{\tau \in \mathbb{R} : \mu[\ell(x) \leq \tau] \geq 1-t\right\}.$$

The proof of Theorem 1.4.1 also relies on the following lemma.

Lemma 1.4.2 (Minkowski sums of regular one-dimensional grids)**.** *If $\mathcal{N}$ is a one-dimensional regular grid with cardinality N and grid spacing constant $d > 0$, then the k-fold Minkowski sum $\sum_{i=1}^{k} \mathcal{N}$ of $\mathcal{N}$ is another one-dimensional regular grid with cardinality $k(N-1)+1$ and the same grid spacing constant d.*

Proof. Any regular one-dimensional grid with cardinality N and grid spacing constant $d > 0$ is representable as the image of $\{1, \ldots, N\}$ under the affine transformation $f(s) = \hat{s}_1 - d + ds$, where $\hat{s}_1$ denotes the smallest element of $\mathcal{N}$. It is immediate to see that the k-fold Minkowski sum of $\mathcal{N}$ is another one-dimensional regular grid with grid spacing constant d. In addition, the cardinality of this Minkowski sum satisfies

$$\begin{aligned} \left|\sum_{i=1}^{k} \mathcal{N}\right| &= \left|\sum_{i=1}^{k} f(\{1,\ldots,N\})\right| = \left|f\left(\sum_{i=1}^{k}\{1,\ldots,N\}\right)\right| \\ &= |f(\{k,\ldots,kN\})| = |\{k,\ldots,kN\}| = k(N-1)+1, \end{aligned}$$

where the second equality holds because f is affine and because the cardinality of any set is invariant under translations. Thus, the claim follows. □

Proof of Theorem 1.4.1. Throughout this proof we exceptionally assume that each arithmetic operation can be performed in unit time irrespective of the bit lengths of the involved operands. We emphasize that everywhere else in the chapter, however, time is measured in the standard Turing machine model of computation. Throughout this proof we further set $I = L^K$ and denote as usual by $\boldsymbol{x}_i$, $i \in \mathcal{I}$, the I different support points of μ. Then, the OT distance between μ and ν_t can be expressed as

$$\begin{aligned} W_c(\mu,\nu_t) &= \min_{\boldsymbol{\pi}\in\Pi(\mu,\nu_t)} \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \left(\|\boldsymbol{x}_i\|^2 + \|\boldsymbol{y}_j\|^2 - 2\boldsymbol{x}_i^\top \boldsymbol{y}_j\right)\pi_{ij} \\ &= \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\|\boldsymbol{x}\|^2\right] + \mathbb{E}_{\boldsymbol{y}\sim\nu_t}\left[\|\boldsymbol{y}\|^2\right] - 2\max_{\boldsymbol{\pi}\in\Pi(\mu,\nu_t)} \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \boldsymbol{x}_i^\top \boldsymbol{y}_j \pi_{ij}. \end{aligned} \tag{1.14}$$

The two expectations in (1.14) can be evaluated in $\mathcal{O}(KL)$ arithmetic operations because

$$\begin{aligned} &\mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\|\boldsymbol{x}\|^2\right] = \sum_{k\in\mathcal{K}} \mathbb{E}_{x_k\sim\mu_k}\left[(x_k)^2\right] = \sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{L}} \mu_k^l (x_k^l)^2 \text{ and} \\ &\mathbb{E}_{\boldsymbol{y}\sim\nu_t}\left[\|\boldsymbol{y}\|^2\right] = t\|\boldsymbol{y}_1\|^2 + (1-t)\|\boldsymbol{y}_2\|^2, \end{aligned}$$

and it is easy to verify that their bit lengths are polynomially bounded in K, L and $\log_2(U)$. Moreover, as in the proof of Lemma 1.3.5, the maximization problem in (1.14) simplifies to

$$\max_{\boldsymbol{\pi}\in\Pi(\mu,\nu_t)} \sum_{i\in\mathcal{I}}\sum_{j\in\mathcal{J}} \boldsymbol{x}_i^\top \boldsymbol{y}_j \pi_{ij} = \left\{\begin{array}{cl} \max\limits_{\boldsymbol{q}_1,\boldsymbol{q}_2\in\mathbb{R}^I_+} & t\sum\limits_{i\in\mathcal{I}} \boldsymbol{x}_i^\top \boldsymbol{y}_1 q_{1,i} + (1-t)\sum\limits_{i\in\mathcal{I}} \boldsymbol{x}_i^\top \boldsymbol{y}_2 q_{2,i} \\ \text{s.t.} & \boldsymbol{1}^\top \boldsymbol{q}_1 = 1,\ \boldsymbol{1}^\top \boldsymbol{q}_2 = 1 \\ & t q_{1,i} + (1-t) q_{2,i} = \mu[\boldsymbol{x} = \boldsymbol{x}_i] \quad \forall i \in \mathcal{I}. \end{array}\right.$$

$$= \sum_{i\in\mathcal{I}} \boldsymbol{x}_i^\top \boldsymbol{y}_2\, \mu[\boldsymbol{x}=\boldsymbol{x}_i] + \left\{\begin{array}{cl} \max\limits_{\boldsymbol{q}\in\mathbb{R}^I_+} & \sum\limits_{i\in\mathcal{I}} \boldsymbol{x}_i^\top(\boldsymbol{y}_1-\boldsymbol{y}_2) q_i \\ \text{s.t.} & \boldsymbol{1}^\top \boldsymbol{q} = t \\ & q_i \le \mu[\boldsymbol{x}=\boldsymbol{x}_i] \quad \forall i \in \mathcal{I}, \end{array}\right. \tag{1.15}$$

where the second equality follows from the variable substitution $\boldsymbol{q} \leftarrow t\boldsymbol{q}_1$ and the subsequent elimination of $\boldsymbol{q}_2$ by using the equations $(1-t)q_{2,i} = \mu[\boldsymbol{x} = \boldsymbol{x}_i] - q_i$ for all $i \in \mathcal{I}$. Observe next that the first sum in (1.15) can again be evaluated using $\mathcal{O}(KL)$ arithmetic operations because

$$\sum_{i\in\mathcal{I}} \boldsymbol{x}_i^\top \boldsymbol{y}_2\, \mu[\boldsymbol{x} = \boldsymbol{x}_i] = \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\boldsymbol{x}^\top \boldsymbol{y}_2\right] = \sum_{k\in\mathcal{K}} \mathbb{E}_{x_k\sim\mu_k}\left[x_k y_{2,k}\right] = \sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{L}} x_k^l \mu_k^l y_{2,k},$$

and the bit length of this sum is polynomially bounded in K, L and $\log_2(U)$. For $t = 0$, the optimal value of the maximization problem in (1.15) vanishes. For $t = 1$, on the other hand, the problem's optimal solution satisfies $q_i = \mu[\boldsymbol{x} = \boldsymbol{x}_i]$ for all $i \in \mathcal{I}$. By using now standard arguments, one readily verifies that the corresponding optimal value can once again be computed in $\mathcal{O}(KL)$ arithmetic operations and has polynomially bounded bit length in K, L and $\log_2(U)$. In the remainder of the proof we may thus assume that $t \in (0,1)$. To solve the maximization problem in (1.15) in this generic case, we first reformulate it as

$$t \cdot \max\left\{\sum_{i\in\mathcal{I}} \ell(\boldsymbol{x}_i)\, \mu[\boldsymbol{x} = \boldsymbol{x}_i]\, q_i \ :\ \boldsymbol{0} \le \boldsymbol{q} \le t\cdot\boldsymbol{1},\ \sum_{i\in\mathcal{I}} \mu[\boldsymbol{x} = \boldsymbol{x}_i]\, q_i = 1\right\} \tag{1.16}$$

by applying the variable substitution $q_i \leftarrow q_i/(t\cdot\mu[\boldsymbol{x} = \boldsymbol{x}_i])$ and defining $\ell(\boldsymbol{x}) = \boldsymbol{x}^\top(\boldsymbol{y}_1 - \boldsymbol{y}_2)$. The maximization problem in (1.16) is then readily recognized as the dual representation of the CVaR of $\ell(\boldsymbol{x})$ at level t; see, *e.g.*, [SDR21, Example 6.16]. The expression (1.16) thus equals $t \cdot \mathrm{CVaR}_t(\ell(\boldsymbol{x}))$.

By assumption, there exists a one-dimensional regular grid $\mathcal{N}$ with cardinality N such that $x_k^l(y_{1,k} - y_{2,k}) \in \mathcal{N}$ for every $k \in \mathcal{K}$ and $l \in \mathcal{L}$. This readily implies that $\ell(\boldsymbol{x}_i) = \boldsymbol{x}_i^\top(\boldsymbol{y}_1 - \boldsymbol{y}_2) \in \mathcal{N}_K = \sum_{k=1}^K \mathcal{N}$. Assume from now on without loss of generality that $\mathcal{N}_K = \{\hat{s}_{K,1}, \ldots, \hat{s}_{K,|\mathcal{N}_K|}\}$ and that the elements of $\mathcal{N}_K$ are sorted in ascending order, that is, $\hat{s}_{K,1} < \cdots < \hat{s}_{K,|\mathcal{N}_K|}$. Also, denote by n_t the unique index satisfying

$$\sum_{n=1}^{n_t} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}] \geq 1 - t > \sum_{n=1}^{n_t-1} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}]. \tag{1.17}$$

By [RU02, Proposition 8], the expression (1.16) can therefore be reformulated as

$$t \cdot \mathrm{CVaR}_t[\ell(\boldsymbol{x})] = \left(\sum_{n=1}^{n_t} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}] - (1-t) \right) \hat{s}_{K,n_t} + \sum_{n=n_t+1}^{|\mathcal{N}_K|} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}] \hat{s}_{K,n}. \tag{1.18}$$

Computing (1.18) thus amounts to evaluating a sum of $\mathcal{O}(|\mathcal{N}_K|)$ terms. We will now prove that evaluating this sum requires $\mathcal{O}(KL\log_2(KL) + KLN + K^2N^2)$ arithmetic operations. To this end, we first show that the grid points $\hat{s}_{K,n}$, $n = 1, \ldots, |\mathcal{N}_K|$, can be computed in time $\mathcal{O}(KL\log_2(KL) + KN)$ (Step 1), then we show that the probabilities $\mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}]$, $n = 1, \ldots, |\mathcal{N}_K|$, can be computed recursively in time $\mathcal{O}(KLN + K^2N^2)$ (Step 2), and finally we use these ingredients to compute the right hand side of (1.18) in time $\mathcal{O}(KN)$ (Step 3).

Step 1. By assumption, the regular grid $\mathcal{N}$ has known cardinality N and is spanned by $\mathcal{M} = \{x_k^l(y_{1,k} - y_{2,k}) : k \in \mathcal{K},\, l \in \mathcal{L}\}$. To compute all elements of $\mathcal{N}$, we first compute all elements of $\mathcal{M}$ in time $\mathcal{O}(KL)$ and sort them in non-decreasing order in time $\mathcal{O}(KL\log_2(KL))$ using merge sort, for example. As $\mathcal{M}$ spans $\mathcal{N}$, the minimum and the maximum of $\mathcal{M}$ coincide with the minimum $\hat{s}_1$ and the maximum $\hat{s}_N$ of $\mathcal{N}$, respectively. Given $\hat{s}_1$ and $\hat{s}_N$, we can then compute the grid spacing constant $d = (\hat{s}_N - \hat{s}_1)/(N-1)$ as well as the elements $\hat{s}_n = \hat{s}_1 + d(n-1)$, $n = 1, \ldots, N$, of $\mathcal{N}$, which requires $\mathcal{O}(N)$ arithmetic operations. The bit lengths of all numbers computed so far are bounded by a polynomial in $\log_2(U)$ and $\log_2(N)$.

It is easy to see that $\mathcal{N}_K = \sum_{k=1}^K \mathcal{N}$ is also a one-dimensional regular grid that has the same grid spacing constant as $\mathcal{N}$ and whose minimum $\hat{s}_{K,1} = K\hat{s}_1$ can be computed in constant time. The elements of $\mathcal{N}_K$ are then obtained by computing $\hat{s}_{K,n} = \hat{s}_{K,1} + d(n-1)$ for all $n = 1, \ldots, |\mathcal{N}_K|$, where $|\mathcal{N}_K| = K(N-1)+1$ thanks to Lemma 1.4.2. This computation requires $\mathcal{O}(KN)$ arithmetic operations, and the bit lengths of all involved numbers are still bounded by a polynomial in $\log_2(U)$ and $\log_2(N)$. This completes Step 1.

Step 2. We now show that the probabilities $\mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}]$ for $n = 1, \ldots, |\mathcal{N}_K|$ can be calculated recursively in time $\mathcal{O}(K^2N^2)$. To this end, we introduce the partial sums $\ell_k(\boldsymbol{x}) = \sum_{m=1}^k x_m(y_{1,m} - y_{2,m})$ for every $k \in \mathcal{K}$ and note that $\ell_K(\boldsymbol{x}) = \ell(\boldsymbol{x})$. For every $k \in \mathcal{K}$, the range of the function $\ell_k(\boldsymbol{x})$ is a subset of the one-dimensional regular grid $\mathcal{N}_k = \sum_{k'=1}^k \mathcal{N}$. The law of total probability then implies that

$$\mu[\ell_k(\boldsymbol{x}) = \hat{s}] = \sum_{\hat{s}' \in \mathcal{N}} \mu\left[\ell_{k-1}(\boldsymbol{x}) = \hat{s} - \hat{s}',\ x_k(y_{1,k} - y_{2,k}) = \hat{s}'\right]$$

$$\forall k \in \mathcal{K} \backslash \{1\},\ \forall \hat{s} \in \mathcal{N}_k,$$

where $\hat{s}_1, \ldots, \hat{s}_N$ denote as usual the elements of $\mathcal{N}$, and where $\mu[\ell_1(\boldsymbol{x}) = \hat{s}] = \mu_1[x_1(y_{1,1} - y_{2,1}) = \hat{s}]$ for all $\hat{s} \in \mathcal{N}_1$. As $\ell_k(\boldsymbol{x}) = \ell_{k-1}(\boldsymbol{x}) + x_k(y_{1,k} - y_{2,k})$, $\ell_{k-1}(\boldsymbol{x})$ is constant in $x_k, \ldots, x_K$ and the components of $\boldsymbol{x}$ are mutually independent under the product distribution $\mu = \otimes_{k \in \mathcal{K}} \mu_k$, we thus have

$$\mu[\ell_k(\boldsymbol{x}) = \hat{s}] = \sum_{\hat{s}' \in \mathcal{N}} \mu\left[\ell_{k-1}(\boldsymbol{x}) = \hat{s} - \hat{s}'\right] \times \mu_k\left[x_k(y_{1,k} - y_{2,k}) = \hat{s}'\right] \qquad (1.19)$$
$$\forall k \in \mathcal{K} \backslash \{1\},\ \forall \hat{s} \in \mathcal{N}_k.$$

The marginal probabilities $\mu_k[x_k(y_{1,k} - y_{2,k}) = \hat{s}']$ for all $k \in \mathcal{K}$ and $\hat{s}' \in \mathcal{N}$ can be pre-computed in time $\mathcal{O}(KLN)$. Given $\mu[\ell_{k-1}(\boldsymbol{x}) = \hat{s}]$, $\hat{s} \in \mathcal{N}_{k-1}$, each probability $\mu[\ell_k(\boldsymbol{x}) = \hat{s}]$, $\hat{s} \in \mathcal{N}_k$, can then be computed in time $\mathcal{O}(N)$ by using (1.19). As $|\mathcal{N}_k| = \mathcal{O}(kN)$ for every $k \in \mathcal{K}$ thanks to Lemma 1.4.2, each iteration $k \in \mathcal{K}$ of the the dynamic programming-type recursion (1.19) requires at most $\mathcal{O}(KN^2)$ arithmetic operations. Finally, as there are $\mathcal{O}(K)$ iterations in total, the sought probabilities $\mu[\ell_K(\boldsymbol{x}) = \hat{s}]$, $\hat{s} \in \mathcal{N}_K$, can be computed in time $\mathcal{O}(K^2N^2)$. An elementary calculation further shows that the bit lengths of these probabilities are bounded by a polynomial in K, N and $\log_2(U)$. This completes Step 2.

Step 3. As all terms appearing in the sum on the right hand side of (1.18) have been pre-computed in Steps 1 and 2, the sum itself can now be evaluated in time $\mathcal{O}(KN)$ thanks to Lemma 1.4.2. Note that the critical index n_t defined in (1.17) can also be computed in time $\mathcal{O}(KN)$. The bit lengths of all numbers involved in these calculations are bounded by a polynomial in K, N and $\log_2(U)$. This completes Step 3.

In summary, the time required for evaluating the CVaR in (1.18) totals $\mathcal{O}(KL \log_2(KL) + KLN + K^2N^2)$, which matches the overall time required for all calculations described in Steps 1, 2 and 3. This computation time dominates the time $\mathcal{O}(KL)$ spent on all preprocessing steps, and thus the claim follows. □

The dynamic programming-type procedure developed in the proof of Theorem 1.3.3 is summarized in Algorithm 2. This procedure outputs the OT distance between μ and ν_t (denoted by W_c). In addition, Algorithm 2 can be used for constructing the optimal transportation plan from μ to ν_t.

Algorithm 2 OT with Independent Marginals

```
Input: $\{\mu_k^l\}_{k\in\mathcal{K},l\in\mathcal{L}},\ \{x_k^l\}_{k\in\mathcal{K},l\in\mathcal{L}},\ \boldsymbol{y}_1,\boldsymbol{y}_2\in\mathbb{R}^K,\ t,\ N$
Initialize $\hat{s}_1 = \min_{k\in\mathcal{K},l\in\mathcal{L}} x_k^l(y_{1,k}-y_{2,k})$ and $\hat{s}_N = \max_{k\in\mathcal{K},l\in\mathcal{K}} x_k^l(y_{1,k}-y_{2,k})$
Set $d=(\hat{s}_N-\hat{s}_1)/(N-1)$ and $\hat{s}_n=\hat{s}_1+d(n-1)\ \forall n=1,\dots,N$
Compute $\mu_k[x_k(y_{1,k}-y_{2,k})=\hat{s}_n]\ \forall k\in\mathcal{K}$ and $n\in\mathcal{N}$
Set $\mu[\ell_1(\boldsymbol{x})=\hat{s}_n]=\mu_1[x_1(y_{1,1}-y_{2,1})=\hat{s}_n]\ \forall n=1,\dots,N$
for $k=2,\dots,K$ do
    for $n=1,\dots,k(N-1)+1$ do
        $\hat{s}_{k,n}=k\hat{s}_1+d(n-1)$
        $\mu[\ell_k(\boldsymbol{x})=\hat{s}_{k,n}]=\sum_{\hat{s}'\in\mathcal{N}}\mu[\ell_{k-1}(\boldsymbol{x})=\hat{s}_{k,n}-\hat{s}']\times\mu_k[x_k(y_{1,k}-y_{2,k})=\hat{s}']$
    end for
end for
Find the index $n_t\in\{1,\dots,K(N-1)+1\}$ satisfying (1.17)
Set
    $\mathrm{CVaR}=\frac{1}{t}\Big[\Big(\sum_{n=1}^{n_t}\mu[\ell_K(\boldsymbol{x})=\hat{s}_{K,n}]-1+t\Big)\hat{s}_{K,n_t} - 2\sum_{n=n_t+1}^{K(N-1)+1}\mu[\ell_K(\boldsymbol{x})=\hat{s}_{K,n}]\hat{s}_{K,n}\Big]$
Set
    $W_c=\sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{L}}\mu_k^l(x_k^l)^2+t\sum_{k\in\mathcal{K}}y_{1,k}^2+(1-t)\sum_{k\in\mathcal{K}}y_{2,k}^2 - 2\sum_{k\in\mathcal{K}}\sum_{l\in\mathcal{L}}x_k^l\mu_k^l y_{2,k}-2t\cdot\mathrm{CVaR}$
Output: $W_c$
```

Remark 4 (Optimal Transportation Plan)**.** *The critical index n_t computed by Algorithm 2 allows us to construct an optimal transportation plan $\boldsymbol{\pi}^\star\in\mathbb{R}_+^{I\times J}$ that solves the linear program (1.1), where $\pi_{i,j}^\star$ denotes the probability mass moved from $\boldsymbol{x}_i$ to $\boldsymbol{y}_j$ for every $i\in\mathcal{I}$ and $j\in\mathcal{J}$. To see this, note that the defining properties of n_t in (1.17) imply that $\mathrm{VaR}_t[\ell(\boldsymbol{x})]=\hat{s}_{K,n_t}$ and $\mu[\ell(\boldsymbol{x})=$*

$\hat{s}_{K,n_t}] > 0$. *We may thus define* $\boldsymbol{\pi}^\star$ *via*

$$\pi^\star_{i,1} = \begin{cases} \mu[\boldsymbol{x} = \boldsymbol{x}_i] & \text{if } \ell(\boldsymbol{x}_i) > \hat{s}_{K,n_t} \\ \dfrac{t - 1 + \sum_{n=1}^{n_t} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}]}{\mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n_t}]} \times \mu[\boldsymbol{x} = \boldsymbol{x}_i] & \text{if } \ell(\boldsymbol{x}_i) = \hat{s}_{K,n_t} \\ 0 & \text{if } \ell(\boldsymbol{x}_i) < \hat{s}_{K,n_t} \end{cases}$$

and $\pi^\star_{i,2} = \mu[\boldsymbol{x} = \boldsymbol{x}_i] - \pi^\star_{i,1}$ *for all* $i \in \mathcal{I}$. *By the first inequality in* (1.17), *we have* $\boldsymbol{\pi}^\star \geq \mathbf{0}$. *In addition, we trivially find* $\pi^\star_{i,1} + \pi^\star_{i,2} = \mu[\boldsymbol{x} = \boldsymbol{x}_i]$ *for all* $i \in \mathcal{I}$, *and we have*

$$\begin{aligned} \sum_{i\in\mathcal{I}} \pi^\star_{i,1} = & \sum_{\substack{i\in\mathcal{I}: \\ \ell(\boldsymbol{x}_i) > \hat{s}_{K,n_t}}} \mu[\boldsymbol{x} = \boldsymbol{x}_i] + \\ & \sum_{\substack{i\in\mathcal{I}: \\ \ell(\boldsymbol{x}_i) = \hat{s}_{K,n_t}}} \frac{t - 1 + \sum_{n=1}^{n_t} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}]}{\mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n_t}]} \times \mu[\boldsymbol{x} = \boldsymbol{x}_i] \\ = & \sum_{n=n_t+1}^{|\mathcal{N}_K|} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}] + t - 1 + \sum_{n=1}^{n_t} \mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}] = t = 1 - \sum_{i\in\mathcal{I}} \pi^\star_{i,2}. \end{aligned}$$

In summary, this shows that $\boldsymbol{\pi}^\star$ *is feasible in the OT problem* (1.1). *Finally, we have*

$$\begin{aligned} \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \pi^\star_{ij} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2 &= \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\|\boldsymbol{x}\|^2\right] + \mathbb{E}_{\boldsymbol{y}\sim\nu_t}\left[\|\boldsymbol{y}\|^2\right] - 2\sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \boldsymbol{x}_i^\top \boldsymbol{y}_j \pi^\star_{ij} \\ &= \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\|\boldsymbol{x}\|^2\right] + \mathbb{E}_{\boldsymbol{y}\sim\nu_t}\left[\|\boldsymbol{y}\|^2\right] - \\ &\qquad 2\,\mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\boldsymbol{x}^\top \boldsymbol{y}_2\right] - 2\sum_{i\in\mathcal{I}} \ell(\boldsymbol{x}_i)\, \pi^\star_{i,1} \\ &= \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\|\boldsymbol{x}\|^2\right] + \mathbb{E}_{\boldsymbol{y}\sim\nu_t}\left[\|\boldsymbol{y}\|^2\right] - \\ &\qquad 2\,\mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\boldsymbol{x}^\top \boldsymbol{y}_2\right] - 2t \cdot \mathit{CVaR}_t[\ell(\boldsymbol{x})], \end{aligned}$$

where the first two equalities follow from (1.14) *and* (1.15), *respectively, while the third equality exploits the definitions of* $\boldsymbol{\pi}^\star$ *and the CVaR. The last expression manifestly matches the output* W_c *of Algorithm* 2. *Hence, we may conclude that* $\boldsymbol{\pi}^\star$ *is indeed optimal in* (1.1). *Note that evaluating* $\pi^\star_{ij}$ *for a fixed* $i \in \mathcal{I}$ *and* $j \in \mathcal{J}$ *requires at most* $\mathcal{O}(NK + KL)$ *arithmetic operations provided that the critical index* n_t *and the probabilities* $\mu[\ell(\boldsymbol{x}) = \hat{s}_{K,n}]$, $n \in \mathcal{N}_K$, *are given. These quantities are indeed computed by Algorithm* 2.

In the following we will identify special instances of the OT problem with independent marginals that can be solved efficiently. Assume first that both μ and ν are supported on $\{0,1\}^K$. This implies that all marginals of μ represent independent Bernoulli distributions. Unlike in Theorem 1.3.3, however, these

Bernoulli distributions may be non-uniform. The following corollary shows that, in this case, the OT problem with independent marginals can be solved in strongly polynomial time.

Corollary 2 (Binary Support)**.** *Suppose that all assumptions of Theorem 1.4.1 hold. If in addition $L = 2$, $x_k^1 = 0$ and $x_k^2 = 1$ for all $k \in \mathcal{K}$, and $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \{0,1\}^K$, then the OT distance between μ and ν_t can be computed in strongly polynomial time.*

Proof. Under the given assumptions, we have $\mathcal{M} = \{x_k^l(y_{1,k} - y_{2,k}) : k \in \mathcal{K},\, l \in \mathcal{L}\} \subseteq \{-1, 0, 1\}$. Hence, Theorem 1.4.1 applies with $\mathcal{N} \subseteq \{-1, 0, 1\}$ and $N \leq 3$, and therefore Algorithm 2 computes $W_c(\mu, \nu_t)$ using $\mathcal{O}(K^2)$ arithmetic operations. As N is constant in K, L and $\log_2(U)$, Remark 3 *(i)* implies that $W_c(\mu, \nu_t)$ can be computed in strongly polynomial time in the Turing machine model. □

By generalizing the proof of Corollary 2 in the obvious way, one can show that the OT problem with independent marginals remains strongly polynomial-time solvable whenever μ and ν_t are supported on a (fixed) bounded subset of the scaled integer lattice $\mathbb{Z}^K/M$ for some (fixed) scaling factor $M \in \mathbb{N}$. If μ and ν_t are supported on a subset of $\mathbb{Z}^K/M$ that may grow with the problem's input size or if the scaling factor M may grow with the input size, then Algorithm 2 ceases to run in polynomial time. We now show, however, that Algorithm 2 stills run in pseudo-polynomial time in these cases.

Corollary 3 (Lattice Support)**.** *Suppose that all assumptions of Theorem 1.4.1 hold. If there exists a positive integer $M \leq U$, such that $x_k^l \in \mathbb{Z}/M$ for all $k \in \mathcal{K}$ and $l \in \mathcal{L}$, while $\boldsymbol{y}_1, \boldsymbol{y}_2 \in \mathbb{Z}^K/M$, then the OT distance between μ and ν_t can be computed in pseudo-polynomial time.*

Proof. Under the given assumptions, we have $\mathcal{M} = \{x_k^l(y_{1,k} - y_{2,k}) : k \in \mathcal{K},\, l \in \mathcal{L}\} \subseteq \mathbb{Z}/M^2$. Therefore, $\mathcal{M}$ spans a one-dimensional regular grid $\mathcal{N} \subseteq \mathbb{Z}/M^2$ with grid spacing constant $d = 1/M^2$ and cardinality

$$\begin{aligned} N &= (\max \mathcal{M} - \min \mathcal{M})\,/d \\ &= \max_{k \in \mathcal{K},\, l \in \mathcal{L}} \left\{ M x_k^l (M y_{1,k} - M y_{2,k}) \right\} - \min_{k \in \mathcal{K},\, l \in \mathcal{L}} \left\{ M x_k^l (M y_{1,k} - M y_{2,k}) \right\}. \end{aligned} \tag{1.20}$$

Recall that $x_k^l = a_k^l / b_k^l$ for some $a_k^l \in \mathbb{Z}$ and $b_k^l \in \mathbb{N}$ with $|a_k^l|, |b_k^l| \leq U$ and that $M \leq U$. We may thus conclude that $|M x_k^l| \leq U^2$ for all $k \in \mathcal{K}$ and $l \in \mathcal{L}$. Similarly, one can show that $|M y_{1,k}| \leq U^2$ and $|M y_{2,k}| \leq U^2$ for all $k \in \mathcal{K}$. By (1.20), we thus have $N \leq 4U^2$, which implies via Theorem 1.4.1 that Algorithm 2 computes $W_c(\mu, \nu_t)$ using $\mathcal{O}(KL \log_2(KL) + K^2 U^4)$ arithmetic operations. We emphasize that the number of arithmetic operations thus grows polynomially with K, L and U but exponentially with $\log_2(U)$. By Remark 3 *(iii)*, $W_c(\mu, \nu_t)$ can therefore be computed in pseudo-polynomial time. □

So far we have discussed methods to solve the OT problem with independent marginals *exactly*. In the remainder of this section we will show that *approximate* solutions can always be computed in pseudo-polynomial time. The following lemma provides a key ingredient for this argument.

Lemma 1.4.3 (Approximating OT Distances)**.** *Consider four discrete probability distributions* $\mu = \sum_{i\in\mathcal{I}} \mu_i \delta_{\boldsymbol{x}_i}$, $\tilde{\mu} = \sum_{i\in\mathcal{I}} \mu_i \delta_{\tilde{\boldsymbol{x}}_i}$, $\nu = \sum_{j\in\mathcal{J}} \nu_j \delta_{\boldsymbol{y}_j}$ *and* $\tilde{\nu} = \sum_{j\in\mathcal{J}} \nu_j \delta_{\tilde{\boldsymbol{y}}_j}$ *supported on a hypercube* $[-U,U]^K$ *for some* $U > 0$*. If* $c(\boldsymbol{x},\boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|^2$ *and there exists* $\varepsilon \geq 0$ *such that* $\|\tilde{\boldsymbol{x}}_i - \boldsymbol{x}_i\|_\infty \leq \varepsilon$ *for all* $i \in \mathcal{I}$ *and* $\|\tilde{\boldsymbol{y}}_j - \boldsymbol{y}_j\|_\infty \leq \varepsilon$ *for all* $j \in \mathcal{J}$*, then we have*

$$|W_c(\mu,\nu) - W_c(\tilde{\mu},\tilde{\nu})| \leq 8KU\varepsilon. \tag{1.21}$$

We emphasize that Lemma 1.4.3 holds for arbitrary discrete distributions μ, $\tilde{\mu}$, ν and $\tilde{\nu}$ provided that $\tilde{\mu}$ and $\tilde{\nu}$ are obtained by perturbing only the support points of μ and ν, respectively, but not the corresponding probabilities. In particular, the lemma holds even if μ and $\tilde{\mu}$ fail to represent product distributions with independent marginals, and even if ν and $\tilde{\nu}$ fail to represent two-point distributions. Note also that, by slight abuse of notation, μ_i, $i \in \mathcal{I}$, represent here the probabilties of the support points of μ and should not be confused with the univariate marginal distributions μ_k, $k \in \mathcal{K}$, in the rest of the chapter.

Proof of Lemma 1.4.3. The elementary identity $|a^2 - b^2| = (a+b)|a-b|$ for any $a, b \in \mathbb{R}_+$ implies that

$$|W_c(\mu,\nu) - W_c(\tilde{\mu},\tilde{\nu})| = \left(W_c(\mu,\nu_t)^{\frac{1}{2}} + W_c(\tilde{\mu},\tilde{\nu})^{\frac{1}{2}}\right) \left|W_c(\mu,\nu)^{\frac{1}{2}} - W_c(\tilde{\mu},\tilde{\nu})^{\frac{1}{2}}\right|. \tag{1.22}$$

By the definition of the OT distance, the first term on the right-hand-side of (1.22) satisfies

$$\begin{aligned} W_c(\mu,\nu)^{\frac{1}{2}} + W_c(\tilde{\mu},\tilde{\nu})^{\frac{1}{2}} &= \left(\min_{\pi\in\Pi(\mu,\nu)} \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2 \pi_{ij}\right)^{\frac{1}{2}} + \\ &\qquad \left(\min_{\tilde{\pi}\in\Pi(\tilde{\mu},\tilde{\nu})} \sum_{i\in\mathcal{I}} \sum_{j\in\mathcal{J}} \|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{y}}_j\|^2 \tilde{\pi}_{ij}\right)^{\frac{1}{2}} \\ &\leq 4\sqrt{K}U, \end{aligned}$$

where the inequality holds because π and $\tilde{\pi}$ are probability distributions and because

$$\|\boldsymbol{x}_i - \boldsymbol{y}_j\|^2 \leq K\|\boldsymbol{x}_i - \boldsymbol{y}_j\|_\infty^2 \leq 4KU^2 \quad \text{and} \quad \|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{y}}_j\|^2 \leq \|\tilde{\boldsymbol{x}}_i - \tilde{\boldsymbol{y}}_j\|_\infty^2 \leq 4KU^2$$

for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$, taking into account that all support points of the four probability distributions μ, $\tilde{\mu}$, ν and $\tilde{\nu}$ fall into the hypercube $[-U,U]^K$. The second term on the right-hand-side of (1.22) satisfies

$$\left|W_c(\mu,\nu)^{\frac{1}{2}} - W_c(\tilde{\mu},\tilde{\nu})^{\frac{1}{2}}\right| \leq \left|W_c(\mu,\nu)^{\frac{1}{2}} - W_c(\tilde{\mu},\nu)^{\frac{1}{2}}\right| + \left|W_c(\tilde{\mu},\nu)^{\frac{1}{2}} - W_c(\tilde{\mu},\tilde{\nu})^{\frac{1}{2}}\right|$$

$$
\begin{aligned}
&\le W_c(\mu, \tilde{\mu})^{\frac{1}{2}} + W_c(\nu, \tilde{\nu})^{\frac{1}{2}} \\
&= \left(\min_{\pi^\mu \in \Pi(\mu, \tilde{\mu})} \sum_{i,i' \in \mathcal{I}} \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_{i'}\|^2 \pi^\mu_{ii'} \right)^{\frac{1}{2}} + \\
&\qquad \left(\min_{\pi^\nu \in \Pi(\nu, \tilde{\nu})} \sum_{j,j' \in \mathcal{J}} \|\boldsymbol{y}_j - \tilde{\boldsymbol{y}}_{j'}\|^2 \pi^\nu_{jj'} \right)^{\frac{1}{2}} \\
&\le \left(\frac{1}{I} \sum_{i \in \mathcal{I}} \|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|^2 \right)^{\frac{1}{2}} + \left(\frac{1}{J} \sum_{j \in \mathcal{J}} \|\boldsymbol{y}_j - \tilde{\boldsymbol{y}}_j\|^2 \right)^{\frac{1}{2}} \\
&\le 2\sqrt{K}\varepsilon,
\end{aligned}
$$

where the second inequality holds because the 2-Wasserstein distance is a metric and thus obeys the triangle inequality [Vil08, § 6], whereas the third inequality holds because π^μ and π^ν with $\pi^\mu_{ii'} = \frac{1}{I}\delta_{ii'}$ for all $i, i' \in \mathcal{I}$ and $\pi^\nu_{jj'} = \frac{1}{J}\delta_{jj'}$ for all $j, j' \in \mathcal{J}$, respectively, are feasible transportation plans. Finally, the last inequality follows from our assumption that $\|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|_\infty \le \varepsilon$ and $\|\boldsymbol{y}_j - \tilde{\boldsymbol{y}}_j\|_\infty \le \varepsilon$, which implies that

$$
\|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|^2 \le K\|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|_\infty^2 \le K\varepsilon^2 \quad \text{and} \quad \|\boldsymbol{y}_j - \tilde{\boldsymbol{y}}_j\|^2 \le K\|\boldsymbol{y}_j - \tilde{\boldsymbol{y}}_j\|_\infty^2 \le K\varepsilon^2
$$

for all $i \in \mathcal{I}$ and $j \in \mathcal{J}$. Substituting the above estimates back into (1.22) finally yields (1.21). □

We now address the approximate solution of OT problems with independent marginals.

Theorem 1.4.4 (Approximate Solutions of the OT Problem with Independent Marginals)**.** *Suppose that $\mu = \otimes_{k\in\mathcal{K}}\mu_k$ with $\mu_k = \sum_{l\in\mathcal{L}} \mu_k^l \delta_{x_k^l}$ for every $k \in \mathcal{K}$ and that $\nu_t = t\delta_{\boldsymbol{y}_1} + (1-t)\delta_{\boldsymbol{y}_2}$, and let $\varepsilon > 0$ be an error tolerance. If $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^2$ and if all probabilities and coordinates of the support points of μ and ν_t are representable as fractions of two integers with absolute values of at most U, then the OT distance between μ and ν_t can be computed to within an absolute error of at most ε by a dynamic programming-type algorithm using $\mathcal{O}(KL\log_2(KL) + K^6U^8/\varepsilon^4)$ arithmetic operations. The bit lengths of all numbers computed by this algorithm are polynomially bounded in K, L, $\log_2(U)$ and $\log_2(\frac{1}{\varepsilon})$.*

Proof. In order to approximate $W_c(\mu, \nu_t)$ to within an absolute accuracy of ε, we define $M = \lceil 8KU/\varepsilon \rceil$ and map all support points of μ and ν to the nearest lattice points in $\mathbb{Z}^K/M$ to construct perturbed probability distributions $\tilde{\mu}$ and $\tilde{\nu}_t$, respectively. Specifically, we construct $\tilde{x}_k^l$ by rounding x_k^l to the nearest point in $\mathbb{Z}/M$ for every $k \in \mathcal{K}$ and $l \in \mathcal{L}$. This requires $\mathcal{O}(KL)$ arithmetic operations. We then define the perturbed marginal distributions $\tilde{\mu}_k = \sum_{l\in\mathcal{L}} \mu_k^l \delta_{\tilde{x}_k^l}$ for all $k \in \mathcal{K}$ and set $\tilde{\mu} = \otimes_{k\in\mathcal{K}}\tilde{\mu}_k$. In addition, we denote by $\tilde{\boldsymbol{x}}_i$, $i \in \mathcal{I}$, the I different support points of $\tilde{\mu}$. Here, it is imperative to use the same orderings for the support points of μ and $\tilde{\mu}$, which implies that $\|\boldsymbol{x}_i - \tilde{\boldsymbol{x}}_i\|_\infty \leq \frac{1}{M} \leq \frac{\varepsilon}{8KU}$ for all $i \in \mathcal{I}$ thanks to the construction of $\tilde{\mu}$. We further construct $\tilde{y}_{j,k}$ by rounding $y_{j,k}$ to the nearest points in $\mathbb{Z}/M$ for every $j \in \mathcal{J} = \{1, 2\}$ and $k \in \mathcal{K}$, and we define $\tilde{\boldsymbol{y}}_j = (\tilde{y}_{j,1}, \ldots, \tilde{y}_{j,K})$ for all $j \in \mathcal{J}$. This construction requires $\mathcal{O}(K)$ arithmetic operations and guarantees that $\|\boldsymbol{y}_j - \tilde{\boldsymbol{y}}_j\|_\infty \leq \frac{1}{M} \leq \frac{\varepsilon}{8KU}$ for all $j \in \mathcal{J}$. Finally, we introduce the perturbed two-point distribution $\tilde{\nu}_t = t\delta_{\tilde{\boldsymbol{y}}_1} + (1-t)\delta_{\tilde{\boldsymbol{y}}_2}$. All support points of μ and ν have rational coordinates that are representable as fractions of two integers with absolute values at most U. Therefore, μ and ν are supported on $[-U, U]^K$. Similarly, as U and M are integers, which implies that U is an integer multiple of $\frac{1}{M}$, and as all support points of $\tilde{\mu}$ and $\tilde{\nu}$ are obtained by mapping the support points of μ and ν to the nearest lattice points in $\mathbb{Z}^K/M$, respectively, the perturbed distributions $\tilde{\mu}$ and $\tilde{\nu}$ must also be supported on $[-U, U]^K$. Lemma 1.4.3 therefore guarantees that $|W_c(\mu, \nu_t) - W_c(\tilde{\mu}, \tilde{\nu}_t)| \leq \varepsilon$.

In the remainder of the proof we will estimate the number of arithmetic operations needed to compute $W_c(\tilde{\mu}, \tilde{\nu}_t)$. Note first that the coordinates of all support points of $\tilde{\mu}$ and $\tilde{\nu}_t$ are fractions of integers with absolute values of at most $\tilde{U} = MU$. To see this, recall that $x_k^l = a_k^l / b_k^l$ for some $a_k^l \in \mathbb{Z}$ and $b_k^l \in \mathbb{N}$ with $|a_k^l|, |b_k^l| \leq U$. Using 'round' to denote the rounding operator that maps any real number to its nearest integer, we can express $\tilde{x}_k^l$ as $\tilde{a}_k^l / \tilde{b}_k^l$ with $\tilde{a}_k^l = \mathrm{round}(M x_k^l) \in \mathbb{Z}$ and $\tilde{b}_k^l = M \in \mathbb{N}$. By construction, we have $|\tilde{a}_k^l| \leq MU = \tilde{U}$ and $\tilde{b}_k^l = M \leq \tilde{U}$ for all $k \in \mathcal{K}$ and $l \in \mathcal{L}$. Similarly, one can show that $\tilde{y}_{j,k}$ is representable as a fraction of two integers with absolute values of at most $\tilde{U}$ for all $j \in \mathcal{J}$ and $k \in \mathcal{K}$. As $M \leq \tilde{U}$, $\tilde{\mu}$ and $\tilde{\nu}$ thus satisfy all assumptions of Corollary 3 with $\tilde{U}$ instead of U, respectively. From the proof of this corollary we may therefore conclude that $W_c(\tilde{\mu}, \tilde{\nu}_t)$ can be computed in $\mathcal{O}(KL \log_2(KL) + K^2 \tilde{U}^4)$ arithmetic operations using Algorithm 2. As $\tilde{U} = MU = \mathcal{O}(KU^2/\varepsilon)$, this establishes the claim about the number of arithmetic operations. From the definitions of $\tilde{U}$ and M and from the analysis of Algorithm 2 in Theorem 1.4.1, it is clear that the bit lengths of all numbers computed by the proposed procedure are indeed polynomially bounded in K, L, $\log_2(U)$ and $\log_2(\frac{1}{\varepsilon})$. This observation completes the proof. □

Theorem 1.4.4 shows that an ε-approximation of $W_c(\mu, \nu_t)$ can be computed with a number of arithmetic operations that grows only polynomially with K, L, U and $\frac{1}{\varepsilon}$ but exponentially with $\log_2(U)$ and $\log_2(\frac{1}{\varepsilon})$. By Remark 3 *(iii)*, approximations of $W_c(\mu, \nu_t)$ can therefore be computed in pseudo-polynomial time.

2. Semi-discrete Optimal Transport

2.1. Introduction

Optimal transport theory has a long and distinguished history in mathematics dating back to the seminal work of [Mon81b] and [Kan42]. While originally envisaged for applications in civil engineering, logistics and economics, optimal transport problems provide a natural framework for comparing probability measures and have therefore recently found numerous applications in statistics and machine learning. Indeed, the minimum cost of transforming a probability measure μ on $\mathcal{X}$ to some other probability measure ν on $\mathcal{Y}$ with respect to a prescribed cost function on $\mathcal{X} \times \mathcal{Y}$ can be viewed as a measure of distance between μ and ν. If $\mathcal{X} = \mathcal{Y}$ and the cost function coincides with (the p^{th} power of) a metric on $\mathcal{X} \times \mathcal{X}$, then the resulting optimal transport distance represents (the p^{th} power of) a Wasserstein metric on the space of probability measures over $\mathcal{X}$ [Vil08]. In the remainder of this chapter we distinguish *discrete*, *semi-discrete* and *continuous* optimal transport problems in which either both, only one or none of the two probability measures μ and ν are discrete, respectively.

In the wider context of machine learning, *discrete* optimal transport problems are nowadays routinely used, for example, in the analysis of mixture models [Kol+17; Ngu+13] as well as in image processing [AMJJ18; Fer+14; KR15; PR17; TPG16], computer vision and graphics [PW08; PW09; RTG00; Sol+14; Sol+15], data-driven bioengineering [Fey+17; Kun+18; Wan+10], clustering [Ho+17], dimensionality reduction [Caz+18; Fla+18; RCP16; Sch16; SC15], domain adaptation [Cou+16; Mur+18], distributionally robust optimization [EK18; Ngu+20], scenario reduction [HR07; Ruj+18], scenario generation [Pfl01; HP07], the assessment of the fairness properties of machine learning algorithms [Gor+19; Taş+20; Taş+21a] and signal processing [Tho+17].

The discrete optimal transport problem represents a tractable linear program that is susceptible to the network simplex algorithm [Orl97]. Alternatively, it can be addressed with dual ascent methods [BT97], the Hungarian algorithm for assignment problems [Kuh55] or customized auction algorithms [Ber81; Ber92]. The currently best known complexity bound for computing an *exact* solution is attained by modern interior-point algorithms. Indeed, if N denotes the number of atoms in μ or in ν, whichever is larger, then the discrete optimal transport problem can be solved in time[1] $\tilde{\mathcal{O}}(N^{2.5})$ with an interior point algorithm by [LS14]. The need to evaluate optimal transport distances between increasingly fine-grained histograms has also motivated efficient approximation schemes. [Bla+18] and [Qua19] show that an ϵ-optimal solution can be found in time $\mathcal{O}(N^2/\epsilon)$ by reducing the discrete optimal transport problem to a matrix scaling or a positive linear programming problem, which can be solved efficiently by a Newton-type algorithm. [JST19] describe a parallelizable primal-dual first-order method that achieves a similar convergence rate.

The tractability of the discrete optimal transport problem can be improved by adding an entropy regularizer to its objective function, which penalizes

[1] We use the soft-O notation $\tilde{\mathcal{O}}(\cdot)$ to hide polylogarithmic factors.

the entropy of the transportation plan for morphing μ into ν. When the weight of the regularizer grows, this problem reduces to the classical Schrödinger bridge problem of finding the most likely random evolution from μ to ν [Sch31]. Generic linear programs with entropic regularizers were first studied by [Fan92]. [CSM94] prove that the optimal values of these regularized problems converge exponentially fast to the optimal values of the corresponding unregularized problems as the regularization weight drops to zero. Non-asymptotic convergence rates for entropeq: semi-disc: dualy regularized linear programs are derived by [Wee18]. [Cut13] was the first to realize that entropic penalties are computationally attractive because they make the discrete optimal transport problem susceptible to a fast matrix scaling algorithm by [Sin67]. This insight has spurred widespread interest in machine learning and led to a host of new applications of optimal transport in color transfer [Chi+18], inverse problems [KR17; Adl+17], texture synthesis [Pey+17], the analysis of crowd evolutions [Pey15] and shape interpolation [Sol+15] to name a few. This surge of applications inspired in turn several new algorithms for the entropy regularized discrete optimal transport problem such as a greedy dual coordinate descent method also known as the Greenkhorn algorithm [AWR17; CK20; AG18]. [DGK18] and [LHJ19a] prove that both the Sinkhorn and the Greenkhorn algorithms are guaranteed to find an ϵ-optimal solution in time $\tilde{\mathcal{O}}(N^2/\epsilon^2)$. In practice, however, the Greenkhorn algorithm often outperforms the Sinkhorn algorithm [LHJ19a]. The runtime guarantee of both algorithms can be improved to $\tilde{\mathcal{O}}(N^{7/3}/\epsilon)$ via a randomization scheme [LHJ19b]. In addition, the regularized discrete optimal transport problem can be addressed by tailoring general-purpose optimization algorithms such as accelerated gradient descent algorithms [DGK18], iterative Bregman projections [Ben+15], quasi-Newton methods [BSR18] or stochastic average gradient descent algorithms [Gen+16]. While the original optimal transport problem induces sparse solutions, the entropy penalty forces the optimal transportation plan of the regularized optimal transport problem to be strictly positive and thus completely dense. In applications where the interpretability of the optimal transportation plan is important, the lack of sparsity could be undesirable; examples include color transfer [PKD07], domain adaptation [Cou+16] or ecological inference [Muz+17]. Hence, there is merit in exploring alternative regularization schemes that retain the attractive computational properties of the entropic regularizer but induce sparsity. Examples that have attracted significant interest include smooth convex regularization and Tikhonov regularization [DPR18; BSR18; Seg+18; ES18], Lasso regularization [LOG16], Tsallis entropy regularization [Muz+17] or group Lasso regularization [Cou+16].

Much like the discrete optimal transport problems, the significantly more challenging *semi-discrete* optimal transport problems emerge in numerous applications including variational inference [Amb+18], blue noise sampling [Qin+17], computational geometry [Lév15], image quantization [DG+12] or deep learning with generative adversarial networks [ACB17; GPC18; Gul+17]. Semi-discrete optimal transport problems are also used in fluid mechanics to simulate incompressible fluids [Goe+15].

Exact solutions of a semi-discrete optimal transport problem can be constructed by solving an incompressible Euler-type partial differential equation discovered by [Bre91]. Any optimal solution is known to partition the support of the non-discrete measure into cells corresponding to the atoms of the discrete measure [AHA98], and the resulting tessellation is usually referred to as a power diagram. [Mir15] uses this insight to solve Monge-Ampère equations with a damped Newton algorithm, and [KMT16] show that a closely related algorithm with a global linear convergence rate lends itself for the numerical solution of generic semi-discrete optimal transport problems. In addition, [Mér11] proposes a quasi-Newton algorithm for semi-discrete optimal transport, which improves a method due to [AHA98] by exploiting Llyod's algorithm to iteratively simplify the discrete measure. If the transportation cost is quadratic, [Bon13] relates the optimal transportation plan to the Knothe-Rosenblatt rearrangement for mapping μ to ν, which is very easy to compute.

As usual, regularization improves tractability. [Gen+16] show that the dual of a semi-discrete optimal transport problem with an entropic regularizer is susceptible to an averaged stochastic gradient descent algorithm that enjoys a convergence rate of $\mathcal{O}(1/\sqrt{T})$, T being the number of iterations. [ANWS22] show that the optimal value of the entropically regularized problem converges to the optimal value of the unregularized problem at a quadratic rate as the regularization weight drops to zero. Improved error bounds under stronger regularity conditions are derived by [Del21].

Continuous optimal transport problems constitute difficult variational problems involving infinitely many variables and constraints. [BB00] recast them as boundary value problems in fluid dynamics, and [PPO14] solve discretized versions of these reformulations using first-order methods. For a comprehensive survey of the interplay between partial differential equations and optimal transport we refer to [Eva97]. As nearly all numerical methods for partial differential equations suffer from a curse of dimensionality, current research focuses on solution schemes for *regularized* continuous optimal transport problems. For instance, [Gen+16] embed their duals into a reproducing kernel Hilbert space to obtain finite-dimensional optimization problems that can be solved with a stochastic gradient descent algorithm. [Seg+18] solve regularized continuous optimal transport problems by representing the transportation plan as a multilayer neural network. This approach results in finite-dimensional optimization problems that are non-convex and offer no approximation guarantees. However, it provides an effective means to compute approximate solutions in high dimensions. Indeed, the optimal value of the entropically regularized continuous optimal transport problem is known to converge to the optimal value of the unregularized problem at a linear rate as the regularization weight drops to zero [Chi+20; CT21; EMR15; Pal19]. Due to a lack of efficient algorithms, applications of continuous optimal transport problems are scarce in the extant literature. [PC19a] provide a comprehensive survey of numerous applications and solution methods for discrete, semi-discrete and continuous optimal transport problems.

This chapter focuses on semi-discrete optimal transport problems. Our main goal is to formally establish that these problems are computationally hard, to propose a unifying regularization scheme for improving their tractability and to develop efficient algorithms for solving the resulting regularized problems, assuming only that we have access to independent samples from the continuous probability measure μ. Our regularization scheme is based on the observation that any *dual* semi-discrete optimal transport problem maximizes the expectation of a piecewise affine function with N pieces, where the expectation is evaluated with respect to μ, and where N denotes the number of atoms of the discrete probability measure ν. We argue that this piecewise affine function can be interpreted as the optimal value of a *discrete choice problem*, which can be smoothed by adding random disturbances to the underlying utility values [Thu27; McF74]. As probabilistic discrete choice problems are routinely studied in economics and psychology, we can draw on a wealth of literature in choice theory to design various smooth (dual) optimal transport problems with favorable numerical properties. For maximal generality we will also study *semi-parametric* discrete choice models where the disturbance distribution is itself subject to uncertainty [NST09; Mis+14; FLW17; ALN18]. Specifically, we aim to evaluate the best-case (maximum) expected utility across a Fréchet ambiguity set containing all disturbance distributions with prescribed marginals. Such models can be addressed with customized methods from modern distributionally robust optimization [NST09]. For Fréchet ambiguity sets, we prove that smoothing the dual objective is equivalent to regularizing the primal objective of the semi-discrete optimal transport problem. The corresponding regularizer penalizes the discrepancy between the chosen transportation plan and the product measure $\mu \otimes \nu$ with respect to a divergence measure constructed from the marginal disturbance distributions. Connections between primal regularization and dual smoothing were previously recognized by [BSR18] and [PC20] in discrete optimal transport and by [Gen+16] in semi-discrete optimal transport. As they are constructed ad hoc or under a specific adversarial noise model, these existing regularization schemes lack the intuitive interpretation offered by discrete choice theory and emerge as special cases of our unifying scheme.

The key contributions of this chapter are summarized below.

i. We study the computational complexity of semi-discrete optimal transport problems. Specifically, we prove that computing the optimal transport distance between two probability measures μ and ν on the same Euclidean space is #P-hard even if only approximate solutions are sought and even if μ is the Lebesgue measure on the standard hypercube and ν is supported on merely two points.

ii. We propose a unifying framework for regularizing semi-discrete optimal transport problems by leveraging ideas from distributionally robust optimization and discrete choice theory [NST09; Mis+14; FLW17; ALN18]. Specifically, we perturb the transportation cost to every atom of the discrete measure ν with a random disturbance, and we assume that the vector of all disturbances is governed by an uncertain probability distribution from within a Fréchet ambiguity set that prescribes the marginal disturbance distributions. Solving the dual optimal transport problem under the least favorable disturbance distribution in the ambiguity set amounts to smoothing the dual and regularizing the primal objective function. We show that numerous known and new regularization schemes emerge as special cases of this framework, and we derive a priori approximation bounds for the resulting regularized optimal transport problems.

iii. We derive new convergence guarantees for an averaged stochastic gradient descent (SGD) algorithm that has only access to a *biased* stochastic gradient oracle. Specifically, we prove that this algorithm enjoys a convergence rate of $\mathcal{O}(1/\sqrt{T})$ for Lipschitz continuous and of $\mathcal{O}(1/T)$ for generalized self-concordant objective functions. We also show that this algorithm lends itself to solving the smooth dual optimal transport problems obtained from the proposed regularization scheme. When the smoothing is based on a semi-parametric discrete choice model with a Fréchet ambiguity set, the algorithm's convergence rate depends on the smoothness properties of the marginal noise distributions, and its per-iteration complexity depends on our ability to compute the optimal choice probabilities. We demonstrate that these choice probabilities can indeed be computed efficiently via bisection or sorting, and in special cases they are even available in closed form. As a byproduct, we show that our algorithm can improve the state-of-the-art $\mathcal{O}(1/\sqrt{T})$ convergence guarantee of [Gen+16] for the semi-discrete optimal transport problem with an *entropic* regularizer.

The rest of this chapter unfolds as follows. In Section 2.2 we study the computational complexity of semi-discrete optimal transport problems, and in Section 2.3 we develop our unifying regularization scheme. In Section 2.4 we analyze the convergence rate of an averaged SGD algorithm with a biased stochastic gradient oracle that can be used for solving smooth dual optimal transport problems, and in Section 2.5 we compare its empirical convergence behavior against the theoretical convergence guarantees.

Notation. We denote by $\|\cdot\|$ the 2-norm, by $[N] = \{1, \ldots, N\}$ the set of all integers up to $N \in \mathbb{N}$ and by $\Delta^d = \{\boldsymbol{x} \in \mathbb{R}^d_+ : \sum_{i=1}^d x_i = 1\}$ the probability simplex in $\mathbb{R}^d$. For a logical statement $\mathcal{E}$ we define $\mathbb{1}_{\mathcal{E}} = 1$ if $\mathcal{E}$ is true and $\mathbb{1}_{\mathcal{E}} = 0$ if $\mathcal{E}$ is false. For any closed set $\mathcal{X} \subseteq \mathbb{R}^d$ we define $\mathcal{M}(\mathcal{X})$ as the family of all Borel measures and $\mathcal{P}(\mathcal{X})$ as its subset of all Borel probability measures on $\mathcal{X}$. For $\mu \in \mathcal{P}(\mathcal{X})$, we denote by $\mathbb{E}_{\boldsymbol{x}\sim\mu}[\cdot]$ the expectation operator under μ and define $\mathcal{L}(\mathcal{X}, \mu)$ as the family of all μ-integrable functions $f : \mathcal{X} \to \mathbb{R}$, that is, $f \in \mathcal{L}(\mathcal{X}, \mu)$ if and only if $\int_{\mathcal{X}} |f(\boldsymbol{x})| \mu(\mathrm{d}\boldsymbol{x}) < \infty$. The Lipschitz modulus of a function $f : \mathbb{R}^d \to \mathbb{R}$ is defined as $\mathrm{lip}(f) = \sup_{\boldsymbol{x},\boldsymbol{x}'} \{|f(\boldsymbol{x}) - f(\boldsymbol{x}')| / \|\boldsymbol{x} - \boldsymbol{x}'\| : \boldsymbol{x} \neq \boldsymbol{x}'\}$. The convex conjugate of $f : \mathbb{R}^d \to [-\infty, +\infty]$ is the function $f^* : \mathbb{R}^d \to [-\infty, +\infty]$ defined through $f^*(\boldsymbol{y}) = \sup_{\boldsymbol{x}\in\mathbb{R}^d} \boldsymbol{y}^\top \boldsymbol{x} - f(\boldsymbol{x})$.

2.2. Hardness of Computing Optimal Transport Distances

If $\mathcal{X}$ and $\mathcal{Y}$ are closed subsets of finite-dimensional Euclidean spaces and $c : \mathcal{X} \times \mathcal{Y} \to [0, +\infty]$ is a lower-semicontinuous cost function, then the Monge-Kantorovich *optimal transport distance* between two probability measures $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ is defined as

$$W_c(\mu, \nu) = \min_{\pi \in \Pi(\mu,\nu)} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\pi} \left[c(\boldsymbol{x}, \boldsymbol{y})\right], \tag{2.1}$$

where $\Pi(\mu, \nu)$ denotes the family of all *couplings* of μ and ν, that is, the set of all probability measures on $\mathcal{X} \times \mathcal{Y}$ with marginals μ on $\mathcal{X}$ and ν on $\mathcal{Y}$. One can show that the minimum in (2.1) is always attained [Vil08, Theorem 4.1]. If $\mathcal{X} = \mathcal{Y}$ is a metric space with metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$ and the transportation cost is defined as $c(\boldsymbol{x}, \boldsymbol{y}) = d^p(\boldsymbol{x}, \boldsymbol{y})$ for some $p \geq 1$, then $W_c(\mu, \nu)^{1/p}$ is termed the p-th Wasserstein distance between μ and ν. The optimal transport problem (2.1) constitutes an infinite-dimensional linear program over measures and admits a strong dual linear program over functions [Vil08, Theorem 5.9].

Proposition 2.2.1 (Kantorovich duality)**.** *The optimal transport distance between $\mu \in \mathcal{P}(\mathcal{X})$ and $\nu \in \mathcal{P}(\mathcal{Y})$ admits the dual representation*

$$W_c(\mu, \nu) = \begin{cases} \sup & \mathbb{E}_{\boldsymbol{y}\sim\nu}\left[\phi(\boldsymbol{y})\right] - \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\psi(\boldsymbol{x})\right] \\ \text{s.t.} & \psi \in \mathcal{L}(\mathcal{X}, \mu),\ \phi \in \mathcal{L}(\mathcal{Y}, \nu) \\ & \phi(\boldsymbol{y}) - \psi(\boldsymbol{x}) \leq c(\boldsymbol{x}, \boldsymbol{y}) \quad \forall \boldsymbol{x} \in \mathcal{X},\ \boldsymbol{y} \in \mathcal{Y}. \end{cases} \tag{2.2}$$

The linear program (2.2) optimizes over the two *Kantorovich potentials* $\psi \in \mathcal{L}(\mathcal{X}, \mu)$ and $\phi \in \mathcal{L}(\mathcal{Y}, \nu)$, but it can be reformulated as the following non-linear program over a single potential function,

$$W_c(\mu, \nu) = \sup_{\phi \in \mathcal{L}(\mathcal{Y},\nu)} \mathbb{E}_{\boldsymbol{y}\sim\nu}\left[\phi(\boldsymbol{y})\right] - \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\phi_c(\boldsymbol{x})\right], \tag{2.3}$$

where $\phi_c : \mathcal{X} \to [-\infty, +\infty]$ is called the *c-transform* of ϕ and is defined through

$$\phi_c(\boldsymbol{x}) = \sup_{\boldsymbol{y} \in \mathcal{Y}} \ \phi(\boldsymbol{y}) - c(\boldsymbol{x}, \boldsymbol{y}) \qquad \forall \boldsymbol{x} \in \mathcal{X}, \tag{2.4}$$

see [Vil03, § 5] for details. The Kantorovich duality is the key enabling mechanism to study the computational complexity of the optimal transport problem (2.1).

Theorem 2.2.2 (Hardness of computing optimal transport distances)**.** *Computing $W_c(\mu, \nu)$ is #P-hard even if $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^p$ for some $p \geq 1$, μ is the Lebesgue measure on the standard hypercube $[0, 1]^d$, and ν is a discrete probability measure supported on only two points.*

To prove Theorem 2.2.2, we will show that computing the optimal transport distance $W_c(\mu, \nu)$ is at least as hard computing the volume of the knapsack polytope $P(\boldsymbol{w}, b) = \{\boldsymbol{x} \in [0,1]^d : \boldsymbol{w}^\top \boldsymbol{x} \leq b\}$ for a given $\boldsymbol{w} \in \mathbb{R}^d_+$ and $b \in \mathbb{R}_+$, which is known to be #P-hard [DF88, Theorem 1]. Specifically, we will leverage the following variant of this hardness result, which establishes that approximating the volume of the knapsack polytope $P(\boldsymbol{w}, b)$ to a sufficiently high accuracy is already #P-hard.

Lemma 2.2.3 ([HKW16, Lemma 1])**.** *Computing the volume of the knapsack polytope $P(\boldsymbol{w}, b)$ for a given $\boldsymbol{w} \in \mathbb{R}^d_+$ and $b \in \mathbb{R}_+$ to within an absolute accuracy of $\delta > 0$ is #P-hard whenever*

$$\delta < \frac{1}{2d!(\|\boldsymbol{w}\|_1 + 2)^d (d+1)^{d+1} \prod_{i=1}^d w_i}. \tag{2.5}$$

Fix now any knapsack polytope $P(\boldsymbol{w}, b)$ encoded by $\boldsymbol{w} \in \mathbb{R}^d_+$ and $b \in \mathbb{R}_+$. Without loss of generality, we may assume that $\boldsymbol{w} \neq \boldsymbol{0}$ and $b > 0$. Indeed, we are allowed to exclude $\boldsymbol{w} = \boldsymbol{0}$ because the volume of $P(\boldsymbol{0}, b)$ is trivially equal to 1. On the other hand, $b = 0$ can be excluded by applying a suitable rotation and translation, which are volume-preserving transformations. In the remainder, we denote by μ the Lebesgue measure on the standard hypercube $[0,1]^d$ and by $\nu_t = t\delta_{\boldsymbol{y}_1} + (1-t)\delta_{\boldsymbol{y}_2}$ a family of discrete probability measures with two atoms at $\boldsymbol{y}_1 = \boldsymbol{0}$ and $\boldsymbol{y}_2 = 2b\boldsymbol{w}/\|\boldsymbol{w}\|^2$, respectively, whose probabilities are parameterized by $t \in [0,1]$. The following preparatory lemma relates the volume of $P(\boldsymbol{w}, b)$ to the optimal transport problem (2.1) and is thus instrumental for the proof of Theorem 2.2.2.

Lemma 2.2.4. *If $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^p$ for some $p \geq 1$, then we have* $\mathrm{Vol}(P(\boldsymbol{w}, b)) = \mathrm{argmin}_{t \in [0,1]} W_c(\mu, \nu_t)$.

Proof. By the definition of the optimal transport distance in (2.1) and our choice of $c(\boldsymbol{x}, \boldsymbol{y})$, we have

$$\begin{aligned}
\min_{t \in [0,1]} W_c(\mu, \nu_t) &= \min_{t \in [0,1]} \min_{\pi \in \Pi(\mu, \nu_t)} \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \pi} \left[\|\boldsymbol{x} - \boldsymbol{y}\|^p\right] \\
&= \min_{t \in [0,1]} \left\{ \begin{array}{cl} \min\limits_{q_1, q_2 \in \mathcal{P}(\mathbb{R}^d)} & t \displaystyle\int_{\mathbb{R}^d} \|\boldsymbol{x} - \boldsymbol{y}_1\|^p q_1(\mathrm{d}\boldsymbol{x}) + \\ & \quad (1-t) \displaystyle\int_{\mathbb{R}^d} \|\boldsymbol{x} - \boldsymbol{y}_2\|^p \, q_2(\mathrm{d}\boldsymbol{x}) \\ \text{s.t.} & t \cdot q_1 + (1-t) \cdot q_2 = \mu, \end{array} \right.
\end{aligned}$$

where the second equality holds because any coupling π of μ and ν_t can be constructed from the marginal probability measure ν_t of $\boldsymbol{y}$ and the probability measures q_1 and q_2 of $\boldsymbol{x}$ conditional on $\boldsymbol{y} = \boldsymbol{y}_1$ and $\boldsymbol{y} = \boldsymbol{y}_2$, respectively, that is, we may write $\pi = t \cdot q_1 \otimes \delta_{\boldsymbol{y}_1} + (1-t) \cdot q_2 \otimes \delta_{\boldsymbol{y}_2}$. The constraint of the inner minimization problem ensures that the marginal probability measure of $\boldsymbol{x}$ under π coincides with μ. By applying the variable transformations $q_1 \leftarrow t \cdot q_1$ and $q_2 \leftarrow (1-t) \cdot q_2$ to eliminate all bilinear terms, we then obtain

$$\min_{t\in[0,1]} W_c(\mu, \nu_t) = \left\{ \begin{array}{cl} \displaystyle\min_{\substack{t\in[0,1] \\ q_1,q_2 \in \mathcal{M}(\mathbb{R}^d)}} & \displaystyle\int_{\mathbb{R}^d} \|\boldsymbol{x} - \boldsymbol{y}_1\|^p q_1(\mathrm{d}\boldsymbol{x}) + \int_{\mathbb{R}^d} \|\boldsymbol{x} - \boldsymbol{y}_2\|^p \, q_2(\mathrm{d}\boldsymbol{x}) \\ \text{s.t.} & \displaystyle\int_{\mathbb{R}^d} q_1(\mathrm{d}\boldsymbol{x}) = t \\ & \displaystyle\int_{\mathbb{R}^d} q_2(\mathrm{d}\boldsymbol{x}) = 1 - t \\ & q_1 + q_2 = \mu. \end{array} \right.$$

Observe next that the decision variable t and the two normalization constraints can be eliminated without affecting the optimal value of the resulting infinite-dimensional linear program because the Borel measures q_1 and q_2 are non-negative and because the constraint $q_1 + q_2 = \mu$ implies that $q_1(\mathbb{R}^d) + q_2(\mathbb{R}^d) = \mu(\mathbb{R}^d) = 1$. Thus, there always exists $t \in [0,1]$ such that $q_1(\mathbb{R}^d) = t$ and $q_2(\mathbb{R}^d) = 1 - t$. This reasoning implies that

$$\min_{t\in[0,1]} W_c(\mu, \nu_t) = \left\{ \begin{array}{cl} \displaystyle\min_{q_1,q_2 \in \mathcal{M}(\mathbb{R}^d)} & \displaystyle\int_{\mathbb{R}^d} \|\boldsymbol{x} - \boldsymbol{y}_1\|^p q_1(\mathrm{d}\boldsymbol{x}) + \int_{\mathbb{R}^d} \|\boldsymbol{x} - \boldsymbol{y}_2\|^p \, q_2(\mathrm{d}\boldsymbol{x}) \\ \text{s.t.} & q_1 + q_2 = \mu. \end{array} \right.$$

The constraint $q_1 + q_2 = \mu$ also implies that q_1 and q_2 are absolutely continuous with respect to μ, and thus

$$\begin{aligned} \min_{t\in[0,1]} W_c(\mu, \nu_t) &= \left\{ \begin{array}{cl} \displaystyle\min_{q_1,q_2 \in \mathcal{M}(\mathbb{R}^d)} & \displaystyle\int_{\mathbb{R}^d} \|\boldsymbol{x} - \boldsymbol{y}_1\|^p \frac{\mathrm{d}q_1}{\mathrm{d}\mu}(\boldsymbol{x}) + \\ & \quad \|\boldsymbol{x} - \boldsymbol{y}_2\|^p \; \tfrac{\mathrm{d}q_2}{\mathrm{d}\mu}(\boldsymbol{x}) \, \mu(\mathrm{d}\boldsymbol{x}) \\ \text{s.t.} & \displaystyle\frac{\mathrm{d}q_1}{\mathrm{d}\mu}(\boldsymbol{x}) + \frac{\mathrm{d}q_2}{\mathrm{d}\mu}(\boldsymbol{x}) = 1 \;\forall \boldsymbol{x} \in [0,1]^d \end{array} \right. \\ &= \int_{\mathbb{R}^d} \min\left\{ \|\boldsymbol{x} - \boldsymbol{y}_1\|^p, \|\boldsymbol{x} - \boldsymbol{y}_2\|^p \right\} \, \mu(\mathrm{d}\boldsymbol{x}), \end{aligned} \tag{2.6}$$

where the second equality holds because at optimality the Radon-Nikodym derivatives must satisfy

$$\frac{\mathrm{d}q_i}{\mathrm{d}\mu}(\boldsymbol{x}) = \left\{ \begin{array}{ll} 1 & \text{if } \|\boldsymbol{x} - \boldsymbol{y}_i\|^p \leq \|\boldsymbol{x} - \boldsymbol{y}_{3-i}\|^p \\ 0 & \text{otherwise} \end{array} \right.$$

for μ-almost every $\boldsymbol{x} \in \mathbb{R}^d$ and for every $i = 1, 2$.

In the second part of the proof we will demonstrate that the minimization problem $\min_{t\in[0,1]} W_c(\mu, \nu_t)$ is solved by $t^\star = \mathrm{Vol}(P(\boldsymbol{w}, b))$. By Proposition 2.2.1 and the definition of the c-transform, we first note that

$$\begin{aligned} W_c(\mu, \nu_{t^\star}) &= \max_{\phi \in \mathcal{L}(\mathbb{R}^d, \nu_{t^\star})} \mathbb{E}_{\boldsymbol{y}\sim\nu_{t^\star}}[\phi(\boldsymbol{y})] - \mathbb{E}_{\boldsymbol{x}\sim\mu}[\phi_c(\boldsymbol{x})] \\ &= \max_{\boldsymbol{\phi}\in\mathbb{R}^2} t^\star \cdot \phi_1 + (1 - t^\star) \cdot \phi_2 - \int_{\mathbb{R}^d} \max_{i=1,2} \{\phi_i - \|\boldsymbol{x} - \boldsymbol{y}_i\|^p\} \, \mu(\mathrm{d}\boldsymbol{x}) \\ &= \max_{\boldsymbol{\phi}\in\mathbb{R}^2} t^\star \cdot \phi_1 + (1 - t^\star) \cdot \phi_2 - \sum_{i=1}^{2} \int_{\mathcal{X}_i(\boldsymbol{\phi})} (\phi_i - \|\boldsymbol{x} - \boldsymbol{y_i}\|^p) \, \mu(\mathrm{d}\boldsymbol{x}), \end{aligned} \tag{2.7}$$

where

$$\mathcal{X}_i(\boldsymbol{\phi}) = \{\boldsymbol{x} \in \mathbb{R}^d : \phi_i - \|\boldsymbol{x} - \boldsymbol{y}_i\|^p \geq \phi_{3-i} - \left\|\boldsymbol{x} - \boldsymbol{y}_{3-i}\right\|^p\} \quad \forall i = 1, 2.$$

The second equality in (2.7) follows from the construction of $\nu_{t^\star}$ as a probability measure with only two atoms at the points $\boldsymbol{y}_i$ for $i = 1, 2$. Indeed, by fixing the corresponding function values $\phi_i = \phi(\boldsymbol{y}_i)$ for $i = 1, 2$, the expectation $\mathbb{E}_{\boldsymbol{y}\sim\nu_{t^\star}}[\phi(\boldsymbol{y})]$ simplifies to $t^\star \cdot \phi_1 + (1 - t^\star) \cdot \phi_2$, while the negative expectation $-\mathbb{E}_{\boldsymbol{x}\sim\mu}[\phi_c(\boldsymbol{x})]$ is maximized by setting $\phi(\boldsymbol{y})$ to a large negative constant for all $\boldsymbol{y} \notin \{\boldsymbol{y}_1, \boldsymbol{y}_2\}$, which implies that

$$\phi_c(\boldsymbol{x}) = \sup_{\boldsymbol{y}\in\mathbb{R}^d} \phi(\boldsymbol{y}) - \|\boldsymbol{x} - \boldsymbol{y}\|^p = \max_{i=1,2} \{\phi_i - \|\boldsymbol{x} - \boldsymbol{y}_i\|^p\} \quad \forall \boldsymbol{x} \in [0, 1]^d.$$

Next, we will prove that any $\boldsymbol{\phi}^\star \in \mathbb{R}^2$ with $\phi_1^\star = \phi_2^\star$ attains the maximum of the unconstrained convex optimization problem on the last line of (2.7). To see this, note that

$$\begin{aligned} \nabla_{\boldsymbol{\phi}} \left[\sum_{i=1}^{2} \int_{\mathcal{X}_i(\boldsymbol{\phi})} (\phi_i - \|\boldsymbol{x} - \boldsymbol{y}_i\|^p) \, \mu(\mathrm{d}\boldsymbol{x}) \right] &= \sum_{i=1}^{2} \int_{\mathcal{X}_i(\boldsymbol{\phi})} \nabla_{\boldsymbol{\phi}} (\phi_i - \|\boldsymbol{x} - \boldsymbol{y}_i\|^p) \, \mu(\mathrm{d}\boldsymbol{x}) \\ &= \begin{bmatrix} \mu(\mathcal{X}_1(\boldsymbol{\phi})) \\ \mu(\mathcal{X}_2(\boldsymbol{\phi})) \end{bmatrix} \end{aligned}$$

virtue of the Reynolds theorem. Thus, the first-order optimality condition[2] $t^\star = \mu(\mathcal{X}_1(\boldsymbol{\phi}))$ is necessary and sufficient for global optimality. Fix now any

[2] Note that the first-order condition $1 - t^\star = \mu(\mathcal{X}_2(\boldsymbol{\phi}))$ for ϕ_2 is redundant in view of the first-order condition $t^\star = \mu(\mathcal{X}_1(\boldsymbol{\phi}))$ for ϕ_1 because μ is the Lebesgue measure on $[0, 1]^d$, whereby $\mu(\mathcal{X}_1(\boldsymbol{\phi}) \cup \mathcal{X}_2(\boldsymbol{\phi})) = \mu(\mathcal{X}_1(\boldsymbol{\phi})) + \mu(\mathcal{X}_2(\boldsymbol{\phi})) = 1$.

$\boldsymbol{\phi}^\star \in \mathbb{R}^2$ with $\phi_1^\star = \phi_2^\star$ and observe that

$$\begin{aligned} t^\star = \mathrm{Vol}(P(\boldsymbol{w}, b)) =& \mu\left(\left\{\boldsymbol{x} \in \mathbb{R}^d : \boldsymbol{w}^\top \boldsymbol{x} \leq b\right\}\right) \\ =& \mu\left(\left\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|^2 \leq \|\boldsymbol{x} - 2b\boldsymbol{w}/\|\boldsymbol{w}\|^2\|^2\right\}\right) \\ =& \mu\left(\left\{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x} - \boldsymbol{y}_1\|^p \leq \|\boldsymbol{x} - \boldsymbol{y}_2\|^p\right\}\right) = \mu(\mathcal{X}_1(\boldsymbol{\phi}^\star)), \end{aligned}$$

where the first and second equalities follow from the definitions of $t^\star$ and the knapsack polytope $P(\boldsymbol{w}, b)$, respectively, the fourth equality holds because $\boldsymbol{y}_1 = \boldsymbol{0}$ and $\boldsymbol{y}_2 = 2b\boldsymbol{w}/\|\boldsymbol{w}\|^2$, and the fifth equality follows from the definition of $\mathcal{X}_1(\boldsymbol{\phi}^\star)$ and our assumption that $\phi_1^\star = \phi_2^\star$. This reasoning implies that $\boldsymbol{\phi}^\star$ attains indeed the maximum of the optimization problem on the last line of (2.7). Hence, we find

$$\begin{aligned} W_c(\mu, \nu_{t^\star}) &= t^\star \cdot \phi_1^\star + (1 - t^\star) \cdot \phi_2^\star - \sum_{i=1}^2 \int_{\mathcal{X}_i(\boldsymbol{\phi}^\star)} (\phi_i^\star - \|\boldsymbol{x} - \boldsymbol{y_i}\|^p)\, \mu(\mathrm{d}\boldsymbol{x}) \\ &= \sum_{i=1}^2 \int_{\mathcal{X}_i(\boldsymbol{\phi}^\star)} \|\boldsymbol{x} - \boldsymbol{y_i}\|^p\, \mu(\mathrm{d}\boldsymbol{x}) = \int_{\mathbb{R}^d} \min_{i=1,2} \{\|\boldsymbol{x} - \boldsymbol{y}_i\|^p\}\, \mu(\mathrm{d}\boldsymbol{x}) \\ &= \min_{t \in [0,1]} W_c(\mu, \nu_t), \end{aligned}$$

where the second equality holds because $\phi_1^\star = \phi_2^\star$, the third equality exploits the definition of $\mathcal{X}_1(\boldsymbol{\phi}^\star)$, and the fourth equality follows from (2.6). We may thus conclude that $t^\star = \mathrm{Vol}(P(\boldsymbol{w}, b))$ solves indeed the minimization problem $\min_{t \in [0,1]} W_c(\mu, \nu_t)$.

Using similar techniques, one can further prove that $\partial_t W_c(\mu, \nu_t)$ exists and is strictly increasing in t, which ensures that $W_c(\mu, \nu_t)$ is strictly convex in t and, in particular, that $t^\star$ is the unique solution of $\min_{t \in [0,1]} W_c(\mu, \nu_t)$. Details are omitted for brevity. □

Proof of Theorem 2.2.2. Lemma 2.2.4 applies under the assumptions of the theorem, and therefore the volume of the knapsack polytope $P(\boldsymbol{w}, b)$ coincides with the unique minimizer of

$$\min_{t \in [0,1]} W_c(\mu, \nu_t). \tag{2.8}$$

From the proof of Lemma 2.2.4 we know that the Wasserstein distance $W_c(\mu, \nu_t)$ is strictly convex in t, which implies that the minimization problem (2.8) constitutes a one-dimensional convex program with a unique minimizer. A near-optimal solution that approximates the exact minimizer to within an absolute accuracy $\delta = (6d!(\|\boldsymbol{w}\|_1 + 2)^d (d+1)^{d+1} \prod_{i=1}^d w_i)^{-1}$ can readily be computed with a binary search method such as Algorithm 5 described in Lemma 2.5.1 (i), which evaluates $g(t) = W_c(\mu, \nu_t)$ at exactly $2L = 2(\lceil \log_2(1/\delta) \rceil + 1)$ test points. Note that δ falls within the interval $(0, 1)$ and satisfies the strict inequality (2.5). Note also that L grows only polynomially with the bit length of $\boldsymbol{w}$ and b; see Appendix 2.5 for details. One readily verifies that all operations in Algorithm 5 except for the computation of $W_c(\mu, \nu_t)$ can be carried out in time polynomial in the bit length of $\boldsymbol{w}$ and b. Thus, if we could compute $W_c(\mu, \nu_t)$ in time polynomial in the bit length of $\boldsymbol{w}$, b and t, then we could efficiently compute the volume of the knapsack polytope $P(\boldsymbol{w}, b)$ to within accuracy δ, which is #P-hard by Lemma 2.2.3. We have thus constructed a polynomial-time Turing reduction from the #P-hard problem of (approximately) computing the volume of a knapsack polytope to computing the Wasserstein distance $W_c(\mu, \nu_t)$. By the definition of the class of #P-hard problems (see, *e.g.*, [VL90, Definition 1]), we may thus conclude that computing $W_c(\mu, \nu_t)$ is #P-hard. □

Corollary 4 (Hardness of computing approximate optimal transport distances)**.** *Computing $W_c(\mu, \nu)$ to within an absolute accuracy of*

$$\varepsilon = \frac{1}{4} \min_{l \in [2^L]} \left\{ |W_c(\mu, \nu_{t_l}) - W_c(\mu, \nu_{t_{l-1}})| : W_c(\mu, \nu_{t_l}) \neq W_c(\mu, \nu_{t_{l-1}}) \right\},$$

where $L = \lceil \log_2(1/\delta) \rceil + 1$, $\delta = (6d!(\|\boldsymbol{w}\|_1 + 2)^d (d+1)^{d+1} \prod_{i=1}^d w_i)^{-1}$ and $t_l = l/2^L$ for all $l = 0, \dots, 2^L$, is #P-hard even if $\mathcal{X} = \mathcal{Y} = \mathbb{R}^d$, $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|^p$ for some $p \geq 1$, μ is the Lebesgue measure on the standard hypercube $[0, 1]^d$, and ν is a discrete probability measure supported on only two points.

Proof. Assume that we have access to an inexact oracle that outputs an approximate optimal transport distance $\widetilde{W}_c(\mu, \nu_t)$ with $|\widetilde{W}_c(\mu, \nu_t) - W_c(\mu, \nu_t)| \leq \varepsilon$ for any fixed $t \in [0, 1]$. By Lemma 2.5.1 (ii), which applies thanks to the definition of ε, we can then find a 2δ-approximation for the unique minimizer of (2.8) using $2L$ oracle calls. Note that $\delta' = 2\delta$ falls within the interval $(0, 1)$ and satisfies the strict inequality (2.5). Recall also that L grows only polynomially with the bit length of $\boldsymbol{w}$ and b; see Appendix 2.5 for details. Thus, if we could compute $\widetilde{W}_c(\mu, \nu_t)$ in time polynomial in the bit length of $\boldsymbol{w}$, b and t, then we could efficiently compute the volume of the knapsack polytope $P(\boldsymbol{w}, b)$ to within accuracy δ', which is #P-hard by Lemma 2.2.3. Computing $W_c(\mu, \nu)$ to within an absolute accuracy of ε is therefore also #P-hard. □

The hardness of optimal transport established in Theorem 2.2.2 and Corollary 4 is predicated on the hardness of numerical integration. A popular technique to reduce the complexity of numerical integration is smoothing, whereby an initial (possibly discontinuous) integrand is approximated with a differentiable one [DKS13]. Smoothness is also a desired property of objective functions when designing scalable optimization algorithms [Bub15]. These observations prompt us to develop a systematic way to smooth the optimal transport problem that leads to efficient approximate numerical solution schemes.

2.3. Smooth Optimal Transport

The semi-discrete optimal transport problem evaluates the optimal transport distance (2.1) between an arbitrary probability measure μ supported on $\mathcal{X}$ and a discrete probability measure $\nu = \sum_{i=1}^N \nu_i \delta_{\boldsymbol{y_i}}$ with atoms $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N \in \mathcal{Y}$ and corresponding probabilities $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_N) \in \Delta^N$ for some $N \geq 2$. In the following, we define the *discrete c-transform* $\psi_c : \mathbb{R}^N \times \mathcal{X} \to [-\infty, +\infty)$ of $\boldsymbol{\phi} \in \mathbb{R}^N$ through

$$\psi_c(\boldsymbol{\phi}, \boldsymbol{x}) = \max_{i \in [N]} \phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i) \quad \forall \boldsymbol{x} \in \mathcal{X}. \tag{2.9}$$

Armed with the discrete c-transform, we can now reformulate the semi-discrete optimal transport problem as a finite-dimensional maximization problem over a single dual potential vector.

Lemma 2.3.1 (Discrete c-transform)**.** *The semi-discrete optimal transport problem is equivalent to*

$$W_c(\mu, \nu) = \sup_{\boldsymbol{\phi} \in \mathbb{R}^N} \boldsymbol{\nu}^\top \boldsymbol{\phi} - \mathbb{E}_{\boldsymbol{x} \sim \mu}[\psi_c(\boldsymbol{\phi}, \boldsymbol{x})]. \tag{2.10}$$

Proof. As $\nu = \sum_{i=1}^N \nu_i \delta_{\boldsymbol{y_i}}$ is discrete, the dual optimal transport problem (2.3) simplifies to

$$\begin{aligned} W_c(\mu, \nu) &= \sup_{\boldsymbol{\phi} \in \mathbb{R}^N} \sup_{\phi \in \mathcal{L}(\mathcal{Y}, \nu)} \left\{ \boldsymbol{\nu}^\top \boldsymbol{\phi} - \mathbb{E}_{\boldsymbol{x} \sim \mu} \left[\phi_c(\boldsymbol{x}) \right] \; : \; \phi(\boldsymbol{y}_i) = \phi_i \;\forall i \in [N] \right\} \\ &= \sup_{\boldsymbol{\phi} \in \mathbb{R}^N} \boldsymbol{\nu}^\top \boldsymbol{\phi} - \inf_{\phi \in \mathcal{L}(\mathcal{Y}, \nu)} \Big\{ \mathbb{E}_{\boldsymbol{x} \sim \mu} \left[\phi_c(\boldsymbol{x}) \right] \; : \; \phi(\boldsymbol{y}_i) = \phi_i \;\forall i \in [N] \Big\}. \end{aligned}$$

Using the definition of the standard c-transform, we can then recast the inner minimization problem as

$$\begin{aligned} &\inf_{\phi \in \mathcal{L}(\mathcal{Y}, \nu)} \left\{ \mathbb{E}_{\boldsymbol{x} \sim \mu} \left[\sup_{\boldsymbol{y} \in \mathcal{Y}} \phi(\boldsymbol{y}) - c(\boldsymbol{x}, \boldsymbol{y}) \right] \; : \; \phi(\boldsymbol{y}_i) = \phi_i \;\forall i \in [N] \right\} \\ &= \; \mathbb{E}_{\boldsymbol{x} \sim \mu} \left[\max_{i \in [N]} \{ \phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i) \} \right] \; = \; \mathbb{E}_{\boldsymbol{x} \sim \mu} \left[\psi_c(\boldsymbol{\phi}, \boldsymbol{x}) \right], \end{aligned}$$

where the first equality follows from setting $\phi(\boldsymbol{y}) = \underline{\phi}$ for all $\boldsymbol{y} \notin \{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\}$ and letting $\underline{\phi}$ tend to $-\infty$, while the second equality exploits the definition of the discrete c-transform. Thus, (2.10) follows. □

The discrete c-transform (2.9) can be viewed as the optimal value of a *discrete choice model*, where a utility-maximizing agent selects one of N mutually exclusive alternatives with utilities $\phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i)$, $i \in [N]$, respectively. Discrete choice models are routinely used for explaining the preferences of travelers selecting among different modes of transportation [BAL85], but they are also used for modeling the choice of residential location [McF78], the interests of end-users in engineering design [WC03] or the propensity of consumers to adopt new technologies [HM13].

In practice, the preferences of decision-makers and the attributes of the different choice alternatives are invariably subject to uncertainty, and it is impossible to specify a discrete choice model that reliably predicts the behavior of multiple individuals. Psychological theory thus models the utilities as random variables [Thu27], in which case the optimal choice becomes random, too. The theory as well as the econometric analysis of probabilistic discrete choice models were pioneered by [McF74].

The availability of a wealth of elegant theoretical results in discrete choice theory prompts us to add a random noise term to each deterministic utility value $\phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i)$ in (2.9). We will argue below that the expected value of the resulting maximal utility with respect to the noise distribution provides a smooth approximation for the c-transform $\psi_c(\boldsymbol{\phi}, \boldsymbol{x})$, which in turn leads to a smooth optimal transport problem that displays favorable numerical properties. For a comprehensive survey of additive random utility models in discrete choice theory we refer to [DM84] and [Dag14]. Generalized semi-parametric discrete choice models where the noise distribution is itself subject to uncertainty are studied by [NST09]. Using techniques from modern distributionally robust optimization, these models evaluate the best-case (maximum) expected utility across an ambiguity set of multivariate noise distributions. Semi-parametric discrete choice models are studied in the context of appointment scheduling [MRZ15], traffic management [AAN16] and product line pricing [Li+19].

We now define the *smooth* (*discrete*) *c-transform* as a best-case expected utility of the type studied in semi-parametric discrete choice theory, that is,

$$\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x}) = \sup_{\theta \in \Theta} \, \mathbb{E}_{\boldsymbol{z} \sim \theta} \left[\max_{i \in [N]} \phi_i - c(\boldsymbol{x}, \boldsymbol{y_i}) + z_i \right], \tag{2.11}$$

where $\boldsymbol{z}$ represents a random vector of perturbations that are independent of $\boldsymbol{x}$ and $\boldsymbol{y}$. Specifically, we assume that $\boldsymbol{z}$ is governed by a Borel probability measure θ from within some ambiguity set $\Theta \subseteq \mathcal{P}(\mathbb{R}^N)$. Note that if Θ is a singleton that contains only the Dirac measure at the origin of $\mathbb{R}^N$, then the smooth c-transform collapses to ordinary c-transform defined in (2.9), which is piecewise affine and thus non-smooth in $\boldsymbol{\phi}$. For many commonly used ambiguity sets, however, we will show below that the smooth c-transform is indeed differentiable in $\boldsymbol{\phi}$. In practice, the additive noise z_i in the transportation cost could originate, for example, from uncertainty about the position $\boldsymbol{y}_i$ of the i-th atom of the discrete distribution ν. This interpretation is justified if $c(\boldsymbol{x}, \boldsymbol{y})$ is approximately affine in $\boldsymbol{y}$ around the atoms $\boldsymbol{y}_i$, $i \in [N]$. The smooth c-transform gives rise to the following *smooth (semi-discrete) optimal transport problem* in dual form.

$$\overline{W}_c(\mu,\nu) = \sup_{\boldsymbol{\phi}\in\mathbb{R}^N} \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\boldsymbol{\nu}^\top\boldsymbol{\phi} - \overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x})\right] \tag{2.12}$$

Note that (2.12) is indeed obtained from the original dual optimal transport problem (2.10) by replacing the original c-transform $\psi_c(\boldsymbol{\phi},\boldsymbol{x})$ with the smooth c-transform $\overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x})$. As smooth functions are susceptible to efficient numerical integration, we expect that (2.12) is easier to solve than (2.10). A key insight of this work is that the smooth *dual* optimal transport problem (2.12) typically has a primal representation of the form

$$\min_{\pi\in\Pi(\mu,\nu)} \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\pi}\left[c(\boldsymbol{x},\boldsymbol{y})\right] + R_\Theta(\pi), \tag{2.13}$$

where $R_\Theta(\pi)$ can be viewed as a regularization term that penalizes the complexity of the transportation plan π. In the remainder of this section we will prove (2.13) and derive $R_\Theta(\pi)$ for different ambiguity sets Θ. We will see that this regularization term is often related to an f-divergence, where $f:\mathbb{R}_+\to\mathbb{R}\cup\{\infty\}$ constitutes a lower-semicontinuous convex function with $f(1)=0$. If τ and ρ are two Borel probability measures on a closed subset $\mathcal{Z}$ of a finite-dimensional Euclidean space, and if τ is absolutely continuous with respect to ρ, then the continuous f-divergence form τ to ρ is defined as $D_f(\tau\parallel\rho)=\int_{\mathcal{Z}} f(\mathrm{d}\tau/\mathrm{d}\rho(\boldsymbol{z}))\rho(\mathrm{d}\boldsymbol{z})$, where $\mathrm{d}\tau/\mathrm{d}\rho$ stands for the Radon-Nikodym derivative of τ with respect to ρ. By slight abuse of notation, if $\boldsymbol{\tau}$ and $\boldsymbol{\rho}$ are two probability vectors in Δ^N and if $\boldsymbol{\rho}>\boldsymbol{0}$, then the discrete f-divergence form $\boldsymbol{\tau}$ to $\boldsymbol{\rho}$ is defined as $D_f(\boldsymbol{\tau}\parallel\boldsymbol{\rho})=\sum_{i=1}^N f(\tau_i/\rho_i)\rho_i$. The correct interpretation of D_f is usually clear from the context.

The following lemma shows that the smooth optimal transport problem (2.13) equipped with an f-divergence regularization term is equivalent to a finite-dimensional convex minimization problem. This result will be instrumental to prove the equivalence of (2.12) and (2.13) for different ambiguity sets Θ.

Lemma 2.3.2 (Strong duality)**.** *If $\boldsymbol{\eta}\in\Delta^N$ with $\boldsymbol{\eta}>\boldsymbol{0}$ and $\eta=\sum_{i=1}^N \eta_i\delta_{\boldsymbol{y}_i}$ is a discrete probability measure on $\mathcal{Y}$, then problem (2.13) with regularization term $R_\Theta(\pi)=D_f(\pi\|\mu\otimes\eta)$ is equivalent to*

$$\sup_{\boldsymbol{\phi}\in\mathbb{R}^N} \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\min_{\boldsymbol{p}\in\Delta^N}\sum_{i=1}^N \phi_i\nu_i - (\phi_i - c(\boldsymbol{x},\boldsymbol{y_i}))p_i + D_f(\boldsymbol{p}\parallel\boldsymbol{\eta})\right]. \tag{2.14}$$

Proof of Lemma 2.3.2. If $\mathbb{E}_{\boldsymbol{x}\sim\mu}[c(\boldsymbol{x},\boldsymbol{y}_i)]=\infty$ for some $i\in[N]$, then both (2.13) and (2.14) evaluate to infinity, and the claim holds trivially. In the remainder of the proof we may thus assume without loss of generality that $\mathbb{E}_{\boldsymbol{x}\sim\mu}[c(\boldsymbol{x},\boldsymbol{y}_i)]<\infty$ for all $i\in[N]$. Using [RW09, Theorem 14.6] to interchange the minimization over $\boldsymbol{p}$ with the expectation over $\boldsymbol{x}$, problem (2.14) can first be reformulated as

$$\begin{array}{lll} \sup\limits_{\boldsymbol{\phi}\in\mathbb{R}^N} & \min\limits_{\boldsymbol{p}\in\mathcal{L}_\infty^N(\mathcal{X},\mu)} & \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\sum_{i=1}^N \phi_i\nu_i - (\phi_i - c(\boldsymbol{x},\boldsymbol{y_i}))p_i(\boldsymbol{x}) + D_f(\boldsymbol{p}(\boldsymbol{x})\|\boldsymbol{\eta})\right] \\ & \text{s.t.} & \boldsymbol{p}(\boldsymbol{x})\in\Delta^N \quad \mu\text{-a.s.}, \end{array}$$

where $\mathcal{L}_\infty^N(\mathcal{X},\mu)$ denotes the Banach space of all Borel-measurable functions from $\mathcal{X}$ to $\mathbb{R}^N$ that are essentially bounded with respect to μ. Interchanging the supremum over $\boldsymbol{\phi}$ with the minimum over $\boldsymbol{p}$ and evaluating the resulting unconstrained linear program over $\boldsymbol{\phi}$ in closed form then yields the dual problem

$$\begin{array}{ll} \min\limits_{\boldsymbol{p}\in\mathcal{L}_\infty^N(\mathcal{X},\mu)} & \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\sum_{i=1}^N c(\boldsymbol{x},\boldsymbol{y_i})p_i(\boldsymbol{x}) + D_f(\boldsymbol{p}(\boldsymbol{x})\,\|\,\boldsymbol{\eta})\right] \\ \text{s.t.} & \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\boldsymbol{p}(\boldsymbol{x})\right] = \boldsymbol{\nu}, \quad \boldsymbol{p}(\boldsymbol{x})\in\Delta^N \quad \mu\text{-a.s.} \end{array} \tag{2.15}$$

Strong duality holds for the following reasons. As c and f are lower-semicontinuous and c is non-negative, we may proceed as in [Sha17, § 3.2] to show that the dual objective function is weakly* lower semicontinuous in $\boldsymbol{p}$. Similarly, as Δ^N is compact, one can use the Banach-Alaoglu theorem to show that the dual feasible set is weakly* compact. Finally, as f is real-valued and $\mathbb{E}_{\boldsymbol{x}\sim\mu}[c(\boldsymbol{x},\boldsymbol{y}_i)]<\infty$ for all $i\in[N]$, the constant solution $\boldsymbol{p}(\boldsymbol{x})=\boldsymbol{\nu}$ is dual feasible for all $\boldsymbol{\nu}\in\Delta^N$. Thus, the dual problem is solvable and has a finite optimal value. This argument remains valid if we add a perturbation $\boldsymbol{\delta}\in H=\{\boldsymbol{\delta}'\in\mathbb{R}^N:\sum_{i=1}^N\delta_i'=0\}$ to the right hand side vector $\boldsymbol{\nu}$ as long as $\boldsymbol{\delta}>-\boldsymbol{\nu}$. The optimal value of the perturbed dual problem is thus pointwise finite as well as convex and—consequently—continuous and locally bounded in $\boldsymbol{\delta}$ at the origin of H. As $\boldsymbol{\nu}>\boldsymbol{0}$, strong duality therefore follows from [Roc74a, Theorem 17 (a)].

Any dual feasible solution $\boldsymbol{p}\in\mathcal{L}_\infty^N(\mathcal{X},\mu)$ gives rise to a Borel probability measure $\pi\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ defined through $\pi(\boldsymbol{y}\in\mathcal{B})=\nu(\boldsymbol{y}\in\mathcal{B})$ for all Borel sets $\mathcal{B}\subseteq\mathcal{Y}$ and $\pi(\boldsymbol{x}\in\mathcal{A}|\boldsymbol{y}=\boldsymbol{y}_i)=\int_{\mathcal{A}}p_i(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x})/\nu_i$ for all Borel sets $\mathcal{A}\subseteq\mathcal{X}$ and $i\in[N]$. This follows from the law of total probability, whereby the joint distribution of $\boldsymbol{x}$ and $\boldsymbol{y}$ is uniquely determined if we specify the marginal distribution of $\boldsymbol{y}$ and the conditional distribution of $\boldsymbol{x}$ given $\boldsymbol{y}=\boldsymbol{y}_i$ for every $i\in[N]$. By construction, the marginal distributions of $\boldsymbol{x}$ and $\boldsymbol{y}$ under π are determined by μ and ν, respectively. Indeed, note that for any Borel set $\mathcal{A}\subseteq\mathcal{X}$ we have

$$\pi(\boldsymbol{x}\in\mathcal{A})=\sum_{i=1}^N\pi(\boldsymbol{x}\in\mathcal{A}|\boldsymbol{y}=\boldsymbol{y}_i)\cdot\pi(\boldsymbol{y}=\boldsymbol{y}_i)=\sum_{i=1}^N\pi(\boldsymbol{x}\in\mathcal{A}|\boldsymbol{y}=\boldsymbol{y}_i)\cdot\nu_i$$

$$= \sum_{i=1}^N \int_{\mathcal{A}} p_i(\boldsymbol{x})\mu(\mathrm{d}\boldsymbol{x}) = \int_{\mathcal{A}} \mu(\mathrm{d}\boldsymbol{x}) = \mu(\boldsymbol{x} \in \mathcal{A}),$$

where the first equality follows from the law of total probability, the second and the third equalities both exploit the construction of π, and the fourth equality holds because $\boldsymbol{p}(\boldsymbol{x}) \in \Delta^N$ μ-almost surely due to dual feasibility. This reasoning implies that π constitutes a coupling of μ and ν (that is, $\pi \in \Pi(\mu, \nu)$) and is thus feasible in (2.13). Conversely, any $\pi \in \Pi(\mu, \nu)$ gives rise to a function $\boldsymbol{p} \in \mathcal{L}_\infty^N(\mathcal{X}, \mu)$ defined through

$$p_i(\boldsymbol{x}) = \nu_i \cdot \frac{\mathrm{d}\pi}{\mathrm{d}(\mu \otimes \nu)}(\boldsymbol{x}, \boldsymbol{y}_i) \quad \forall i \in [N].$$

By the properties of the Randon-Nikodym derivative, we have $p_i(\boldsymbol{x}) \geq 0$ μ-almost surely for all $i \in [N]$. In addition, for any Borel set $\mathcal{A} \subseteq \mathcal{X}$ we have

$$\begin{aligned}\int_{\mathcal{A}} \sum_{i=1}^N p_i(\boldsymbol{x})\, \mu(\mathrm{d}\boldsymbol{x}) &= \int_{\mathcal{A}} \sum_{i=1}^N \nu_i \cdot \frac{\mathrm{d}\pi}{\mathrm{d}(\mu \otimes \nu)}(\boldsymbol{x}, \boldsymbol{y}_i)\, \mu(\mathrm{d}\boldsymbol{x}) \\ &= \int_{\mathcal{A} \times \mathcal{Y}} \frac{\mathrm{d}\pi}{\mathrm{d}(\mu \otimes \nu)}(\boldsymbol{x}, \boldsymbol{y})\, (\mu \otimes \nu)(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \\ &= \int_{\mathcal{A} \times \mathcal{Y}} \pi(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) = \int_{\mathcal{A}} \mu(\mathrm{d}\boldsymbol{x}),\end{aligned}$$

where the second equality follows from Fubini's theorem and the definition of $\nu = \sum_{i=1}^N \nu_i \delta_{\boldsymbol{y}_i}$, while the fourth equality exploits that the marginal distribution of $\boldsymbol{x}$ under π is determined by μ. As the above identity holds for all Borel sets $\mathcal{A} \subseteq \mathcal{X}$, we find that $\sum_{i=1}^N p_i(\boldsymbol{x}) = 1$ μ-almost surely. Similarly, we have

$$\begin{aligned}\mathbb{E}_{\boldsymbol{x} \sim \mu}\left[p_i(\boldsymbol{x})\right] &= \int_{\mathcal{X}} \nu_i \cdot \frac{\mathrm{d}\pi}{\mathrm{d}(\mu \otimes \nu)}(\boldsymbol{x}, \boldsymbol{y}_i)\, \mu(\mathrm{d}\boldsymbol{x}) \\ &= \int_{\mathcal{X} \times \{\boldsymbol{y}_i\}} \frac{\mathrm{d}\pi}{\mathrm{d}(\mu \otimes \nu)}(\boldsymbol{x}, \boldsymbol{y})\, (\mu \otimes \nu)(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \\ &= \int_{\mathcal{X} \times \{\boldsymbol{y}_i\}} \pi(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) = \int_{\{\boldsymbol{y}_i\}} \nu(\mathrm{d}\boldsymbol{y}) = \nu_i\end{aligned}$$

for all $i \in [N]$. In summary, $\boldsymbol{p}$ is feasible in (2.15). Thus, we have shown that every probability measure π feasible in (2.13) induces a function $\boldsymbol{p}$ feasible in (2.15) and vice versa. We further find that the objective value of $\boldsymbol{p}$ in (2.15) coincides with the objective value of the corresponding π in (2.13). Specifically, we have

$$\mathbb{E}_{\boldsymbol{x} \sim \mu}\left[\sum_{i=1}^N c(\boldsymbol{x}, \boldsymbol{y_i})\, p_i(\boldsymbol{x}) + D_f(\boldsymbol{p}(\boldsymbol{x}) \| \boldsymbol{\eta})\right] = \int_{\mathcal{X}} \sum_{i=1}^N c(\boldsymbol{x}, \boldsymbol{y}_i) p_i(\boldsymbol{x})\, \mu(\mathrm{d}\boldsymbol{x}) +$$

$$\begin{aligned}
&\int_{\mathcal{X}} \sum_{i=1}^N f\left(\frac{p_i(\boldsymbol{x})}{\eta_i}\right) \eta_i \, \mu(\mathrm{d}\boldsymbol{x}) \\
&= \int_{\mathcal{X}} \sum_{i=1}^N c(\boldsymbol{x}, \boldsymbol{y}_i) \cdot \nu_i \cdot \frac{\mathrm{d}\pi}{\mathrm{d}(\mu \otimes \nu)}(\boldsymbol{x}, \boldsymbol{y}_i) \, \mu(\mathrm{d}\boldsymbol{x}) + \\
&\qquad \int_{\mathcal{X}} \sum_{i=1}^N f\left(\frac{\nu_i}{\eta_i} \cdot \frac{\mathrm{d}\pi}{\mathrm{d}(\mu \otimes \nu)}(\boldsymbol{x}, \boldsymbol{y}_i)\right) \cdot \eta_i \, \mu(\mathrm{d}\boldsymbol{x}) \\
&= \int_{\mathcal{X} \times \mathcal{Y}} c(\boldsymbol{x}, \boldsymbol{y}) \frac{\mathrm{d}\pi}{\mathrm{d}(\mu \otimes \nu)}(\boldsymbol{x}, \boldsymbol{y}) \, (\mu \otimes \nu)(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) + \\
&\qquad \int_{\mathcal{X} \times \mathcal{Y}} f\left(\frac{\mathrm{d}\pi}{\mathrm{d}(\mu \otimes \eta)}(\boldsymbol{x}, \boldsymbol{y})\right) (\mu \otimes \eta)(\mathrm{d}\boldsymbol{x}, \mathrm{d}\boldsymbol{y}) \\
&= \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y}) \sim \pi} \left[c(\boldsymbol{x}, \boldsymbol{y})\right] + D_f(\pi \| \mu \otimes \eta),
\end{aligned}$$

where the first equality exploits the definition of the discrete f-divergence, the second equality expresses the function $\boldsymbol{p}$ in terms of the corresponding probability measure π, the third equality follows from Fubini's theorem and uses the definitions $\nu = \sum_{i=1}^N \nu_i \delta_{\boldsymbol{y}_i}$ and $\eta = \sum_{i=1}^N \eta_i \delta_{\boldsymbol{y}_i}$, and the fourth equality follows from the definition of the continuous f-divergence. In summary, we have thus shown that (2.13) is equivalent to (2.15), which in turn is equivalent to (2.14). This observation completes the proof. □

Proposition 2.3.3 (Approximation bound)**.** *If $\boldsymbol{\eta} \in \Delta^N$ with $\boldsymbol{\eta} > \boldsymbol{0}$ and $\eta = \sum_{i=1}^N \eta_i \delta_{\boldsymbol{y}_i}$ is a discrete probability measure on $\mathcal{Y}$, then problem (2.13) with regularization term $R_\Theta(\pi) = D_f(\pi \| \mu \otimes \eta)$ satisfies*

$$|\overline{W}_c(\mu, \nu) - W_c(\mu, \nu)| \leq \max \left\{ \left| \min_{\boldsymbol{p} \in \Delta^N} D_f(\boldsymbol{p} \| \boldsymbol{\eta}) \right|, \left| \max_{i \in [N]} \left\{ f\left(\frac{1}{\eta_i}\right) \eta_i + f(0) \sum_{k \neq i} \eta_k \right\} \right| \right\}.$$

Proof. By Lemma 2.3.2, problem (2.13) is equivalent to (2.14). Note that the inner optimization problem in (2.14) can be viewed as an f-divergence regularized linear program with optimal value $\boldsymbol{\nu}^\top \boldsymbol{\phi} - \ell(\boldsymbol{\phi}, \boldsymbol{x})$, where

$$\ell(\boldsymbol{\phi}, \boldsymbol{x}) = \max_{\boldsymbol{p} \in \Delta^N} \sum_{i=1}^N (\phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i)) p_i - D_f(\boldsymbol{p} \| \boldsymbol{\eta}).$$

Bounding $D_f(\boldsymbol{p} \| \boldsymbol{\eta})$ by its minimum and its maximum over $\boldsymbol{p} \in \Delta^N$ then yields the estimates

$$\psi_c(\boldsymbol{\phi}, \boldsymbol{x}) - \max_{\boldsymbol{p} \in \Delta^N} D_f(\boldsymbol{p} \| \boldsymbol{\eta}) \leq \ell(\boldsymbol{\phi}, \boldsymbol{x}) \leq \psi_c(\boldsymbol{\phi}, \boldsymbol{x}) - \min_{\boldsymbol{p} \in \Delta^N} D_f(\boldsymbol{p} \| \boldsymbol{\eta}). \tag{2.16}$$

Here, $\psi_c(\boldsymbol{\phi}, \boldsymbol{x})$ stands as usual for the discrete c-transform defined in (2.9), which can be represented as

$$\psi_c(\boldsymbol{\phi}, \boldsymbol{x}) = \max_{\boldsymbol{p} \in \Delta^N} \sum_{i=1}^{N} (\phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i)) p_i. \tag{2.17}$$

Multiplying (2.16) by -1, adding $\boldsymbol{\nu}^\top \boldsymbol{\phi}$, averaging over $\boldsymbol{x}$ using the probability measure μ and maximizing over $\boldsymbol{\phi} \in \mathbb{R}^N$ further implies via (2.10) and (2.14) that

$$W_c(\mu, \nu) + \min_{\boldsymbol{p} \in \Delta^N} D_f(\boldsymbol{p} \| \boldsymbol{\eta}) \leq \overline{W}_c(\mu, \nu) \leq W_c(\mu, \nu) + \max_{\boldsymbol{p} \in \Delta^N} D_f(\boldsymbol{p} \| \boldsymbol{\eta}). \tag{2.18}$$

As $D_f(\boldsymbol{p} \| \boldsymbol{\eta})$ is convex in $\boldsymbol{p}$, its maximum is attained at a vertex of Δ^N [Hof81, Theorem 1], that is,

$$\max_{\boldsymbol{p} \in \Delta^N} D_f(\boldsymbol{p} \| \boldsymbol{\eta}) = \max_{i \in [N]} \left\{ f\left(\frac{1}{\eta_i}\right) \eta_i + f(0) \sum_{k \neq i} \eta_k \right\}.$$

The claim then follows by substituting the above formula into (2.18) and rearranging terms. □

In the following we discuss three different classes of ambiguity sets Θ for which the dual smooth optimal transport problem (2.12) is indeed equivalent to the primal reguarized optimal transport problem (2.13).

2.3.1. Generalized Extreme Value Distributions

Assume first that the ambiguity set Θ represents a singleton that accommodates only one single Borel probability measure θ on $\mathbb{R}^N$ defined through

$$\theta(\boldsymbol{z} \leq \boldsymbol{s}) = \exp\left(-G\left(\exp(-s_1), \ldots, \exp(-s_N)\right)\right) \quad \forall \boldsymbol{s} \in \mathbb{R}^N, \tag{2.19}$$

where $G : \mathbb{R}^N \to \mathbb{R}_+$ is a smooth generating function with the following properties. First, G is homogeneous of degree $1/\lambda$ for some $\lambda > 0$, that is, for any $\alpha \neq 0$ and $\boldsymbol{s} \in \mathbb{R}^N$ we have $G(\alpha \boldsymbol{s}) = \alpha^{1/\lambda} G(\boldsymbol{s})$. In addition, $G(\boldsymbol{s})$ tends to infinity as s_i grows for any $i \in [N]$. Finally, the partial derivative of G with respect to k distinct arguments is non-negative if k is odd and non-positive if k is even. These properties ensure that the noise vector $\boldsymbol{z}$ follows a generalized extreme value distribution in the sense of [Tra09, § 4.1].

Proposition 2.3.4 (Entropic regularization)**.** *Assume that Θ is a singleton ambiguity set that contains only a generalized extreme value distribution with $G(\boldsymbol{s}) = \exp(-e) N \sum_{i=1}^N \eta_i s_i^{1/\lambda}$ for some $\lambda > 0$ and $\boldsymbol{\eta} \in \Delta^N$, $\boldsymbol{\eta} > \boldsymbol{0}$, where e stands for Euler's constant. Then, the components of $\boldsymbol{z}$ follow independent*

Gumbel distributions with means $\lambda \log(N\eta_i)$ and variances $\lambda^2\pi^2/6$ for all $i \in [N]$, while the smooth c-transform (2.11) reduces to the log-partition *function*

$$\overline{\psi}(\boldsymbol{\phi}, \boldsymbol{x}) = \lambda \log \left(\sum_{i=1}^{N} \eta_i \exp\left(\frac{\phi_i - c(\boldsymbol{x}, \boldsymbol{y_i})}{\lambda} \right) \right). \tag{2.20}$$

In addition, the smooth dual optimal transport problem (2.12) is equivalent to the regularized primal optimal transport problem (2.13) with $R_\Theta(\pi) = D_f(\pi\|\mu \otimes \eta)$, where $f(s) = \lambda s \log(s)$ and $\eta = \sum_{i=1}^N \eta_i \delta_{\boldsymbol{y}_i}$.

Note that the log-partition function (2.20) constitutes indeed a smooth approximation for the maximum function in the definition (2.9) of the discrete c-transform. As λ decreases, this approximation becomes increasingly accurate. It is also instructive to consider the special case where $\mu = \sum_{i=1}^M \mu_i \delta_{\boldsymbol{x}_i}$ is a discrete probability measure with atoms $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_M \in \mathcal{X}$ and corresponding vector of probabilities $\boldsymbol{\mu} \in \Delta^M$. In this case, any coupling $\pi \in \Pi(\mu, \nu)$ constitutes a discrete probability measure $\pi = \sum_{i=1}^M \sum_{j=1}^N \pi_{ij} \delta_{(\boldsymbol{x}_i, \boldsymbol{y}_j)}$ with matrix of probabilities $\boldsymbol{\pi} \in \Delta^{M \times N}$. If $f(x) = s\log(s)$, then the continuous f-divergence reduces to

$$\begin{aligned} D_f(\pi\|\mu \otimes \eta) &= \sum_{i=1}^M \sum_{j=1}^N \pi_{ij} \log(\pi_{ij}) - \sum_{i=1}^M \sum_{j=1}^N \pi_{ij} \log(\mu_i) - \sum_{i=1}^M \sum_{j=1}^N \pi_{ij} \log(\eta_j) \\ &= \sum_{i=1}^M \sum_{j=1}^N \pi_{ij} \log(\pi_{ij}) - \sum_{i=1}^M \mu_i \log(\mu_i) - \sum_{j=1}^N \nu_j \log(\eta_j), \end{aligned}$$

where the second equality holds because π is a coupling of μ and ν. Thus, $D_f(\pi\|\mu \otimes \eta)$ coincides with the negative entropy of the probability matrix $\boldsymbol{\pi}$ offset by a constant that is independent of $\boldsymbol{\pi}$. For $f(s) = s\log(s)$ the choice of $\boldsymbol{\eta}$ has therefore no impact on the minimizer of the smooth optimal transport problem (2.13), and we simply recover the celebrated entropic regularization proposed by [Cut13; Gen+16; RW18; PC19a] and [Cla+21].

Proof of Proposition 2.3.4. Substituting the explicit formula for the generating function G into (2.19) yields

$$\begin{aligned} \theta(\boldsymbol{z} \le \boldsymbol{s}) &= \exp\left(-\exp(-e) N \sum_{i=1}^N \eta_i \exp\left(-\frac{s_i}{\lambda} \right) \right) \\ &= \prod_{i=1}^N \exp\left(-\exp(-e) N \eta_i \exp\left(-\frac{s_i}{\lambda} \right) \right) \\ &= \prod_{i=1}^N \exp\left(-\exp\left(-\frac{s_i - \lambda(\log(N\eta_i) - e)}{\lambda} \right) \right), \end{aligned}$$

where e stands for Euler's constant. The components of the noise vector $\boldsymbol{z}$ are thus independent under θ, and z_i follows a Gumbel distribution with location parameter $\lambda(\log(N\eta_i) - e)$ and scale parameter λ for every $i \in [N]$. Therefore, z_i has mean $\lambda \log(N\eta_i)$ and variance $\lambda^2\pi^2/6$.

If the ambiguity set Θ contains only one single probability measure θ of the form (2.19), then Theorem 5.2 of [McF81] readily implies that the smooth c-transform (2.11) simplifies to

$$\overline{\psi}(\boldsymbol{\phi}, \boldsymbol{x}) = \lambda \log G\left(\exp(\phi_1 - c(\boldsymbol{x}, \boldsymbol{y}_1)), \ldots, \exp(\phi_N - c(\boldsymbol{x}, \boldsymbol{y}_N))\right) + \lambda e. \quad (2.21)$$

The closed-form expression for the smooth c-transform in (2.20) follows immediately by substituting the explicit formula for the generating function G into (2.21). One further verifies that (2.20) can be reformulated as

$$\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x}) = \max_{\boldsymbol{p} \in \Delta^N} \sum_{i=1}^N (\phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i))p_i - \lambda \sum_{i=1}^N p_i \log\left(\frac{p_i}{\eta_i}\right). \quad (2.22)$$

Indeed, solving the underlying Karush-Kuhn-Tucker conditions analytically shows that the optimal value of the nonlinear program (2.22) coincides with the smooth c-transform (2.20). In the special case where $\eta_i = 1/N$ for all $i \in [N]$, the equivalence of (2.20) and (2.22) has already been recognized by [ADPT88]. Substituting the representation (2.22) of the smooth c-transform into the dual smooth optimal transport problem (2.12) yields (2.14) with $f(s) = \lambda s \log(s)$. By Lemma 2.3.2, problem (2.12) is thus equivalent to the regularized primal optimal transport problem (2.13) with $R_\Theta(\pi) = D_f(\pi \| \mu \otimes \eta)$, where $\eta = \sum_{i=1}^N \eta_i \delta_{\boldsymbol{y}_i}$. □

2.3.2. Chebyshev Ambiguity Sets

Assume next that Θ constitutes a Chebyshev ambiguity set comprising all Borel probability measures on $\mathbb{R}^N$ with mean vector $\mathbf{0}$ and positive definite covariance matrix $\lambda\boldsymbol{\Sigma}$ for some $\boldsymbol{\Sigma} \succ \mathbf{0}$ and $\lambda > 0$. Formally, we thus set $\Theta = \{\theta \in \mathcal{P}(\mathbb{R}^N) : \mathbb{E}_\theta[\boldsymbol{z}] = \mathbf{0},\ \mathbb{E}_\theta[\boldsymbol{z}\boldsymbol{z}^\top] = \lambda\boldsymbol{\Sigma}\}$. In this case, [ALN18, Theorem 1] implies that the smooth c-transform (2.11) can be equivalently expressed as

$$\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x}) = \max_{\boldsymbol{p} \in \Delta^N} \sum_{i=1}^N (\phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i))p_i + \lambda \operatorname{tr}\left((\boldsymbol{\Sigma}^{1/2}(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top)\boldsymbol{\Sigma}^{1/2})^{1/2}\right), \quad (2.23)$$

where $\operatorname{diag}(\boldsymbol{p}) \in \mathbb{R}^{N\times N}$ represents the diagonal matrix with $\boldsymbol{p}$ on its main diagonal. Note that the maximum in (2.23) evaluates the convex conjugate of the extended real-valued regularization function

$$V(\boldsymbol{p}) = \begin{cases} -\lambda \operatorname{tr}\left((\boldsymbol{\Sigma}^{1/2}(\operatorname{diag}(\boldsymbol{p}) - \boldsymbol{p}\boldsymbol{p}^\top)\boldsymbol{\Sigma}^{1/2})^{1/2}\right) & \text{if } \boldsymbol{p} \in \Delta^N \\ \infty & \text{if } \boldsymbol{p} \notin \Delta^N \end{cases}$$

at the point $(\phi_i - c(\boldsymbol{x}, \boldsymbol{y_i}))_{i\in[N]}$. As $\boldsymbol{\Sigma} \succ \boldsymbol{0}$ and $\lambda > 0$, [ALN18, Theorem 3] implies that $V(\boldsymbol{p})$ is strongly convex over its effective domain Δ^N. By [RW09, Proposition 12.60], the smooth discrete c-transform $\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x})$ is therefore indeed differentiable in $\boldsymbol{\phi}$ for any fixed $\boldsymbol{x}$. It is further known that problem (2.23) admits an exact reformulation as a tractable semidefinite program; see [Mis+12, Proposition 1]. If $\boldsymbol{\Sigma} = \boldsymbol{I}$, then the regularization function $V(\boldsymbol{p})$ can be re-expressed in terms of a discrete f-divergence, which implies via Lemma 2.3.2 that the smooth optimal transport problem is equivalent to the original optimal transport problem regularized with a continuous f-divergence.

Proposition 2.3.5 (Chebyshev regularization)**.** *If Θ is the Chebyshev ambiguity set of all Borel probability measures with mean $\boldsymbol{0}$ and covariance matrix $\lambda\boldsymbol{I}$ with $\lambda > 0$, then the smooth c-transform (2.11) simplifies to*

$$\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x}) = \max_{\boldsymbol{p}\in\Delta^N} \sum_{i=1}^N (\phi_i - c(\boldsymbol{x}, \boldsymbol{y_i}))p_i + \lambda \sum_{i=1}^N \sqrt{p_i(1-p_i)}. \tag{2.24}$$

In addition, the smooth dual optimal transport problem (2.12) is equivalent to the regularized primal optimal transport problem (2.13) with $R_\Theta(\pi) = D_f(\pi\|\mu\otimes\eta) + \lambda\sqrt{N-1}$, where $\eta = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{y}_i}$ and

$$f(s) = \begin{cases} -\lambda\sqrt{s(N-s)} + \lambda s\sqrt{N-1} & \text{if } 0 \le s \le N \\ +\infty & \text{if } s > N. \end{cases} \tag{2.25}$$

Proof. The relation (2.24) follows directly from (2.23) by replacing $\boldsymbol{\Sigma}$ with $\boldsymbol{I}$. Next, one readily verifies that $-\sum_{i\in[N]}\sqrt{p_i(1-p_i)}$ can be re-expressed as the discrete f-divergence $D_f(\boldsymbol{p}\|\boldsymbol{\eta})$ from $\boldsymbol{p}$ to $\boldsymbol{\eta} = (\frac{1}{N}, \dots, \frac{1}{N})$, where $f(s) = -\lambda\sqrt{s(N-s)} + \lambda\sqrt{N-1}$. This implies that (2.24) is equivalent to

$$\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x}) = \max_{\boldsymbol{p}\in\Delta^N} \sum_{i=1}^N (\phi_i - c(\boldsymbol{x}, \boldsymbol{y_i}))p_i - D_f(\boldsymbol{p}\|\boldsymbol{\eta}).$$

Substituting the above representation of the smooth c-transform into the dual smooth optimal transport problem (2.12) yields (2.14) with

$$f(s) = -\lambda\sqrt{s(N-s)} + \lambda s\sqrt{N-1}.$$

By Lemma 2.3.2, (2.12) thus reduces to the regularized primal optimal transport problem (2.13) with $R_\Theta(\pi) = D_f(\pi\|\mu\otimes\eta)$, where $\eta = \frac{1}{N}\sum_{i=1}^N \delta_{\boldsymbol{y}_i}$. □

Note that the function $f(s)$ defined in (2.25) is indeed convex, lower-semi-continuous and satisfies $f(1) = 0$. Therefore, it induces a standard f-divergence. Proposition 2.3.5 can be generalized to arbitrary diagonal matrices $\boldsymbol{\Sigma}$, but the emerging f-divergences are rather intricate and not insightful. Hence, we do not show this generalization. We were not able to generalize Proposition 2.3.5 to non-diagonal matrices $\boldsymbol{\Sigma}$.

2.3.3. Marginal Ambiguity Sets

We now investigate the class of marginal ambiguity sets of the form

$$\Theta = \Big\{\theta \in \mathcal{P}(\mathbb{R}^N) \,:\, \theta(z_i \leq s) = F_i(s) \;\forall s \in \mathbb{R},\; \forall i \in [N]\Big\}, \tag{2.26}$$

where F_i stands for the cumulative distribution function of the uncertain disturbance z_i, $i \in [N]$. Marginal ambiguity sets completely specify the marginal distributions of the components of the random vector $\boldsymbol{z}$ but impose no restrictions on their dependence structure (*i.e.*, their copula). Sometimes marginal ambiguity sets are also referred to as Fréchet ambiguity sets [Fré51]. We will argue below that the marginal ambiguity sets explain most known as well as several new regularization methods for the optimal transport problem. In particular, they are more expressive than the extreme value distributions as well as the Chebyshev ambiguity sets in the sense that they induce a richer family of regularization terms. Below we denote by $F_i^{-1} : [0,1] \to \mathbb{R}$ the (left) quantile function corresponding to F_i, which is defined through

$$F_i^{-1}(t) = \inf\{s : F_i(s) \geq t\} \quad \forall t \in \mathbb{R}.$$

We first prove that if Θ constitutes a marginal ambiguity set, then the smooth c-transform (2.11) admits an equivalent reformulation as the optimal value of a finite convex program.

Proposition 2.3.6 (Smooth c-transform for marginal ambiguity sets)**.** *If Θ is a marginal ambiguity set of the form (2.26), and if the underlying cumulative distribution functions F_i, $i \in [N]$, are continuous, then the smooth c-transform (2.11) can be equivalently expressed as*

$$\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x}) = \max_{\boldsymbol{p} \in \Delta^N} \sum_{i=1}^N (\phi_i - c(\boldsymbol{x}, \boldsymbol{y_i}))p_i + \sum_{i=1}^N \int_{1-p_i}^1 F_i^{-1}(t)\mathrm{d}t \tag{2.27}$$

for all $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\phi} \in \mathbb{R}^N$. In addition, the smooth c-transform is convex and differentiable with respect to $\boldsymbol{\phi}$, and $\nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x})$ represents the unique solution of the convex maximization problem (2.27).

Recall that the smooth c-transform (2.11) can be viewed as the best-case utility of a semi-parametric discrete choice model. Thus, (2.27) follows from [NST09, Theorem 1]. To keep this chapter self-contained, we provide a new proof of Proposition 2.3.6, which exploits a natural connection between the smooth c-transform induced by a marginal ambiguity set and the conditional value-at-risk (CVaR).

Proof of Proposition 2.3.6. Throughout the proof we fix $\boldsymbol{x} \in \mathcal{X}$ and $\boldsymbol{\phi} \in \mathbb{R}^N$, and we introduce the nominal utility vector $\boldsymbol{u} \in \mathbb{R}^N$ with components $u_i =$

$\phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i)$ in order to simplify notation. In addition, it is useful to define the binary function $\boldsymbol{r} : \mathbb{R}^N \to \{0,1\}^N$ with components

$$r_i(\boldsymbol{z}) = \begin{cases} 1 & \text{if } i = \min \operatorname*{argmax}_{j\in[N]} \; u_j + z_j, \\ 0 & \text{otherwise.} \end{cases}$$

For any fixed $\theta \in \Theta$, we then have

$$\mathbb{E}_{\boldsymbol{z}\sim\theta}\Big[\max_{i\in[N]} u_i + z_i\Big] = \mathbb{E}_{\boldsymbol{z}\sim\theta}\Big[\sum_{i=1}^N (u_i + z_i) r_i(\boldsymbol{z})\Big] = \sum_{i=1}^N u_i p_i + \sum_{i=1}^N \mathbb{E}_{\boldsymbol{z}\sim\theta}\left[z_i q_i(z_i)\right],$$

where $p_i = \mathbb{E}_{\boldsymbol{z}\sim\theta}[r_i(\boldsymbol{z})]$ and $q_i(z_i) = \mathbb{E}_{\boldsymbol{z}\sim\theta}[r_i(\boldsymbol{z})|z_i]$ almost surely with respect to θ. From now on we denote by θ_i the marginal probability distribution of the random variable z_i under θ. As θ belongs to a marginal ambiguity set of the form (2.26), we thus have $\theta_i(z_i \le s) = F_i(s)$ for all $s \in \mathbb{R}$, that is, θ_i is uniquely determined by the cumulative distribution function F_i. The above reasoning then implies that

$$\begin{aligned} \overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x}) &= \sup_{\theta\in\Theta} \; \mathbb{E}_{\boldsymbol{z}\sim\theta}\Big[\max_{i\in[N]} u_i + z_i\Big] \\ &= \left\{ \begin{array}{lll} \sup & \displaystyle\sum_{i=1}^N u_i p_i + \sum_{i=1}^N \mathbb{E}_{\boldsymbol{z}\sim\theta}\left[z_i q_i(z_i)\right] & \\ \text{s.t.} & \theta \in \Theta, \; \boldsymbol{p} \in \Delta^N, \; \boldsymbol{q} \in \mathcal{L}^N(\mathbb{R}) & \\ & \mathbb{E}_{\boldsymbol{z}\sim\theta}\left[r_i(\boldsymbol{z})\right] = p_i & \forall i \in [N] \\ & \mathbb{E}_{\boldsymbol{z}\sim\theta}[r_i(\boldsymbol{z})|z_i] = q_i(z_i) \quad \theta\text{-a.s.} & \forall i \in [N] \end{array} \right. \qquad (2.28) \\ &\le \left\{ \begin{array}{lll} \sup & \displaystyle\sum_{i=1}^N u_i p_i + \sum_{i=1}^N \mathbb{E}_{z_i\sim\theta_i}\left[z_i q_i(z_i)\right] & \\ \text{s.t.} & \boldsymbol{p} \in \Delta^N, \; \boldsymbol{q} \in \mathcal{L}^N(\mathbb{R}) & \\ & \mathbb{E}_{z_i\sim\theta_i}\left[q_i(z_i)\right] = p_i & \forall i \in [N] \\ & 0 \le q_i(z_i) \le 1 \quad \theta_i\text{-a.s.} & \forall i \in [N]. \end{array} \right. \qquad (2.29) \end{aligned}$$

The inequality can be justified as follows. One may first add the redundant expectation constraints $p_i = \mathbb{E}_{z_i\sim\theta}[q_i(z_i)]$ and the redundant θ_i-almost sure constraints $0 \le q_i(z_i) \le 1$ to the maximization problem over θ, $\boldsymbol{p}$ and $\boldsymbol{q}$ without affecting the problem's optimal value. Next, one may remove the constraints that express p_i and $q_i(z_i)$ in terms of $r_i(\boldsymbol{z})$. The resulting relaxation provides an upper bound on the original maximization problem. Note that all remaining expectation operators involve integrands that depend on $\boldsymbol{z}$ only through z_i for some $i \in [N]$, and therefore the expectations with respect to the joint probability measure θ can all be simplified to expectations with respect to one of the marginal probability measures θ_i. As neither the objective nor the constraints of the resulting problem depend on θ, we may finally remove θ from the list of decision variables without affecting the problem's optimal value. For any fixed $\boldsymbol{p} \in \Delta^N$, the upper bounding problem (2.29) gives rise the following N subproblems indexed by $i \in [N]$.

$$\sup_{q_i \in \mathcal{L}(\mathbb{R})} \left\{ \mathbb{E}_{z_i \sim \theta_i}\left[z_i q_i(z_i)\right] : \mathbb{E}_{z_i \sim \theta_i}\left[q_i(z_i)\right] = p_i,\ 0 \le q_i(z_i) \le 1 \ \theta_i\text{-a.s.} \right\} \tag{2.30a}$$

If $p_i > 0$, the optimization problem (2.30a) over the functions $q_i \in \mathcal{L}(\mathbb{R})$ can be recast as an optimization problem over probability measures $\tilde{\theta}_i \in \mathcal{P}(\mathbb{R})$ that are absolutely continuous with respect to θ_i,

$$\sup_{\tilde{\theta}_i \in \mathcal{P}(\mathbb{R})} \left\{ p_i \, \mathbb{E}_{z_i \sim \tilde{\theta}_i}\left[z_i\right] : \frac{\mathrm{d}\tilde{\theta}_i}{\mathrm{d}\theta_i}(z_i) \le \frac{1}{p_i} \ \theta_i\text{-a.s.} \right\}, \tag{2.30b}$$

where $\mathrm{d}\tilde{\theta}_i/\mathrm{d}\theta_i$ denotes as usual the Radon-Nikodym derivative of $\tilde{\theta}_i$ with respect to θ_i. Indeed, if q_i is feasible in (2.30a), then $\tilde{\theta}_i$ defined through $\tilde{\theta}_i[\mathcal{B}] = \frac{1}{p_i}\int_B q_i(z_i)\theta_i(\mathrm{d}z_i)$ for all Borel sets $B \subseteq \mathbb{R}$ is feasible in (2.30b) and attains the same objective function value. Conversely, if $\tilde{\theta}_i$ is feasible in (2.30b), then $q_i(z_i) = p_i \, \mathrm{d}\tilde{\theta}_i/\mathrm{d}\theta_i(z_i)$ is feasible in (2.30a) and attains the same objective function value. Thus, (2.30a) and (2.30b) are indeed equivalent. By [FS04, Theorem 4.47], the optimal value of (2.30b) is given by $p_i \, \theta_i\text{-CVaR}_{p_i}(z_i) = \int_{1-p_i}^1 F_i^{-1}(t)\mathrm{d}t$, where $\theta_i\text{-CVaR}_{p_i}(z_i)$ denotes the CVaR of z_i at level p_i under θ_i.

If $p_i = 0$, on the other hand, then the optimal value of (2.30a) and the integral $\int_{1-p_i}^1 F_i^{-1}(t)\mathrm{d}t$ both evaluate to zero. Thus, the optimal value of the subproblem (2.30a) coincides with $\int_{1-p_i}^1 F_i^{-1}(t)\mathrm{d}t$ irrespective of p_i. Substituting this optimal value into (2.29) finally yields the explicit upper bound

$$\sup_{\theta \in \Theta} \mathbb{E}_{z \sim \theta}\left[\max_{i \in [N]} u_i + z_i\right] \le \sup_{\boldsymbol{p} \in \Delta^N} \sum_{i=1}^N u_i p_i + \sum_{i=1}^N \int_{1-p_i}^1 F_i^{-1}(t)\mathrm{d}t. \tag{2.31}$$

Note that the objective function of the upper bounding problem on the right hand side of (2.31) constitutes a sum of the strictly concave and differentiable univariate functions $u_i p_i + \int_{1-p_i}^1 F_i^{-1}(t)$. Indeed, the derivative of the i^{th} function with respect to p_i is given by $u_i + F_i^{-1}(1-p_i)$, which is strictly increasing in p_i because F_i is continuous by assumption. The upper bounding problem in (2.31) is thus solvable as it has a compact feasible set as well as a differentiable objective function. Moreover, the solution is unique thanks to the strict concavity of the objective function. In the following we denote this unique solution by $\boldsymbol{p}^\star$.

It remains to be shown that there exists a distribution $\theta^\star \in \Theta$ that attains the upper bound in (2.31). To this end, we define the functions $q_i^\star(z_i) = \mathbb{1}_{\{z_i > F_i^{-1}(1-p_i^\star)\}}$ for all $i \in [N]$. By [FS04, Remark 4.48], $q_i^\star(z_i)$ is optimal in (2.30a) for $p_i = p_i^\star$. In other words, we have $\mathbb{E}_{z_i \sim \theta_i}[q_i^\star(z_i)] = p_i^\star$ and $\mathbb{E}_{z_i \sim \theta_i}[z_i q_i^\star(z_i)] = \int_{1-p_i^\star}^1 F_i^{-1}(t)\mathrm{d}t$. In addition, we also define the Borel measures θ_i^+ and θ_i^- through

$$\theta_i^+(B) = \theta_i(B|z_i > F_i^{-1}(1-p_i^\star)) \quad \text{and} \quad \theta_i^-(B) = \theta_i(B|z_i \le F_i^{-1}(1-p_i^\star))$$

for all Borel sets $B \subseteq \mathbb{R}$, respectively. By construction, θ_i^+ is supported on $(F_i^{-1}(1-p_i^\star), \infty)$, while θ_i^- is supported on $(-\infty, (F_i^{-1}(1-p_i^\star)]$. The law of total probability further implies that $\theta_i = p_i^\star \theta_i^+ + (1-p_i^\star)\theta_i^-$. In the remainder of the proof we will demonstrate that the maximization problem on the left hand side of (2.31) is solved by the mixture distribution

$$\theta^\star = \sum_{j=1}^N p_j^\star \cdot \left(\otimes_{k=1}^{j-1} \theta_k^-\right) \otimes \theta_j^+ \otimes \left(\otimes_{k=j+1}^N \theta_k^-\right).$$

This will show that the inequality in (2.31) is in fact an equality, which in turn implies that the smooth c-transform is given by (2.27). We first prove that $\theta^\star \in \Theta$. To see this, note that for all $i \in [N]$ we have

$$\theta^\star(z_i \le s) = p_i^\star \theta_i^+(z_i \le s) + (\textstyle\sum_{j \ne i} p_j^\star)\theta_i^-(z_i \le s) = \theta_i(z_i \le s) = F_i(s),$$

where the second equality exploits the relation $\sum_{j \ne i} p_j^\star = 1 - p_i^\star$. This observation implies that $\theta^\star \in \Theta$. Next, we prove that $\theta^\star$ attains the upper bound in (2.31). By the definition of the binary function $\boldsymbol{r}$, we have

$$\begin{aligned}
\mathbb{E}_{\boldsymbol{z} \sim \theta^\star}\left[\max_{i \in [N]} u_i + z_i\right] &= \mathbb{E}_{\boldsymbol{z} \sim \theta^\star}\left[(u_i + z_i) r_i(\boldsymbol{z})\right] \\
&= \mathbb{E}_{z_i \sim \theta_i}\left[(u_i + z_i)\mathbb{E}_{\boldsymbol{z} \sim \theta^\star}\left[r_i(\boldsymbol{z})|z_i\right]\right] \\
&= \mathbb{E}_{z_i \sim \theta_i}\left[(u_i + z_i)\,\theta^\star\Big(i = \min \operatorname*{argmax}_{j \in [N]}\ u_j + z_j \big| z_i\Big)\right] \\
&= \mathbb{E}_{z_i \sim \theta_i}\left[(u_i + z_i)\,\theta^\star\left(z_j < u_i + z_i - u_j\ \forall j \ne i \big| z_i\right)\right],
\end{aligned}$$

where the third equality holds because $r_i(\boldsymbol{z}) = 1$ if and only if

$$i = \min \operatorname*{argmax}_{j\in[N]} u_j + z_j,$$

and the fourth equality follows from the assumed continuity of the marginal distribution functions F_i, $i \in [N]$, which implies that $\theta^\star(z_j = u_i + z_i - u_j \ \forall j \neq i|z_i) = 0$ θ_i-almost surely for all $i, j \in [N]$. Hence, we find

$$\begin{aligned}\mathbb{E}_{\boldsymbol{z}\sim\theta^\star}\left[\max_{i\in[N]} u_i + z_i\right] &= p_i^\star\, \mathbb{E}_{z_i\sim\theta_i^+}\left[(u_i+z_i)\,\theta^\star\left(z_j < u_i + z_i - u_j \ \forall j\neq i|z_i\right)\right]\\ &\quad + (1-p_i^\star)\,\mathbb{E}_{z_i\sim\theta_i^-}\left[(u_i+z_i)\,\theta^\star\left(z_j < u_i + z_i - u_j \ \forall j\neq i|z_i\right)\right]\\ &= p_i^\star\, \mathbb{E}_{z_i\sim\theta_i^+}\Big[(u_i+z_i)\Big(\prod_{j\neq i}\theta_j^-(z_j < z_i + u_i - u_j)\Big)\Big] \qquad (2.32a)\\ &\quad + \sum_{j\neq i} p_j^\star\, \mathbb{E}_{z_i\sim\theta_i^-}\Big[(u_i+z_i)\Big(\prod_{k\neq i,j}\theta_k^-(z_k < z_i + u_i - u_k)\Big)\theta_j^+(z_j < z_i + u_i - u_j)\Big], \qquad (2.32b)\end{aligned}$$

where the first equality exploits the relation $\theta_i = p_i^\star\theta_i^+ + (1-p_i^\star)\theta_i^-$, while the second equality follows from the definition of $\theta^\star$. The expectations in (2.32) can be further simplified by using the stationarity conditions of the upper bounding problem in (2.31), which imply that the partial derivatives of the objective function with respect to the decision variables p_i, $i \in [N]$, are all equal at $\boldsymbol{p} = \boldsymbol{p}^\star$. Thus, $\boldsymbol{p}^\star$ must satisfy

$$u_i + F_i^{-1}(1-p_i^\star) = u_j + F_j^{-1}(1-p_j^\star) \quad \forall i,j\in[N]. \qquad (2.33)$$

Consequently, for every $z_i > F_i^{-1}(1-p_i^\star)$ and $j \neq i$ we have

$$\begin{aligned}\theta_j^-(z_j < z_i + u_i - u_j) &\geq \theta_j^-(z_j \leq F_i^{-1}(1-p_i^\star) + u_i - u_j)\\ &= \theta_j^-(z_j \leq F_j^{-1}(1-p_j^\star)) = 1,\end{aligned}$$

where the first equality follows from (2.33), and the second equality holds because θ_j^- is supported on $(-\infty, F_j^{-1}(1-p_j^\star)]$. As no probability can exceed 1, the above reasoning implies that $\theta_j^-(z_j < z_i + u_i - u_j) = 1$ for all $z_i > F_i^{-1}(1-p_i^\star)$ and $j \neq i$. Noting that $q_i^\star(z_i) = \mathbb{1}_{\{z_i > F_i^{-1}(1-p_i^\star)\}}$ represents the characteristic function of the set $(F_i^{-1}(1-p_i^\star), \infty)$ covering the support of θ_i^+, the term (2.32a) can thus be simplified to

$$p_i^\star\, \mathbb{E}_{z_i\sim\theta_i^+}\Big[(u_i+z_i)\Big(\prod_{j\neq i}\theta_j^-(z_j < z_i + u_i - u_j)\Big)q_i^\star(z_i)\Big] = \mathbb{E}_{z_i\sim\theta_i}\left[(u_i+z_i)q_i^\star(z_i)\right].$$

Similarly, for any $z_i \leq F_i^{-1}(1-p_i^\star)$ and $j \neq i$ we have

$$\begin{aligned}\theta_j^+(z_j < z_i + u_i - u_j) &\leq \theta_j^+(z_j < F_i^{-1}(1-p_i^\star) + u_i - u_j) \\ &= \theta_j^+(z_j < F_j^{-1}(1-p_j^\star)) = 0,\end{aligned}$$

where the two equalities follow from (2.33) and the observation that θ_j^+ is supported on $(F_j^{-1}(1-p_j^\star), \infty)$, respectively. As probabilities are non-negative, the above implies that $\theta_j^+(z_j < z_i + u_i - u_j) = 0$ for all $z_i \leq F_i^{-1}(1-p_i^\star)$ and $j \neq i$. Hence, as θ_i^- is supported on $(-\infty, F_i^{-1}(1-p_i^\star)]$, the term (2.32b) simplifies to

$$\begin{aligned}\sum_{j\neq i} p_j^\star \mathbb{E}_{z_i\sim\theta_i^-}\Big[(u_i+z_i)\Big(\prod_{k\neq i,j} \theta_k^-(z_k < z_i + u_i - u_k)\Big) \\ \theta_j^+(z_j < z_i + u_i - u_j)\mathbb{1}_{\{z_i \leq F_i^{-1}(1-p_i^\star)\}}\Big] = 0.\end{aligned}$$

By combining the simplified reformulations of (2.32a) and (2.32b), we finally obtain

$$\mathbb{E}_{\boldsymbol{z}\sim\theta^\star}\Big[\max_{i\in[N]} u_i + z_i\Big] = \sum_{i=1}^N \mathbb{E}_{z_i\sim\theta_i}\left[(u_i+z_i)q_i^\star(z_i)\right] = \sum_{i=1}^N u_i p_i^\star + \sum_{i=1}^N \int_{1-p_i^\star}^1 F_i^{-1}(t)\mathrm{d}t,$$

where the last equality exploits the relations

$$\mathbb{E}_{z_i\sim\theta_i}[q_i^\star(z_i)] = p_i^\star \text{ and } \mathbb{E}_{z_i\sim\theta_i}[z_i q_i^\star(z_i)] = \int_{1-p_i^\star}^1 F_i^{-1}(t)\mathrm{d}t$$

derived in the first part of the proof. We have thus shown that the smooth c-transform is given by (2.27).

Finally, by the envelope theorem [Fue00, Theorem 2.16], the gradient of $\nabla_{\boldsymbol{\phi}}\overline{\psi}(\boldsymbol{\phi}, \boldsymbol{x})$ exists and coincides with the unique maximizer $\boldsymbol{p}^\star$ of the upper bounding problem in (2.27). □

The next theorem reveals that the smooth dual optimal transport problem (2.12) with a marginal ambiguity set corresponds to a regularized primal optimal transport problem of the form (2.13).

Theorem 2.3.7 (Fréchet regularization). *Suppose that Θ is a marginal ambiguity set of the form (2.26) and that the marginal cumulative distribution functions are defined through*

$$F_i(s) = \min\{1, \max\{0, 1 - \eta_i F(-s)\}\} \tag{2.34}$$

for some probability vector $\boldsymbol{\eta} \in \Delta^N$ and strictly increasing function $F : \mathbb{R} \to \mathbb{R}$ with $\int_0^1 F^{-1}(t)\mathrm{d}t = 0$. Then, the smooth dual optimal transport problem (2.12) is equivalent to the regularized primal optimal transport problem (2.13) with $R_\Theta = D_f(\pi\|\mu\otimes\eta)$, where $f(s) = \int_0^s F^{-1}(t)\mathrm{d}t$ and $\eta = \sum_{i=1}^N \eta_i\delta_{y_i}$.

The function $f(s)$ introduced in Theorem 2.3.7 is smooth and convex because its derivative $\mathrm{d}f(s)/\mathrm{d}s = F^{-1}(s)$ is strictly increasing, and $f(1) = \int_0^1 F^{-1}(t)\mathrm{d}t = 0$ by assumption. Therefore, this function induces a standard f-divergence. From now on we will refer to F as the *marginal generating function*.

Proof of Theorem 2.3.7. By Proposition 2.3.6, the smooth dual optimal transport problem (2.12) is equivalent to

$$\overline{W}_c(\mu,\nu) = \sup_{\boldsymbol{\phi}\in\mathbb{R}^N} \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\min_{\boldsymbol{p}\in\Delta^N}\sum_{i=1}^N \phi_i\nu_i - \sum_{i=1}^N(\phi_i - c(\boldsymbol{x},\boldsymbol{y_i}))p_i - \sum_{i=1}^N\int_{1-p_i}^1 F_i^{-1}(t)\mathrm{d}t\right].$$

As F is strictly increasing, we have $F_i^{-1}(s) = -F^{-1}((1-s)/\eta_i)$ for all $s\in(0,1)$. Thus, we find

$$f(s) = \int_0^s F^{-1}(t)\mathrm{d}t = -\frac{1}{\eta_i}\int_1^{1-s\eta_i} F^{-1}\left(\frac{1-z}{\eta_i}\right)\mathrm{d}z = -\frac{1}{\eta_i}\int_{1-s\eta_i}^1 F_i^{-1}(z)\mathrm{d}z, \tag{2.35}$$

where the second equality follows from the variable substitution $z \leftarrow 1-\eta_i t$. This integral representation of $f(s)$ then allows us to reformulate the smooth dual optimal transport problem as

$$\overline{W}_c(\mu,\nu) = \sup_{\boldsymbol{\phi}\in\mathbb{R}^N} \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\min_{\boldsymbol{p}\in\Delta^N}\sum_{i=1}^N \phi_i\nu_i - \sum_{i=1}^N(\phi_i - c(\boldsymbol{x},\boldsymbol{y_i}))p_i + \sum_{i=1}^N \eta_i\, f\left(\frac{p_i}{\eta_i}\right)\right],$$

which is manifestly equivalent to problem (2.14) thanks to the definition of the discrete f-divergence. Lemma 2.3.2 finally implies that the resulting instance of (2.14) is equivalent to the regularized primal optimal transport problem (2.13) with regularization term $R_\Theta(\pi) = D_f(\pi\|\mu\otimes\eta)$. Hence, the claim follows. □

Theorem 2.3.7 imposes relatively restrictive conditions on the marginals of $\boldsymbol{z}$. Indeed, it requires that all marginal distribution functions F_i, $i\in[N]$, must be generated by a single marginal generating function F through the relation (2.34). The following examples showcase, however, that the freedom to select F offers significant flexibility in designing various (existing as well as new) regularization schemes. Details of the underlying derivations are relegated to Appendix 2.5.

Example 2.3.8 (Exponential distribution model)**.** *Suppose that* Θ *is a marginal ambiguity set with (shifted) exponential marginals of the form* (2.34) *induced by the generating function* $F(s) = \exp(s/\lambda - 1)$ *with* $\lambda > 0$. *Then the smooth dual optimal transport problem* (2.12) *is equivalent to the regularized optimal transport problem* (2.13) *with an entropic regularizer of the form* $R_\Theta(\pi) =$

$D_f(\pi \| \mu \otimes \eta)$, where $f(s) = \lambda s \log(s)$, while the smooth c-transform (2.11) reduces to the log-partition function (2.20). This example shows that entropic regularizers are not only induced by singleton ambiguity sets containing a generalized extreme value distribution (see Section 2.3.1) but also by marginal ambiguity sets with exponential marginals.

Example 2.3.9 (Uniform distribution model)**.** *Suppose that Θ is a marginal ambiguity set with uniform marginals of the form (2.34) induced by the generating function $F(s) = s/(2\lambda) + 1/2$ with $\lambda > 0$. In this case the smooth dual optimal transport problem (2.12) is equivalent to the regularized optimal transport problem (2.13) with a χ^2-divergence regularizer of the form $R_\Theta(\pi) = D_f(\pi \| \mu \otimes \eta)$, where $f(s) = \lambda(s^2 - s)$. Such regularizers were previously investigated by [BSR18] and [Seg+18] under the additional assumption that η_i is independent of $i \in [N]$, yet their intimate relation to noise models with uniform marginals remained undiscovered until now. In addition, the smooth c-transform (2.11) satisfies*

$$\overline{\psi}(\boldsymbol{\phi}, \boldsymbol{x}) = \lambda + \lambda \operatorname*{spmax}_{i \in [N]} \frac{\phi_i - c(\boldsymbol{x}, \boldsymbol{y_i})}{\lambda},$$

*where the sparse maximum operator '*spmax*' inspired by [MA16] is defined through*

$$\operatorname*{spmax}_{i \in [N]} \; u_i = \max_{\boldsymbol{p} \in \Delta^N} \sum_{i=1}^N u_i p_i - p_i^2/\eta_i \qquad \forall \boldsymbol{u} \in \mathbb{R}^N. \tag{2.36}$$

The envelope theorem [Fue00, Theorem 2.16] ensures that $\operatorname{spmax}_{i\in[N]} u_i$ is smooth and that its gradient with respect to $\boldsymbol{u}$ is given by the unique solution $\boldsymbol{p}^\star$ of the maximization problem on the right hand side of (2.36). We note that $\boldsymbol{p}^\star$ has many zero entries due to the sparsity-inducing nature of the problem's simplicial feasible set. In addition, we have $\lim_{\lambda \downarrow 0} \lambda \operatorname{spmax}_{i\in[N]} u_i/\lambda = \max_{i\in[N]} u_i$. Thus, the sparse maximum can indeed be viewed as a smooth approximation of the ordinary maximum. In marked contrast to the more widely used LogSumExp function, however, the sparse maximum has a sparse gradient. Proposition 2.6.1 in Appendix 2.6 shows that $\boldsymbol{p}^\star$ can be computed efficiently by sorting.

Example 2.3.10 (Pareto distribution model)**.** *Suppose that Θ is a marginal ambiguity set with (shifted) Pareto distributed marginals of the form (2.34) induced by the generating function $F(s) = (s(q-1)/(\lambda q) + 1/q)^{1/(q-1)}$ with $\lambda, q > 0$. Then the smooth dual optimal transport problem (2.12) is equivalent* *to the regularized optimal transport problem* (2.13) *with a Tsallis divergence regularizer of the form* $R_\Theta(\pi) = D_f(\pi\|\mu\otimes\eta)$*, where* $f(s) = \lambda(s^q - s)/(q-1)$*. Such regularizers were investigated by [Muz+17] under the additional assumption that* η_i *is independent of* $i \in [N]$*. The Pareto distribution model encapsulates the exponential model (in the limit* $q \to 1$*) and the uniform distribution model (for* $q = 2$*) as special cases. The smooth c-transform admits no simple closed-form representation under this model.*

Example 2.3.11 (Hyperbolic cosine distribution model)**.** *Suppose that* Θ *is a marginal ambiguity set with hyperbolic cosine distributed marginals of the form* (2.34) *induced by the generating function* $F(s) = \sinh(s/\lambda - k)$ *with* $k = \sqrt{2} - 1 - \mathit{arcsinh}(1)$ *and* $\lambda > 0$*. Then the marginal probability density functions are given by scaled and truncated hyperbolic cosine functions, and the smooth dual optimal transport problem* (2.12) *is equivalent to the regularized optimal transport problem* (2.13) *with a hyperbolic divergence regularizer of the form* $R_\Theta(\pi) = D_f(\pi\|\mu\otimes\eta)$*, where* $f(s) = \lambda(s\,\mathit{arcsinh}(s) - \sqrt{s^2+1} + 1 + ks)$*. Hyperbolic divergences were introduced by [GHS20] in order to unify several gradient descent algorithms.*

Example 2.3.12 (t-distribution model)**.** *Suppose that* Θ *is a marginal ambiguity set where the marginals are determined by* (2.34)*, and assume that the generating function is given by*

$$F(s) = \frac{N}{2}\left(1 + \frac{s - \sqrt{N-1}}{\sqrt{\lambda^2 + (s - \sqrt{N-1})^2}}\right)$$

for some $\lambda > 0$*. In this case one can show that all marginals constitute t-distributions with* 2 *degrees of freedom. In addition, one can show that the smooth dual optimal transport problem* (2.12) *is equivalent to the Chebyshev regularized optimal transport problem described in Proposition 2.3.5.*

To close this section, we remark that different regularization schemes differ as to how well they approximate the original (unregularized) optimal transport problem. Proposition 2.3.3 provides simple error bounds that may help in selecting suitable regularizers. For the entropic regularization scheme associated with the exponential distribution model of Example 2.3.8, for example, the error bound evaluates to $\max_{i\in[N]} \lambda \log(1/\eta_i)$, while for the χ^2-divergence regularization scheme associated with the uniform distribution model of Example 2.3.9, the error bound is given by $\max_{i\in[N]} \lambda(1/\eta_i - 1)$. In both cases, the error is minimized by setting $\eta_i = 1/N$ for all $i \in [N]$. Thus, the error bound grows logarithmically with N for entropic regularization and linearly with N for

χ^2-divergence regularization. Different regularization schemes also differ with regard to their computational properties, which will be discussed in Section 2.4.

2.4. Numerical Solution of Smooth Optimal Transport Problems

The smooth semi-discrete optimal transport problem (2.12) constitutes a stochastic optimization problem and can therefore be addressed with a stochastic gradient descent (SGD) algorithm. In Section 2.4.1 we first derive new convergence guarantees for an averaged gradient descent algorithm that has only access to a biased stochastic gradient oracle. This algorithm outputs the uniform average of the iterates (instead of the last iterate) as the recommended candidate solution. We prove that if the objective function is Lipschitz continuous, then the suboptimality of this candidate solution is of the order $\mathcal{O}(1/\sqrt{T})$, where T stands for the number of iterations. An improvement in the non-leading terms is possible if the objective function is additionally smooth. We further prove that a convergence rate of $\mathcal{O}(1/T)$ can be obtained for generalized self-concordant objective functions. In Section 2.4.2 we then show that the algorithm of Section 2.4.1 can be used to efficiently solve the smooth semi-discrete optimal transport problem (2.12) corresponding to a marginal ambiguity set of the type (2.26). As a byproduct, we prove that the convergence rate of the averaged SGD algorithm for the semi-discrete optimal transport problem with *entropic* regularization is of the order $\mathcal{O}(1/T)$, which improves the $\mathcal{O}(1/\sqrt{T})$ guarantee of [Gen+16].

2.4.1. Averaged Gradient Descent Algorithm with Biased Gradient Oracles

Consider a general convex minimization problem of the form

$$\min_{\boldsymbol{\phi}\in\mathbb{R}^n} \; h(\boldsymbol{\phi}), \tag{2.37}$$

where the objective function $h:\mathbb{R}^n\to\mathbb{R}$ is convex and differentiable. We assume that problem (2.37) admits a minimizer $\boldsymbol{\phi}^\star$. We study the convergence behavior of the inexact gradient descent algorithm

$$\boldsymbol{\phi}_t = \boldsymbol{\phi}_{t-1} - \gamma \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1}), \tag{2.38}$$

where $\gamma > 0$ is a fixed step size, $\boldsymbol{\phi}_0$ is a given deterministic initial point and the function $\boldsymbol{g}_t:\mathbb{R}^n\to\mathbb{R}^n$ is an inexact gradient oracle that returns for every fixed $\boldsymbol{\phi}\in\mathbb{R}^n$ a random estimate of the gradient of h at $\boldsymbol{\phi}$. Note that we allow the gradient oracle to depend on the iteration counter t, which allows us to account for increasingly accurate gradient estimates. In contrast to the previous sections, we henceforth model all random objects as measurable functions on an abstract filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_t)_{t\geq 0}, \mathbb{P})$, where $\mathcal{F}_0 = \{\emptyset, \Omega\}$ represents the trivial σ-field, while the gradient oracle $\boldsymbol{g}_t(\boldsymbol{\phi})$ is $\mathcal{F}_t$-measurable for all $t \in \mathbb{N}$ and $\boldsymbol{\phi} \in \mathbb{R}^n$. In order to avoid clutter, we use $\mathbb{E}[\cdot]$ to denote the expectation operator with respect to $\mathbb{P}$, and all inequalities and equalities involving random variables are understood to hold $\mathbb{P}$-almost surely.

In the following we analyze the effect of averaging in inexact gradient descent algorithms. We will show that after T iterations with a constant step size $\gamma = \mathcal{O}(1/\sqrt{T})$, the objective function value of the uniform average of all iterates generated by (2.38) converges to the optimal value of (2.37) at a sublinear rate. Specifically, we will prove that the rate of convergence varies between $\mathcal{O}(1/\sqrt{T})$ and $\mathcal{O}(1/T)$ depending on properties of the objective function. Our convergence analysis will rely on several regularity conditions.

Assumption 1 (Regularity conditions)**.** *Different combinations of the following regularity conditions will enable us to establish different convergence guarantees for the averaged inexact gradient descent algorithm.*

(i) ***Biased gradient oracle:*** *There exists tolerances $\varepsilon_t > 0$, $t \in \mathbb{N} \cup \{0\}$, such that*

$$\left\| \mathbb{E}\left[\boldsymbol{g}_t(\boldsymbol{\phi}_{t-1}) \middle| \mathcal{F}_{t-1} \right] - \nabla h(\boldsymbol{\phi}_t) \right\| \leq \varepsilon_{t-1} \quad \forall t \in \mathbb{N}.$$

(ii) ***Bounded gradients:*** *There exists $R > 0$ such that*

$$\|\nabla h(\boldsymbol{\phi})\| \leq R \quad \textit{and} \quad \|\boldsymbol{g}_t(\boldsymbol{\phi})\| \leq R \quad \forall \boldsymbol{\phi} \in \mathbb{R}^n,\ \forall t \in \mathbb{N}.$$

(iii) ***Generalized self-concordance:*** *The function h is M-generalized self-concordant for some $M > 0$, that is, h is three times differentiable, and for any $\boldsymbol{\phi}, \boldsymbol{\phi}' \in \mathbb{R}^n$ the function $u(s) = h(\boldsymbol{\phi} + s(\boldsymbol{\phi}' - \boldsymbol{\phi}))$ satisfies the inequality*

$$\left| \frac{\mathrm{d}^3 u(s)}{\mathrm{d}s^3} \right| \leq M \|\boldsymbol{\phi} - \boldsymbol{\phi}'\| \frac{\mathrm{d}^2 u(s)}{\mathrm{d}s^2} \quad \forall s \in \mathbb{R}.$$

(iv) ***Lipschitz continuous gradient:*** *The function h is L-smooth for some $L > 0$, that is, we have*

$$\|\nabla h(\boldsymbol{\phi}) - \nabla h(\boldsymbol{\phi}')\| \leq L \|\boldsymbol{\phi} - \boldsymbol{\phi}'\| \quad \forall \boldsymbol{\phi}, \boldsymbol{\phi}' \in \mathbb{R}^n.$$

(v) ***Bounded second moments:*** *There exists $\sigma > 0$ such that*

$$\mathbb{E}\left[\left\| \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1}) - \nabla h(\boldsymbol{\phi}_{t-1}) \right\|^2 | \mathcal{F}_{t-1} \right] \leq \sigma^2 \quad \forall t \in \mathbb{N}.$$

The averaged gradient descent algorithm with biased gradient oracles lends itself to solving both deterministic as well as stochastic optimization problems. In deterministic optimization, the gradient oracles $\boldsymbol{g}_t$ are deterministic and output inexact gradients satisfying $\|\boldsymbol{g}_t(\boldsymbol{\phi}) - \nabla h(\boldsymbol{\phi})\| \leq \varepsilon_t$ for all $\boldsymbol{\phi} \in \mathbb{R}^n$, where the tolerances ε_t bound the errors associated with the numerical computation of the gradients. A vast body of literature on deterministic optimization focuses on exact gradient oracles for which these tolerances can be set to 0. Inexact deterministic gradient oracles with bounded error tolerances are investigated by [NB00] and [d'A08]. In this case exact convergence to $\boldsymbol{\phi}^\star$ is not possible. If the error bounds decrease to 0, however, [LT93; SRB11] and [FS12] show that adaptive gradient descent algorithms are guaranteed to converge to $\boldsymbol{\phi}^\star$.

In stochastic optimization, the objective function is representable as $h(\boldsymbol{\phi}) = \mathbb{E}[H(\boldsymbol{\phi}, \boldsymbol{x})]$, where the marginal distribution of the random vector $\boldsymbol{x}$ under $\mathbb{P}$ is given by μ, while the integrand $H(\boldsymbol{\phi}, \boldsymbol{x})$ is convex and differentiable in $\boldsymbol{\phi}$ and μ-integrable in $\boldsymbol{x}$. In this setting it is convenient to use gradient oracles of the form $\boldsymbol{g}_t(\boldsymbol{\phi}) = \nabla_{\boldsymbol{\phi}} H(\boldsymbol{\phi}, \boldsymbol{x}_t)$ for all $t \in \mathbb{N}$, where the samples $\boldsymbol{x}_t$ are drawn independently from μ. As these oracles output unbiased estimates for $\nabla h(\boldsymbol{\phi})$, all tolerances ε_t in Assumptions 1 (i) may be set to 0. SGD algorithms with unbiased gradient oracles date back to the seminal paper by [RM51]. Nowadays, averaged SGD algorithms with Polyak-Ruppert averaging figure among the most popular variants of the SGD algorithm [Rup88; PJ92; Nem+09]. For general convex objective functions the best possible convergence rate of any averaged SGD algorithm run over T iterations amounts to $\mathcal{O}(1/\sqrt{T})$, but it improves to $\mathcal{O}(1/T)$ if the objective function is strongly convex; see for example [NV08; Nem+09; SS+09; DS09; Xia09; MB11; SS+11; LJSB12]. While smoothness plays a critical role to achieve acceleration in deterministic optimization, it only improves the constants in the convergence rate in stochastic optimization [SST10; Dek+12; Lan12; CDO18; Kav+19]. In fact, [Tsy03] demonstrates that smoothness does not provide any acceleration in general, that is, the best possible convergence rate of any averaged SGD algorithm can still not be improved beyond $\mathcal{O}(1/\sqrt{T})$. Nevertheless, a substantial acceleration is possible when focusing on special problem classes such as linear or logistic regression problems [Bac14; BM13; HKL14]. In these special cases, the improvement in the convergence rate is facilitated by a generalized self-concordance property of the objective function [Bac10]. Self-concordance was originally introduced in the context of Newton-type interior point methods [NN94] and later generalized to facilitate the analysis of probabilistic models [Bac10] and second-order optimization algorithms [STD19].

In the following we analyze the convergence properties of the averaged SGD algorithm when we have only access to an *inexact* stochastic gradient oracle, in which case the tolerances ε_t cannot be set to 0. To our best knowledge, inexact stochastic gradient oracles have only been considered by [CDO18; HSL20] and [AS20]. Specifically, [HSL20] use sequential semidefinite programs to analyze the convergence rate of the averaged SGD algorithm when μ has a finite support. In contrast, we do not impose any restrictions on the support of μ. [CDO18] and [AS20], on the other hand, study the convergence behavior of accelerated gradient descent algorithms for smooth stochastic optimization problems under the assumption that $\boldsymbol{\phi}$ ranges over a compact domain. The proposed algorithms necessitate a projection onto the compact feasible set in each iteration. In contrast, our convergence analysis does not rely on any compactness assumptions. We note that compactness assumptions have been critical for the convergence analysis of the averaged SGD algorithm in the context of convex stochastic optimization [Nem+09; Dek+12; Bub15; CDO18]. By leveraging a trick due to [Bac14], however, we can relax this assumption provided that the objective function is Lipschitz continuous.

Proposition 2.4.1. *Consider the inexact gradient descent algorithm* (2.38) *with constant step size* $\gamma > 0$*. If Assumptions* 1 *(i)–(ii) hold with* $\varepsilon_t \leq \bar{\varepsilon}/(2\sqrt{1+t})$ *for some* $\bar{\varepsilon} \geq 0$*, then we have for all* $p \in \mathbb{N}$ *that*

$$\mathbb{E}\left[\left(h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right) - h(\boldsymbol{\phi}^\star)\right)^p\right]^{1/p} \leq \frac{\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2}{\gamma T} + 20\gamma\left(R + \bar{\varepsilon}\right)^2 p.$$

If additionally Assumption 1 *(iii) holds and if* $G = \max\{M, R + \bar{\varepsilon}\}$*, then we have for all* $p \in \mathbb{N}$ *that*

$$\mathbb{E}\left[\left\|\nabla h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)\right\|^{2p}\right]^{1/p} \leq \frac{G^2}{T}\Bigg(10\sqrt{p} + \frac{4p}{\sqrt{T}} + 80G^2\gamma\sqrt{T}p + \frac{2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2}{\gamma\sqrt{T}} + \frac{3\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|}{G\gamma\sqrt{T}}\Bigg)^2.$$

The proof of Proposition 2.4.1 relies on two lemmas. In order to state these lemmas concisely, we define the L_p-norm, of a random variable $\boldsymbol{z} \in \mathbb{R}^n$ for any $p > 0$ through $\|\boldsymbol{z}\|_{L_p} = \left(\mathbb{E}\left[\|\boldsymbol{z}\|^p\right]\right)^{1/p}$. For any random variables $\boldsymbol{z}, \boldsymbol{z}' \in \mathbb{R}^n$ and $p \geq 1$, Minkowski's inequality [BLM13, § 2.11] then states that

$$\|\boldsymbol{z} + \boldsymbol{z}'\|_{L_p} \leq \|\boldsymbol{z}\|_{L_p} + \|\boldsymbol{z}'\|_{L_p}. \tag{2.39}$$

Another essential tool for proving Proposition 2.4.1 is the Burkholder-Rosenthal-Pinelis (BRP) inequality [Pin94, Theorem 4.1], which we restate below without proof to keep this chapter self-contained.

Lemma 2.4.2 (BRP inequality)**.** *Let* $\boldsymbol{z}_t$ *be an* $\mathcal{F}_t$*-measurable random variable for every* $t \in \mathbb{N}$ *and assume that* $p \geq 2$*. For any* $t \in [T]$ *with* $\mathbb{E}[\boldsymbol{z}_t|\mathcal{F}_{t-1}] = 0$ *and* $\|\boldsymbol{z}_t\|_{L_p} < \infty$ *we then have*

$$\left\|\max_{t\in[T]}\left\|\sum_{k=1}^{t}\boldsymbol{z}_k\right\|\right\|_{L_p} \leq \sqrt{p}\left\|\sum_{t=1}^{T}\mathbb{E}[\|\boldsymbol{z}_t\|^2|\mathcal{F}_{t-1}]\right\|_{L_{p/2}}^{1/2} + p\left\|\max_{t\in[T]}\|\boldsymbol{z}_t\|\right\|_{L_p}.$$

The following lemma reviews two useful properties of generalized self-concordant functions.

Lemma 2.4.3. *[Generalized self-concordance] Assume that the objective function h of the convex optimization problem* (2.37) *is M-generalized self-concordant in the sense of Assumption 1 (iii) for some $M > 0$.*

(i) [Bac14, Appendix D.2] For any sequence $\boldsymbol{\phi}_0, \ldots, \boldsymbol{\phi}_{T-1} \in \mathbb{R}^n$, we have

$$\left\| \nabla h \left(\frac{1}{T} \sum_{t=1}^{T} \boldsymbol{\phi}_{t-1} \right) - \frac{1}{T} \sum_{t=1}^{T} \nabla h(\boldsymbol{\phi}_{t-1}) \right\| \leq 2M \left(\frac{1}{T} \sum_{t=1}^{T} h(\boldsymbol{\phi}_{t-1}) - h(\boldsymbol{\phi}^\star) \right).$$

(ii) [Bac14, Lemma 9] For any $\boldsymbol{\phi} \in \mathbb{R}^n$ with $\|\nabla h(\boldsymbol{\phi})\| \leq 3\kappa/(4M)$, where κ is the smallest eigenvalue of $\nabla^2 h(\boldsymbol{\phi}^\star)$, and $\boldsymbol{\phi}^\star$ is the optimizer of (2.37)*, we have $h(\boldsymbol{\phi}) - h(\boldsymbol{\phi}^\star) \leq 2\|\nabla h(\boldsymbol{\phi})\|^2/\kappa$.*

Armed with Lemmas 2.4.2 and 2.4.3, we are now ready to prove Proposition 2.4.1.

Proof of Proposition 2.4.1. The first claim generalizes Proposition 5 by [Bac14] to inexact gradient oracles. By the assumed convexity and differentiability of the objective function h, we have

$$\begin{aligned} h(\boldsymbol{\phi}_{k-1}) &\leq h(\boldsymbol{\phi}_\star) + \nabla h(\boldsymbol{\phi}_{k-1})^\top (\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}_\star) && (2.40)\\ &= h(\boldsymbol{\phi}_\star) + \boldsymbol{g}_k(\boldsymbol{\phi}_{k-1})^\top (\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}_\star) && (2.41)\\ &\quad + \left(\nabla h(\boldsymbol{\phi}_{k-1}) - \boldsymbol{g}_k(\boldsymbol{\phi}_{k-1})\right)^\top (\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}_\star). \end{aligned}$$

In addition, elementary algebra yields the recursion

$$\|\boldsymbol{\phi}_k - \boldsymbol{\phi}^\star\|^2 = \|\boldsymbol{\phi}_k - \boldsymbol{\phi}_{k-1}\|^2 + \|\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}^\star\|^2 + 2(\boldsymbol{\phi}_k - \boldsymbol{\phi}_{k-1})^\top (\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}^\star).$$

Thanks to the update rule (2.38), this recursion can be re-expressed as

$$\boldsymbol{g}_k(\boldsymbol{\phi}_{k-1})^\top (\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}^\star) = \frac{1}{2\gamma} \left(\gamma^2 \|\boldsymbol{g}_k(\boldsymbol{\phi}_{k-1})\|^2 + \|\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}^\star\|^2 - \|\boldsymbol{\phi}_k - \boldsymbol{\phi}^\star\|^2\right),$$

where $\gamma > 0$ is an arbitrary step size. Combining the above identity with (2.40) then yields

$$\begin{aligned} &h(\boldsymbol{\phi}_{k-1})\\ &\leq h(\boldsymbol{\phi}_\star) + \frac{1}{2\gamma} \left(\gamma^2 \|\boldsymbol{g}_k(\boldsymbol{\phi}_{k-1})\|^2 + \|\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}^\star\|^2 - \|\boldsymbol{\phi}_k - \boldsymbol{\phi}^\star\|^2\right)\\ &\qquad + \left(\nabla h(\boldsymbol{\phi}_{k-1}) - \boldsymbol{g}_k(\boldsymbol{\phi}_{k-1})\right)^\top (\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}_\star) \end{aligned}$$

$$\leq h(\boldsymbol{\phi}_\star) + \frac{1}{2\gamma}\left(\gamma^2 R^2 + \|\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}^\star\|^2 - \|\boldsymbol{\phi}_k - \boldsymbol{\phi}^\star\|^2\right) \\ + \left(\nabla h(\boldsymbol{\phi}_{k-1}) - \boldsymbol{g}_k(\boldsymbol{\phi}_{k-1})\right)^\top (\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}_\star),$$

where the last inequality follows from Assumption 1 (ii). Summing this inequality over k then shows that

$$2\gamma \sum_{k=1}^{t} \left(h(\boldsymbol{\phi}_{k-1}) - h(\boldsymbol{\phi}_\star)\right) + \|\boldsymbol{\phi}_t - \boldsymbol{\phi}^\star\|^2 \leq A_t, \tag{2.42}$$

where

$$A_t = t\gamma^2 R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \sum_{k=1}^{t} B_k \quad \text{and} \\ B_t = 2\gamma\left(\nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})\right)^\top (\boldsymbol{\phi}_{t-1} - \boldsymbol{\phi}_\star)$$

for all $t \in \mathbb{N}$. Note that the term on the left hand side of (2.42) is non-negative because $\boldsymbol{\phi}^\star$ is a global minimizer of h, which implies that the random variable A_t is also non-negative for all $t \in \mathbb{N}$. For later use we further define $A_0 = \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2$. The estimate (2.42) for $t = T$ then implies via the convexity of h that

$$h\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\phi}_{t-1}\right) - h(\boldsymbol{\phi}_\star) \leq \frac{A_T}{2\gamma T}, \tag{2.43}$$

where we dropped the non-negative term $\|\boldsymbol{\phi}_T - \boldsymbol{\phi}^\star\|^2/(2\gamma T)$ without invalidating the inequality. In the following we analyze the L_p-norm of A_T in order to obtain the desired bounds from the proposition statement. To do so, we distinguish three different regimes for $p \in \mathbb{N}$, and we show that the L_p-norm of the non-negative random variable A_T is upper bounded by an affine function of p in each of these regimes.

Case I ($p \geq T/4$): By using the update rule (2.38) and Assumption 1 (ii), one readily verifies that

$$\|\boldsymbol{\phi}_k - \boldsymbol{\phi}^\star\| \leq \|\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}^\star\| + \|\boldsymbol{\phi}_k - \boldsymbol{\phi}_{k-1}\| \leq \|\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}^\star\| + \gamma R.$$

Iterating the above recursion k times then yields the conservative estimate $\|\boldsymbol{\phi}_k - \boldsymbol{\phi}^\star\| \leq \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\| + k\gamma R$. By definitions of A_t and B_t for $t \in \mathbb{N}$, we thus have

$$A_t = t\gamma^2 R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 2\gamma \textstyle\sum_{k=1}^{t} \left(\nabla h(\boldsymbol{\phi}_{k-1}) - \boldsymbol{g}_k(\boldsymbol{\phi}_{k-1})\right)^\top (\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}_\star) \\ \leq t\gamma^2 R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 4\gamma R \textstyle\sum_{k=1}^{t} \|\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}_\star\|$$

$$
\begin{aligned}
&\le t\gamma^2R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 4\gamma R \textstyle\sum_{k=1}^t \left(\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\| + (k-1)\gamma R\right)\\
&\le t\gamma^2R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 4t\gamma R\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\| + 2t^2\gamma^2R^2\\
&\le t\gamma^2R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 4t^2\gamma^2R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 2t^2\gamma^2R^2\\
&\le 7t^2\gamma^2R^2 + 2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2,
\end{aligned}
$$

where the first two inequalities follow from Assumption 1 (ii) and the conservative estimate derived above, respectively, while the fourth inequality holds because $2ab \le a^2 + b^2$ for all $a, b \in \mathbb{R}$. As $A_t \ge 0$, the random variable A_t is bounded and satisfies $|A_t| \le 2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 7t^2\gamma^2R^2$ for all $t \in \mathbb{N}$, which implies that

$$
\|A_T\|_{L_p} \le 2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 7T^2\gamma^2R^2 \le 2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 28T\gamma^2R^2p, \tag{2.44}
$$

where the last inequality holds because $p \ge T/4$. Note that the resulting upper bound is affine in p.

Case II ($2 \le p \le T/4$)**:** The subsequent analysis relies on the simple bounds

$$
\max\nolimits_{t\in[T]} \varepsilon_{t-1} \le \tfrac{\bar{\varepsilon}}{2} \quad \text{and} \quad \textstyle\sum_{t=1}^T \varepsilon_{t-1} \le \bar{\varepsilon}\sqrt{T}, \tag{2.45}
$$

which hold because $\varepsilon_t \le \bar{\varepsilon}/(2\sqrt{1+t})$ by assumption and because $\sum_{t=1}^T 1/\sqrt{t} \le 2\sqrt{T}$, which can be proved by induction. In addition, it proves useful to introduce the martingale differences $\bar{B}_t = B_t - \mathbb{E}[B_t|\mathcal{F}_{t-1}]$ for all $t \in \mathbb{N}$. By the definition of A_t and the subadditivity of the supremum operator, we then have

$$
\begin{aligned}
\max_{t\in[T+1]} A_{t-1} &= \max_{t\in[T+1]} \left\{(t-1)\gamma^2R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \sum_{k=1}^{t-1} \mathbb{E}[B_k|\mathcal{F}_{k-1}] + \sum_{k=1}^{t-1} \bar{B}_k\right\}\\
&\le T\gamma^2R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \max_{t\in[T]} \sum_{k=1}^{t} \mathbb{E}[B_k|\mathcal{F}_{k-1}] + \max_{t\in[T]} \sum_{k=1}^{t} \bar{B}_k.
\end{aligned}
$$

As $p \ge 2$, Minkowski's inequality (2.39) thus implies that

$$
\left\|\max_{t\in[T+1]} A_{t-1}\right\|_{L_p} \le T\gamma^2R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \left\|\max_{t\in[T]} \sum_{k=1}^{t} \mathbb{E}[B_k|\mathcal{F}_{k-1}]\right\|_{L_p} \tag{2.46}
$$

$$
+ \left\|\max_{t\in[T]} \sum_{k=1}^{t} \bar{B}_k\right\|_{L_p}. \tag{2.47}
$$

In order to bound the penultimate term in (2.46), we first note that

$$
|\mathbb{E}[B_k|\mathcal{F}_{k-1}]| = 2\gamma \left|\mathbb{E}\left[\left(\nabla h(\boldsymbol{\phi}_{k-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{k-1})\right)|\mathcal{F}_{k-1}\right]^\top (\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}_\star)\right|
$$

$$\begin{aligned}
&\leq 2\gamma \|\mathbb{E}\left[\left(\nabla h(\boldsymbol{\phi}_{k-1}) - \boldsymbol{g}_k(\boldsymbol{\phi}_{k-1})\right)|\mathcal{F}_{k-1}\right]\| \|\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}_\star\| \\
&\leq 2\gamma\varepsilon_{k-1}\|\boldsymbol{\phi}_{k-1} - \boldsymbol{\phi}_\star\| \leq 2\gamma\varepsilon_{k-1}\sqrt{A_{k-1}} \qquad (2.48)
\end{aligned}$$

for all $k \in \mathbb{N}$, where the second inequality holds due to Assumption 1 (i), and the last inequality follows from (2.42). This in turn implies that for all $t \in [T]$ we have

$$\begin{aligned}
\left|\sum_{k=1}^{t} \mathbb{E}[B_k|\mathcal{F}_{k-1}]\right| &\leq 2\gamma \sum_{k=1}^{t} \varepsilon_{k-1}\sqrt{A_{k-1}} \\
&\leq 2\gamma \left(\sum_{k=1}^{t} \varepsilon_{k-1}\right)\left(\max_{k\in[t]} \sqrt{A_{k-1}}\right) \\
&\leq 2\gamma\bar{\varepsilon}\sqrt{t}\max_{k\in[t]}\sqrt{A_{k-1}},
\end{aligned}$$

where the last inequality exploits (2.45). Therefore, the penultimate term in (2.46) satisfies

$$\left\|\max_{t\in[T]}\sum_{k=1}^{t}\mathbb{E}[B_k|\mathcal{F}_{k-1}]\right\|_{L_p} \leq 2\gamma\bar{\varepsilon}\sqrt{T}\left\|\max_{t\in[T+1]}\sqrt{A_{t-1}}\right\|_{L_p} = 2\gamma\bar{\varepsilon}\sqrt{T}\left\|\max_{t\in[T+1]}A_{t-1}\right\|_{L_{p/2}}^{1/2}, \qquad (2.49)$$

where the equality follows from the definition of the L_p-norm.

Next, we bound the last term in (2.46) by using the BRP inequality of Lemma 2.4.2. To this end, note that

$$\begin{aligned}
|\bar{B}_t| &\leq |B_t| + |\mathbb{E}[B_t|\mathcal{F}_{t-1}]| \\
&\leq 2\gamma\|\boldsymbol{\phi}_{t-1} - \boldsymbol{\phi}_\star\|\|\nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})\| + 2\gamma\varepsilon_{t-1}\sqrt{A_{t-1}} \\
&\leq 2\gamma\sqrt{A_{t-1}}\left(\|\nabla h(\boldsymbol{\phi}_{t-1})\| + \|\boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})\|\right) + 2\gamma\varepsilon_{t-1}\sqrt{A_{t-1}} \\
&\leq 2\gamma(2R + \varepsilon_{t-1})\sqrt{A_{t-1}}
\end{aligned}$$

for all $t \in \mathbb{N}$, where the second inequality exploits the definition of B_t and (2.48), the third inequality follows from (2.42), and the last inequality holds because of Assumption 1 (ii). Hence, we obtain

$$\begin{aligned}
\left\|\max\nolimits_{t\in[T]}|\bar{B}_t|\right\|_{L_p} &\leq 2\gamma\left(2R + \max_{t\in[T]}\varepsilon_{t-1}\right)\left\|\max_{t\in[T]}\sqrt{A_{t-1}}\right\|_{L_p} \\
&\leq (4\gamma R + \gamma\bar{\varepsilon})\left\|\max_{t\in[T+1]}A_{t-1}\right\|_{L_{p/2}}^{1/2},
\end{aligned}$$

where the second inequality follows from (2.45) and the definition of the L_p-norm. In addition, we have

$$\begin{aligned}
\left\|\sum_{t=1}^{T}\mathbb{E}[\bar{B}_t^2|\mathcal{F}_{t-1}]\right\|_{L_{p/2}}^{1/2} &= \left\|\sqrt{\sum_{t=1}^{T}\mathbb{E}[\bar{B}_t^2|\mathcal{F}_{t-1}]}\right\|_{L_p} \\
&\leq 2\gamma\left\|\sqrt{\sum_{t=1}^{T}(2R+\varepsilon_{t-1})^2 A_{t-1}}\right\|_{L_p} \\
&\leq 2\gamma\left(\sum_{t=1}^{T}(2R+\varepsilon_{t-1})^2\right)^{1/2}\left\|\max_{t\in[T+1]} A_{t-1}^{1/2}\right\|_{L_p} \\
&\leq 2\gamma\left(2R\sqrt{T}+\sqrt{\sum_{t=1}^{T}\varepsilon_{t-1}^2}\right)\left\|\max_{t\in[T+1]} A_{t-1}^{1/2}\right\|_{L_p} \\
&\leq \left(4\gamma R\sqrt{T}+\gamma\bar{\varepsilon}\sqrt{T}\right)\left\|\max_{t\in[T+1]} A_{t-1}\right\|_{L_{p/2}}^{1/2},
\end{aligned}$$

where the first inequality exploits the upper bound on $|\bar{B}_t|$ derived above, which implies that $\mathbb{E}[\bar{B}_t^2|\mathcal{F}_{t-1}] \leq 4\gamma^2(2R+\varepsilon_{t-1})^2 A_{t-1}$. The last three inequalities follow from the Hölder inequality, the triangle inequality for the Euclidean norm and the two inequalities in (2.45), respectively. Recalling that $p \geq 2$, we may then apply the BRP inequality of Lemma 2.4.2 to the martingale differences $\bar{B}_t$, $t \in [T]$, and use the bounds derived in the last two display equations in order to conclude that

$$\left\|\max_{t\in[T]}\left|\sum_{k=1}^{t}\bar{B}_k\right|\right\|_{L_p} \leq \left(4\gamma R\sqrt{pT}+\gamma\bar{\varepsilon}\sqrt{pT}+\gamma\bar{\varepsilon}p+4\gamma Rp\right)\left\|\max_{t\in[T+1]} A_{t-1}\right\|_{L_{p/2}}^{1/2}. \tag{2.50}$$

Substituting (2.49) and (2.50) into (2.46), we thus obtain

$$\begin{aligned}
\left\|\max_{t\in[T+1]} A_{t-1}\right\|_{L_p} &\leq T\gamma^2R^2+\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+\Bigg(4\gamma R\left(\sqrt{pT}+p\right)+ \\
&\qquad \gamma\bar{\varepsilon}\left(\sqrt{pT}+p+2\sqrt{T}\right)\Bigg)\left\|\max_{t\in[T+1]} A_{t-1}\right\|_{L_{p/2}}^{1/2} \\
&\leq T\gamma^2R^2+\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+6\gamma\left(R+\bar{\varepsilon}\right)\sqrt{pT}\left\|\max_{t\in[T+1]} A_{t-1}\right\|_{L_{p/2}}^{1/2},
\end{aligned}$$

where the second inequality holds because $p \leq T/4$ by assumption, which implies that $\sqrt{pT}+p \leq 1.5\sqrt{pT}$ and $\sqrt{pT}+p+2\sqrt{T} \leq 6\sqrt{pT}$. As Jensen's inequality ensures that $\|\boldsymbol{z}\|_{L_{p/2}} \leq \|\boldsymbol{z}\|_{L_p}$ for any random variable $\boldsymbol{z}$ and $p > 0$, the following inequality holds for all $2 \leq p \leq T/4$.

$$\left\| \max_{t\in[T+1]} A_{t-1} \right\|_{L_p} \leq T\gamma^2 R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 6\gamma\left(R + \bar{\varepsilon}\right)\sqrt{pT} \left\| \max_{t\in[T+1]} A_{t-1} \right\|_{L_p}^{1/2}$$

To complete the proof of Case II, we note that for any numbers $a, b, c \geq 0$ the inequality $c \leq a + 2b\sqrt{c}$ is equivalent to $\sqrt{c} \leq b + \sqrt{b^2 + a}$ and therefore also to $c \leq (b + \sqrt{b^2 + a})^2 \leq 4b^2 + 2a$. Identifying a with $T\gamma^2 R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2$, b with $3\gamma\left(R + \bar{\varepsilon}\right)\sqrt{pT}$ and c with $\|\max_{t\in[T+1]} A_{t-1}\|_{L_p}$ then allows us to translate the inequality in the last display equation to

$$\|A_T\|_{L_p} \leq \left\| \max_{t\in[T+1]} A_{t-1} \right\|_{L_p} \leq 2T\gamma^2 R^2 + 2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 36\gamma^2\left(R + \bar{\varepsilon}\right)^2 pT. \tag{2.51}$$

Thus, for any $2 \leq p \leq T/4$, we have again found an upper bound on $\|A_T\|_{L_p}$ that is affine in p.

Case III ($p = 1$)**:** Recalling the definition of $A_T \geq 0$, we find that

$$\begin{aligned}
\|A_T\|_{L_1} = \mathbb{E}[A_T] &= T\gamma^2 R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \mathbb{E}\left[\sum_{t=1}^{T} \mathbb{E}[B_t|\mathcal{F}_{t-1}]\right] \\
&\leq T\gamma^2 R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \left\| \max_{t\in[T]} \sum_{k=1}^{t} \mathbb{E}[B_k|\mathcal{F}_{k-1}] \right\|_{L_1} \\
&\leq T\gamma^2 R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 2\gamma\bar{\varepsilon}\sqrt{T} \left\| \max_{t\in[T+1]} A_{t-1} \right\|_{L_{1/2}}^{1/2} \\
&\leq T\gamma^2 R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 2\gamma\bar{\varepsilon}\sqrt{T} \left\| \max_{t\in[T+1]} A_{t-1} \right\|_{L_2}^{1/2},
\end{aligned}$$

where the second inequality follows from the estimate (2.49), which holds indeed for all $p \in \mathbb{N}$, while the last inequality follows from Jensen's inequality. By the second inequality in (2.51) for $p = 2$, we thus find

$$\begin{aligned}
\|A_T\|_{L_1} &\leq T\gamma^2 R^2 + \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 && \text{(2.52a)} \\
&\quad + 2\bar{\varepsilon}\gamma\sqrt{T} \cdot \sqrt{2T\gamma^2 R^2 + 2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 72\gamma^2 (R + \bar{\varepsilon})^2 T} && \text{(2.52b)} \\
&\leq 2T\gamma^2 R^2 + 2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 36\gamma^2 (R + \bar{\varepsilon})^2 T + 2\bar{\varepsilon}^2\gamma^2 T, && \text{(2.52c)}
\end{aligned}$$

where the last inequality holds because $2ab \leq 2a^2 + b^2/2$ for all $a, b \in \mathbb{R}$.

We now combine the bounds derived in Cases I, II and III to obtain a universal bound on $\|A_T\|_{L_p}$ that holds for all $p \in \mathbb{N}$. Specifically, one readily verifies that the bound

$$\|A_T\|_{L_p} \leq 2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 40\gamma^2 \left(R + \bar{\varepsilon}\right)^2 pT, \tag{2.53}$$

is more conservative than each of the bounds (2.44), (2.51) and (2.52), and thus it holds indeed for any $p \in \mathbb{N}$. Combining this universal bound with (2.43) proves the first inequality from the proposition statement.

In order to prove the second inequality, we need to extend [Bac14, Proposition 7] to biased gradient oracles. To this end, we first note that

$$\begin{aligned}\left\|\nabla h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)\right\| &\leq \left\|\nabla h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right) - \frac{1}{T}\sum_{t=1}^{T}\nabla h(\boldsymbol{\phi}_{t-1})\right\| \\ &\quad + \left\|\frac{1}{T}\sum_{t=1}^{T}\nabla h(\boldsymbol{\phi}_{t-1})\right\| \\ &\leq 2M\left(\frac{1}{T}\sum_{t=1}^{T}h(\boldsymbol{\phi}_{t-1}) - h(\boldsymbol{\phi}^\star)\right) + \\ &\quad \left\|\frac{1}{T}\sum_{t=1}^{T}\nabla h(\boldsymbol{\phi}_{t-1})\right\| \\ &\leq \frac{M}{T\gamma}A_T + \left\|\frac{1}{T}\sum_{t=1}^{T}\nabla h(\boldsymbol{\phi}_{t-1})\right\|,\end{aligned}$$

where the second inequality follows from Lemma 2.4.3 (i), and the third inequality holds due to (2.42). By Minkowski's inequality (2.39), we thus have for any $p \geq 1$ that

$$\begin{aligned}\left\|\nabla h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)\right\|_{L_{2p}} &\leq \frac{M}{T\gamma}\|A_T\|_{L_{2p}} + \left\|\frac{1}{T}\sum_{t=1}^{T}\nabla h(\boldsymbol{\phi}_{t-1})\right\|_{L_{2p}} \\ &\leq \frac{2M}{T\gamma}\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 80M\gamma\left(R + \bar{\varepsilon}\right)^2 p + \\ &\quad \left\|\frac{1}{T}\sum_{t=1}^{T}\nabla h(\boldsymbol{\phi}_{t-1})\right\|_{L_{2p}},\end{aligned}$$

where the last inequality follows from the universal bound (2.53). In order to estimate the last term in the above expression, we recall that the update rule (2.38) is equivalent to $\boldsymbol{g}_t(\boldsymbol{\phi}_{t-1}) = \left(\boldsymbol{\phi}_{t-1} - \boldsymbol{\phi}_t\right)/\gamma$, which in turn implies that $\sum_{t=1}^T \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1}) = (\boldsymbol{\phi}_0 - \boldsymbol{\phi}_T)/\gamma$. Hence, for any $p \geq 1$, we have

$$\begin{aligned}\left\| \frac{1}{T}\sum_{t=1}^T \nabla h(\boldsymbol{\phi}_{t-1}) \right\|_{L_{2p}} &= \left\| \frac{1}{T}\sum_{t=1}^T \Big(\nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})\Big) + \frac{\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star}{T\gamma} + \frac{\boldsymbol{\phi}^\star - \boldsymbol{\phi}_T}{T\gamma} \right\|_{L_{2p}} \\ &\leq \left\| \frac{1}{T}\sum_{t=1}^T \nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1}) \right\|_{L_{2p}} + \\ &\qquad \frac{1}{T\gamma}\left\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\right\| + \frac{1}{T\gamma}\left\|\boldsymbol{\phi}^\star - \boldsymbol{\phi}_T\right\|_{L_{2p}} \\ &\leq \left\| \frac{1}{T}\sum_{t=1}^T \nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1}) \right\|_{L_{2p}} + \\ &\qquad \frac{1}{T\gamma}\left\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\right\| + \frac{1}{T\gamma}\left\|A_T\right\|_{L_p}^{1/2} \\ &\leq \left\| \frac{1}{T}\sum_{t=1}^T \nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1}) \right\|_{L_{2p}} + \\ &\qquad \frac{1+\sqrt{2}}{T\gamma}\left\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\right\| + \frac{2\sqrt{10}\,(R+\bar{\varepsilon})\,\sqrt{p}}{\sqrt{T}},\end{aligned}$$

where the first inequality exploits Minkowski's inequality (2.39), the second inequality follows from (2.42), which implies that $\|\boldsymbol{\phi}^\star - \boldsymbol{\phi}_T\| \leq \sqrt{A_T}$, and the definition of the L_p-norm. The last inequality in the above expression is a direct consequence of the universal bound (2.53) and the inequality $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b \geq 0$. Next, define for any $t \in \mathbb{N}$ a martingale difference of the form

$$\boldsymbol{C}_t = \frac{1}{T}\Big(\nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1}) - \mathbb{E}[\nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})|\mathcal{F}_{t-1}]\Big).$$

Note that these martingale differences are bounded because

$$\begin{aligned}\|\boldsymbol{C}_t\| &\leq \frac{1}{T}\Big(\|\nabla h(\boldsymbol{\phi}_{t-1})\| + \|\boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})\| + \|\mathbb{E}[\nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})|\mathcal{F}_{t-1}]\|\Big) \\ &\leq \frac{2R + \varepsilon_{t-1}}{T} \leq \frac{2R+\bar{\varepsilon}}{T},\end{aligned}$$

and thus the BRP inequality of Lemma 2.4.2 implies that

$$\left\| \sum_{t=1}^T \boldsymbol{C}_t \right\|_{L_{2p}} \leq \sqrt{2p}\,\frac{2R+\bar{\varepsilon}}{\sqrt{T}} + 2p\,\frac{2R+\bar{\varepsilon}}{T}.$$

Recalling the definition of the martingale differences $\boldsymbol{C}_t$, $t \in \mathbb{N}$, this bound allows us to conclude that

$$\frac{1}{T}\left\| \sum_{t=1}^T \nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1}) \right\|_{L_{2p}} \leq \left\| \sum_{t=1}^T \boldsymbol{C}_t \right\|_{L_{2p}} +$$

$$\begin{aligned}
&\frac{1}{T}\left\|\sum_{t=1}^{T}\mathbb{E}[\nabla h(\boldsymbol{\phi}_{t-1})-\boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})|\mathcal{F}_{t-1}]\right\|_{L_{2p}}\\
&\leq \sqrt{2p}\,\frac{2R+\bar{\varepsilon}}{\sqrt{T}}+2p\,\frac{2R+\bar{\varepsilon}}{T}+\frac{\bar{\varepsilon}}{\sqrt{T}}\\
&\leq 2\sqrt{2p}\,\frac{R+\bar{\varepsilon}}{\sqrt{T}}+4p\,\frac{R+\bar{\varepsilon}}{T},
\end{aligned}$$

where the second inequality exploits Assumption 1 (i) as well as the second inequality in (2.45). Combining all inequalities derived above and observing that $2\sqrt{2}+2\sqrt{10}<10$ finally yields

$$\begin{aligned}
\left\|\nabla h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)\right\|_{L_{2p}} &\leq \frac{2M}{T\gamma}\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+80M\gamma\left(R+\bar{\varepsilon}\right)^2p\\
&\quad+2\sqrt{2p}\,\frac{R+\bar{\varepsilon}}{\sqrt{T}}+4p\,\frac{R+\bar{\varepsilon}}{T}\\
&\quad+\frac{1+\sqrt{2}}{T\gamma}\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|+\frac{2\sqrt{10}\left(R+\bar{\varepsilon}\right)\sqrt{p}}{\sqrt{T}}\\
&\leq\frac{G}{\sqrt{T}}\Bigg(10\sqrt{p}+\frac{4p}{\sqrt{T}}+80G^2\gamma\sqrt{T}p+\\
&\qquad\frac{2}{\gamma\sqrt{T}}\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+\frac{3}{G\gamma\sqrt{T}}\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|\Bigg),
\end{aligned}$$

where $G=\max\{M,R+\bar{\varepsilon}\}$. This proves the second inequality from the proposition statement. □

The following corollary follows immediately from the proof of Proposition 2.4.1.

Corollary 5. *Consider the inexact gradient descent algorithm* (2.38) *with constant step size $\gamma>0$. If Assumptions 1 (i)–(ii) hold with $\varepsilon_t\leq\bar{\varepsilon}/(2\sqrt{1+t})$ for some $\bar{\varepsilon}\geq 0$, then we have*

$$\begin{aligned}
&\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\nabla h(\boldsymbol{\phi}_t)-\boldsymbol{g}_t(\boldsymbol{\phi}_t)\right)^\top\left(\boldsymbol{\phi}_t-\boldsymbol{\phi}_\star\right)\right]\\
&\quad\leq\frac{\bar{\varepsilon}}{\sqrt{T}}\sqrt{2\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+74\gamma^2(R+\bar{\varepsilon})^2T}.
\end{aligned}$$

Proof of Corollary 5. Defining B_t as in the proof of Proposition 2.4.1, we find

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\nabla h(\boldsymbol{\phi}_t)-\boldsymbol{g}_t(\boldsymbol{\phi}_t)\right)^\top\left(\boldsymbol{\phi}_t-\boldsymbol{\phi}_\star\right)\right]=\frac{1}{2\gamma T}\mathbb{E}\left[\sum_{t=1}^{T}\mathbb{E}[B_t|\mathcal{F}_{t-1}]\right]$$

$$\leq \frac{\bar{\varepsilon}}{\sqrt{T}}\sqrt{2T\gamma^2R^2 + 2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 72\gamma^2(R+\bar{\varepsilon})^2T},$$

where the inequality is an immediate consequence of the reasoning in Case (III) in the proof of Proposition 2.4.1. The claim then follows from the trivial inequality $R + \bar{\varepsilon} \geq R$. □

Armed with Proposition 2.4.1 and Corollary 5, we are now ready to prove the main convergence result.

Theorem 2.4.4. *Consider the inexact gradient descent algorithm* (2.38) *with constant step size $\gamma > 0$. If Assumptions 1 (i)–(ii) hold with $\varepsilon_t \leq \bar{\varepsilon}/(2\sqrt{1+t})$ for some $\bar{\varepsilon} \geq 0$, then the following statements hold.*

(i) If $\gamma = 1/(2(R+\bar{\varepsilon})^2\sqrt{T})$, then we have

$$\mathbb{E}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)\right] - h(\boldsymbol{\phi}^\star) \leq \frac{(R+\bar{\varepsilon})^2}{\sqrt{T}}\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \frac{1}{4\sqrt{T}} + \frac{\bar{\varepsilon}}{\sqrt{T}}\sqrt{2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \frac{37}{2(R+\bar{\varepsilon})^2}}.$$

(ii) If $\gamma = 1/(2(R+\bar{\varepsilon})^2\sqrt{T}+L)$ and the Assumptions 1 (iv)–(v) hold in addition to the blanket assumptions mentioned above, then we have

$$\mathbb{E}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t}\right)\right] - h(\boldsymbol{\phi}^\star) \leq \frac{L}{2T}\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \frac{(R+\bar{\varepsilon})^2}{\sqrt{T}}\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \frac{\sigma^2}{4(R+\bar{\varepsilon})^2\sqrt{T}} + \frac{\bar{\varepsilon}}{\sqrt{T}}\sqrt{2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \frac{37}{2(R+\bar{\varepsilon})^2}}.$$

(iii) If $\gamma = 1/(2G^2\sqrt{T})$ with $G = \max\{M, R+\bar{\varepsilon}\}$, the smallest eigenvalue κ of $\nabla^2 h(\boldsymbol{\phi}^\star)$ is strictly positive and Assumption 1 (iii) holds in addition to the blanket assumptions mentioned above, then we have

$$\mathbb{E}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)\right] - h(\boldsymbol{\phi}^\star) \leq \frac{G^2}{\kappa T}\left(4G\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\| + 20\right)^4.$$

The proof of Theorem 2.4.4 relies on the following concentration inequalities due to [Bac14].

Lemma 2.4.5. *[Concentration inequalities]*

(i) *[Bac14, Lemma 11]: If there exist $a, b > 0$ and a random variable $\boldsymbol{z} \in \mathbb{R}^n$ with $\|\boldsymbol{z}\|_{L_p} \leq a + bp$ for all $p \in \mathbb{N}$, then we have*

$$\mathbb{P}\left[\|\boldsymbol{z}\| \geq 3bs + 2a\right] \leq 2\exp(-s) \quad \forall s \geq 0.$$

(ii) *[Bac14, Lemma 12]: If there exist $a, b, c > 0$ and a random variable $\boldsymbol{z} \in \mathbb{R}^n$ with $\|\boldsymbol{z}\|_{L_p} \leq (a\sqrt{p} + bp + c)^2$ for all $p \in [T]$, then we have*

$$\mathbb{P}\left[\|\boldsymbol{z}\| \geq (2a\sqrt{s} + 2bs + 2c)^2\right] \leq 4\exp(-s) \quad \forall s \leq T.$$

Proof of Theorem 2.4.4. Define A_t as in the proof of Proposition 2.4.1. Then, we have

$$\mathbb{E}\left[h\left(\frac{1}{T}\sum_{t=0}^{T-1}\boldsymbol{\phi}_t\right) - h(\boldsymbol{\phi}^\star)\right] \leq \frac{\mathbb{E}[A_T]}{2\gamma T} = \frac{\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2}{2\gamma T} + \frac{\gamma R^2}{2} + \tag{2.54}$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\left(\nabla h(\boldsymbol{\phi}_t) - \boldsymbol{g}_t(\boldsymbol{\phi}_t)\right)^\top(\boldsymbol{\phi}_t - \boldsymbol{\phi}_\star)\right]$$

$$\leq \frac{\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2}{2\gamma T} + \tag{2.55}$$

$$\frac{\gamma R^2}{2} + \frac{\bar{\varepsilon}}{\sqrt{T}}\sqrt{2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 74\gamma^2(R + \bar{\varepsilon})^2 T}, \tag{2.56}$$

where the two inequalities follow from (2.43) and from Corollary 5, respectively. Setting the step size to $\gamma = 1/(2(R + \bar{\varepsilon})^2\sqrt{T})$ then completes the proof of assertion (i).

Assertion (ii) generalizes [Dek+12, Theorem 1]. By the L-smoothness of $h(\boldsymbol{\phi})$, we have

$$h(\boldsymbol{\phi}_t) \leq h(\boldsymbol{\phi}_{t-1}) + \nabla h(\boldsymbol{\phi}_{t-1})^\top(\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}) \tag{2.57}$$

$$+ \frac{L}{2}\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\|^2$$

$$= h(\boldsymbol{\phi}_{t-1}) + \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})^\top(\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}) + (\nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1}))^\top(\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}) + \tag{2.58}$$

$$\frac{L}{2}\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\|^2$$

$$\leq h(\boldsymbol{\phi}_{t-1}) + \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})^\top(\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}) \tag{2.59}$$

$$+ \frac{\zeta}{2}\|\nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})\|^2 + \frac{L + 1/\zeta}{2}\|\boldsymbol{\phi}_t - \boldsymbol{\phi}_{t-1}\|^2, \tag{2.60}$$

where the last inequality exploits the Cauchy-Schwarz inequality together with the elementary inequality $2ab \leq \zeta a^2 + b^2/\zeta$, which holds for all $a, b \in \mathbb{R}$ and $\zeta > 0$. Next, note that the iterates satisfy the recursion

$$\|\boldsymbol{\phi}_{t-1} - \boldsymbol{\phi}^\star\|^2 = \|\boldsymbol{\phi}_{t-1} - \boldsymbol{\phi}_t\|^2 + \|\boldsymbol{\phi}_t - \boldsymbol{\phi}^\star\|^2 + 2(\boldsymbol{\phi}_{t-1} - \boldsymbol{\phi}_t)^\top (\boldsymbol{\phi}_t - \boldsymbol{\phi}^\star),$$

which can be re-expressed as

$$\boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})^\top (\boldsymbol{\phi}_t - \boldsymbol{\phi}^\star) = \frac{1}{2\gamma}\left(\|\boldsymbol{\phi}_{t-1} - \boldsymbol{\phi}^\star\|^2 - \|\boldsymbol{\phi}_{t-1} - \boldsymbol{\phi}_t\|^2 - \|\boldsymbol{\phi}_t - \boldsymbol{\phi}^\star\|^2\right)$$

by using the update rule (2.38). In the remainder of the proof we assume that $0 < \gamma < 1/L$. Substituting the above equality into (2.60) and setting $\zeta = \gamma/(1-\gamma L)$ then yields

$$\begin{aligned} h(\boldsymbol{\phi}_t) \leq h(\boldsymbol{\phi}_{t-1}) + \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})^\top (\boldsymbol{\phi}^\star - \boldsymbol{\phi}_{t-1}) + \frac{\gamma}{2(1-\gamma L)} \|\nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})\|^2 \\ + \frac{1}{2\gamma}\left(\|\boldsymbol{\phi}_{t-1} - \boldsymbol{\phi}^\star\|^2 - \|\boldsymbol{\phi}_t - \boldsymbol{\phi}^\star\|^2\right). \end{aligned}$$

By the convexity of h, we have $h(\boldsymbol{\phi}^\star) \geq h(\boldsymbol{\phi}_{t-1}) + \nabla h(\boldsymbol{\phi}_{t-1})^\top (\boldsymbol{\phi}^\star - \boldsymbol{\phi}_{t-1})$, which finally implies that

$$\begin{aligned} h(\boldsymbol{\phi}_t) \leq h(\boldsymbol{\phi}^\star) &+ \left(\nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})\right)^\top (\boldsymbol{\phi}_{t-1} - \boldsymbol{\phi}^\star) \\ &+ \frac{\gamma}{2(1-\gamma L)} \|\nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})\|^2 \\ &+ \frac{1}{2\gamma}\left(\|\boldsymbol{\phi}_{t-1} - \boldsymbol{\phi}^\star\|^2 - \|\boldsymbol{\phi}_t - \boldsymbol{\phi}^\star\|^2\right). \end{aligned}$$

Averaging the above inequality over t and taking expectations then yields the estimate

$$\begin{aligned} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T h(\boldsymbol{\phi}_t)\right] - h(\boldsymbol{\phi}^\star) \leq \frac{\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2}{2\gamma T} + \\ \frac{\gamma}{2(1-\gamma L)} \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \|\nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})\|^2\right] + \\ \mathbb{E}\left[\frac{1}{T}\sum_{t=1}^T \left(\nabla h(\boldsymbol{\phi}_{t-1}) - \boldsymbol{g}_t(\boldsymbol{\phi}_{t-1})\right)^\top (\boldsymbol{\phi}_{t-1} - \boldsymbol{\phi}_\star)\right] \\ \leq \frac{\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2}{2\gamma T} + \frac{\gamma\sigma^2}{2(1-\gamma L)} + \\ \frac{\bar{\varepsilon}}{\sqrt{T}} \sqrt{2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 74\gamma^2 (R + \bar{\varepsilon})^2 T}, \end{aligned}$$

where the second inequality exploits Assumption 1 (v) and Corollary 5. Using Jensen's inequality to move the average over t inside h, assertion (ii) then follows by setting $\gamma = 1/(2(R+\bar{\varepsilon})^2\sqrt{T}+L)$ and observing that $\gamma/(1-\gamma L) = 1/(2(R+\bar{\varepsilon})^2\sqrt{T})$.

To prove assertion (iii), we distinguish two different cases.

Case I: Assume first that $4G^2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 6G\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\| \leq \kappa\sqrt{T}/(8G^2)$, where $G = \max\{M, R+\bar{\varepsilon}\}$ and κ denotes the smallest eigenvalue of $\nabla^2 h(\boldsymbol{\phi}^\star)$. By a standard formula for the expected value of a non-negative random variable, we find

$$\begin{aligned}
\mathbb{E}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right) - h(\boldsymbol{\phi}^\star)\right] &= \int_0^\infty \mathbb{P}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right) - h(\boldsymbol{\phi}^\star) \geq u\right]\mathrm{d}u \\
&= \int_0^{u_1} \mathbb{P}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right) - h(\boldsymbol{\phi}^\star) \geq u\right]\mathrm{d}u \\
&\quad + \int_{u_1}^{u_2} \mathbb{P}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right) - h(\boldsymbol{\phi}^\star) \geq u\right]\mathrm{d}u \qquad (2.61)\\
&\quad + \int_{u_2}^{\infty} \mathbb{P}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right) - h(\boldsymbol{\phi}^\star) \geq u\right]\mathrm{d}u,
\end{aligned}$$

where $u_1 = \frac{8G^2}{\kappa T}(4G^2\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+6G\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|)^2$ and $u_2 = \frac{8G^2}{\kappa T}(\frac{\kappa\sqrt{T}}{4G^2}+4G^2\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+6G\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|)^2$. The first of the three integrals in (2.61) is trivially upper bounded by u_1. Next, we investigate the third integral in (2.61), which is easier to bound from above than the second one. By combining the first inequality in Proposition 2.4.1 for $\gamma = 1/(2G^2\sqrt{T})$ with the trivial inequality $G \geq R+\bar{\varepsilon}$, we find

$$\left\|h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right) - h(\boldsymbol{\phi}^\star)\right\|_{L_p} \leq \frac{2G^2}{\sqrt{T}}\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \frac{10}{\sqrt{T}}p \quad \forall p \in \mathbb{N}.$$

Lemma 2.4.5 (i) with $a = 2G^2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2/\sqrt{T}$ and $b = 10/\sqrt{T}$ thus implies that

$$\mathbb{P}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right) - h(\boldsymbol{\phi}^\star) \geq \frac{30}{\sqrt{T}}s + \frac{4G^2}{\sqrt{T}}\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2\right] \leq 2\exp(-s) \quad \forall s \geq 0. \tag{2.62}$$

We also have

$$u_2 - \frac{4G^2}{\sqrt{T}}\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 \geq u_2 - \frac{\kappa}{8G^2} \geq \frac{8G^2}{\kappa T}\left(\frac{\kappa\sqrt{T}}{4G^2}\right)^2 - \frac{\kappa}{8G^2} = \frac{3\kappa}{8G^2} \geq 0, \tag{2.63}$$

where the first inequality follows from the basic assumption underlying Case I, while the second inequality holds due to the definition of u_2. By (2.62) and (2.63), the third integral in (2.61) satisfies

$$\begin{aligned}
&\int_{u_2}^{\infty} \mathbb{P}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right) - h(\boldsymbol{\phi}^\star) \geq u\right] \mathrm{d}u \\
&= \int_{u_2 - \frac{4G^2}{\sqrt{T}}\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2}^{\infty} \mathbb{P}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right) - h(\boldsymbol{\phi}^\star) \geq u + \frac{4G^2}{\sqrt{T}}\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2\right] \mathrm{d}u \\
&\leq 2\int_{u_2 - \frac{4G^2}{\sqrt{T}}\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2}^{\infty} \exp\left(-\frac{\sqrt{T}u}{30}\right) \mathrm{d}u \\
&= \frac{60}{\sqrt{T}}\exp\left(-\frac{\sqrt{T}}{30}\left(u_2 - \frac{4G^2}{\sqrt{T}}\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2\right)\right) \\
\leq\; &\frac{60}{\sqrt{T}}\exp\left(-\frac{\kappa\sqrt{T}}{80G^2}\right) \leq \frac{2400G^2}{\kappa T},
\end{aligned}$$

where the first inequality follows from the concentration inequality (2.62) and the insight from (2.63) that $u_2 - \frac{4G^2}{\sqrt{T}}\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 \geq 0$. The second inequality exploits again (2.63), and the last inequality holds because $\exp(-x) \leq 1/(2x)$ for all $x > 0$. We have thus found a simple upper bound on the third integral in (2.61). It remains to derive an upper bound on the second integral in (2.61). To this end, we first observe that the second inequality in Proposition 2.4.1 for $\gamma = 1/(2G^2\sqrt{T})$ translates to

$$\begin{aligned}
&\left\|\left\|\nabla h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)\right\|^2\right\|_{L_p} \\
&\leq \frac{G^2}{T}\left(10\sqrt{p} + \frac{4p}{\sqrt{T}} + 40p + 4G^2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 6G\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|\right)^2 \quad \forall p \in \mathbb{N}.
\end{aligned}$$

Lemma 2.4.5 (ii) with $a = 10G/\sqrt{T}$, $b = 4G/T + 40G/\sqrt{T}$ and $c = 4G^3\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2/\sqrt{T} + 6G^2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|/\sqrt{T}$ thus gives rise to the concentration inequality

$$\mathbb{P}\Bigg[\;\left\|\nabla h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)\right\|^2$$

$$\geq \frac{4G^2}{T}\left(10\sqrt{s}+\frac{4s}{\sqrt{T}}+40s+4G^2\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+6G\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|\right)^2\Bigg]$$
$$\leq 4\exp(-s),$$

which holds only for small deviations $s \leq T$. However, this concentration inequality can be simplified to

$$\mathbb{P}\Bigg[\ \left\|\nabla h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)\right\|$$
$$\geq \frac{2G}{\sqrt{T}}\left(12\sqrt{s}+40s+4G^2\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+6G\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|\right)\Bigg] \leq 4\exp(-s),$$

which remains valid for all deviations $s \geq 0$. To see this, note that if $s \leq T/4$, then the simplified concentration inequality holds because $4s/T \leq 2\sqrt{s/T}$. Otherwise, if $s > T/4$, then the simplified concentration inequality holds trivially because the probability on the left hand vanishes. Indeed, this is an immediate consequence of Assumption 1 (ii), which stipulates that the norm of the gradient of h is bounded by R, and of the elementary estimate $24G\sqrt{s/T} > G \geq R$, which holds for all $s > T/4$.

In the following, we restrict attention to those deviations $s \geq 0$ that are small in the sense that

$$12\sqrt{s}+40s \leq \frac{\kappa\sqrt{T}}{4G^2}. \tag{2.64}$$

Assume now for the sake of argument that the event inside the probability in the simplified concentration inequality does *not* occur, that is, assume that

$$\left\|\nabla h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)\right\| \tag{2.65}$$
$$< \frac{2G}{\sqrt{T}}\left(12\sqrt{s}+40s+4G^2\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+6G\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|\right). \tag{2.66}$$

By (2.64) and the assumption of Case I, (2.65) implies that

$$\|\nabla h(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1})\| < 3\kappa/(4G) < 3\kappa/(4M).$$

Hence, we may apply Lemma 2.4.3 (ii) to conclude that $h(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}) - h(\boldsymbol{\phi}^\star) \leq \frac{2}{\kappa}\|\nabla h(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1})\|^2$. Combining this inequality with (2.65) then yields

$$h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)-h(\boldsymbol{\phi}^\star)<\frac{8G^2}{\kappa T}\left(12\sqrt{s}+40s+4G^2\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+6G\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|\right)^2. \tag{2.67}$$

By the simplified concentration inequality derived above, we may thus conclude that

$$\begin{aligned} 4\exp(-s) \geq \mathbb{P}\Bigg[& \left\|\nabla h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)\right\| \geq \\ & \frac{2G}{\sqrt{T}}\left(12\sqrt{s}+40s+4G^2\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+6G\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|\right)\Bigg] \end{aligned} \tag{2.68}$$

$$\begin{aligned} \geq \mathbb{P}\Bigg[& h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)-h(\boldsymbol{\phi}^\star) \geq \qquad (2.69) \\ & \frac{8G^2}{\kappa T}\left(12\sqrt{s}+40s+4G^2\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+6G\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|\right)^2\Bigg] \end{aligned} \tag{2.70}$$

for any $s \geq 0$ that satisfies (2.64), where the second inequality holds because (2.65) implies (2.67) or, equivalently, because the negation of (2.67) implies the negation of (2.65). The resulting concentration inequality (2.70) now enables us to construct an upper bound on the second integral in (2.61). To this end, we define the function

$$\ell(s)=\frac{8G^2}{\kappa T}\left(12\sqrt{s}+40s+4G^2\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+6G\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|\right)^2$$

for all $s \geq 0$, and set $\bar{s}=((9/400+\kappa\sqrt{T}/(160G^2))^{\frac{1}{2}}-3/20)^2$. Note that $s \geq 0$ satisfies the inequality (2.64) if and only if $s \leq \bar{s}$ and that $\ell(0)=u_1$ as well as $\ell(\bar{s})=u_2$. By substituting u with $\ell(s)$ and using the concentration inequality (2.70) to bound the integrand, we find that the second integral in (2.61) satisfies

$$\begin{aligned} & \int_{u_1}^{u_2}\mathbb{P}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)-h(\boldsymbol{\phi}^\star)\geq u\right]\mathrm{d}u \\ & =\int_{0}^{\bar{s}}\mathbb{P}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)-h(\boldsymbol{\phi}^\star)\geq \ell(s)\right]\frac{\mathrm{d}\ell(s)}{\mathrm{d}s}\mathrm{d}s \end{aligned}$$

$$
\begin{aligned}
&\leq \int_0^{\bar{s}} 4\mathrm{e}^{-s} \, \frac{\mathrm{d}}{\mathrm{d}s}\left(\frac{8G^2}{\kappa T} \left(12\sqrt{s} + 40s + \tau\right)^2 \right) \mathrm{d}s \\
&\leq \frac{32G^2}{\kappa T} \int_0^{\infty} \mathrm{e}^{-s} \left(144 + 3200s + 1440s^{1/2} + 80\tau + 12\tau s^{-1/2}\right) \mathrm{d}s \\
&= \frac{32G^2}{\kappa T} \big(144 + 3200\Gamma(2) + 1440\Gamma(3/2) + 80\tau + 12\tau\Gamma(1/2)\big) \\
&\leq \frac{32G^2}{\kappa T} (4621 + 102\tau),
\end{aligned}
$$

where τ is is a shorthand for $4G^2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 6G\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|$, and Γ denotes the Gamma function with $\Gamma(2) = 1$, $\Gamma(1/2) = \sqrt{\pi}$ and $\Gamma(3/2) = \sqrt{\pi}/2$; see for example [Rud64, Chapter 8]. The last inequality is obtained by rounding all fractional numbers up to the next higher integer. Combining the upper bounds for the three integrals in (2.61) finally yields

$$
\begin{aligned}
\mathbb{E}\left[h\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\phi}_{t-1} \right) - h(\boldsymbol{\phi}^\star) \right] &\leq \frac{8G^2}{\kappa T} \left(\tau^2 + 18484 + 408\tau + 300\right) \\
&= \frac{8G^2}{\kappa T} \Big(16G^4\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^4 + 48G^3\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^3 + \\
&\qquad 1668G^2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 \\
&\qquad + 2448G\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\| + 18784 \Big) \\
&\leq \frac{G^2}{\kappa T} (4G\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\| + 20)^4.
\end{aligned}
$$

This complete the proof of assertion (iii) in Case I.

Case II: Assume now that $4G^2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + 6G\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\| > \kappa\sqrt{T}/(8G^2)$, where G is defined as before. Since h has bounded gradients, the inequality (2.56) remains valid. Setting the step size to $\gamma = 1/(2G^2\sqrt{T})$ and using the trivial inequalities $G \geq R + \bar{\varepsilon} \geq R$, we thus obtain

$$
\begin{aligned}
\mathbb{E}\left[h\left(\frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\phi}_{t-1} \right) \right] - h(\boldsymbol{\phi}^\star) &\leq \frac{G^2}{\sqrt{T}} \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \frac{1}{4\sqrt{T}} + \\
&\qquad \frac{\bar{\varepsilon}}{\sqrt{T}} \sqrt{2\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \frac{37}{2G^2}} \\
&\leq \frac{G^2}{\sqrt{T}} \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\|^2 + \frac{2G}{\sqrt{T}} \|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\| + \frac{5}{\sqrt{T}},
\end{aligned}
$$

where the second inequality holds because $G \geq \bar{\varepsilon}$ and $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for all $a, b \geq 0$. Multiplying the right hand side of the last inequality by $G^2(32G^2\|\boldsymbol{\phi}_0^\star - \boldsymbol{\phi}^\star\|^2 + 48G\|\boldsymbol{\phi}_0^\star - \boldsymbol{\phi}^\star\|)/(\kappa\sqrt{T})$, which is strictly larger than 1 by the basic assumption underlying Case II, we then find

$$\begin{aligned}&\mathbb{E}\left[h\left(\frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\phi}_{t-1}\right)\right]-h(\boldsymbol{\phi}^\star)\\ &\leq \frac{G^2}{\kappa T}\left(G^2\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|^2+2G\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|+5\right)\left(32G^2\|\boldsymbol{\phi}_0^\star-\boldsymbol{\phi}^\star\|^2+48G\|\boldsymbol{\phi}_0^\star-\boldsymbol{\phi}^\star\|\right)\\ &\leq \frac{G^2}{\kappa T}(4G\|\boldsymbol{\phi}_0-\boldsymbol{\phi}^\star\|+20)^4.\end{aligned}$$

This observation completes the proof. □

2.4.2. Smooth Optimal Transport Problems with Marginal Ambiguity Sets

The smooth optimal transport problem (2.12) can be viewed as an instance of a stochastic optimization problem, that is, a convex maximization problem akin to (2.37), where the objective function is representable as $h(\boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{x}\sim\mu}[\boldsymbol{\nu}^\top\boldsymbol{\phi}-\overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x})]$. Throughout this section we assume that the smooth (discrete) c-transform $\overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x})$ defined in (2.11) is induced by a marginal ambiguity set of the form (2.26) with continuous marginal distribution functions. By Proposition 2.3.6, the integrand $\boldsymbol{\nu}^\top\boldsymbol{\phi}-\overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x})$ is therefore concave and differentiable in $\boldsymbol{\phi}$. We also assume that $\overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x})$ is μ-integrable in $\boldsymbol{x}$, that we have access to an oracle that generates independent samples from μ and that problem (2.12) is solvable.

The following proposition establishes several useful properties of the smooth c-transform.

Proposition 2.4.6 (Properties of the smooth c-transform)**.** *If Θ is a marginal ambiguity set of the form* (2.26) *with cumulative distribution functions F_i, $i\in[N]$, then $\overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x})$ has the following properties for all $\boldsymbol{x}\in\mathcal{X}$.*

(i) ***Bounded gradient:*** *If F_i, $i\in[N]$, are continuous, then we have*

$$\|\nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x})\|\leq 1$$

for all $\boldsymbol{\phi}\in\mathbb{R}^N$.

(ii) ***Lipschitz continuous gradient:*** *If F_i, $i\in[N]$, are Lipschitz continuous with Lipschitz constant $L>0$, then $\overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x})$ is L-smooth with respect to $\boldsymbol{\phi}$ in the sense of Assumption 1 (iv).*

(iii) ***Generalized self-concordance:*** *If F_i, $i\in[N]$, are twice differentiable on the interiors of their respective supports and if there is $M>0$ with*

$$\sup_{s\in F_i^{-1}(0,1)}\frac{|\mathrm{d}^2F_i(s)/\mathrm{d}s^2|}{\mathrm{d}F_i(s)/\mathrm{d}s}\leq M, \tag{2.71}$$

then $\overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x})$ is M-generalized self-concordant with respect to $\boldsymbol{\phi}$ in the sense of Assumption 1 (iii).

Proof. As for (i), Proposition 2.3.6 implies that $\nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x}) \in \Delta^N$, and thus we have $\|\nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x})\| \leq 1$. As for (ii), note that the convex conjugate of the smooth c-transform with respect to $\boldsymbol{\phi}$ is given by

$$\begin{aligned}
\overline{\psi}_c^*(\boldsymbol{p},\boldsymbol{x}) &= \sup_{\boldsymbol{\phi}\in\mathbb{R}^N} \boldsymbol{p}^\top\boldsymbol{\phi} - \overline{\psi}(\boldsymbol{\phi},\boldsymbol{x}) \\
&= \sup_{\boldsymbol{\phi}\in\mathbb{R}^N} \inf_{\boldsymbol{q}\in\Delta^N} \sum_{i=1}^N p_i\phi_i - (\phi_i - c(\boldsymbol{x},\boldsymbol{y_i}))q_i - \int_{1-q_i}^1 F_i^{-1}(t)\mathrm{d}t \\
&= \inf_{\boldsymbol{q}\in\Delta^N} \sup_{\boldsymbol{\phi}\in\mathbb{R}^N} \sum_{i=1}^N p_i\phi_i - (\phi_i - c(\boldsymbol{x},\boldsymbol{y_i}))q_i - \int_{1-q_i}^1 F_i^{-1}(t)\mathrm{d}t \\
&= \begin{cases} \displaystyle\sum_{i=1}^N c(\boldsymbol{x},\boldsymbol{y_i})p_i - \int_{1-p_i}^1 F_i^{-1}(t)\mathrm{d}t & \text{if } \boldsymbol{p}\in\Delta^N \\ +\infty & \text{otherwise,} \end{cases}
\end{aligned}$$

where the second equality follows again from Proposition 2.3.6, and the interchange of the infimum and the supremum is allowed by Sion's classical minimax theorem. In the following we first prove that $\overline{\psi}_c^*(\boldsymbol{p},\boldsymbol{x})$ is $1/L$-strongly convex in $\boldsymbol{p}$, that is, the function $\overline{\psi}_c^*(\boldsymbol{p},\boldsymbol{x}) - \|\boldsymbol{p}\|^2/(2L)$ is convex in $\boldsymbol{p}$ for any fixed $\boldsymbol{x}\in\mathcal{X}$. To this end, recall that F_i is assumed to be Lipschitz continuous with Lipschitz constant L. Thus, we have

$$L \geq \sup_{\substack{s_1,s_2\in\mathbb{R}\\ s_1\neq s_2}} \frac{|F_i(s_1) - F_i(s_2)|}{|s_1 - s_2|} = \sup_{\substack{s_1,s_2\in\mathbb{R}\\ s_1> s_2}} \frac{F_i(s_1) - F_i(s_2)}{s_1 - s_2} \geq \sup_{\substack{p_i,q_i\in(0,1)\\ p_i>q_i}} \frac{p_i - q_i}{F_i^{-1}(p_i) - F_i^{-1}(q_i)},$$

where the second inequality follows from restricting s_1 and s_2 to the preimage of $(0,1)$ with respect to F_i. Rearranging terms in the above inequality then yields

$$-F_i^{-1}(1-q_i) - q_i/L \leq -F_i^{-1}(1-p_i) - p_i/L$$

for all $p_i, q_i \in (0,1)$ with $q_i < p_i$. Consequently, the function $-F_i^{-1}(1-p_i) - p_i/L$ is non-decreasing and its primitive $-\int_{1-p_i}^1 F_i^{-1}(t)\mathrm{d}t - p_i^2/(2L)$ is convex in p_i on the interval $(0,1)$. This implies that

$$\overline{\psi}_c^*(\boldsymbol{p},\boldsymbol{x}) - \frac{\|\boldsymbol{p}\|_2^2}{2L} = \sum_{i=1}^N c(\boldsymbol{x},\boldsymbol{y_i})p_i - \int_{1-p_i}^1 F_i^{-1}(t)\mathrm{d}t - \frac{p_i^2}{2L}$$

constitutes a sum of convex univariate functions for every fixed $\boldsymbol{x}\in\mathcal{X}$. Thus, $\overline{\psi}_c^*(\boldsymbol{p},\boldsymbol{x})$ is $1/L$-strongly convex in $\boldsymbol{p}$. By [KSST09, Theorem 6], however, any convex function whose conjugate is $1/L$-strongly convex is guaranteed to be L-smooth. This observation completes the proof of assertion (ii). As for assertion (iii), choose any $\boldsymbol{\phi}, \boldsymbol{\varphi} \in \mathbb{R}^N$ and $\boldsymbol{x} \in \mathcal{X}$, and introduce the auxiliary function

$$\begin{aligned} u(s) &= \overline{\psi}_c\left(\boldsymbol{\phi} + s(\boldsymbol{\varphi} - \boldsymbol{\phi}), \boldsymbol{x}\right) && (2.72)\\ &= \max_{\boldsymbol{p} \in \Delta^N} \sum_{i=1}^N \left(\phi_i + s(\varphi_i - \phi_i) - c(\boldsymbol{x}, \boldsymbol{y_i})\right)p_i + \int_{1-p_i}^1 F_i^{-1}(t)\mathrm{d}t. && (2.73) \end{aligned}$$

For ease of exposition, in the remainder of the proof we use prime symbols to designate derivatives of univariate functions. A direct calculation then yields

$$\begin{aligned} u'(s) &= (\boldsymbol{\varphi} - \boldsymbol{\phi})^\top \nabla_{\boldsymbol{\phi}} \overline{\psi}\left(\boldsymbol{\phi} + s(\boldsymbol{\varphi} - \boldsymbol{\phi}), \boldsymbol{x}\right) \quad \text{and}\\ u''(s) &= (\boldsymbol{\varphi} - \boldsymbol{\phi})^\top \nabla^2_{\boldsymbol{\phi}} \overline{\psi}\left(\boldsymbol{\phi} + s(\boldsymbol{\varphi} - \boldsymbol{\phi}), \boldsymbol{x}\right)(\boldsymbol{\varphi} - \boldsymbol{\phi}). \end{aligned}$$

By Proposition 2.3.6, $\boldsymbol{p}^\star(s) = \nabla_{\boldsymbol{\phi}} \overline{\psi}_c\left(\boldsymbol{\phi} + s(\boldsymbol{\varphi} - \boldsymbol{\phi}), \boldsymbol{x}\right)$ represents the unique solution of the maximization problem in (2.72). In addition, by [STD19, Proposition 6], the Hessian of the smooth c-transform with respect to $\boldsymbol{\phi}$ can be computed from the Hessian of its convex conjugate as follows.

$$\begin{aligned} \nabla^2_{\boldsymbol{\phi}} \overline{\psi}_c\left(\boldsymbol{\phi} + s(\boldsymbol{\varphi} - \boldsymbol{\phi}), \boldsymbol{x}\right) &= \left(\nabla^2_{\boldsymbol{p}} \overline{\psi}_c^*(\boldsymbol{p}^\star(s), \boldsymbol{x})\right)^{-1}\\ &= \mathrm{diag}\left([F_1'(F_1^{-1}(1 - p_1^\star(s))), \ldots, F_N'(F_N^{-1}(1 - p_N^\star(s)))]\right) \end{aligned}$$

Hence, the first two derivatives of the auxiliary function $u(s)$ simplify to

$$u'(s) = \sum_{i=1}^N (\varphi_i - \phi_i)p_i^\star(s) \quad \text{and} \quad u''(s) = \sum_{i=1}^N (\varphi_i - \phi_i)^2 F_i'(F_i^{-1}(1 - p_i^\star(s))).$$

Similarly, the above formula for the Hessian of the smooth c-transform can be used to show that $(p_i^\star)'(s) = (\varphi_i - \phi_i)F_i'(F_i^{-1}(1 - p_i^\star(s)))$ for all $i \in [N]$. The third derivative of $u(s)$ therefore simplifies to

$$\begin{aligned} u'''(s) &= -\sum_{i=1}^N (\varphi_i - \phi_i)^2 \frac{F_i''(F_i^{-1}(1 - p_i^\star(s)))}{F_i'(F_i^{-1}(1 - p_i^\star(s)))} (p_i^\star)'(s)\\ &= -\sum_{i=1}^N (\varphi_i - \phi_i)^3 F_i''(F_i^{-1}(1 - p_i^\star(s))). \end{aligned}$$

This implies via Hölder's inequality that

$$|u'''(s)| = \left| \sum_{i=1}^N (\varphi_i - \phi_i)^2 \, F_i'(F_i^{-1}(1 - p_i^\star(s))) \, \frac{F_i''(F_i^{-1}(1 - p_i^\star(s)))}{F_i'(F_i^{-1}(1 - p_i^\star(s)))} \, (\varphi_i - \phi_i) \right|$$

$$\leq \left(\sum_{i=1}^{N} (\varphi_i - \phi_i)^2 \, F_i'(F_i^{-1}(1 - p_i^\star(s))) \right) \left(\max_{i \in [N]} \left| \frac{F_i''(F_i^{-1}(1 - p_i^\star(s)))}{F_i'(F_i^{-1}(1 - p_i^\star(s)))} \, (\varphi_i - \phi_i) \right| \right).$$

Notice that the first term in the above expression coincides with $u''(s)$, and the second term satisfies

$$\begin{aligned} \max_{i \in [N]} \left| \frac{F_i''(F_i^{-1}(1 - p_i^\star(s)))}{F_i'(F_i^{-1}(1 - p_i^\star(s)))} \, (\varphi_i - \phi_i) \right| &\leq \max_{i \in [N]} \left| \frac{F_i''(F_i^{-1}(1 - p_i^\star(s)))}{F_i'(F_i^{-1}(1 - p_i^\star(s)))} \right| \, \|\boldsymbol{\varphi} - \boldsymbol{\phi}\|_\infty \\ &\leq M \|\boldsymbol{\varphi} - \boldsymbol{\phi}\|, \end{aligned}$$

where the first inequality holds because $\max_{i \in [N]} |a_i b_i| \leq \|\boldsymbol{a}\|_\infty \|\boldsymbol{b}\|_\infty$ for all $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^N$, and the second inequality follows from the definition of M and the fact that the 2-norm provides an upper bound on the ∞-norm. Combining the above results shows that $|u'''(s)| \leq M \|\boldsymbol{\varphi} - \boldsymbol{\phi}\| u''(s)$ for all $s \in \mathbb{R}$. The claim now follows because $\boldsymbol{\phi}, \boldsymbol{\varphi} \in \mathbb{R}^N$ and $\boldsymbol{x} \in \mathcal{X}$ were chosen arbitrarily. □

In the following we use the averaged SGD algorithm of Section 2.4.1 to solve the smooth optimal transport problem (2.12). A detailed description of this algorithm in pseudocode is provided in Algorithm 3. This algorithm repeatedly calls a sub-routine for estimating the gradient of $\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x})$ with respect to $\boldsymbol{\phi}$. By Proposition 2.3.6, this gradient coincides with the unique solution $\boldsymbol{p}^\star$ of the convex maximization problem (2.27). In addition, from the proof of Proposition 2.3.6 it is clear that its components are given by

$$p_i^\star = \theta^\star \left[i = \min \operatorname*{argmax}_{j \in [N]} \phi_j - c(\boldsymbol{x}, \boldsymbol{y}_j) + z_j \right] \quad \forall i \in [N],$$

where $\theta^\star$ represents an optimizer of the semi-parametric discrete choice problem (2.11). Therefore, $\boldsymbol{p}^\star$ can be interpreted as a vector of choice probabilities under the best-case probability measure $\theta^\star$. Sometimes these choice probabilities are available in closed form. This is the case, for instance, in the exponential distribution model of Example 2.3.8, which is equivalent to the generalized extreme value distribution model of Section 2.3.1. Indeed, in this case $\boldsymbol{p}^\star$ is given by a softmax of the utility values $\phi_i - c(\boldsymbol{x}, \boldsymbol{y_i})$, $i \in [N]$, *i.e.*,

$$p_i^\star = \frac{\eta_i \exp\left((\phi_i - c(\boldsymbol{x}, \boldsymbol{y_i}))/\lambda\right)}{\sum_{j=1}^N \eta_j \exp\left((\phi_j - c(\boldsymbol{x}, \boldsymbol{y_j}))/\lambda\right)} \quad \forall i \in [N]. \tag{2.74}$$

Note that these particular choice probabilities are routinely studied in the celebrated multinomial logit choice model [BAL85, § 5.1]. The choice probabilities are also available in closed form in the uniform distribution model of Example 2.3.9. As the derivation of $\boldsymbol{p}^\star$ is somewhat cumbersome in this case, we relegate it to Appendix 2.6. For general marginal ambiguity sets with continuous marginal distribution functions, we propose a bisection method to compute the gradient of the smooth c-transform numerically up to any prescribed accuracy; see Algorithm 4.

Theorem 2.4.7 (Biased gradient oracle)**.** *If Θ is a marginal ambiguity set of the form* (2.26) *and the cumulative distribution function F_i is continuous for every $i \in [N]$, then, for any $\boldsymbol{x} \in \mathcal{X}$, $\boldsymbol{\phi} \in \mathbb{R}^N$ and $\varepsilon > 0$, Algorithm 4 outputs $\boldsymbol{p} \in \mathbb{R}^N$ with $\|\boldsymbol{p}\| \leq 1$ and $\|\nabla_{\boldsymbol{\phi}} \overline{\psi}(\boldsymbol{\phi}, \boldsymbol{x}) - \boldsymbol{p}\| \leq \varepsilon$.*

Proof. Thanks to Proposition 2.3.6, we can recast the smooth c-transform in dual form as

$$\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x}) = \min_{\substack{\boldsymbol{\zeta} \in \mathbb{R}_+^N \\ \tau \in \mathbb{R}}} \sup_{\boldsymbol{p} \in \mathbb{R}^N} \sum_{i=1}^N (\phi_i - c(\boldsymbol{x}, \boldsymbol{y_i}))p_i + \\ \sum_{i=1}^N \int_{1-p_i}^1 F_i^{-1}(t)\mathrm{d}t + \tau \left(\sum_{i=1}^N p_i - 1 \right) + \sum_{i=1}^N \zeta_i p_i.$$

Strong duality and dual solvability hold because we may construct a Slater point for the primal problem by setting $p_i = 1/N$, $i \in [N]$. By the Karush-Kuhn-Tucker optimality conditions, $\boldsymbol{p}^\star$ and $(\tau^\star, \boldsymbol{\zeta}^\star)$ are therefore optimal in the primal and dual problems, respectively, if and only if we have

$$\begin{array}{lll} \sum_{i=1}^N p_i^\star = 1,\ p_i^\star \geq 0 & \forall i \in [N] & \text{(primal feasibility)} \\ \zeta_i^\star \geq 0 & \forall i \in [N] & \text{(dual feasibility)} \\ \zeta_i^\star p_i^\star = 0 & \forall i \in [N] & \text{(complementary slackness)} \\ \phi_i - c(\boldsymbol{x}, \boldsymbol{y_i}) + F_i^{-1}(1-p_i^\star) + \tau^\star + \zeta_i^\star = 0 & \forall i \in [N] & \text{(stationarity).} \end{array}$$

If $p_i^\star > 0$, then the complementary slackness and stationarity conditions imply that $\zeta_i^\star = 0$ and that $\phi_i - c(\boldsymbol{x}, \boldsymbol{y_i}) + F_i^{-1}(1 - p_i^\star) + \tau^\star = 0$, respectively. Thus, we have $p_i^\star = 1 - F_i(c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \tau^\star)$. If $p_i^\star = 0$, on the other hand, then similar arguments show that $\zeta_i^\star \geq 0$ and $\phi_i - c(\boldsymbol{x}, \boldsymbol{y_i}) + F_i^{-1}(1) + \tau^\star \leq 0$. These two inequalities are equivalent to $1 - F_i(c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \tau^\star) \leq 0$. As all values of F_i are smaller or equal to 1, the last equality must in in fact hold as an equality. Combining the insights gained so far thus yields $p_i^\star = 1 - F_i(c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \tau^\star)$, which holds for all $i \in [N]$ irrespective of the sign of $p_i^\star$. Primal feasibility therefore ensures that $\sum_{i=1}^N 1 - F_i(c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \tau^\star) = 1$. Finding the unique optimizer $\boldsymbol{p}^\star$ of (2.27) (*i.e.*, finding the gradient of $\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x})$) is therefore tantamount to finding a root $\tau^\star$ of the univariate equation

$$\sum_{i=1}^N 1 - F_i(c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \tau) = 1. \tag{2.75}$$

Note the function on the left hand side of (2.75) is continuous and non-decreasing in τ because of the continuity (by assumption) and monotonicity (by definition) of the cumulative distribution functions F_i, $i \in [N]$. Hence, the root finding problem can be solved efficiently via bisection. To complete the proof, we first show that the interval between the constants $\underline{\tau}$ and $\overline{\tau}$ defined in Algorithm 4 is guaranteed to contain $\tau^\star$. Specifically, we will demonstrate that evaluating the function on the left hand side of (2.75) at $\underline{\tau}$ or $\overline{\tau}$ yields a number that is not larger or not smaller than 1, respectively. For $\tau = \underline{\tau}$ we have

$$\begin{aligned}1 - F_i(c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \underline{\tau}) = 1 - F_i\Bigg(& c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \\ & \min_{j\in[N]} \left\{ c\left(\boldsymbol{x}, \boldsymbol{y_j}\right) - \phi_j - F_j^{-1}(1-1/N) \right\} \Bigg) \\ \leq 1 - F_i\left(F_i^{-1}(1-1/N)\right) = 1/N \qquad & \forall i \in [N],\end{aligned}$$

where the inequality follows from the monotonicity of F_i. Summing the above inequality over all $i \in [N]$ then yields the desired inequality $\sum_{i=1}^N 1 - F_i(c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \underline{\tau}) \leq 1$. Similarly, for $\tau = \overline{\tau}$ we have

$$\begin{aligned}1 - F_i(c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \overline{\tau}) = 1 - F_i\Bigg(& c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \\ & \max_{j\in[N]} \left\{ c\left(\boldsymbol{x}, \boldsymbol{y_j}\right) - \phi_j - F_j^{-1}(1-1/N) \right\} \Bigg) \\ \geq 1 - F_i\left(F_i^{-1}(1-1/N)\right) = 1/N \qquad & \forall i \in [N].\end{aligned}$$

We may thus conclude that $\sum_{i=1}^N 1 - F_i(c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \overline{\tau}) \geq 1$. Therefore, $[\underline{\tau}, \overline{\tau}]$ constitutes a valid initial search interval for the bisection algorithm. Note that the function $1 - F_i(c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \tau)$, which defines p_i in terms of τ, is uniformly continuous in τ throughout $\mathbb{R}$. This follows from [Bil95, Problem 14.8] and our assumption that F_i is continuous. The uniform continuity ensures that the tolerance

$$\delta(\varepsilon) = \min_{i\in N} \left\{ \max_{\delta} \left\{ \delta : |F_i(t_1) - F_i(t_2)| \leq \varepsilon/\sqrt{N} \;\; \forall t_1, t_2 \in \mathbb{R} \text{ with } |t_1 - t_2| \leq \delta \right\} \right\} \tag{2.76}$$

is strictly positive for every $\varepsilon > 0$. As the length of the search interval is halved in each iteration, Algorithm 4 outputs a near optimal solution τ with $|\tau - \tau^\star| \leq \delta(\varepsilon)$ after $\lceil \log_2((\overline{\tau} - \underline{\tau})/\delta(\varepsilon)) \rceil$ iterations. Moreover, the construction of $\delta(\varepsilon)$ guarantees that $|1 - F_i(c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \tau) - p_i^\star| \leq \varepsilon/\sqrt{N}$ for all τ with $|\tau - \tau^\star| \leq \delta(\varepsilon)$. Therefore, the output $\boldsymbol{p} \in \mathbb{R}_+^N$ of Algorithm 4 satisfies $|p_i - p_i^\star| \leq \varepsilon/\sqrt{N}$ for each $i \in [N]$, which in turn implies that $\|\boldsymbol{p} - \boldsymbol{p}^\star\| \leq \varepsilon$. By construction, finally, Algorithm 4 outputs $\boldsymbol{p} \geq \boldsymbol{0}$ with $\sum_{i \in [N]} p_i < 1$, which ensures that $\|p\| \leq 1$. Thus, the claim follows. □

If all cumulative distribution functions F_i, $i \in [N]$, are Lipschitz continuous with a common Lipschitz constant $L > 0$, then the uniform continuity parameter $\delta(\varepsilon)$ required in Algorithm 4 can simply be set to $\delta(\varepsilon) = \varepsilon/(L\sqrt{N})$. We are now ready to prove that Algorithm 3 offers different convergence guarantees depending on the continuity and smoothness properties of the marginal cumulative distribution functions.

Algorithm 3 Averaged SGD

```
Input: $\gamma, T, \bar{\varepsilon}$
Set $\boldsymbol{\phi}_0 \leftarrow \boldsymbol{0}$
for $t = 1, 2, \ldots, T$ do
    Sample $\boldsymbol{x_t}$ from $\mu$
    Choose $\varepsilon_{t-1} \in (0, \bar{\varepsilon}/(2\sqrt{t})]$
    Set $\boldsymbol{p} \leftarrow$ Bisection$(\boldsymbol{x_t}, \boldsymbol{\phi}_{t-1}, \varepsilon_{t-1})$
    Set $\boldsymbol{\phi}_t \leftarrow \boldsymbol{\phi}_{t-1} + \gamma(\boldsymbol{\nu} - \boldsymbol{p})$
end for
Output: $\underline{\boldsymbol{\phi}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\phi}_{t-1}$ and $\bar{\boldsymbol{\phi}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\phi}_t$
```

Algorithm 4 Bisection method to approximate $\nabla_{\boldsymbol{\phi}} \overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x})$

```
Input: $\boldsymbol{x}, \boldsymbol{\phi}, \varepsilon$
Set $\overline{\tau} \leftarrow \max_{i \in [N]} \; \{c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - F_i^{-1}(1 - 1/N)\}$
Set $\underline{\tau} \leftarrow \min_{i \in [N]} \; \{c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - F_i^{-1}(1 - 1/N)\}$
Evaluate $\delta(\varepsilon)$ as defined in (2.76)
for $k = 1, 2, \ldots, \lceil \log_2((\overline{\tau} - \underline{\tau})/\delta(\varepsilon)) \rceil$ do
    Set $\tau \leftarrow (\overline{\tau} + \underline{\tau})/2$
    Set $p_i \leftarrow 1 - F_i(c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \tau)$ for $i \in [N]$
    if $\sum_{i \in [N]} p_i > 1$ then $\overline{\tau} \leftarrow \tau$ else $\underline{\tau} \leftarrow \tau$
end for
Output: $\boldsymbol{p}$ with $p_i = 1 - F_i(c(\boldsymbol{x}, \boldsymbol{y_i}) - \phi_i - \underline{\tau})$, $i \in [N]$
```

Corollary 6. *Use $h(\boldsymbol{\phi}) = \mathbb{E}_{\boldsymbol{x} \sim \mu}[\boldsymbol{\nu}^\top \boldsymbol{\phi} - \overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x})]$ as a shorthand for the objective function of the smooth optimal transport problem (2.12), and let $\boldsymbol{\phi}^\star$ be a* *maximizer of* (2.12)*. If* Θ *is a marginal ambiguity set of the form* (2.26) *with distribution functions* F_i*,* $i \in [N]$*, then for any* $T \in \mathbb{N}$ *and* $\bar{\varepsilon} \geq 0$*, the outputs* $\underline{\boldsymbol{\phi}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\phi}_{t-1}$ *and* $\bar{\boldsymbol{\phi}}_T = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{\phi}_t$ *of Algorithm* 3 *satisfy the following inequalities.*

(i) If $\gamma = 1/(2(2+\bar{\varepsilon})\sqrt{T})$ *and* F_i *is continuous for every* $i \in [N]$*, then we have*

$$\overline{W}_c(\mu,\nu) - \mathbb{E}\left[h\big(\underline{\boldsymbol{\phi}}_T\big)\right] \leq \frac{(2+\bar{\varepsilon})^2}{\sqrt{T}}\|\boldsymbol{\phi}^\star\|^2 + \frac{1}{4\sqrt{T}} + \frac{\bar{\varepsilon}}{\sqrt{T}}\sqrt{2\|\boldsymbol{\phi}^\star\|^2 + \frac{37}{2(2+\bar{\varepsilon})^2}}.$$

(ii) If $\gamma = 1/(2\sqrt{T}+L)$ *and* F_i *is Lipschitz continuous with Lipschitz constant* $L > 0$ *for every* $i \in [N]$*, then we have*

$$\overline{W}_c(\mu,\nu) - \mathbb{E}\left[h\big(\bar{\boldsymbol{\phi}}_T\big)\right] \leq \frac{L}{2T}\|\boldsymbol{\phi}^\star\|^2 + \frac{(2+\bar{\varepsilon})^2}{\sqrt{T}}\|\boldsymbol{\phi}^\star\|^2 + \frac{\bar{\varepsilon}^2+2}{4(2+\bar{\varepsilon})^2\sqrt{T}} + \frac{\bar{\varepsilon}}{\sqrt{T}}\sqrt{2\|\boldsymbol{\phi}^\star\|^2 + \frac{37}{2(2+\bar{\varepsilon})^2}}.$$

(iii) If $\gamma = 1/(2G^2\sqrt{T})$ *with* $G = \max\{M, 2+\bar{\varepsilon}\}$*,* F_i *satisfies the generalized self-concordance condition* (2.71) *with* $M > 0$ *for every* $i \in [N]$*, and the smallest eigenvalue* κ *of* $-\nabla^2_{\boldsymbol{\phi}} h(\boldsymbol{\phi}^\star)$ *is strictly positive, then we have*

$$\overline{W}_c(\mu,\nu) - \mathbb{E}\left[h\big(\underline{\boldsymbol{\phi}}_T\big)\right] \leq \frac{G^2}{T\kappa}\left(4G\|\boldsymbol{\phi}_0 - \boldsymbol{\phi}^\star\| + 20\right)^4.$$

Proof. Recall that problem (2.12) can be viewed as an instance of the convex minimization problem (2.37) provided that its objective function is inverted. Throughout the proof we denote by $\boldsymbol{p}_t(\boldsymbol{\phi}_t, \boldsymbol{x}_t)$ the inexact estimate for $\nabla_{\boldsymbol{\phi}}\overline{\psi}(\boldsymbol{\phi}_t, \boldsymbol{x}_t)$ output by Algorithm 4 in iteration t of the averaged SGD algorithm. Note that

$$\begin{aligned}
&\left\|\mathbb{E}\left[\boldsymbol{\nu} - \boldsymbol{p}_t(\boldsymbol{\phi}_{t-1}, \boldsymbol{x}_t)\middle|\mathcal{F}_{t-1}\right] - \nabla h(\boldsymbol{\phi}_{t-1})\right\| \\
&= \left\|\mathbb{E}\left[\boldsymbol{p}_t(\boldsymbol{\phi}_{t-1}, \boldsymbol{x}_t) - \nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi}_{t-1}, \boldsymbol{x}_t)\right]\right\| \\
&\leq \mathbb{E}\left[\left\|\boldsymbol{p}_t(\boldsymbol{\phi}_{t-1}, \boldsymbol{x}_t) - \nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi}_{t-1}, \boldsymbol{x}_t)\right\|\right] \\
&\leq \varepsilon_{t-1} \leq \frac{\bar{\varepsilon}}{2\sqrt{t}},
\end{aligned}$$

where the two inequalities follow from Jensen's inequality and the choice of ε_{t-1} in Algorithm 3, respectively. The triangle inequality and Proposition 2.4.6 (i) further imply that

$$\|\nabla h(\boldsymbol{\phi})\| = \mathbb{E}\left[\left\|\boldsymbol{\nu} - \nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x})\right\|\right] \leq \|\boldsymbol{\nu}\| + \mathbb{E}\left[\left\|\nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x})\right\|\right] \leq 2.$$

Assertion (i) thus follows from Theorem 2.4.4 (i) with $R = 2$. As for assertion (ii), we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|\boldsymbol{\nu} - \boldsymbol{p}_t(\boldsymbol{\phi}_{t-1},\boldsymbol{x}_t) - \nabla h(\boldsymbol{\phi}_{t-1})\right\|^2 |\mathcal{F}_{t-1}\right]\\
&= \mathbb{E}\left[\left\|\boldsymbol{p}_t(\boldsymbol{\phi}_{t-1},\boldsymbol{x}_t) - \mathbb{E}\left[\nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi}_{t-1},\boldsymbol{x})\right]\right\|^2 |\mathcal{F}_{t-1}\right]\\
&= \mathbb{E}\Bigg[\Bigg\|\boldsymbol{p}_t(\boldsymbol{\phi}_{t-1},\boldsymbol{x}_t) - \nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi}_{t-1},\boldsymbol{x}) +\\
&\qquad \nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi}_{t-1},\boldsymbol{x}) - \mathbb{E}\left[\nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi}_{t-1},\boldsymbol{x})\right]\Bigg\|^2\Bigg|\mathcal{F}_{t-1}\Bigg]\\
&\leq \mathbb{E}\Bigg[2\Big\|\boldsymbol{p}_t(\boldsymbol{\phi}_{t-1},\boldsymbol{x}_t) - \nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi}_{t-1},\boldsymbol{x})\Big\|^2 +\\
&\qquad 2\Big\|\nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi}_{t-1},\boldsymbol{x}) - \mathbb{E}\Big[\nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi}_{t-1},\boldsymbol{x})\Big]\Big\|^2\Bigg|\mathcal{F}_{t-1}\Bigg]\\
&\leq 2\varepsilon_{t-1}^2 + 2 \leq \bar{\varepsilon}^2 + 2,
\end{aligned}$$

where the second inequality holds because $\nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi}_{t-1},\boldsymbol{x}) \in \Delta^N$ and because $\|\nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi}_{t-1},\boldsymbol{x})\|_2^2 \leq 1$, while the last inequality follows from the choice of ε_{t-1} in Algorithm 1. As $\overline{\psi}(\boldsymbol{\phi},\boldsymbol{x})$ is L-smooth with respect to $\boldsymbol{\phi}$ by virtue of Proposition 2.4.6 (ii), we further have

$$\begin{aligned}
\|\nabla h(\boldsymbol{\phi}) - \nabla h(\boldsymbol{\phi}')\| &= \left\|\mathbb{E}\left[\nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi},\boldsymbol{x}) - \nabla_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi}',\boldsymbol{x})\right]\right\|\\
&\leq L\|\boldsymbol{\phi} - \boldsymbol{\phi}'\| \ \forall \boldsymbol{\phi},\boldsymbol{\phi}' \in \mathbb{R}^n.
\end{aligned}$$

Assertion (ii) thus follows from Theorem 2.4.4 (ii) with $R = 2$ and $\sigma = \sqrt{\bar{\varepsilon}^2 + 2}$. As for assertion (iii), finally, we observe that h is M-generalized self-concordant thanks to Proposition 2.4.6 (iii). Then, assertion (iii) thus follows from Theorem 2.4.4 (iii) with $R = 2$. □

One can show that the objective function of the smooth optimal transport problem (2.12) with marginal exponential noise distributions as described in Example 2.3.8 is generalized self-concordant. Hence, the convergence rate of Algorithm 3 for the exponential distribution model of Example 2.3.8 is of the order $\mathcal{O}(1/T)$, which improves the state-of-the-art $\mathcal{O}(1/\sqrt{T})$ guarantee established by [Gen+16].

2.5. Numerical Experiments

All experiments are run on a 2.6 GHz 6-Core Intel Core i7, and all optimization problems are implemented in MATLAB R2020a. The corresponding codes are available at https://github.com/RAO-EPFL/Semi-Discrete-Smooth-OT.git.

We now aim to assess the empirical convergence behavior of Algorithm 3 and to showcase the effects of regularization in semi-discrete optimal transport. To this end, we solve the original dual optimal transport problem (2.10) as well as its smooth variant (2.12) with a Fréchet ambiguity set corresponding to the exponential distribution model of Example 2.3.8, to the uniform distribution model of Example 2.3.9 and to the hyperbolic cosine distribution model of Example 2.3.11. Recall from Theorem 2.3.7 that any Fréchet ambiguity set is uniquely determined by a marginal generating function F and a probability vector $\boldsymbol{\eta}$. As for the exponential distribution model of Example 2.3.8, we set $F(s) = \exp(10s - 1)$ and $\eta_i = 1/N$ for all $i \in [N]$. In this case problem (2.12) is equivalent to the regularized primal optimal transport problem (2.13) with an entropic regularizer, and the gradient $\nabla_{\boldsymbol{\phi}} \overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x})$, which is known to coincide with the vector $\boldsymbol{p}^\star$ of optimal choice probabilities in problem (2.27), admits the closed-form representation (2.74). We can therefore solve problem (2.12) with a variant of Algorithm 3 that calculates $\nabla_{\boldsymbol{\phi}} \overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x})$ exactly instead of approximately via bisection. As for the uniform distribution model of Example 2.3.9, we set $F(s) = s/20 + 1/2$ and $\eta_i = 1/N$ for all $i \in [N]$. In this case problem (2.12) is equivalent to the regularized primal optimal transport problem (2.13) with a χ^2-divergence regularizer, and the vector $\boldsymbol{p}^\star$ of optimal choice probabilities can be computed exactly and highly efficiently by sorting thanks to Proposition 2.6.1 in the appendix. We can therefore again solve problem (2.12) with a variant of Algorithm 3 that calculates $\nabla_{\boldsymbol{\phi}} \overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x})$ exactly. As for the hyperbolic cosine model of Example 2.3.11, we set $F(s) = \sinh(10s - k)$ with $k = \sqrt{2} - 1 - \operatorname{arcsinh}(1)$ and $\eta_i = 1/N$ for all $i \in [N]$. In this case problem (2.12) is equivalent to the regularized primal optimal transport problem (2.13) with a hyperbolic divergence regularizer. However, the vector $\boldsymbol{p}^\star$ is not available in closed form, and thus we use Algorithm 4 to compute $\boldsymbol{p}^\star$ approximately. Lastly, note that the original dual optimal transport problem (2.10) can be interpreted as an instance of (2.12) equipped with a degenerate singleton ambiguity set that only contains the Dirac measure at the origin of $\mathbb{R}^N$. In this case $\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x}) = \psi_c(\boldsymbol{\phi}, \boldsymbol{x})$ fails to be smooth in $\boldsymbol{\phi}$, but an exact subgradient $\boldsymbol{p}^\star \in \partial_{\boldsymbol{\phi}} \overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x})$ is given by

$$p_i^\star = \begin{cases} 1 & \text{if } i = \min \operatorname*{argmax}_{i \in [N]} \ \phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i), \\ 0 & \text{otherwise.} \end{cases}$$

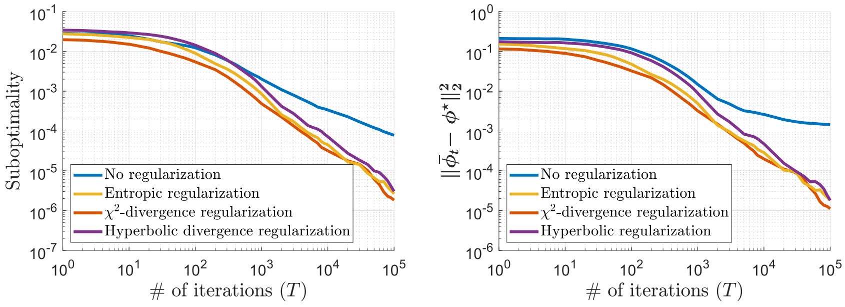
We can therefore solve problem (2.10) with a variant of Algorithm 3 that has access to exact subgradients (instead of gradients) of $\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x})$. Note that the maximizer $\boldsymbol{\phi}^\star$ of (2.10) may not be unique. In our experiments, we force Algorithm 3 to converge to the maximizer with minimal Euclidean norm by adding a vanishingly small Tikhonov regularization term to $\psi_c(\boldsymbol{\phi}, \boldsymbol{x})$. Thus, we set $\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x}) = \psi_c(\boldsymbol{\phi}, \boldsymbol{x}) + \varepsilon\|\boldsymbol{\phi}\|_2^2$ for some small regularization weight $\varepsilon > 0$, in which case $\boldsymbol{p}^\star + 2\varepsilon\boldsymbol{\phi} \in \partial_{\boldsymbol{\phi}}\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x})$ is an exact subgradient.

In the following we set μ to the standard Gaussian measure on $\mathcal{X} = \mathbb{R}^2$ and ν to the uniform measure on 10 independent samples drawn uniformly from $\mathcal{Y} = [-1, 1]^2$. We further set the transportation cost to $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x}-\boldsymbol{y}\|_\infty$. Under these assumptions, we use Algorithm 3 to solve the original as well as the three smooth optimal transport problems approximately for $T = 1, \dots, 10^5$. For each fixed T the step size is selected in accordance with Corollary 6. We emphasize that Corollary 6 (i) remains valid if $\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x})$ fails to be smooth in $\boldsymbol{\phi}$ and we have only access to subgradients; see [NV08, Corollary 1]. Denoting by $\bar{\boldsymbol{\phi}}_T$ the output of Algorithm 3, we record the suboptimality

$$\overline{W}_c(\mu, \nu) - \mathbb{E}_{\boldsymbol{x}\sim\mu}\left[\boldsymbol{\nu}^\top \bar{\boldsymbol{\phi}}_T - \overline{\psi}_c(\bar{\boldsymbol{\phi}}_T, \boldsymbol{x})\right]$$

of $\bar{\boldsymbol{\phi}}_T$ in (2.12) as well as the discrepancy $\|\bar{\boldsymbol{\phi}}_T - \boldsymbol{\phi}^\star\|_2^2$ of $\bar{\boldsymbol{\phi}}_T$ to the exact maximizer $\boldsymbol{\phi}^\star$ of problem (2.12) as a function of T. In order to faithfully measure the convergence rate of $\bar{\boldsymbol{\phi}}_T$ and its suboptimality, we need to compute $\boldsymbol{\phi}^\star$ as well as $\overline{W}_c(\mu, \nu)$ to within high accuracy. This is only possible if the dimension of $\mathcal{X}$ is small (*e.g.*, if $\mathcal{X} = \mathbb{R}^2$ as in our numerical example); even though Algorithm 3 can efficiently solve optimal transport problems in high dimensions. We obtain high-quality approximations for $\overline{W}_c(\mu, \nu)$ and $\boldsymbol{\phi}^\star$ by solving the finite-dimensional optimal transport problem between ν and the discrete distribution that places equal weight on $10 \times T$ samples drawn independently from μ. Note that only the first T of these samples are used by Algorithm 3. The proposed high-quality approximations of the entropic and χ^2-divergence regularized optimal transport problems are conveniently solved via Nesterov's accelerated gradient descent method, where the suboptimality gap of the t^{th} iterate is guaranteed to decay as $\mathcal{O}(1/t^2)$ under the step size rule advocated in [Nes83, Theorem 1]. To our best knowledge, Nesterov's accelerated gradient descent algorithm is not guaranteed to converge with inexact gradients. For the hyperbolic divergence regularized optimal transport problem, we thus use Algorithm 3 with $50 \times T$ iterations to obtain an approximation for $\overline{W}_c(\mu, \nu)$ and $\boldsymbol{\phi}^\star$. In contrast, we model the high-quality approximation of the original optimal transport problem (2.10) in YALMIP [L04] and solve it with MOSEK. If this problem has multiple maximizers, we report the one with minimal Euclidean norm.

Figure 2.1 shows how the suboptimality of $\bar{\boldsymbol{\phi}}_T$ and the discrepancy between $\bar{\boldsymbol{\phi}}_T$ and the exact maximizer decay with T, both for the original as well as for the entropic, the χ^2-divergence and hyperbolic divergence regularized optimal transport problems, averaged across 20 independent simulation runs. Figure 2.1a suggests that the suboptimality decays as $\mathcal{O}(1/\sqrt{T})$ for the original optimal transport problem, which is in line with the theoretical guarantees by [NV08, Corollary 1], and as $\mathcal{O}(1/T)$ for the entropic, the χ^2-divergence and the hyperbolic divergence regularized optimal transport problems, which is consistent with the theoretical guarantees established in Corollary 6. Indeed,



(a)

(b)

Figure 2.1: Suboptimality (a) and discrepancy to $\boldsymbol{\phi}^\star$ (b) of the outputs $\bar{\boldsymbol{\phi}}_T$ of Algorithm 3 for the original (blue), the entropic regularized (orange), the χ^2-divergence regularized (red) and the hyperbolic divergence regularized (purple) optimal transport problems.

entropic regularization can be explained by the exponential distribution model of Example 2.3.8, where the exponential distribution functions F_i satisfy the generalized self-concordance condition (2.71) with $M = 1/\lambda$. Similarly, χ^2-divergence regularization can be explained by the uniform distribution model of Example 2.3.9, where the uniform distribution functions F_i satisfy the generalized self-concordance condition with any $M > 0$. Finally, hyperbolic divergence regularization can be explained by the hyperbolic cosine distribution model of Example 2.3.11, where the hyperbolic cosine functions F_i satisfy the generalized self-concordance condition with $M = 1/\lambda$. In all cases the smallest eigenvalue of $-\nabla^2_{\boldsymbol{\phi}} \mathbb{E}_{\boldsymbol{x}\sim\mu}[\boldsymbol{\nu}^\top \boldsymbol{\phi}^\star - \overline{\psi}_c(\boldsymbol{\phi}^\star, \boldsymbol{x})]$, which we estimate when solving the high-quality approximations of the two smooth optimal transport problems, is strictly positive. Therefore, Corollary 6 (iii) is indeed applicable and guarantees that the suboptimality gap is bounded above by $\mathcal{O}(1/T)$. Finally, Figure 2.1b suggests that $\|\bar{\boldsymbol{\phi}}_T - \boldsymbol{\phi}^\star\|_2^2$ converges to 0 at rate $\mathcal{O}(1/T)$ for the entropic, the χ^2-divergence and the hyperbolic divergence regularized optimal transport problems, which is consistent with [Bac14, Proposition 10].

Appendix

Approximating the Minimizer of a Strictly Convex Function. The following lemma is key ingredient for the proofs of Theorem 2.2.2 and Corollary 4.

Lemma 2.5.1. *Assume that $g : [0,1] \to \mathbb{R}_+$ is a strictly convex function with unique minimizer $t^\star \in [0,1]$, and define $L = \lceil \log_2(1/\delta) \rceil + 1$ for some prescribed tolerance $\delta \in (0,1)$. Then, the following hold.*

(i) *Given an oracle that evaluates g exactly, we can compute a δ-approximation for $t^\star$ with $2L$ oracle calls.*

(ii) *Given an oracle that evaluates g inexactly to within an absolute accuracy*

$$\varepsilon = \frac{1}{4} \min_{l \in [2^L]} \{ |g(t_l) - g(t_{l-1})| : g(t_l) \neq g(t_{l-1}) \} \text{ with } t_l = \frac{l}{2^L} \; \forall l = 0, \ldots, 2^L,$$

we can compute a 2δ-approximation for $t^\star$ with $2L$ oracle calls.

Proof. Consider the uniform grid $\{t_0, \ldots, t_{2^L}\}$, and note that the difference 2^{-L} between consecutive grid points is strictly smaller than δ. Next, introduce a piecewise affine function $\bar{g} : [0, 1] \to \mathbb{R}_+$ that linearly interpolates g between consecutive grid points. By construction, $\bar{g}$ is affine on the interval $[t_{l-1}, t_l]$ with slope a_l/δ and $a_l = g(t_l) - g(t_{l-1})$ for all $l \in [2^L]$. In addition, $\bar{g}$ is continuous and inherits convexity from g. As g is strictly convex, it is easy to verify that $\bar{g}$ has also a kink at every inner grid point t_l for $l \in [2^L - 1]$, and therefore the distance between the unique minimizer $t^\star$ of g and any minimizer of $\bar{g}$ is at most $2^{-L} < \delta$. In order to compute a δ-approximation for $t^\star$, it thus suffices to find a minimizer of $\bar{g}$.

Next, define $\boldsymbol{a} = (a_0, \ldots, a_{2^L})$ with $a_0 = -\infty$. As $\bar{g}$ has a kink at every inner grid point, we may conclude that the array $\boldsymbol{a}$ is sorted in ascending order, that is, $a_l > a_{l-1}$ for all $l \in [2^L]$. This implies that at most one element of $\boldsymbol{a}$ can vanish. In the following, define $l^\star = \max\{l \in \{0\} \cup [2^L] : a_l \leq 0\}$. If $l^\star = 0$, then $\bar{g}$ is uniquely minimized by $t_{l^\star} = 0$, and $t^\star$ must fall within the interval $[t_0, t_1]$. If $l^\star > 0$ and $a_{l^\star} < 0$, on the other hand, then $\bar{g}$ is uniquely minimized by $t_{l^\star}$, and $t^\star$ must fall within the interval $[t_{l^\star-1}, t_{l^\star+1}]$. If $l^\star > 0$ and $a_{l^\star} = 0$, finally, then every point in $[t_{l^\star-1}, t_{l^\star}]$ minimizes $\bar{g}$, and $t^\star$ must also fall within $[t_{l^\star-1}, t_{l^\star}]$. In any case, $t_{l^\star}$ provides a δ-approximation for $t^\star$. In the remainder of the proof we show that the index $l^\star$ can be computed efficiently by Algorithm 5. This bisection scheme maintains lower and upper bounds $\underline{l}$ and $\bar{l}$ on the sought index $l^\star$, respectively, and reduces the search interval between $\underline{l}$ and $\bar{l}$ by a factor of two in each iteration. Thus, Algorithm 5 computes $l^\star$ in exactly L iterations [Cor+09, § 12].

We remark that l is guaranteed to be an integer and thus represents a valid index in each iteration of the algorithm because $\underline{l}$ and $\bar{l}$ are initialized as 0 and 2^L, respectively. Note also that if we have access to an oracle for evaluating g exactly, then any element a_l of the array $\boldsymbol{a}$ can be computed with merely two oracle calls. Algorithm 5 thus computes $l^\star$ with $2L$ oracle calls in total. Hence, assertion (i) follows.

Algorithm 5 Binary search algorithm

```
Input: An array $\boldsymbol{a} \in \mathbb{R}^{2^L}$ sorted in ascending order
Initialize $\underline{l} = 0$ and $\bar{l} = 2^L$
while $\underline{l} < \bar{l}$ do
    Set $l \leftarrow (\bar{l} + \underline{l})/2$
    if $a_l \leq 0$ then $\underline{l} \leftarrow l$ else $\bar{l} \leftarrow l$
end while
if $a_{\underline{l}} \leq 0$ then $l \leftarrow \underline{l}$ else $l \leftarrow \bar{l}$
Output: $l^\star \leftarrow l$
```

As for assertion (ii), assume now that we have only access to an inexact oracle that outputs, for any fixed $t \in [0,1]$, an approximate function value $\widetilde{g}(t)$ with $|\widetilde{g}(t) - g(t)| \leq \varepsilon$, where ε is defined as in the statement of the lemma. Reusing the notation from the first part of the proof, one readily verifies that $\varepsilon = \frac{1}{4}\min_{l \in [2^L]}\{a_l : a_l \neq 0\}$. Next, set $\widetilde{a}_0 = -\infty$, and define $\widetilde{a}_l = \widetilde{g}(t_l) - \widetilde{g}(t_{l-1})$ for all $l \in [2^L]$. Therefore, $\widetilde{\boldsymbol{a}} = (\widetilde{a}_0, \ldots, \widetilde{a}_{2^L})$ can be viewed as an approximation of $\boldsymbol{a}$. Moreover, Algorithm 5 with input $\widetilde{\boldsymbol{a}}$ computes an approximation $\tilde{l}^\star$ of $l^\star$ in L iterations. Next, we will show that $|\tilde{l}^\star - l^\star| \leq 1$ even though $\widetilde{\boldsymbol{a}}$ is not necessarily sorted in ascending order. To see this, note that $|a_l| \geq 4\varepsilon$ for all $l \in [2^L]$ with $a_l \neq 0$ by the definition of ε. By the triangle inequality and the assumptions about the inexact oracle, we further have

$$|a_l - \widetilde{a}_l| \leq |\widetilde{g}(t_l) - g(t_l)| + |\widetilde{g}(t_{l-1}) - g(t_{l-1})| \leq 2\varepsilon \quad \forall l \in [2^L].$$

This reasoning reveals that $\widetilde{a}_l$ has the same sign as a_l for every $l \in [2^L]$ with $a_l \neq 0$. In addition, it implies that $t_{\tilde{l}^\star}$ approximates $t^\star$ irrespective of whether or not the array $\boldsymbol{a}$ has a vanishing element. Indeed, if no element of $\boldsymbol{a}$ vanishes, then $\widetilde{a}_l$ has the same sign as a_l for all $l \in [2^L]$. As Algorithm 5 only checks signs, this implies that $\tilde{l}^\star = l^\star$ and that $t_{\tilde{l}^\star}$ provides a δ-approximation for $t^\star$ as in assertion (i). If one element of $\boldsymbol{a}$ vanishes, on the other hand, then $\widetilde{a}_l$ has the same sign as a_l for all $l \in [2^L]$ with $l \neq l^\star$. As Algorithm 5 only checks signs, this implies that $|\tilde{l}^\star - l^\star| \leq 1$. Recalling that $|t^\star - t_{l^\star}| \leq \delta$, we thus have

$$|t_{\tilde{l}^\star} - t^\star| \leq |t_{\tilde{l}^\star} - t_{l^\star}| + |t^\star - t_{l^\star}| \leq 2\delta.$$

In either case, $t_{\tilde{l}^\star}$ provides a 2δ-approximation for $t^\star$. As any element of the array $\widetilde{\boldsymbol{a}}$ can be evaluated with two oracle calls, Algorithm 5 computes $\tilde{l}^\star$ with

$2L$ oracle calls in total. Hence, assertion (ii) follows. □

Efficiency of Binary Search

We adopt the conventions of [Sch98, § 2.1] to measure the size of a computational problem, which is needed to reason about the problem's complexity. Specifically, the size of a scalar $x \in \mathbb{R}$ is defined as

$$\mathrm{size}(x) = \begin{cases} 1 + \lceil \log_2 (|p|+1) \rceil + \lceil \log_2 (q+1) \rceil & \text{if } x = \frac{p}{q} \text{ with } p \in \mathbb{Z} \text{ and} \\ & q \in \mathbb{N} \text{ are relatively prime,} \\ \infty & \text{if } x \text{ is irrational,} \end{cases}$$

where we reserve one bit to encode the sign of x. The size of a real vector is defined as the sum of the sizes of its components plus its dimension. Thus, the input size of an instance $\boldsymbol{w} \in \mathbb{R}^d_+$ and $b \in \mathbb{R}_+$ of the knapsack problem described in Lemma 2.2.3 amounts to

$$\mathrm{size}(\boldsymbol{w}, b) = d + 1 + \sum_{i=1}^{d} \mathrm{size}(w_i) + \mathrm{size}(b).$$

In the following we will prove that the number of iterations

$$L = \left\lceil \log_2(6) + \log_2 d! + +d \log_2(\|\boldsymbol{w}\|_1 + 2) + (d+1)\log_2(d+1) + \sum_{i=1}^{d} \log_2(w_i) \right\rceil + 1$$

of the bisection algorithm used in the proof of Theorem 2.2.2 is upper bounded by a polynomial in $\mathrm{size}(\boldsymbol{w}, b)$. The claim holds trivially if any component of $(\boldsymbol{w}, b)$ is irrational. Below we may thus assume that $w_i = p_i/q_i$ and $b = p_{d+1}/q_{d+1}$, where $p_i \in \mathbb{Z}_+$ and $q_i \in \mathbb{N}$ are relatively prime for every $i \in [d+1]$. This implies that

$$\mathrm{size}(\boldsymbol{w}, b) = 2d + 1 + \sum_{i=1}^{d+1} \lceil \log_2 (p_i + 1) \rceil + \lceil \log_2 (q_i + 1) \rceil.$$

In order to show that L is bounded by a polynomial in $\mathrm{size}(\boldsymbol{w}, b)$, we first note that

$$\log_2 d! \le \log_2 d^d \le d^2 \le \mathrm{size}(\boldsymbol{w}, b)^2 \text{ and } (d+1)\log_2(d+1) \le (d+1)^2 \le \mathrm{size}(\boldsymbol{w}, b)^2. \tag{2.77}$$

This follows from the properties of the logarithm and the definition of the size function. Similarly, we find

$$\begin{aligned} d \log_2(2 + \|\boldsymbol{w}\|_1) &= d \log_2 \left(2 + \sum_{i=1}^{d} p_i/q_i \right) \\ &\le d \log_2((d+1) \max\{2, \max_{i \in [d]}\{p_i/q_i\}\}) \end{aligned}$$

$$
\begin{aligned}
&= d\log_2(d+1) + d\max\{1, \max_{i\in[d]}\{\log_2(p_i) - \log_2(q_i)\}\} \\
&\le (d+1)\log_2(d+1) + d\max\{1, \max_{i\in[d]}\{\log_2(p_i) + \log_2(q_i)\}\} \\
&\le \operatorname{size}(\boldsymbol{w}, b)^2 + \operatorname{size}(\boldsymbol{w}, b)\max_{i\in[d]}\{\log_2(p_i+1) + \log_2(q_i+1)\} \\
&\le 2\operatorname{size}(\boldsymbol{w}, b)^2,
\end{aligned}
$$

where the first inequality follows from the monotonicity of the logarithm, the second inequality holds because $\log_2(q_i) \ge 0$ for all $q_i \in \mathbb{N}$, and the third inequality exploits the second bound in (2.77) as well as the trivial estimates $d \le \operatorname{size}(\boldsymbol{w}, b)$ and $1 = \log_2 2 \le \log_2(q_i+1)$ for all $q_i \in \mathbb{N}$. The last inequality, finally, follows from the observation that

$$
\max_{i\in[d]}\{\log_2(p_i+1) + \log_2(q_i+1)\} \le \sum_{i=1}^{d} \log_2(p_i+1) + \log_2(q_i+1) \le \operatorname{size}(\boldsymbol{w}, b).
$$

Using a similar reasoning, we find

$$
\sum_{i=1}^{d} \log_2(w_i) = \sum_{i=1}^{d} \log_2(p_i/q_i) \le \sum_{i=1}^{d} \log_2(p_i) + \log_2(q_i) \le \operatorname{size}(\boldsymbol{w}, b),
$$

and thus all terms in the definition of L grow at most quadratically with $\operatorname{size}(\boldsymbol{w}, b)$. Hence, the number of iterations L of the bisection algorithm is indeed bounded by a polynomial in $\operatorname{size}(\boldsymbol{w}, b)$.

Derivations for the Examples of Marginal Ambiguity Sets

Example 2.5.2 (Exponential distribution model)**.** *If the marginal generating function in* (2.34) *is set to $F(s) = \exp(s/\lambda - 1)$ for some $\lambda > 0$, then the marginal distribution function F_i for any $i \in [N]$ reduces to*

$$
F_i(s) = \min\{1, \max\{0, 1 - \eta_i \exp(-s/\lambda - 1)\}\},
$$

which characterizes a (shifted) exponential distribution. Defining f as in Theorem 2.3.7, we then obtain

$$
f(s) = \int_0^s F^{-1}(t)\mathrm{d}t = \lambda \int_0^s (\log(t) + 1)\mathrm{d}t = \lambda s \log(s),
$$

where the third equality exploits the standard convention that $0\log(0) = 0$. The proof of Theorem 2.3.7 further implies that $\int_{1-p_i}^1 F_i^{-1}(t)\mathrm{d}t = -\eta_i f(p_i/\eta_i)$ for all $i \in [N]$; see (2.35)*. By Proposition 2.3.6 we thus have*

$$
\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x}) = \max_{\boldsymbol{p}\in\Delta^N} \sum_{i=1}^{N} (\phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i))p_i - \lambda \sum_{i=1}^{N} p_i \log\left(\frac{p_i}{\eta_i}\right).
$$

Next, assign Lagrange multipliers τ *and* $\boldsymbol{\zeta}$ *to the simplex constraints* $\sum_{i=1}^N p_i = 1$ *and* $\boldsymbol{p} \geq \mathbf{0}$*, respectively. If we can find* $\boldsymbol{p}^\star$, $\tau^\star$ *and* $\boldsymbol{\zeta}^\star$ *that satisfy the Karush-Kuhn-Tucker optimality conditions*

$$\begin{array}{lll} \sum_{i=1}^N p_i^\star = 1,\;\; p_i^\star \geq 0 & \forall i \in [N] & \textit{(primal feasibility)} \\ \zeta_i^\star \geq 0 & \forall i \in [N] & \textit{(dual feasibility)} \\ \zeta_i^\star p_i^\star = 0 & \forall i \in [N] & \textit{(complementary} \\ & & \quad \textit{slackness)} \\ \phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i) - \lambda \log\left(\frac{p_i}{\eta_i}\right) - \lambda - \tau^\star + \zeta_i^\star = 0 & \forall i \in [N] & \textit{(stationarity)}, \end{array}$$

then $\boldsymbol{p}^\star$ *is optimal in the above maximization problem. To see that* $\boldsymbol{p}^\star$, $\tau^\star$ *and* $\boldsymbol{\zeta}^\star$ *exist, we use the stationarity condition to conclude that* $p_i^\star = \eta_i \exp((\phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i) - \lambda - \tau^\star + \zeta_i^\star)/\lambda) > 0$*. As* $\eta_i > 0$*, we have* $\zeta_i^\star = 0$ *by complementary slackness. We may then conclude that* $p_i^\star = \eta_i \exp((\phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i) - \lambda - \tau^\star)/\lambda)$ *for all* $i \in [N]$*, which implies via primal feasibility that* $\sum_{i=1}^N \eta_i \exp((\phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i) - \lambda - \tau^\star)/\lambda) = 1$*. Solving this equation for* $\tau^\star$ *and substituting the resulting formula for* $\tau^\star$ *back into the formula for* $p_i^\star$ *yields*

$$\tau^\star = \lambda \log\left(\sum_{i=1}^N \eta_i \exp\left(\frac{\phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i) - \lambda}{\lambda}\right)\right) \quad \textit{and}$$

$$p_i^\star = \frac{\eta_i \exp\left((\phi_i - c(\boldsymbol{x}, \boldsymbol{y}_{\boldsymbol{i}}))/\lambda\right)}{\sum_{j=1}^N \eta_j \exp\left((\phi_j - c(\boldsymbol{x}, \boldsymbol{y}_{\boldsymbol{j}}))/\lambda\right)} \quad \forall i \in [N].$$

The vector $\boldsymbol{p}^\star$ *constructed in this way constitutes an optimal solution for the maximization problem that defines* $\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x})$*. Evaluating the objective function value of* $\boldsymbol{p}^\star$ *in this problem finally confirms that the smooth c-transform coincides with the log-partition function* (2.20)*.*

Example 2.5.3 (Uniform distribution model)**.** *If the marginal generating function in* (2.34) *is set to* $F(s) = s/(2\lambda) + 1/2$ *for some* $\lambda > 0$*, then the marginal distribution function* F_i *for any* $i \in [N]$ *reduces to*

$$F_i(s) = \min\{1, \max\{0, 1 + \eta_i s/(2\lambda) - \eta_i/2\}\},$$

which characterizes a uniform distribution. Defining f *as in Theorem 2.3.7, we then obtain*

$$f(s) = \int_0^s F^{-1}(t)\mathrm{d}t = \lambda \int_0^s (2t - 1)\mathrm{d}t = \lambda(s^2 - s).$$

The proof of Theorem 2.3.7 further implies that $\int_{1-p_i}^{1} F_i^{-1}(t)\mathrm{d}t = -\eta_i f(p_i/\eta_i)$ *for all* $i \in [N]$*; see* (2.35)*. By Proposition 2.3.6, the smooth c-transform thus simplifies to*

$$\begin{aligned}\overline{\psi}_c(\boldsymbol{\phi}, \boldsymbol{x}) &= \max_{\boldsymbol{p}\in\Delta^N} \sum_{i=1}^{N} (\phi_i - c(\boldsymbol{x}, \boldsymbol{y}_i))p_i - \lambda \sum_{i=1}^{N} \frac{p_i^2}{\eta_i} + \lambda \\ &= \lambda + \lambda \operatorname*{spmax}_{i\in[N]} \frac{\phi_i - c(\boldsymbol{x}, \boldsymbol{y_i})}{\lambda},\end{aligned}$$

where the last equality follows from the definition of the sparse maximum operator in (2.36)*.*

Example 2.5.4 (Pareto distribution model)**.** *If the marginal generating function in* (2.34) *is set to* $F(s) = (s(q-1)/(\lambda q) + 1/q)^{1/(q-1)}$ *for some* $\lambda, q > 0$*, then the marginal distribution function* F_i *for any* $i \in [N]$ *reduces to*

$$F_i(s) = \min\left\{1, \max\left\{0, 1 - \eta_i \left(\frac{s(1-q)}{\lambda q} + \frac{1}{q}\right)^{\frac{1}{q-1}}\right\}\right\},$$

which characterizes a Pareto distribution. Defining f *as in Theorem 2.3.7, we then obtain*

$$f(s) = \int_0^s F^{-1}(t)\mathrm{d}t = \frac{\lambda}{q-1} \int_0^s (qt^{q-1} - 1)\mathrm{d}t = \lambda \frac{s^q - s}{q-1}.$$

Example 2.5.5 (Hyperbolic cosine distribution model)**.** *If the marginal generating function in* (2.34) *is set to* $F(s) = \sinh(s/\lambda - k)$ *for some* $\lambda > 0$ *and for* $k = \sqrt{2} - 1 - \mathit{arcsinh}(1)$*, then the marginal distribution function* F_i *for any* $i \in [N]$ *reduces to*

$$F_i(s) = \min\left\{1, \max\left\{0, 1 + \eta_i \sinh(s/\lambda + k)\right\}\right\},$$

which characterizes a hyperbolic cosine distribution. Defining f *as in Theorem 2.3.7, we then obtain*

$$\begin{aligned} f(s) &= \int_0^s F^{-1}(t)\mathrm{d}t \\ &= \lambda \int_0^s (\mathit{arcsinh}(s) + k)\mathrm{d}t \qquad = \lambda(s\, \mathit{arcsinh}(s) - \sqrt{s^2+1} + 1 + ks).\end{aligned}$$

Example 2.5.6 (t-distribution model)**.** *If the marginal generating function in* (2.34) *is set to*

$$F(s) = \frac{N}{2}\left(1 + \frac{s - \sqrt{N-1}}{\sqrt{\lambda^2 + (s - \sqrt{N-1})^2}}\right)$$

for some $\lambda, q > 0$*, then the marginal distribution function* F_i *for any* $i \in [N]$ *reduces to*

$$F_i(s) = \min\left\{1, \max\left\{0, 1 - \frac{\eta_i N}{2}\left(1 - \frac{s + \sqrt{N-1}}{\sqrt{\lambda^2 + (s + \sqrt{N-1})^2}}\right)\right\}\right\},$$

which characterizes a t*-distribution with* 2 *degrees of freedom. Defining* f *as in Theorem 2.3.7, we then find*

$$\begin{aligned} f(s) = \int_0^s F^{-1}(t)\mathrm{d}t &= \lambda \int_0^s \left(\frac{2s - N}{2\sqrt{s(N-s)}} + \sqrt{N-1}\right)\mathrm{d}t \\ &= -\lambda\sqrt{s(N-s)} + \lambda s\sqrt{N-1}. \end{aligned}$$

2.6. The Sparse Maximum Function

The following proposition, which is a simple extension of [MA16, Proposition 1], suggests that the solution of (2.36) can be computed by a simple sorting algorithm.

Proposition 2.6.1. *Given* $\boldsymbol{u} \in \mathbb{R}^N$*, let* σ *be a permutation of* $[N]$ *with* $u_{\sigma(1)} \geq u_{\sigma(2)} \geq \cdots \geq u_{\sigma(N)}$*, and set*

$$k = \max\left\{ j \in [N] : 2 + \left(\sum_{i=1}^{j} \eta_{\sigma(i)}\right) u_{\sigma(j)} > \sum_{i=1}^{j} \eta_{\sigma(i)} u_{\sigma(i)} \right\} \quad \textit{and}$$

$$\tau^\star = \frac{\left(\sum_{i=1}^{k} \eta_{\sigma(i)} u_{\sigma(i)}\right) - 2}{\sum_{i=1}^{k} \eta_{\sigma(i)}}.$$

Then $p_i^\star = \eta_i[u_i - \tau^\star]_+/2$*,* $i \in [N]$*, is optimal in* (2.36)*, where* $[\cdot]_+ = \max\{0, \cdot\}$ *stands for the ramp function.*

Proof. Assign Lagrange multipliers τ and $\boldsymbol{\zeta}$ to the simplex constraints $\sum_{i=1}^N p_i = 1$ and $\boldsymbol{p} \geq \boldsymbol{0}$ in problem (2.36), respectively. If we can find $\boldsymbol{p}^\star$, $\tau^\star$ and $\boldsymbol{\zeta}^\star$ that satisfy the Karush-Kuhn-Tucker conditions

$$\begin{array}{lll} \sum_{i=1}^N p_i^\star = 1,\ p_i^\star \geq 0 & \forall i \in [N] & \text{(primal feasibility)} \\ \zeta_i^\star \geq 0 & \forall i \in [N] & \text{(dual feasibility)} \\ \zeta_i^\star p_i^\star = 0 & \forall i \in [N] & \text{(complementary slackness)} \\ u_i - \frac{2p_i^\star}{\eta_i} - \tau^\star + \zeta_i^\star = 0 & \forall i \in [N] & \text{(stationarity)}, \end{array}$$

then $\boldsymbol{p}^\star$ is optimal in (2.36). In the following, we show that $\boldsymbol{p}^\star$, $\tau^\star$ and $\boldsymbol{\zeta}^\star$ exist. Note first that if $p_i^\star > 0$, then $\zeta_i^\star = 0$ by complementary slackness and

$p_i^\star = \eta_i(u_i - \tau^\star)/2$ by stationarity. On the other hand, if $p_i^\star = 0$, then $\zeta_i^\star \geq 0$ by dual feasibility and $u_i - \tau^\star \leq 0$ by stationarity. In both cases we have $p_i^\star = \eta_i[u_i - \tau^\star]_+/2$ for all $i \in [N]$, which implies that $\sum_{i=1}^N \eta_i[u_i - \tau^\star]_+ = 2$ by primal feasibility. It thus remains to show that $\tau^\star$ as defined in the proposition statement solves this nonlinear scalar equation. To this end, note that by the definitions of the permutation σ and the index k we have

$$u_{\sigma(j)} \geq u_{\sigma(k)} > \frac{(\sum_{i=1}^k \eta_{\sigma(i)} u_{\sigma(i)}) - 2}{\sum_{i=1}^k \eta_{\sigma(i)}} = \tau^\star$$

for all $j \leq k$. The definition of the index k further implies that

$$2 + \left(\sum_{i=1}^{k+1} \eta_{\sigma(i)}\right) u_{\sigma(k+1)} \leq \sum_{i=1}^{k+1} \eta_{\sigma(i)} u_{\sigma(i)}.$$

A simple reordering, dividing both sides of the above inequality by $\sum_{i=1}^k \eta_{\sigma(i)}$, and using the definition of $\tau^\star$ then yields $u_{\sigma(k+1)} \leq \tau^\star$. In addition, by the definition of the permutation σ, we have $u_{\sigma(j)} \leq u_{\sigma(k+1)}$ for all $j > k$. Hence, we conclude that $u_{\sigma(j)} \leq \tau^\star$ for all $j > k$. One can then show that

$$\sum_{i=1}^N \eta_i[u_i - \tau^\star]_+ = \sum_{i=1}^k \eta_{\sigma(i)}(u_{\sigma(i)} - \tau^\star) = 2,$$

as desired. Therefore, problem (2.36) is indeed solved by $p_i^\star = \eta_i[u_i - \tau^\star]_+/2$, $i \in [N]$. □

Part II

Static Decision Making

3. Distributionally Robust Domain Adaptation

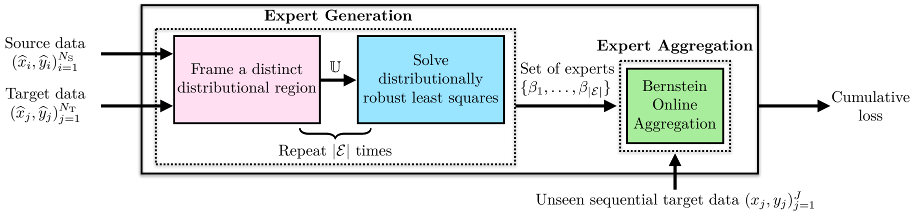

Expert Generation
Source data
$(\widehat{x}_i, \widehat{y}_i)_{i=1}^{N_\mathrm{S}}$
Target data
$(\widehat{x}_j, \widehat{y}_j)_{j=1}^{N_\mathrm{T}}$
Frame a distinct distributional region
$\mathbb{U}$
Solve distributionally robust least squares
Repeat $|\mathcal{E}|$ times
Set of experts
$\{\beta_1, \dots, \beta_{|\mathcal{E}|}\}$
Expert Aggregation
Bernstein Online Aggregation
Cumulative loss
Unseen sequential target data $(x_j, y_j)_{j=1}^{J}$

Figure 3.1: The architecture of our framework for supervised domain adaptation when the unseen target test samples arrive sequentially.

3.1. Introduction

A natural approach to improving predictive performance in data-scarce tasks involves translating informative signals from a data-abundant source domain to the data-scarce target domain. This transfer of knowledge is commonly referred to as domain adaptation or transfer learning, and it is increasingly applied in a wide range of settings, see for example [WC20; CW18; WKW16] and [Red+19].

We consider the supervised domain adaptation setting with scarce labeled target data. The key challenge here is the absence of meaningful data to tune any parameters. However, in many practically relevant applications, new data will arrive sequentially to enrich the information on the target domain. In this case, many online algorithms can be utilized to adaptively learn the best predictor on the target domain, which also guarantee optimal asymptotic regrets [LS20].

In this chapter, we take a pragmatic approach to resolve a specific setup of the domain adaptation problem. We assume access to a scarce labelled target data, and the future target data arrives sequentially. For example, consider understanding the dynamics of ride-sharing platforms requires insights about the demand and supply from both sides of the market. These insights are signalled through the ride fares, which can be explained by characteristics such as the travel distances and the origin-destination pairs of the trips, the time of the day as well as the weather conditions. The capability to correctly predict ride fares directly translates into improved profit forecasts, and thus it vitally supports the growth of new-coming platforms. In a competitive market, a follower (e.g., Lyft) needs to target a slightly different market segment than the leader (e.g., Uber) who had entered earlier. Thus, the demand and supply characteristics for the follower may differ from those of the leader. Nevertheless, as both platforms provide on-demand transportation, it is reasonable to assume that their supply and demand dynamics are similar. The follower, who possesses limited data, can query demand on the leader's platform to collect data in order to leap forward in its predictive precision. Our approach to solve this problem is illustrated in Figure 3.1 and it consists of two components:

1. **Expert Generation Module:** This module generates a set of competitive experts $\mathcal{E}$ by fine-tuning the explanatory power of the source domain data and harnessing the signal guidance from the scarce target domain data.

2. **Expert Aggregation Module:** Acting on the sequential arrival of the

unseen target data, this module aggregates the predictive capability of the generated experts via an online aggregation mechanism. In this work we will use the Bernstein Online Aggregation mechanism.

We will propose two ways to generate the experts. The first approach generates experts corresponding to optimal decisions along a path, with the intention to interpolate between the source and the target distributions. We will consider two types of trajectories, guided by either the Kullback-Leibler or the Wasserstein divergence. The second approach generates distribution regions around both the source and the target. The intersection of these regions is used to generate distributionally robust experts. The geometrical intuition is to find the "direction" induced by the aforementioned divergences, in which the source data can explain the target data. Once the experts are deployed, the aggregation mechanism is executed without re-adapting the experts.

Our ultimate goal is to ensure a competitive performance in the short term and not in the asymptotic regime when the number of test samples from the target domain tends to infinity. Indeed, as soon as the target sample size is sufficient, training the machine learning model on all available target data becomes more attractive. From a short term horizon benchmark, our approach offers an appealing *warm start* for online training procedure, and it may also lead to a faster convergence rate depending on the underlying algorithm.

Contributions. We explore the expert generation problem in the context of supervised domain adaptation.

- We introduce a novel framework to synthesize a family of robust least squares experts by altering various moment-based distribution sets. These sets gradually interpolate from the source information to the target information, capturing different belief levels on the explanatory power of the source domain onto the target domain.

- We present two intuitive strategies to construct the sets of moment information, namely the "Interpolate, then Robustify" and the "Surround, then Intersect" strategies. Both strategies are simply characterized by two parameters representing the aforementioned explanatory power of belief of the source domain and the level of desired robustness.

- We show that when the moment information is prescribed using a Kullback-Leibler or a Wasserstein-type divergence, the experts are efficiently formed by solving convex optimization problems, that can even be solved by a first-order gradient descent algorithm or off-the-shelf solvers.

This chapter is structured as follows. Section 3.2 delineates the problem setup and describes in details two common strategies to generate experts: the convex combination and the reweighting strategies. Section 3.3 introduces our framework to generate experts, while Section 3.4 and 3.5 dive into details about our "Interpolate, then Robustify" and our "Surround, then Intersect" strategies, respectively. Section 3.6 demonstrates experimentally that the proposed robust strategies systematically outperform non-robust interpolations of the empirical least squares estimators.

Literature Review. Domain adaptation arises in various applications including natural language processing [Søg13; Li12; JZ07; BMP06], survival analysis [Li+16] and computer vision [WD18; Csu17]. Domain adaptation methods can be classified into three categories. Unsupervised domain adaptation only requires unlabelled target data, but in large amounts [Ghi+16; Bak+13; GL15; Wan+20a; Lon+16; BD+07; Cou+17]. Semi-supervised domain adaptation requires labelled target data [Yao+15; KSD10; SNB05; LPHLS12; Sah+11; Mat+20; Sun+11]. Finally, supervised domain adaptation only requires scarce labelled target data [Mot+17b; Mot+17a; Tze+15; KTP17]. If the target data is scarce and label information is available, supervised domain adaptation outperforms unsupervised domain adaptation [Mot+17b]. The domain adaptation literature further ramifies by imposing different distributional assumptions into covariate shift [Shi00; Sug+08] or label shift [LWS18; Azi+19].

The domain adaptation literature for regression problems focuses primarily on instance-based reweighting strategies [GV14; Sug+08; GV14; Hua+06; CM14; Che+16], which aim to minimize some distance between the source and target distributions. Most of the instance-based methods solve an optimization problem to find the weights of the instances [GV14; CMM19], which may be computationally expensive when data is abundant. Other approaches rely on deep learning models to minimize the discrepancy between the domain distributions [Zha+18; Ric+20]. The literature on regression for domain adaptation also extends towards boosting-based methods [PS10], and deep learning methods [Sal+19].

We also uses ideas and techniques from robust optimization and adversarial training, which have attracted considerable attention in machine learning [ND16; Gao+18; BKM19; Ngu+19a]. Robust optimization for least squares problem with uncertain data was studied in [GL97]. Distributionally robust optimization with moment ambiguity sets was proposed in [DY10] and extended in [GS10] and [Kuh+19]. Ambiguity sets prescribed by divergences were previously used to robustify Bayes classification [Ngu+19b; NSB20].

Our work is also similar to [Che+16] that consider unsupervised domain adaptation regression, and [Wan+20a] that consider robust domain adaption for the classification setting.

Notation. We use I_d to denotes the identity matrix in $\mathbb{R}^d$. The set of p-by-p positive (semi-)definite matrices is denoted by $\mathbb{S}^p_{++}$ ($\mathbb{S}^p_{+}$). All proofs are relegated to the Appendix.

3.2. Problem Statement and Background

We consider a generic linear regression setting, in which X is a d-dimensional covariate and Y is a univariate response variable. In the context of supervised domain adaptation, we have access to the source domain data $(\widehat{x}_i, \widehat{y}_i)_{i=1}^{N_{\rm S}}$ consisting of $N_{\rm S}$ labelled samples drawn from the source distribution. In addition, we are given a limited number of $N_{\rm T}$ labelled samples $(\widehat{x}_j, \widehat{y}_j)_{j=1}^{N_{\rm T}}$ from the target distribution. Our goal is to predict the responses of the test samples $(x_j, y_j)_{j=1}^J$, which are drawn from the target distribution and arrive sequentially. To this end, we will construct several experts.

In the linear regression setting, each expert is characterized by a vector $\beta \in \mathbb{R}^d$. Given a covariate-response pair $(x, y) \in \mathbb{R}^d \times \mathbb{R}$, we use the square loss function to measure the mismatch between the expert's prediction $\beta^\top x$ and the actual response y. Using the target domain data $(\widehat{x}_i, \widehat{y}_i)_{i=1}^{N_\mathrm{T}}$, one approach is to solve the ridge regression problem

$$\min_{\beta \in \mathbb{R}^d} \frac{1}{N_\mathrm{T}} \sum_{j=1}^{N_\mathrm{T}} (\beta^\top \widehat{x}_j - \widehat{y}_j)^2 + \eta \|\beta\|_2^2$$

for some $\eta \geq 0$ to obtain the empirical target predictor

$$\widehat{\beta}_\mathrm{T} = \left(\frac{1}{N_\mathrm{T}} \sum_{j=1}^{N_\mathrm{T}} \widehat{x}_j \widehat{x}_j^\top + \eta I_d \right)^{-1} \left(\frac{1}{N_\mathrm{T}} \sum_{j=1}^{N_\mathrm{T}} \widehat{x}_j \widehat{y}_j \right).$$

When N_T is small, however, the empirical target predictor may perform poorly on the future target data $(x_j, y_j)_{j=1}^J$.

If the source domain distribution is sufficiently close to the target domain distribution, it is expedient to exploit the available information in the source domain data to construct better predictors for the target domain data. With this promise, one can synthesize several predictors to form an ensemble of experts, and one can apply an online aggregation scheme to predict on the unseen target data. We now first describe several interpolation schemes to generate experts.

Convex Combination Strategy. Denote by $\widehat{\beta}_\mathrm{S}$ the empirical source predictor, which is obtained by solving the ridge regression problem on the source data. The convex combination strategy generates predictors by forming convex combinations between $\widehat{\beta}_\mathrm{S}$ and $\widehat{\beta}_\mathrm{T}$. More precisely, for any $\lambda \in [0, 1]$ a new predictor is synthesized by setting

$$\widehat{\beta}_\lambda = \lambda \beta_\mathrm{S} + (1 - \lambda) \beta_\mathrm{T}.$$

The parameter λ represents our *belief* in the explanatory power of the source domain data: if $\lambda = 0$, the source domain has no power to explain the target domain, and we recover $\widehat{\beta}_0 = \beta_\mathrm{T}$, the empirical target predictor. If $\lambda = 1$, the source domain has an absolute predictive power on the target domain, and it is beneficial to use $\widehat{\beta}_1 = \widehat{\beta}_\mathrm{S}$ because the sample size N_S is large. Discretizing λ in the range $[0, 1]$ forms a family of experts $\mathcal{E}$.

Reweighting Strategy. Reweighting samples is a common strategy in domain adaptation, transfer learning and adversarial training. [GV14] synthesize experts, for example, by solving

$$\min_{\beta \in \mathbb{R}^d} \sum_{i=1}^{N_\mathrm{S}} w_{h,i} (\beta^\top \widehat{x}_i - \widehat{y}_i)^2 + \sum_{j=1}^{N_\mathrm{T}} (\beta^\top \widehat{x}_j - \widehat{y}_j)^2 + \eta \|\beta\|_2^2$$

for some non-negative weights $w_{h,i}$ determined via a Gaussian kernel with bandwidth $h > 0$ of the form

$$w_{h,i} = \sum_{l=1}^{N_{\mathrm{S}}} \alpha_l \exp\left(-\frac{\|\widehat{x}_i - \widehat{x}_l\|_2^2 + (\widehat{y}_i - \widehat{y}_l)^2}{h^2}\right)$$

for $i = 1, \ldots, N_{\mathrm{S}}$. Here, the parameter vector $\alpha \in \mathbb{R}_+^{N_{\mathrm{S}}}$ solves the exponential cone optimization problem

$$\begin{array}{ll} \max & \displaystyle\sum_{j=1}^{N_{\mathrm{T}}} \log\Big(\sum_{l=1}^{N_{\mathrm{S}}} \alpha_l \exp\Big(-\frac{\|\widehat{x}_j - \widehat{x}_l\|_2^2 + (\widehat{y}_j - \widehat{y}_l)^2}{h^2}\Big)\Big) \\ \text{s.t.} & \displaystyle\sum_{i=1}^{N_{\mathrm{S}}}\sum_{l=1}^{N_{\mathrm{S}}} \alpha_l \exp\left(-\frac{\|\widehat{x}_i - \widehat{x}_l\|_2^2 + (\widehat{y}_i - \widehat{y}_l)^2}{h^2}\right) = N_{\mathrm{S}}. \end{array}$$

The predictor β_h, parametrized by the kernel weight h, that solves the reweighted ridge regression problem has the form

$$\Big(\sum_{j=1}^{N_{\mathrm{T}}} \widehat{x}_j \widehat{x}_j^\top + \sum_{i=1}^{N_{\mathrm{S}}} w_i \widehat{x}_i \widehat{x}_i^\top + \eta I_d\Big)^{-1} \Big(\sum_{j=1}^{N_{\mathrm{T}}} \widehat{x}_j \widehat{y}_j + \sum_{i=1}^{N_{\mathrm{S}}} w_i \widehat{x}_i \widehat{y}_i\Big).$$

Discretizing the bandwidth h forms a family of experts $\mathcal{E}$.

Bernstein Online Aggregation (BOA). We now give a brief overview on the BOA algorithm, which is a recursive expert aggregation procedure for sequential prediction [CBL06]. For a given set of experts $\mathcal{E} = \{\beta_1, \ldots, \beta_{|\mathcal{E}|}\}$ and an incumbent weight $\pi_{k,j-1}$ for expert k at time $j-1$, this algorithm aggregates the individual expert's predictions linearly based on the arrival of the input data (x_j, y_j) as $\sum_{k=1}^{|\mathcal{E}|} \pi_{k,j} \beta_k^\top x_j$. The weights of the experts are updated using the exponential rule

$$\pi_{k,j} = \frac{\exp(-\upsilon(1 + \upsilon L_{k,j}) L_{k,j}) \pi_{k,j-1}}{\sum_{k=1}^{|\mathcal{E}|} \exp(-\upsilon(1 + \upsilon L_{kj}) L_{k,j}) \pi_{k,j-1}},$$

where $\upsilon > 0$ is the learning rate and $L_{k,j} = (\beta_k^\top x_j - y_j)^2 - \sum_{k=1}^{|\mathcal{E}|}(\beta_k^\top x_j - y_j)^2 \pi_{k,j-1}$. This algorithm is initialized with weights $\pi_{k,0} \geq 0$ satisfying $\sum_{k=1}^{|\mathcal{E}|} \pi_{k,0} = 1$. The cumulative loss for the stream of test data $(x_j, y_j)_{j=1}^J$ is

$$\sum_{j=1}^{J} \left(\sum_{k=1}^{|\mathcal{E}|} \pi_{k,j} \beta_k^\top x_j - y_j\right)^2. \tag{3.1}$$

For the square loss, the BOA procedure is optimal for the model selection aggregation problem, that is, the excess risk of its batch version achieves the fast rate of convergence $\log(|\mathcal{E}|)/J$ in deviation; see [Win17].

3.3. Predictor Generation via Distributionally Robust Linear Regression

We now specify our framework to generate the set of competitive experts $\mathcal{E}$ for future prediction. Our construction is based on the premises that the source domain carries the explanatory power on the target domain to a certain extent and that the scarce target data can provide directional guidance to pull information from the source data. Moreover, we also leverage ideas from distributionally robust optimization and adversarial training, which have been shown to significantly improve the out-of-sample predictive performance [DN18; MEK18; BKM19; Gao20; Lam19].

With this in mind, our expert generation scheme blends two elements: a distributional probing strategy and a robust estimation procedure. The distributional probing strategy frames the distribution set $\mathbb{B}$, and then each expert is constructed by solving a distributionally robust least squares estimation problem of the form

$$\inf_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2], \tag{3.2}$$

where $\mathbb{Q}$ is a joint distribution over (X, Y). Generating a collection of distribution sets $\mathbb{B}$ in a systematic manner and solving (3.2) for each such set will form a family of experts $\mathcal{E}$.

In a purely data-driven setting with no additional information, it is attractive to probe into the distributional regions *in between* the empirical source distribution $\widehat{\mathbb{P}}_{\mathrm{S}} = N_{\mathrm{S}}^{-1} \sum_{i=1}^{N_{\mathrm{S}}} \delta_{(\widehat{x}_i, \widehat{y}_i)}$ and the empirical target distribution $\widehat{\mathbb{P}}_{\mathrm{T}} = N_{\mathrm{T}}^{-1} \sum_{j=1}^{N_{\mathrm{T}}} \delta_{(\widehat{x}_j, \widehat{y}_j)}$. Because probability distributions reside in infinite-dimensional spaces, framing $\mathbb{B}$ in between $\widehat{\mathbb{P}}_{\mathrm{S}}$ and $\widehat{\mathbb{P}}_{\mathrm{T}}$ is a non-trivial task. Fortunately, because the expected square loss only depends on the first two moments of the joint distribution of (X, Y), it suffices to prescribe $\mathbb{B}$ using a finite parametrization of distributional moments. To this end, let $p = d + 1$ represent the dimension of the joint vector (X, Y). For a given set $\mathbb{U}$ on the space of mean vectors and covariance matrices $\mathbb{R}^p \times \mathbb{S}_+^p$, we consider $\mathbb{B}$ as the lifted distribution set that contains all distributions whose moments belong to $\mathbb{U}$, that is,

$$\mathbb{B} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu, \Sigma),\ (\mu, \Sigma) \in \mathbb{U}\},$$

where $\mathcal{M}(\mathbb{R}^p)$ denotes the set of all distributions on $\mathbb{R}^p$, and the notation $\mathbb{Q} \sim (\mu, \Sigma)$ expresses that $\mathbb{Q}$ has mean μ and covariance matrix Σ. It is convenient to construct the moment information set $\mathbb{U}$ using a divergence on $\mathbb{R}^p \times \mathbb{S}_+^p$.

Definition 1 (Divergence)**.** *A divergence ψ on $\mathbb{R}^p \times \mathbb{S}_+^p$ satisfies the following properties:*

- *non-negativity: for any (μ, Σ), $(\widehat{\mu}, \widehat{\Sigma}) \in \mathbb{R}^p \times \mathbb{S}_+^p$, we have $\psi((\mu, \Sigma) \parallel (\widehat{\mu}, \widehat{\Sigma})) \geq 0$,*

- *indiscernability:* $\psi((\mu,\Sigma)\,\|\,(\widehat{\mu},\widehat{\Sigma}))=0$ *implies* $(\mu,\Sigma)=(\widehat{\mu},\widehat{\Sigma})$.

In this chapter, we will explore two divergences in the space of mean vectors and covariance matrices that are motivated by popular measures of dissimilarity between distributions. The divergence $\mathbb{D}$ is motivated by the Kullback-Leibler (KL) divergence.

Definition 2 (Kullback-Leibler-type divergence)**.** *The divergence* $\mathbb{D}$ *from tuple* $(\mu,\Sigma)\in\mathbb{R}^p\times\mathbb{S}^p_{++}$ *to tuple* $(\widehat{\mu},\widehat{\Sigma})\in\mathbb{R}^p\times\mathbb{S}^p_{++}$ *amounts to*

$$\begin{aligned}&\mathbb{D}\big((\mu,\Sigma)\parallel(\widehat{\mu},\widehat{\Sigma})\big)\triangleq\\&(\widehat{\mu}-\mu)^\top\widehat{\Sigma}^{-1}(\widehat{\mu}-\mu)+\operatorname{Tr}\left[\Sigma\widehat{\Sigma}^{-1}\right]-\log\det(\Sigma\widehat{\Sigma}^{-1})-p.\end{aligned}$$

In fact $\mathbb{D}$ is equivalent to the KL divergence between two non-degenerate Gaussian distributions $\mathcal{N}(\mu,\Sigma)$ and $\mathcal{N}(\widehat{\mu},\widehat{\Sigma})$ (up to a factor of 2). As a consequence, $\mathbb{D}$ is non-negative, and it collapses to 0 if and only if $\Sigma=\widehat{\Sigma}$ and $\mu=\widehat{\mu}$. We can also show that $\mathbb{D}$ is affine-invariant. However, we emphasize that $\mathbb{D}$ is not symmetric and $\mathbb{D}\big((\mu,\Sigma)\parallel(\widehat{\mu},\widehat{\Sigma})\big)\neq\mathbb{D}\big((\widehat{\mu},\widehat{\Sigma})\parallel(\mu,\Sigma)\big)$ in general.

We also study the divergence $\mathbb{W}$ which is motivated by the Wasserstein distance.

Definition 3 (Wasserstein-type divergence)**.** *The divergence* $\mathbb{W}$ *between two tuples* $(\mu,\Sigma)\in\mathbb{R}^p\times\mathbb{S}^p_{+}$ *and* $(\widehat{\mu},\widehat{\Sigma})\in\mathbb{R}^p\times\mathbb{S}^p_{+}$ *amounts to*

$$\mathbb{W}\big((\mu,\Sigma)\,\|\,(\widehat{\mu},\widehat{\Sigma})\big)\triangleq\|\mu-\widehat{\mu}\|_2^2+\operatorname{Tr}\left[\Sigma+\widehat{\Sigma}-2\big(\widehat{\Sigma}^{\frac{1}{2}}\Sigma\widehat{\Sigma}^{\frac{1}{2}}\big)^{\frac{1}{2}}\right].$$

The divergence $\mathbb{W}$ coincides with the *squared* type-2 Wasserstein distance between two Gaussian distributions $\mathcal{N}(\mu,\Sigma)$ and $\mathcal{N}(\widehat{\mu},\widehat{\Sigma})$ [GS84]. One can readily show that $\mathbb{W}$ is non-negative, and it vanishes if and only if $(\mu,\Sigma)=(\widehat{\mu},\widehat{\Sigma})$. Thus, $\mathbb{W}$ is a symmetric divergence.

In Sections 3.4 and 3.5 we examine in detail two strategies to frame $\mathbb{U}$ and its corresponding distribution set $\mathbb{B}$ in a principled manner, and we devise optimization techniques to solve the resulting robust estimation problems.

3.4. "Interpolate, then Robustify" Strategy

"Interpolate, then Robustify" (IR) is an intuitive strategy to systematically probe into distributional regions between $\widehat{\mathbb{P}}_{\mathrm{S}}$ and $\widehat{\mathbb{P}}_{\mathrm{T}}$. Let $(\widehat{\mu}_{\mathrm{S}},\widehat{\Sigma}_{\mathrm{S}})$ be the empirical mean vector and covariance matrix of $\widehat{\mathbb{P}}_{\mathrm{S}}$, that is,

$$\widehat{\mu}_{\mathrm{S}}=\frac{1}{N_{\mathrm{S}}}\sum_{i=1}^{N_{\mathrm{S}}}\begin{pmatrix}\widehat{x}_i\\\widehat{y}_i\end{pmatrix},\ \widehat{\Sigma}_{\mathrm{S}}=\frac{1}{N_{\mathrm{S}}}\sum_{i=1}^{N_{\mathrm{S}}}\begin{pmatrix}\widehat{x}_i\\\widehat{y}_i\end{pmatrix}\begin{pmatrix}\widehat{x}_i\\\widehat{y}_i\end{pmatrix}^\top-\widehat{\mu}_{\mathrm{S}}\widehat{\mu}_{\mathrm{S}}^\top,$$

and let $(\widehat{\mu}_{\mathrm{T}},\widehat{\Sigma}_{\mathrm{T}})$ be defined analogously for $\widehat{\mathbb{P}}_{\mathrm{T}}$. The IR strategy applies repeatedly the following two steps to generate distribution sets. First, interpolate between $(\widehat{\mu}_{\rm S}, \widehat{\Sigma}_{\rm S})$ and $(\widehat{\mu}_{\rm T}, \widehat{\Sigma}_{\rm T})$ to obtain a new pair $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ parametrized by $\lambda \in [0,1]$. Second, construct a moment set $\mathbb{U}_{\lambda,\rho}$ as a ball of radius ρ circumscribing the pair $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$, then lift the moment set $\mathbb{U}_{\lambda,\rho}$ to the corresponding distribution set $\mathbb{B}_{\lambda,\rho}$. More specifically, $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ is the ψ-barycenter between $(\widehat{\mu}_{\rm S}, \widehat{\Sigma}_{\rm S})$ and $(\widehat{\mu}_{\rm T}, \widehat{\Sigma}_{\rm T})$, which is obtained by solving

$$\begin{aligned} \min_{\mu\in\mathbb{R}^p, \Sigma\in\mathbb{S}_+^p} \ & \lambda\psi((\mu,\Sigma) \,\|\, (\widehat{\mu}_{\rm S}, \widehat{\Sigma}_{\rm S})) + \\ & (1-\lambda)\psi((\mu,\Sigma) \,\|\, (\widehat{\mu}_{\rm T}, \widehat{\Sigma}_{\rm T})). \end{aligned} \tag{3.3}$$

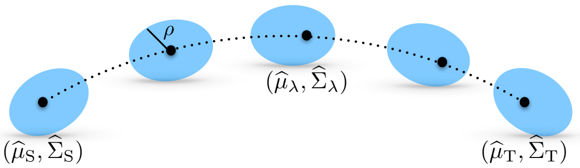

ρ
$(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$
$(\widehat{\mu}_{\rm S}, \widehat{\Sigma}_{\rm S})$
$(\widehat{\mu}_{\rm T}, \widehat{\Sigma}_{\rm T})$

Figure 3.2: The dashed curve shows the barycenter interpolations parametrized by $\lambda \in [0,1]$. Ellipses represent $\mathbb{U}_{\lambda,\rho}$ at different λ.

Then, we employ the divergence ψ to construct an uncertainty set $\mathbb{U}_{\lambda,\rho}$ in the mean-covariance matrix space as

$$\mathbb{U}_{\lambda,\rho} \triangleq \left\{ (\mu,\Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p : \psi((\mu,\Sigma) \,\|\, (\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)) \le \rho \right\}.$$

The outlined procedure is illustrated in Figure 3.2. An expert is now obtained by solving the distributionally robust least squares problem (3.2) with respect to the distribution set

$$\mathbb{B}_{\lambda,\rho} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu,\Sigma), (\mu,\Sigma) \in \mathbb{U}_{\lambda,\rho}\}.$$

Notice that in this strategy the parameter $\lambda \in [0,1]$ characterizes the explanatory power of the source domain to the target domain: if $\lambda = 0$, then $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda) = (\widehat{\mu}_{\rm T}, \widehat{\Sigma}_{\rm T})$, and if $\lambda = 1$, then $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda) = (\widehat{\mu}_{\rm S}, \widehat{\Sigma}_{\rm S})$. Thus, as λ decreases, $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ is moving farther away from the source information $(\widehat{\mu}_{\rm S}, \widehat{\Sigma}_{\rm S})$, and $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ is pulled towards the target information $(\widehat{\mu}_{\rm T}, \widehat{\Sigma}_{\rm T})$.

The choice of the divergence ψ influences both the barycenter problem (3.3) and the formation of the set $\mathbb{U}_{\lambda,\rho}$. Next, we study the special case of the IR strategy with the KL-type divergence and the Wasserstein-type divergence.

3.4.1. Kullback-Leibler-type Divergence

The KL-type divergence $\mathbb{D}$ in Definition 2 is not symmetric. Hence, it is worthwhile to note that the barycenter problem (3.3) optimizes over (μ,Σ) being placed in the first argument of $\mathbb{D}$, and that the set $\mathbb{U}_{\lambda,\rho}$ is also defined with the pair (μ, Σ) being placed in the first argument. Under the divergence $\mathbb{D}$, the barycenter $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ admits a closed form expression. This fact is well-known in the field of KL fusion of Gaussian distributions [Bat+13].

Proposition 3.4.1 (KL barycenter)**.** *Suppose that ψ is the KL-type divergence. If $\widehat{\Sigma}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{T}} \succ 0$, then $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ is the minimizer of the barycenter problem (3.3) with*

$$\begin{aligned}\widehat{\Sigma}_\lambda &= (\lambda\widehat{\Sigma}_{\mathrm{S}}^{-1} + (1-\lambda)\widehat{\Sigma}_{\mathrm{T}}^{-1})^{-1} \succ 0,\\ \widehat{\mu}_\lambda &= \widehat{\Sigma}_\lambda\big(\lambda\widehat{\Sigma}_{\mathrm{S}}^{-1}\widehat{\mu}_{\mathrm{S}} + (1-\lambda)\widehat{\Sigma}_{\mathrm{T}}^{-1}\widehat{\mu}_{\mathrm{T}}\big).\end{aligned}$$

For a given $\lambda \in [0,1]$ and $\rho \geq 0$, the corresponding IR-KL expert is obtained by solving

$$\min_{\beta\in\mathbb{R}^d} \left\{ f_{\lambda,\rho}(\beta) \triangleq \sup_{\mathbb{Q}\in\mathbb{B}_{\lambda,\rho}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2] \right\}. \tag{3.4}$$

Problem (3.4) can be efficiently solved using a gradient-descent algorithm. To do this, the next proposition establishes the relevant properties of $f_{\lambda,\rho}$.

Proposition 3.4.2 (Properties of $f_{\lambda,\rho}$)**.** *The function $f_{\lambda,\rho}$ is convex and continuously differentiable with*

$$\nabla f_{\lambda,\rho}(\beta) = \frac{2\kappa^\star\left(\omega_2\widehat{\Sigma}_\lambda w + (\kappa^\star - \omega_1)(\widehat{\Sigma}_\lambda + \widehat{\mu}_\lambda\widehat{\mu}_\lambda^\top)w\right)_{1:d}}{(\kappa^\star - \omega_1)^2},$$

where $w = [\beta^\top, -1]^\top$, $\omega_1 = w^\top\widehat{\Sigma}_\lambda w$, $\omega_2 = (w^\top\widehat{\mu})^2$ and $\kappa^\star \in (\omega_1, \omega_1\big(1 + 2\rho + \sqrt{1+4\rho\,\omega_2}\big)/(2\rho)]$ is the unique solution of the equation

$$\rho = (\kappa - \omega_1)^{-2}\omega_1\omega_2 + (\kappa-\omega_1)^{-1}\omega_1 + \log(1 - \kappa^{-1}\omega_1).$$

Furthermore, $f_{\lambda,\rho}$ is locally smooth at any $\beta \in \mathbb{R}^d$, i.e.*, there exist constants $C_\beta, \epsilon_\beta > 0$ such that for any $\beta' \in \mathbb{R}^d$ with $\|\beta' - \beta\|_2 \leq \epsilon_\beta$, we have $\|\nabla f_{\lambda,\rho}(\beta') - \nabla f_{\lambda,\rho}(\beta)\|_2 \leq C_\beta \|\beta' - \beta\|_2$.*

Thanks to Proposition 3.4.2, we can apply the adaptive gradient method to solve problem (3.4) to global optimality, and the algorithm enjoys a sublinear rate $|f_{\lambda,\rho}(\bar{\beta}^k) - f_{\lambda,\rho}(\beta^\star_{\lambda,\rho})| \leq O(k^{-1})$, where $\bar{\beta}^k$ is a certain average of the iterates, and $\beta^\star_{\lambda,\rho}$ is an optimal solution of (3.4). The algorithm and its guarantees are detailed in [MM19].

3.4.2. Wasserstein-type Divergence

Under the divergence $\mathbb{W}$ in Definition 3, problem (3.3) resembles the Wasserstein barycenter in the space of Gaussian distributions. The result from [AC11, §6.2] implies that the barycenter $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ admits a closed form expression following the McCann's interpolant [McC97, Example 1.7].

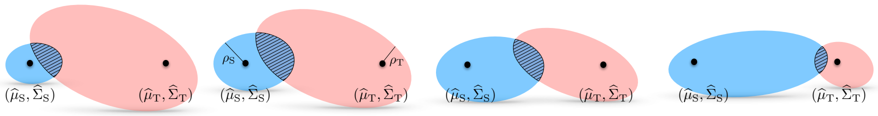

$(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})$
$(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$
ρ_{S}
ρ_{T}
$(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})$
$(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$
$(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})$
$(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$
$(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})$
$(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$

Figure 3.3: Varying $(\rho_{\mathrm{S}}, \rho_{\mathrm{T}})$ frames different moment sets $\mathbb{U}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}$ (hatched regions). The radius ρ_{S} increases from left to right.

Proposition 3.4.3 (Wasserstein interpolation)**.** *Suppose that ψ is the Wasserstein-type divergence. If $\widehat{\Sigma}_{\mathrm{S}} \succ 0$, then $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ is the minimizer of problem (3.3) with*

$$\begin{aligned}
\widehat{\mu}_\lambda &= \lambda \widehat{\mu}_{\mathrm{S}} + (1-\lambda)\widehat{\mu}_{\mathrm{T}}, \\
\widehat{\Sigma}_\lambda &= (\lambda I_p + (1-\lambda)L)\widehat{\Sigma}_{\mathrm{S}}(\lambda I_p + (1-\lambda)L),
\end{aligned}$$

where $L = \widehat{\Sigma}_{\mathrm{T}}^{\frac{1}{2}}(\widehat{\Sigma}_{\mathrm{T}}^{\frac{1}{2}}\widehat{\Sigma}_{\mathrm{S}}\widehat{\Sigma}_{\mathrm{T}}^{\frac{1}{2}})^{-\frac{1}{2}}\widehat{\Sigma}_{\mathrm{T}}^{\frac{1}{2}}$.

For a given $\lambda \in [0, 1]$ and $\rho \geq 0$, we obtain the corresponding IR-Wasserstein expert by solving a conic program using off-the-shelf solvers such as [MOS19].

Proposition 3.4.4 (IR-Wasserstein expert)**.** *Suppose that ψ is the Wasserstein-type divergence. Problem (3.2) with $\mathbb{B} \equiv \mathbb{B}_{\lambda,\rho}$ is equivalent to the second order cone program*

$$\min_{\beta \in \mathbb{R}^d} \quad \left\| (\widehat{\Sigma}_\lambda + \widehat{\mu}_\lambda \widehat{\mu}_\lambda^\top)^{\frac{1}{2}} \begin{bmatrix} \beta \\ -1 \end{bmatrix} \right\|_2 + \sqrt{\rho} \left\| \begin{bmatrix} \beta \\ -1 \end{bmatrix} \right\|_2 .$$

3.5. “Surround, then Intersect” Strategy

“Surround, then Intersect” (SI) probes naturally into the distributional space by intersecting two balls centered at the empirical moments. More specifically, this strategy circumscribes $(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})$ (respectively, $(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$) with a ball of radius ρ_{S} (respectively, ρ_{T}) using the ψ-divergence. Consequentially, the moment information set $\mathbb{U}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}$ in the mean vector-covariance matrix space is defined as

$$\mathbb{U}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}} \triangleq \left\{ \begin{array}{l} (\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p \text{ such that:} \\ \psi((\mu, \Sigma) \,\|\, (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) \leq \rho_{\mathrm{S}} \\ \psi((\mu, \Sigma) \,\|\, (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})) \leq \rho_{\mathrm{T}} \\ \Sigma + \mu\mu^\top \succeq \varepsilon I_p \end{array} \right\},$$

where the small constant $\varepsilon > 0$ improves numerical stability. This construction is graphically illustrated in Figure 3.3. An expert is now obtained by solving the distributionally robust least squares problem (3.2) subject to the distributional set

$$\mathbb{B}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}} = \{\mathbb{Q} \in \mathcal{M}(\mathbb{R}^p) : \mathbb{Q} \sim (\mu, \Sigma), \ (\mu, \Sigma) \in \mathbb{U}_{\rho_{\mathrm{S}}, \rho_{\mathrm{T}}}\}.$$

Note that $\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ is well-defined only if the radii $(\rho_{\mathrm{S}}, \rho_{\mathrm{T}})$ are sufficiently large so that the intersection of the two balls becomes non-empty. A sensible approach to set these parameters is to fix ρ_{S} and to find a sufficiently large ρ_{T} so that $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ is non-empty. In this way, the SI strategy characterizes the explanatory power of the source domain to the target domain by the radius ρ_{S}: if $\rho_{\mathrm{S}} = 0$ then $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ becomes a singleton $\{(\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})\}$, representing the *belief* that the source domain possess absolute explanatory power onto the target domain. As ρ_{S} increases, $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ is gradually pulled towards the empirical target moments $(\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})$. Next, we study the special case of the SI strategy with the KL-type divergence and the Wasserstein-type divergence.

3.5.1. Kullback-Leibler-type Divergence

Recall that $\mathbb{D}$ is asymmetric and (μ, Σ) is the first argument of $\mathbb{D}$ in the definition of $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$. We first study conditions on ρ_{T} under which the ambiguity set $\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ is non-empty.

Proposition 3.5.1 (Minimum radius)**.** *Suppose that ψ is the KL-type divergence. For any $\rho_S > 0$ the sets $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ and $\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ are non-empty if $\rho_{\mathrm{T}} \geq \mathbb{D}((\widehat{\mu}_{\gamma^\star}, \widehat{\Sigma}_{\gamma^\star}) \parallel (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}))$, where $\gamma^\star$ is a maximizer of*

$$\begin{array}{ll}\sup & \mathbb{D}((\widehat{\mu}_\gamma, \widehat{\Sigma}_\gamma) \| (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) + \mathbb{D}((\widehat{\mu}_\gamma, \widehat{\Sigma}_\gamma) \| (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}})) - \gamma \rho_{\mathrm{S}} \\ \text{s.t.} & \gamma \in \mathbb{R}_+, \widehat{\Sigma}_\gamma = (1+\gamma)(\gamma \widehat{\Sigma}_{\mathrm{S}}^{-1} + \widehat{\Sigma}_{\mathrm{T}}^{-1})^{-1} \in \mathbb{S}_+^p, \\ & \widehat{\mu}_\gamma = \widehat{\Sigma}_\gamma(\gamma \widehat{\Sigma}_{\mathrm{S}}^{-1} \widehat{\mu}_{\mathrm{S}} + \widehat{\Sigma}_{\mathrm{T}}^{-1} \widehat{\mu}_{\mathrm{T}})/(1+\gamma) \in \mathbb{R}^p\end{array}$$

The above optimization problem is effectively one-dimensional and can therefore be solved by bisection on γ. The next theorem asserts that the SI-KL experts are formed by solving a semidefinite program.

Theorem 3.5.2 (SI-KL Expert)**.** *Suppose that ψ is the KL-type divergence and $\mathbb{B} \equiv \mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ is non-empty. Then $\beta^\star = (M_{XX}^\star)^{-1} M_{XY}^\star$ solves problem (3.2), where $(M_{XX}^\star, M_{XY}^\star)$ is a solution of the convex semidefinite program*

$$\begin{array}{ll}\sup & \tau \\ \text{s.t.} & M_{XX} \in \mathbb{R}^{d \times d},\ M_{XY} \in \mathbb{R}^{d \times 1},\ M_{YY} \in \mathbb{R} \\ & \tau \in \mathbb{R}_+,\ \mu \in \mathbb{R}^p,\ M \in \mathbb{S}_{++}^p, t \in \mathbb{R}_+ \\ & \widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1} \widehat{\mu}_k - 2 \widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1} \mu + \operatorname{Tr}\left[M \widehat{\Sigma}_k^{-1}\right] - \\ & \log\det(M \widehat{\Sigma}_k^{-1}) - \log(1-t) - p \leq \rho_k \ \ \forall k \in \{\mathrm{S}, \mathrm{T}\} \\ & \begin{bmatrix} M & \mu \\ \mu^\top & t \end{bmatrix} \succeq 0,\ \begin{bmatrix} M_{XX} & M_{XY} \\ M_{XY}^\top & M_{YY} - \tau \end{bmatrix} \succeq 0 \\ & M = \begin{bmatrix} M_{XX} & M_{XY} \\ M_{XY}^\top & M_{YY} \end{bmatrix} \succeq \varepsilon I_p.\end{array}$$

3.5.2. Wasserstein-type Divergence

The space $\mathbb{R}^p \times \mathbb{S}_+^p$ can be endowed with a distance inherited from the Wasserstein distance between Gaussian distribution. For any $\rho_{\mathrm{S}} > 0$, the minimum radius for ρ_{T} that makes $\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ non-empty is known in closed form.

Proposition 3.5.3 (Minimum radius)**.** *Suppose that ψ is the Wasserstein-type divergence. For any $\rho_{\mathrm{S}} > 0$ the sets $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ and $\mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ are non-empty if*

$$\rho_{\mathrm{T}} \geq \left(\sqrt{\mathbb{W}((\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}}) \parallel (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}))} - \sqrt{\rho_{\mathrm{S}}} \right)^2 .$$

The next theorem asserts that the SI-Wasserstein experts are constructed by solving a semidefinite program.

Theorem 3.5.4 (SI-Wasserstein expert)**.** *Suppose that ψ is the Wasserstein-type divergence and $\mathbb{B} \equiv \mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ is non-empty. Then $\beta^\star = (M_{XX}^\star)^{-1} M_{XY}^\star$ solves problem (3.2), where $(M_{XX}^\star, M_{XY}^\star)$ is a solution of the linear semidefinite program*

$$\begin{array}{ll}
\sup & \tau \\
\text{s.t.} & M_{XX} \in \mathbb{R}^{d\times d}, M_{XY} \in \mathbb{R}^{d\times 1}, M_{YY} \in \mathbb{R} \\
& \tau \in \mathbb{R}_+, \mu \in \mathbb{R}^p, M, H \in \mathbb{S}_+^p, C_{\mathrm{S}}, C_{\mathrm{T}} \in \mathbb{R}^{p\times p} \\
& \left.\begin{array}{l} \|\widehat{\mu}_k\|_2^2 - 2\widehat{\mu}_k^\top \mu + \mathrm{Tr}\left[M + \widehat{\Sigma}_k - 2C_k\right] \leq \rho_k \\ \begin{bmatrix} H & C_k \\ C_k^\top & \widehat{\Sigma}_k \end{bmatrix} \succeq 0 \end{array}\right\} k \in \{\mathrm{S}, \mathrm{T}\} \\
& \begin{bmatrix} M - H & \mu \\ \mu^\top & \end{bmatrix} \succeq 0 \\
& \begin{bmatrix} M_{XX} & M_{XY} \\ M_{XY}^\top & M_{YY} - \tau \end{bmatrix} \succeq 0, \; M = \begin{bmatrix} M_{XX} & M_{XY} \\ M_{XY}^\top & M_{YY} \end{bmatrix} \succeq \varepsilon I_p .
\end{array}$$

3.6. Numerical Experiments

The second-order cone and semidefinite programs are modelled in MATLAB via YALMIP [Lö4] and solved with [MOS19]. All experiments are run on an Intel i7-8700 CPU (3.2 GHz) computer with 16GB RAM. The corresponding codes are available at `https://github.com/RAO-EPFL/DR-DA.git`.

We now aim to assess the performance of experts and demonstrate the effects of robustness. In all experiments we generate the set $\mathcal{E} = \{\beta_1, \ldots, \beta_{|\mathcal{E}|}\}$ of experts with $|\mathcal{E}| = 10$.

We consider four family of robust experts generated by:

- IR-KL: with $\rho = \mathbb{D}((\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}) \parallel (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}}))/(3|\mathcal{E}|)$ and λ is spaced from 1 to 0 in exponentially increasing steps.[1]

- IR-WASS: with $\rho = \mathbb{W}((\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}) \| (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}}))/(3|\mathcal{E}|)$ and λ is spaced from 1 to 0 in exponentially increasing steps.

- SI-KL: with ρ_{S} spaced from 10^{-3} to $\mathbb{D}((\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}) \parallel (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}}))-1$ in exponentially increasing steps. For a given ρ_{S}, ρ_{T} is set to the sum of the minimum target radius satisfying the condition of Proposition 3.5.1 and $\rho_{\mathrm{S}}/2$.[2]

- SI-WASS: with ρ_S spaced from 10^{-4} to $\mathbb{W}((\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}) \parallel (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}}))$ in increasing exponential steps. For a given ρ_{S}, ρ_{T} is set to the sum of the minimum radius that satisfies the condition in Proposition 3.5.3 and $\rho_{\mathrm{S}}/2$.

We benchmark against the Convex Combination (CC) and Reweighting (RW) experts in Section 3.2 generated by

- CC-L: with λ equally spaced in $[0, 1]$, thus provides uniformly spaced distributional regions in between domains.

- CC-TL: with λ equally spaced in $[0, 0.5]$, thus distributional regions are formed around the target domain.

- CC-SL: with λ equally spaced in $[0.5, 1]$, thus distributional regions are formed around the source domain.

- CC-TE: with λ spaced from 0 to 1 in exponentially increasing steps, thus the constructed distributional regions are concentrated towards the target domain.

- CC-SE: with λ spaced from 1 to 0 in exponentially increasing steps, thus the constructed distributional regions are concentrated towards the source domain.

- RWS: with h equally spaced in $[0.5, 10]$.

We consider a family of sequential empirical ridge regression estimators generated by training for each J over

- LSE-T, the union of the target dataset $(\widehat{x}_j, \widehat{y}_j)_{j=1}^{N_{\mathrm{T}}}$, and the sequentially arriving target test data $(x_j, y_j)_{j=1}^{J-1}$,

- LSE-T&S, the union of the source data $(\widehat{x}_i, \widehat{y}_i)_{i=1}^{N_{\mathrm{S}}}$, the target data $(\widehat{x}_j, \widehat{y}_j)_{j=1}^{N_{\mathrm{T}}}$ and the sequentially arriving target test data $(x_j, y_j)_{j=1}^{J-1}$.

Note that both LSE-T and LSE-T&S predictors dynamically incorporate the new data to adapt the prediction. Thereby, they have an unfair advantage in the long run over the other experts that are trained only once at the beginning with N_{T} samples from the test domain.

The main reason behind using exponential step sizes originates from the asymmetric nature of $\mathbb{D}$. For simplicity, we also use it for experts with $\mathbb{W}$. To

[1] We say that λ is spaced from a to b in K exponentially increasing steps if $\lambda_1 = a$ and $\lambda_{k+1} = \lambda_k - (a-b)\exp(k)/\sum_{i=1}^{K-1}\exp(i)$ for all $k \in \{2, \dots, K-1\}$.

[2] If $d \geq 15$, then the minimum value of ρ_{S} is set to 5 to improve numerical stability.

ensure fairness in the competition between experts, we vary the parameters of the non-robust experts also in exponential steps.

We compare the performance of our model against the above non-robust benchmarks on 5 Kaggle datasets:[3]

- **Uber&Lyft** contains $d = 38$ features of Uber and Lyft cab rides in Boston including the distances, date and time of the hailing, a weather summary for that day. The prediction target is the price of the ride. We divide the dataset based on the company, Uber (source) and Lyft (target).
- **US Births (2018)** has $d = 36$ predictive features of child births in the United States in the year of 2018 including the gender of the infant, mother's weight gain, and mother's per-pregnancy body mass index. The task is to predict the weight of the infants. We divide the dataset based on gender: male (source) and female (target).
- **Life Expectancy** contains $d = 19$ predictive features, and the target variable is the life expectancy at birth. The dataset is divided into two subgroups: developing (source) and developed (target) countries.
- **House Prices in King Country** contains $d = 14$ predictive variables, the target variable is the transaction price of the houses. We split the dataset into two domains: houses built in $[1950, 2000)$ (source) and $[2000, 2010]$ (target).
- **California Housing Prices** has $d = 9$ predictive features, the target variable is the price of houses. We divide this dataset into houses with less than an hour drive to the ocean shore (source) and houses in inland (target).

We use all samples from the source domain for training, and we form the target training set by drawing $N_{\mathrm{T}} = d$ samples from the target dataset. Later, we randomly sample $J = 1000$ data points from the remaining target samples to form the sequentially arriving target test samples. Note that the performance of the experts is sensitive to the data, and thus we replicate this procedure 100 times. We set the regularization parameter of the ridge regression problem to $\eta = 10^{-6}$ and the learning rate of the BOA algorithm to $\upsilon = 0.5$. We measure the performance of the experts by the cumulative loss (3.1) calculated for every J.

Table 3.1 shows the average cumulative loss of each aggregated expert obtained by the BOA algorithm for all datasets and for $J = \{5, 10, 50, 100\}$ across 100 independent runs. In each row, the minimum loss is normalized to 1, and the remaining entries are presented by the multiplicative factor of the minimum value. This result suggests that the IR-WASS and SI-WASS experts perform favorably over the competitors in that their cumulative loss at each time step is substantially lower than that of most other competitors. Figure 3.4 demonstrates how the average cumulative loss in (3.1) grows over time for the Uber&Lyft dataset. Figure 3.4 shows that the loss of LSE-T&S is initially constant at a high level, which highlights the discrepancy between the two domain distributions.

[3] Descriptions and download links are provided in the appendix.

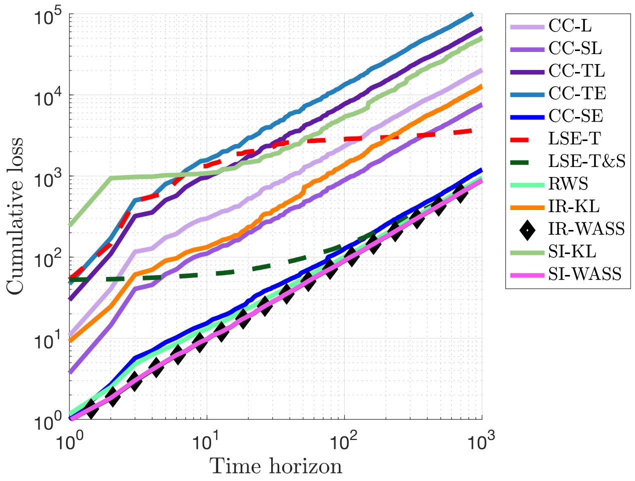


Figure 3.4: Cumulative loss averaged over 100 runs, Uber&Lyft.

The growth rate of LSE-T decays faster than that of other experts, and the time when LSE-T saturates indicates when the combined target domain data alone is sufficient to construct a single, competitive predictor without using any source domain data.

Data Set	Time	IR-KL	IR-WASS	SI-KL	SI-WASS	CC-L	CC-TL	CC-SL	CC-TE	CC-SE	RWS	LSE-T	LSE-T&S
Uber&Lyft	5	17.65	**1.00**	199.28	1.01	34.04	98.43	12.03	155.71	1.74	1.45	119.65	11.08
	10	13.67	**1.00**	111.52	1.01	30.85	99.22	11.40	161.72	1.58	1.34	137.15	6.32
	50	13.39	**1.00**	60.29	1.01	25.87	85.06	9.72	147.45	1.42	1.16	57.85	2.12
	100	15.24	**1.00**	59.06	1.01	26.01	85.77	9.91	148.49	1.41	1.12	31.25	1.57
US Births (2018)	5	79.83	1.02	44.71	**1.00**	64.99	257.60	25.13	432.09	2.07	4.50	727.88	39.17
	10	115.47	1.02	39.35	**1.00**	45.59	195.14	18.33	339.11	1.60	3.29	524.39	19.28
	50	107.40	1.01	40.04	**1.00**	42.74	192.46	13.12	361.51	1.31	2.00	191.27	5.20
	100	117.03	1.01	53.13	**1.00**	45.35	208.65	12.94	397.33	1.22	1.75	104.75	3.19
Life Expectancy	5	33.18	**1.00**	6.24	1.03	17.24	77.06	7.38	125.71	1.46	1.15	255.08	20.72
	10	25.59	**1.00**	5.45	1.02	12.49	60.19	5.50	104.00	1.40	1.15	167.15	10.73
	50	19.81	**1.00**	8.70	1.01	7.57	44.00	3.10	84.98	1.38	1.10	39.83	3.15
	100	19.02	**1.00**	8.25	1.005	6.82	41.40	2.68	83.60	1.38	1.08	20.42	2.10
House Prices in KC	5	1.58	**1.00**	1.21	1.01	3.98	8.87	2.12	13.31	1.29	1.23	11.75	3.70
	10	1.52	**1.00**	1.20	1.01	3.58	7.77	2.02	11.70	1.27	1.23	6.93	2.25
	50	1.34	**1.00**	1.31	1.01	2.79	6.52	1.86	10.37	1.27	1.20	3.91	1.30
	100	1.34	**1.00**	1.30	1.01	2.65	6.54	1.91	10.74	1.27	1.18	2.72	1.12
California Housing	5	63.33	1.05	3.31	**1.00**	27.63	102.82	9.60	181.52	1.35	1.17	96.43	54.34
	10	68.08	1.04	2.42	**1.00**	20.57	91.86	6.23	169.87	1.19	1.17	45.64	24.76
	50	70.08	1.01	1.97	**1.00**	11.79	81.72	2.49	170.18	1.05	1.13	10.17	5.63
	100	72.80	1.003	1.90	**1.00**	9.71	79.19	1.83	173.96	1.04	1.14	5.81	3.39

Table 3.1: Normalized cumulative loss values averaged over 100 independent runs.

Appendix

Proof of Section 3.4

Proof of Proposition 3.4.1. Note that optimization problem (3.3) constitutes an unbounded convex optimization problem when ψ is the Kullback-Leibler-type divergence of Definition 1. Let $g(\mu, \Sigma) \triangleq \lambda \mathbb{D}((\mu, \Sigma) \parallel (\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}})) + (1-\lambda)\mathbb{D}((\mu, \Sigma) \parallel (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}))$, then, the first order optimality condition reads

$$\begin{aligned}
\nabla_\mu g(\mu, \Sigma) &= 2\lambda \widehat{\Sigma}_{\mathrm{S}}^{-1}(\mu - \widehat{\mu}_{\mathrm{S}}) + 2(1-\lambda)\widehat{\Sigma}_{\mathrm{T}}^{-1}(\mu - \widehat{\mu}_{\mathrm{T}}) = 0,\\
\nabla_\Sigma g(\mu, \Sigma) &= \lambda \widehat{\Sigma}_{\mathrm{S}}^{-1} - \lambda \Sigma^{-1} + (1-\lambda)\widehat{\Sigma}_{\mathrm{T}}^{-1} - (1-\lambda)\Sigma^{-1} = 0.
\end{aligned}$$

One can then show $(\widehat{\mu}_\lambda, \widehat{\Sigma}_\lambda)$ provided in statement of Proposition 3.4.1 solves the system of equalities above. □

Below we prove Proposition 3.4.2. In the proof of Proposition 3.4.2 and its auxiliary lemmas, Lemma 3.6.1 and Lemma 3.6.2, we omit the subscripts λ and ρ to avoid clutter.

Lemma 3.6.1 (Dual problem). *Fix $(\widehat{\mu}, \widehat{\Sigma}) \in \mathbb{R}^p \times \mathbb{S}_{++}^p$ and $\rho \geq 0$. For any symmetric matrix $H \in \mathbb{S}^p$, the optimization problem*

$$\left\{\begin{array}{ll}
\sup\limits_{\mu, \Sigma} & \mathrm{Tr}\left[H(\Sigma + \mu\mu^\top)\right]\\
\text{s.t.} & \mathrm{Tr}\left[\Sigma\widehat{\Sigma}^{-1}\right] - \log\det(\Sigma\widehat{\Sigma}^{-1}) - p + (\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu}) \leq \rho,\\
& \Sigma \succ 0
\end{array}\right. \tag{3.5a}$$

admits the dual formulation

$$\left\{\begin{array}{ll}
\inf & \kappa(\rho - \widehat{\mu}^\top \widehat{\Sigma}^{-1}\widehat{\mu}) + \kappa^2 \widehat{\mu}^\top \widehat{\Sigma}^{-1}[\kappa\widehat{\Sigma}^{-1} - H]^{-1}\widehat{\Sigma}^{-1}\widehat{\mu} - \\
& \qquad \kappa \log\det(I - \widehat{\Sigma}^{\frac{1}{2}} H \widehat{\Sigma}^{\frac{1}{2}}/\kappa)\\
\text{s.t.} & \kappa \geq 0,\ \kappa\widehat{\Sigma}^{-1} \succ H.
\end{array}\right. \tag{3.5b}$$

Proof of Lemma 3.6.1. For any $\mu \in \mathbb{R}^p$ such that $(\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu}) \leq \rho$, denote the set $\mathcal{S}_\mu$ as

$$\mathcal{S}_\mu \triangleq \left\{\Sigma \in \mathbb{S}_{++}^p : \mathrm{Tr}\left[\Sigma\widehat{\Sigma}^{-1}\right] - \log\det\Sigma \leq \rho_\mu\right\},$$

where $\rho_\mu \in \mathbb{R}$ is defined as $\rho_\mu \triangleq \rho + p - \log\det\widehat{\Sigma} - (\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu})$. Using these auxiliary notations, problem (3.5a) can be re-expressed as a nested program of the form

$$\begin{array}{ll}
\sup\limits_{\mu} & \mu^\top H \mu + \sup\limits_{\Sigma \in \mathcal{S}_\mu} \ \mathrm{Tr}\left[H\Sigma\right]\\
\text{s.t.} & (\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu}) \leq \rho,
\end{array}$$

where we emphasize that the constraint on μ is redundant, but it is added to ensure the feasibility of the inner supremum over Σ for every feasible value of μ of the outer problem. We now proceed to reformulate the supremum subproblem over Σ.

Assume momentarily that $H \neq 0$ and that μ satisfies $(\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu}) < \rho$. In this case, one can verify that $\widehat{\Sigma}$ is a Slater point of the convex set $\mathcal{S}_\mu$. Using a duality argument, we find

$$\begin{aligned}\sup_{\Sigma \in \mathcal{S}_\mu} \ \mathrm{Tr}\left[H\Sigma\right] &= \sup_{\Sigma \succ 0} \inf_{\phi \geq 0} \ \mathrm{Tr}\left[H\Sigma\right] + \phi\big(\rho_\mu - \mathrm{Tr}\left[\widehat{\Sigma}^{-1}\Sigma\right] + \log\det\Sigma\big) \\ &= \inf_{\phi \geq 0} \ \left\{\phi\rho_\mu + \sup_{\Sigma \succ 0} \ \left\{ \mathrm{Tr}\left[(H - \phi\widehat{\Sigma}^{-1})\Sigma\right] + \phi\log\det\Sigma\right\}\right\},\end{aligned}$$

where the last equality follows from strong duality [Ber09b, Proposition 5.3.1]. If $H - \phi\widehat{\Sigma}^{-1} \not\prec 0$, then the inner supremum problem becomes unbounded. To see this, let $\sigma \in \mathbb{R}_+$ be the maximum eigenvalue of $H - \phi\widehat{\Sigma}^{-1}$ with the corresponding eigenvector v, then the sequence $(\Sigma_k)_{k\in\mathbb{N}}$ with $\Sigma_k = I + kvv^\top$ attains the asymptotic maximum objective value of $+\infty$. If $H - \phi\widehat{\Sigma}^{-1} \prec 0$ then the inner supremum problem admits the unique optimal solution

$$\Sigma^\star(\phi) = \phi(\phi\widehat{\Sigma}^{-1} - H)^{-1}, \tag{3.6}$$

which is obtained by solving the first-order optimality condition. By placing this optimal solution into the objective function and arranging terms, we have

$$\sup_{\Sigma \in \mathcal{S}_\mu} \ \mathrm{Tr}\left[H\Sigma\right] = \inf_{\substack{\phi \geq 0 \\ \phi\widehat{\Sigma}^{-1} \succ H}} \ \phi\big(\rho - (\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu})\big) - \phi\log\det(I - \widehat{\Sigma}^{\frac{1}{2}} H \widehat{\Sigma}^{\frac{1}{2}}/\phi). \tag{3.7}$$

We now argue that the above equality also holds when μ is chosen such that $(\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu}) = \rho$. In this case, $\mathcal{S}_\mu$ collapses into a singleton $\{\widehat{\Sigma}\}$, and the left-hand side supremum problem attains the value $\mathrm{Tr}\left[H\widehat{\Sigma}\right]$. The right-hand side infimum problem becomes

$$\inf_{\substack{\phi \geq 0 \\ \phi\widehat{\Sigma}^{-1} \succ H}} \ - \phi\log\det(I - \widehat{\Sigma}^{\frac{1}{2}} H \widehat{\Sigma}^{\frac{1}{2}}/\phi).$$

One can show using the l'Hopital rule that

$$\lim_{\phi \uparrow +\infty} \ - \phi\log\det(I - \widehat{\Sigma}^{\frac{1}{2}} H \widehat{\Sigma}^{\frac{1}{2}}/\phi) = \mathrm{Tr}\left[H\widehat{\Sigma}\right],$$

which implies that the equality holds. Furthermore, when $H = 0$, the left-hand side of (3.7) evaluates to 0, while the infimum problem on the right-hand side of (3.7) also attains the optimal value of 0 asymptotically as ϕ decreases to 0. This implies that (3.7) holds for all $H \in \mathbb{S}^p$ and for any μ satisfying $(\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu}) \leq \rho$.

The above line of argument shows that problem (3.5a) can now be expressed as the following maximin problem

$$\sup_{\mu: (\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu}) \leq \rho} \inf_{\substack{\phi \geq 0 \\ \phi \widehat{\Sigma}^{-1} \succ H}} \mu^\top H \mu + \phi\big(\rho - (\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu})\big) - \\ \phi \log\det(I - \widehat{\Sigma}^{\frac{1}{2}} H \widehat{\Sigma}^{\frac{1}{2}}/\phi).$$

For any $\phi \geq 0$ such that $\phi\widehat{\Sigma}^{-1} \succ H$, the objective function is concave in μ. For any μ, the objective function is convex in ϕ. Furthermore, the feasible set of μ is convex and compact, and the feasible set of ϕ is convex. As a consequence, we can apply Sion's minimax theorem [Sio58] to interchange the supremum and the infimum operators, and problem (3.5a) is equivalent to

$$\inf_{\substack{\phi \geq 0 \\ \phi \widehat{\Sigma}^{-1} \succ H}} \left\{ \begin{array}{l} \phi\rho - \phi \log\det(I - \widehat{\Sigma}^{\frac{1}{2}} H \widehat{\Sigma}^{\frac{1}{2}}/\phi) \\ \qquad + \sup\limits_{\mu: (\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu}) \leq \rho} \mu^\top H \mu - \phi(\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1}(\mu - \widehat{\mu}) \end{array} \right\}.$$

For any ϕ which is feasible for the outer problem, the inner supremum problem is a convex quadratic optimization problem because $\phi\widehat{\Sigma}^{-1} \succ H$. Using a strong duality argument, the value of the inner supremum equals to the value of

$$\begin{aligned} &\inf_{\nu \geq 0} \left\{ \nu\rho - (\nu + \phi)\widehat{\mu}^\top \widehat{\Sigma}^{-1}\widehat{\mu} + \sup_{\mu} \mu^\top (H - (\phi + \nu)\widehat{\Sigma}^{-1})\mu + 2(\nu + \phi)(\widehat{\Sigma}^{-1}\widehat{\mu})^\top \mu \right\} \\ = &\inf_{\nu \geq 0} \ \nu\rho - (\nu + \phi)\widehat{\mu}^\top \widehat{\Sigma}^{-1}\widehat{\mu} + (\nu + \phi)^2 (\widehat{\Sigma}^{-1}\widehat{\mu})^\top [(\phi + \nu)\widehat{\Sigma}^{-1} - H]^{-1}(\widehat{\Sigma}^{-1}\widehat{\mu}), \end{aligned}$$

where the equality follows from the fact that the unique optimal solution in the variable μ is given by

$$(\phi + \nu)[(\phi + \nu)\widehat{\Sigma}^{-1} - H]^{-1}\widehat{\Sigma}^{-1}\widehat{\mu}. \tag{3.8}$$

By combining two layers of infimum problem and using a change of variables $\kappa \leftarrow \phi + \nu$, problem (3.5a) can now be written as

$$\left\{ \begin{array}{ll} \inf & \kappa(\rho - \widehat{\mu}^\top \widehat{\Sigma}^{-1}\widehat{\mu}) + \kappa^2 \widehat{\mu}^\top \widehat{\Sigma}^{-1}[\kappa\widehat{\Sigma}^{-1} - H]^{-1}\widehat{\Sigma}^{-1}\widehat{\mu} \\ & -\phi \log\det(I - \widehat{\Sigma}^{\frac{1}{2}} H \widehat{\Sigma}^{\frac{1}{2}}/\phi) \\ \text{s.t.} & \phi \geq 0, \ \phi\widehat{\Sigma}^{-1} \succ H, \ \kappa - \phi \geq 0. \end{array} \right. \tag{3.9}$$

We now proceed to eliminate the multiplier ϕ from the above problem. To this end, rewrite the above optimization problem as

$$\begin{array}{ll} \inf & \kappa(\rho - \widehat{\mu}^\top \widehat{\Sigma}^{-1}\widehat{\mu}) + \kappa^2 \widehat{\mu}^\top \widehat{\Sigma}^{-1}[\kappa\widehat{\Sigma}^{-1} - H]^{-1}\widehat{\Sigma}^{-1}\widehat{\mu} + g(\kappa) \\ \text{s.t.} & \kappa \geq 0, \ \kappa\widehat{\Sigma}^{-1} \succ H, \end{array}$$

where $g(\kappa)$ is defined for every feasible value of κ as

$$g(\kappa) \triangleq \begin{cases} \inf & -\phi \log\det(I - \widehat{\Sigma}^{\frac{1}{2}} H \widehat{\Sigma}^{\frac{1}{2}}/\phi) \\ \text{s.t.} & \phi \geq 0,\ \phi\widehat{\Sigma}^{-1} \succ H,\ \phi \leq \kappa. \end{cases} \tag{3.10}$$

Let $g_0(\phi)$ denote the objective function of the above optimization, which is independent of κ. Let $\sigma_1, \ldots, \sigma_p$ be the eigenvalues of $\widehat{\Sigma}^{\frac{1}{2}} H \widehat{\Sigma}^{\frac{1}{2}}$, we can write the function g directly using the eigenvalues $\sigma_1, \ldots, \sigma_p$ as

$$g_0(\phi) = -\phi \sum_{i=1}^{p} \log(1 - \sigma_i/\phi).$$

It is easy to verify by basic algebra manipulation that the gradient of g_0 satisfies

$$\nabla g_0(\phi) = \sum_{i=1}^{p} \left[\log\left(\frac{\phi}{\phi - \sigma_i} \right) - \frac{\phi}{\phi - \sigma_i} \right] + p \leq 0,$$

which implies that the value of ϕ that solves (3.10) is κ, and thus $g(\kappa) = -\kappa \log\det(I - \widehat{\Sigma}^{\frac{1}{2}} H \widehat{\Sigma}^{\frac{1}{2}}/\kappa)$. Substituting ϕ by κ in problem (3.9) leads to the desired claim. □

Lemma 3.6.2 (Optimal solution attaining $f(\beta)$)**.** *For any $(\widehat{\mu}, \widehat{\Sigma}) \in \mathbb{R}^p \times \mathbb{S}^p_{++}$, $\rho \in \mathbb{R}_{++}$ and $w \in \mathbb{R}^p$, $f(\beta)$ equals to the optimal value of the optimization problem*

$$\begin{cases} \sup\limits_{\mu, \Sigma \succ 0} & w^\top (\Sigma + \mu\mu^\top) w \\ \text{s.t.} & \operatorname{Tr}\left[\Sigma\widehat{\Sigma}^{-1}\right] - \log\det(\Sigma\widehat{\Sigma}^{-1}) - p + (\mu - \widehat{\mu})^\top \widehat{\Sigma}^{-1} (\mu - \widehat{\mu}) \leq \rho, \end{cases} \tag{3.11a}$$

which admits the unique optimal solution

$$\Sigma^\star = \kappa^\star (\kappa^\star \widehat{\Sigma}^{-1} - ww^\top)^{-1}, \qquad \mu^\star = \Sigma^\star \widehat{\Sigma}^{-1} \widehat{\mu}, \tag{3.11b}$$

with $\kappa^\star > w^\top \widehat{\Sigma} w$ being the unique solution of the nonlinear equation

$$\rho = \frac{(w^\top \widehat{\mu})^2 w^\top \widehat{\Sigma} w}{(\kappa - w^\top \widehat{\Sigma} w)^2} + \frac{w^\top \widehat{\Sigma} w}{\kappa - w^\top \widehat{\Sigma} w} + \log\Big(1 - \frac{w^\top \widehat{\Sigma} w}{\kappa}\Big). \tag{3.11c}$$

Moreover, we have $\kappa^\star \leq w^\top \widehat{\Sigma} w \big(1 + 2\rho + \sqrt{1 + 4\rho (w^\top \widehat{\mu})^2}\big)/(2\rho)$.

Proof of Lemma 3.6.2. First, note that

$$f(\beta) = \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}} \left[(\beta^\top X - Y)^2 \right] = \sup_{\mathbb{Q} \in \mathbb{B}} \mathbb{E}_{\mathbb{Q}} \left[w^\top \xi \xi^\top w \right] = \sup_{(\mu, \Sigma) \in \mathbb{U}} w^\top \left(\Sigma + \mu\mu^\top \right) w,$$

which, by the definition of $\mathbb{U}$ and definition (2), equals to the optimal value of problem (3.11a).

From the duality result in Lemma 3.6.1, problem (3.11a) is equivalent to

$$\begin{array}{ll} \inf & \kappa(\rho - \widehat{\mu}^\top \widehat{\Sigma}^{-1}\widehat{\mu}) + (\kappa\widehat{\Sigma}^{-1}\widehat{\mu})^\top[\kappa\widehat{\Sigma}^{-1} - ww^\top]^{-1}(\kappa\widehat{\Sigma}^{-1}\widehat{\mu}) \\ & -\kappa \log\det(I - \widehat{\Sigma}^{\frac{1}{2}} ww^\top \widehat{\Sigma}^{\frac{1}{2}}/\kappa) \\ \text{s.t.} & \kappa \geq 0,\ \kappa\widehat{\Sigma}^{-1} \succ ww^\top. \end{array}$$

Applying [Ber09a, Fact 2.16.3], we have the equalities

$$\begin{aligned} \det(I - \widehat{\Sigma}^{\frac{1}{2}} ww^\top \widehat{\Sigma}^{\frac{1}{2}}/\kappa) &= 1 - w^\top \widehat{\Sigma} w/\kappa \\ (\kappa\widehat{\Sigma}^{-1} - ww^\top)^{-1} &= \kappa^{-1}\widehat{\Sigma} + \kappa^{-2}\big(1 - w^\top\widehat{\Sigma}w/\kappa\big)^{-1}\widehat{\Sigma}ww^\top\widehat{\Sigma}, \end{aligned}$$

and thus by some algebraic manipulations we can rewrite

$$f(\beta) = \begin{cases} \inf & \kappa\rho + \frac{\kappa(w^\top\widehat{\mu})^2}{\kappa - w^\top\widehat{\Sigma}w} - \kappa\log\big(1 - w^\top\widehat{\Sigma}w/\kappa\big) \\ \text{s.t.} & \kappa > w^\top\widehat{\Sigma}w. \end{cases} \tag{3.12}$$

Let f_0 be the objective function of the above optimization problem. The gradient of f_0 satisfies

$$\nabla f_0(\kappa) = \rho - \frac{(w^\top\widehat{\mu})^2 w^\top\widehat{\Sigma}w}{(\kappa - w^\top\widehat{\Sigma}w)^2} - \frac{w^\top\widehat{\Sigma}w}{\kappa - w^\top\widehat{\Sigma}w} - \log\Big(1 - \frac{w^\top\widehat{\Sigma}w}{\kappa}\Big).$$

By the above expression of $\nabla f_0(\kappa)$ and the strict convexity of $f_0(\kappa)$, the value $\kappa^\star$ that solves (3.11c) is also the unique minimizer of (3.12). In other words, $f_0(\kappa) = f(\beta)$.

We now proceed to show that $(\mu^\star, \Sigma^\star)$ defined as in (3.11b) is feasible and optimal. First, we prove feasibility of $(\mu^\star, \Sigma^\star)$. By direct computation,

$$(\mu^\star - \widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu^\star - \widehat{\mu}) = \widehat{\mu}^\top(\widehat{\Sigma}^{-1}\Sigma^\star - I)\widehat{\Sigma}^{-1}(\Sigma^\star\widehat{\Sigma}^{-1} - I)\widehat{\mu} = \frac{(\widehat{\mu}^\top w)^2 w^\top\widehat{\Sigma}w}{(\kappa^\star - w^\top\widehat{\Sigma}w)^2}. \tag{3.13a}$$

Moreover, because $\Sigma^\star\widehat{\Sigma}^{-1} = I + (\kappa^\star - w^\top\widehat{\Sigma}w)^{-1}\widehat{\Sigma}ww^\top$, we have

$$\mathrm{Tr}\left[\Sigma^\star\widehat{\Sigma}^{-1}\right] - \log\det(\Sigma^\star\widehat{\Sigma}^{-1}) - p = (\kappa^\star - w^\top\widehat{\Sigma}w)^{-1}w^\top\widehat{\Sigma}w + \log\big(1 - \frac{w^\top\widehat{\Sigma}w}{\kappa^\star}\big). \tag{3.13b}$$

Combining (3.13a) and (3.13b), we have

$$\mathrm{Tr}\left[\Sigma^\star\widehat{\Sigma}^{-1}\right] - \log\det(\Sigma^\star\widehat{\Sigma}^{-1}) - p + (\mu^\star - \widehat{\mu})^\top\widehat{\Sigma}^{-1}(\mu^\star - \widehat{\mu}) = \rho,$$

where the first equality follows from the definition of $\mathbb{D}$, and the second equality follows from the fact that $\kappa^\star$ solves (3.11c). This shows the feasibility of $(\mu^\star, \Sigma^\star)$.

Next, we prove the optimality of $(\mu^\star, \Sigma^\star)$. Through a tedious computation, one can show that

$$\begin{aligned}
& w^\top(\Sigma^\star + (\mu^\star)(\mu^\star)^\top)w = w^\top(\Sigma^\star + \Sigma^\star\widehat{\Sigma}^{-1}\widehat{\mu}\widehat{\mu}^\top\widehat{\Sigma}^{-1}\Sigma^\star)w \\
=& w^\top\widehat{\Sigma}w\Big(1 + \frac{w^\top\widehat{\Sigma}w}{\kappa^\star - w^\top\widehat{\Sigma}w}\Big) + (\widehat{\mu}^\top w)^2\Big(1 + \frac{2w^\top\widehat{\Sigma}w}{\kappa^\star - w^\top\widehat{\Sigma}w}\Big) + \frac{(w^\top\widehat{\mu})^2(w^\top\widehat{\Sigma}w)^2}{(\kappa^\star - w^\top\widehat{\Sigma}w)^2} \\
=& \frac{\kappa^\star w^\top\widehat{\Sigma}w}{\kappa^\star - w^\top\widehat{\Sigma}w} + \frac{(\kappa^\star)^2(\widehat{\mu}^\top w)^2}{(\kappa^\star - w^\top\widehat{\Sigma}w)^2} \\
=& \frac{\kappa^\star w^\top\widehat{\Sigma}w}{\kappa^\star - w^\top\widehat{\Sigma}w} + \frac{\kappa^\star(\widehat{\mu}^\top w)^2 w^\top\widehat{\Sigma}w}{(\kappa^\star - w^\top\widehat{\Sigma}w)^2} + \frac{\kappa^\star(\widehat{\mu}^\top w)^2}{\kappa^\star - w^\top\widehat{\Sigma}w} \\
=& \kappa^\star\rho - \kappa^\star\log\big(1 - \frac{w^\top\widehat{\Sigma}w}{\kappa^\star}\big) + \frac{\kappa^\star(\widehat{\mu}^\top w)^2}{\kappa^\star - w^\top\widehat{\Sigma}w} = f_0(\kappa^\star) = f(\beta),
\end{aligned}$$

where the antepenultimate equality follows from the fact that $\kappa^\star$ solves (3.11c), and the last equality holds because $\kappa^\star$ is the minimizer of (3.12). Therefore, $(\mu^\star, \Sigma^\star)$ is optimal to problem (3.11a). The uniqueness of $(\mu^\star, \Sigma^\star)$ now follows from the unique solution of Σ and μ with respect to the dual variables from (3.6) and (3.8), respectively.

It now remains to show the upper bound on $\kappa^\star$. Towards that end, we note that for any $\kappa > w^\top\widehat{\Sigma}w$,

$$\begin{aligned}
0 &= \rho - \frac{(w^\top\widehat{\mu})^2 w^\top\widehat{\Sigma}w}{(\kappa^\star - w^\top\widehat{\Sigma}w)^2} - \frac{w^\top\widehat{\Sigma}w}{\kappa^\star - w^\top\widehat{\Sigma}w} - \log\Big(1 - \frac{w^\top\widehat{\Sigma}w}{\kappa^\star}\Big) \\
&> \rho - \frac{(w^\top\widehat{\mu})^2 w^\top\widehat{\Sigma}w}{(\kappa^\star - w^\top\widehat{\Sigma}w)^2} - \frac{w^\top\widehat{\Sigma}w}{\kappa^\star - w^\top\widehat{\Sigma}w}.
\end{aligned}$$

Solving the above quadratic inequality in the variable $\kappa^\star - w^\top\widehat{\Sigma}w$ yields the desired bound. This completes the proof. □

We are now ready to prove Proposition 3.4.2.

Proof of Proposition 3.4.2. The convexity of f follows immediately by noting that it is the pointwise supremum of the family of convex functions $\mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2]$ parametrized by $\mathbb{Q}$.

To prove the continuously differentiability and the formula for the gradient, recall the expression (3.12) for the function $f(\beta)$:

$$f(\beta) = \left\{ \begin{array}{ll} \inf & \kappa\rho + \frac{\kappa(w^\top\widehat{\mu})^2}{\kappa - w^\top\widehat{\Sigma}w} - \kappa\log\big(1 - w^\top\widehat{\Sigma}w/\kappa\big) \\ \text{s.t.} & \kappa > w^\top\widehat{\Sigma}w. \end{array} \right. \tag{3.14}$$

Problem (3.14) has only one constraint. Therefore, LICQ (hence MFCQ) always holds, which implies that the Lagrange multiplier ζ_β of problem (3.14) is unique for any β. Also, it is easy to see that the constraint of problem (3.14) is never binding. So, $\zeta_\beta = 0$ for any β. The Lagrangian function $L_\beta : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is given by

$$L_\beta(\kappa, \zeta) = \rho\kappa + \frac{\omega_2 \kappa}{\kappa - \omega_1} - \kappa \log\left(1 - \frac{\omega_1}{\kappa}\right) + \zeta(\omega_1 - \kappa),$$

where $\omega_1 = w^\top \widehat{\Sigma} w$ and $\omega_2 = (w^\top \widehat{\mu})^2$. The first derivative with respect to κ is

$$\frac{\mathrm{d}L_\beta}{\mathrm{d}\kappa}(\kappa, \zeta) = \rho - \frac{\omega_1 \omega_2}{(\kappa - \omega_1)^2} - \log\left(1 - \frac{\omega_1}{\kappa}\right) - \frac{\omega_1}{\kappa - \omega_1} - \zeta.$$

The second derivative with respect to κ is

$$\frac{\mathrm{d}^2 L_\beta}{\mathrm{d}\kappa^2}(\kappa, \zeta) = \frac{\omega_1}{(\kappa - \omega_1)^3}\left(2\omega_2 + \frac{\omega_1}{\kappa}(\kappa - \omega_1)\right).$$

From the proof of Lemma 3.6.2, we have that the minimizer κ_β of problem (3.14) is precisely the $\kappa^\star$ defined by equation (3.11c) (below we write κ_β instead of $\kappa^\star$ to emphasize and keep track of the dependence on β). Therefore, for any β, the minimizer κ_β exists and is unique. So, there exists some constant $\eta_\beta > 0$ such that

$$\frac{\mathrm{d}^2 L_\beta}{\mathrm{d}\kappa^2}(\kappa_\beta, \zeta_\beta) \geq \eta_\beta > 0.$$

Therefore, for any β, the strong second order condition at κ_β holds (see [Sti18, Definition 6.2]). By [Sti18, Theorem 6.7],

$$\nabla f(\beta) = \nabla_\beta L_\beta(\kappa_\beta, \zeta_\beta) = \nabla_\beta L_\beta(\kappa_\beta, 0) \quad \forall \beta \in \mathbb{R}^d. \tag{3.15}$$

Then we compute

$$\begin{aligned}\nabla_w L_\beta(\kappa, \zeta) &= \nabla_w \left[\frac{\kappa(w^\top \widehat{\mu})^2}{\kappa - w^\top \widehat{\Sigma} w} - \kappa \log\left(1 - \frac{w^\top \widehat{\Sigma} w}{\kappa}\right) + \zeta(w^\top \widehat{\Sigma} w - \kappa)\right] \\ &= \frac{2\kappa\omega_2}{(\kappa - \omega_1)^2}\widehat{\Sigma} w + \frac{2\kappa}{(\kappa - \omega_1)}\widehat{\mu}\widehat{\mu}^\top w + \frac{2\kappa}{(\kappa - \omega_1)}\widehat{\Sigma} w + 2\zeta \widehat{\Sigma} w.\end{aligned}$$

Hence,

$$\nabla_\beta L_\beta(\kappa, \zeta) = \frac{dw}{d\beta}^\top \cdot \nabla_w L_\beta(\kappa, \zeta) = [I_d \; \mathbf{0}_d] \cdot \nabla_w L_\beta(\kappa, \zeta),$$

which, when combined with (3.15), yields the desired gradient formula

$$\nabla f(\beta) = \frac{2\kappa_\beta \left(\omega_2 \widehat{\Sigma} w + (\kappa_\beta - \omega_1)(\widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^\top) w\right)_{1:d}}{(\kappa_\beta - \omega_1)^2}.$$

By [Sti18, Theorem 6.5], the function $\beta \mapsto \kappa_\beta$ is locally Lipschitz continuous, *i.e.*, for any $\beta \in \mathbb{R}^d$, there exists $c_\beta, \epsilon_\beta > 0$ such that if $\|\beta' - \beta\|_2 \le \epsilon_\beta$, then

$$|\kappa_{\beta'} - \kappa_\beta| \le c_\beta \|\beta' - \beta\|_2 .$$

Note that ω_1 and ω_2 are both locally Lipschitz continuous in β. Also, it is easy to see that $\kappa_\beta > \omega_1$ for any β. Thus, $\nabla f(\beta)$ is locally Lipschitz continuous in β. □

Proof of 3.4.3. Noting that problem (3.3) is the barycenter problem between two Gaussian distributions with respect to the Wasserstein distance, the proof then directly follows from [AC11, §6.2] and [McC97, Example 1.7]. □

Proof of Proposition 3.4.4. Again we omit the subscripts λ and ρ. Reminding that $\xi = (X, Y)$, we find

$$\begin{aligned}
&\sup_{\mathbb{Q}\in\mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2] = \sup_{\mathbb{Q}\in\mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(w^\top \xi)^2] \\
&= \left\{ \begin{array}{ll} \inf & \kappa\big(\rho - \|\widehat{\mu}\|_2^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big]\big) + z + \operatorname{Tr}\big[Z\big] \\ \text{s.t.} & \kappa \in \mathbb{R}_+,\ z \in \mathbb{R}_+,\ Z \in \mathbb{S}_+^p \\ & \begin{bmatrix} \kappa I - ww^\top & \kappa \widehat{\Sigma}^{\frac{1}{2}} \\ \kappa \widehat{\Sigma}^{\frac{1}{2}} & Z \end{bmatrix} \succeq 0,\ \begin{bmatrix} \kappa I - ww^\top & \kappa \widehat{\mu} \\ \kappa \widehat{\mu}^\top & z \end{bmatrix} \succeq 0 \end{array} \right. \\
&= \left\{ \begin{array}{ll} \inf & \kappa\big(\rho - \|\widehat{\mu}\|_2^2 - \operatorname{Tr}\big[\widehat{\Sigma}\big]\big) + \kappa^2 \widehat{\mu}^\top (\kappa I - ww^\top)^{-1} \widehat{\mu} + \kappa^2 \operatorname{Tr}\big[\widehat{\Sigma}(\kappa I - ww^\top)^{-1}\big] \\ \text{s.t.} & \kappa \ge \|w\|_2^2, \end{array} \right.
\end{aligned} \tag{3.16}$$

where the second equality follows from [Kuh+19, Lemma 2]. By applying [Ber09a, Fact 2.16.3], we find

$$(\kappa I - ww^\top)^{-1} = \kappa^{-1} I + \kappa^{-2}\big(1 - \|w\|_2^2/\kappa\big)^{-1} ww^\top . \tag{3.17}$$

Combining (3.16) and (3.17), we get

$$\sup_{\mathbb{Q}\in\mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2] = \left\{ \begin{array}{ll} \inf & \kappa\rho + \kappa w^\top (\widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^\top) w / (\kappa - \|w\|_2^2) \\ \text{s.t.} & \kappa \ge \|w\|_2^2. \end{array} \right.$$

One can verify through the first-order optimality condition that the optimal solution $\kappa^\star$ is

$$\kappa^\star = \|w\|_2 \left(\|w\|_2 + \sqrt{\frac{w^\top (\widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^\top) w}{\rho}} \right),$$

and by replacing this value $\kappa^\star$ into the objective function, we find

$$\sup_{\mathbb{Q}\in\mathbb{B}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2] = \left(\sqrt{w^\top(\widehat{\Sigma} + \widehat{\mu}\widehat{\mu}^\top)w} + \sqrt{\rho}\|w\|_2\right)^2,$$

which then completes the proof. □

Proof of Section 3.5

Lemma 3.6.3 (Compactness)**.** *For $k \in \{\mathrm{S}, \mathrm{T}\}$, the set*

$$\mathbb{V}_k = \{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : M - \mu\mu^\top \in \mathbb{S}^p_{++}, \mathbb{D}((\mu, M - \mu\mu^\top) \parallel (\widehat{\mu}_k, \widehat{\Sigma}_k)) \le \rho_k\}$$

is convex and compact. Furthermore, the set

$$\mathbb{V} \triangleq \{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++} : (\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}\}$$

is also convex and compact.

Proof of Lemma 3.6.3. For any $(\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++}$ such that $M - \mu\mu^\top \in \mathbb{S}^p_{++}$, we find

$$\begin{aligned}
&\mathbb{D}\big((\mu, M - \mu\mu^\top) \parallel (\widehat{\mu}_k, \widehat{\Sigma}_k)\big)\\
=&(\mu - \widehat{\mu}_k)^\top \widehat{\Sigma}_k^{-1}(\mu - \widehat{\mu}_k) + \mathrm{Tr}\left[(M - \mu\mu^\top)\widehat{\Sigma}^{-1}\right] - \log\det((M - \mu\mu^\top)\widehat{\Sigma}_k^{-1}) - p\\
=&\widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1}\widehat{\mu}_k - 2\widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1}\mu + \mathrm{Tr}\left[M\widehat{\Sigma}_k^{-1}\right] - \log\det(M\widehat{\Sigma}_k^{-1}) - \log(1 - \mu^\top M^{-1}\mu) - p,
\end{aligned} \tag{3.18}$$

where in the last expression, we have used the determinant formula [Ber09a, Fact 2.16.3] to rewrite

$$\det(M - \mu\mu^\top) = (1 - \mu^\top M^{-1}\mu)\det M.$$

Because $M - \mu\mu^\top \in \mathbb{S}^p_{++}$, one can show that $1 - \mu^\top M^{-1}\mu > 0$ by invoking the Schur complement, and as such, the logarithm term in the last expression is well-defined. Moreover, we can write

$$\mathbb{V}_k = \left\{ (\mu, M) : \begin{array}{l} (\mu, M) \in \mathbb{R}^p \times \mathbb{S}^p_{++},\ M - \mu\mu^\top \in \mathbb{S}^p_{++},\ \exists t \in \mathbb{R}_+ : \\ \widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1}\widehat{\mu}_k - 2\widehat{\mu}_k^\top \widehat{\Sigma}_k^{-1}\mu + \mathrm{Tr}\left[M\widehat{\Sigma}_k^{-1}\right] \\ - \log\det(M\widehat{\Sigma}_k^{-1}) - \log(1 - t) - p \le \rho \\ \begin{bmatrix} M & \mu \\ \mu^\top & t \end{bmatrix} \succeq 0 \end{array} \right\}, \tag{3.19}$$

which is a convex set. Notice that by Schur complement, the semidefinite constraint is equivalent to $t \ge \mu^\top M^{-1}\mu$.

Next, we show that $\mathbb{V}_k$ is compact. Denote by $\mathbb{U}_k = \{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p : \mathbb{D}((\mu, \Sigma) \parallel (\widehat{\mu}_k, \widehat{\Sigma}_k)) \le \rho_k\}$. Then, it is easy to see that $\mathbb{V}_k$ is the image of $\mathbb{U}_k$ under the continuous mapping $(\mu, \Sigma) \mapsto (\mu, \Sigma + \mu\mu^\top)$. Therefore, it suffices to prove the compactness of $\mathbb{U}_k$. Towards that end, we note that

$$\mathbb{D}\big((\mu, \Sigma) \parallel (\widehat{\mu}_k, \widehat{\Sigma}_k)\big) = (\widehat{\mu}_k - \mu)^\top \widehat{\Sigma}_k^{-1} (\widehat{\mu}_k - \mu) + \operatorname{Tr}\big[\Sigma \widehat{\Sigma}_k^{-1}\big] - \log\det(\Sigma \widehat{\Sigma}_k^{-1}) - p$$

is a continuous and coercive function in (μ, Σ). Thus, as a level set of $\mathbb{D}\big((\mu, \Sigma) \parallel (\widehat{\mu}_k, \widehat{\Sigma}_k)\big)$, $\mathbb{U}_k$ is closed and bounded, and hence compact.

To prove the last claim, by the definitions of $\mathbb{V}$ and $\mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}$ we write

$$\begin{aligned}\mathbb{V} &= \{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}_{++}^p : (\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}\} \\ &= \{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}_{++}^p : (\mu, M) \in \mathbb{V}_{\mathrm{S}}\} \cap \{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}_{++}^p : (\mu, M) \in \mathbb{V}_{\mathrm{T}}\} \cap \qquad (3.20) \\ &\{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}_{++}^p : M \succeq \varepsilon I\}. \qquad (3.21)\end{aligned}$$

The convexity of $\{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}_{++}^p : (\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}\}$ then follows from the convexity of the three sets in (3.21). Furthermore, from the first part of the proof, we know that both $\{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}_{++}^p : (\mu, M) \in \mathbb{V}_{\mathrm{S}}\}$ and $\{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}_{++}^p : (\mu, M) \in \mathbb{V}_{\mathrm{T}}\}$ are compact sets, so is their intersection. Also, the last set $\{(\mu, M) \in \mathbb{R}^p \times \mathbb{S}_{++}^p : M \succeq \varepsilon I\}$ in (3.21) is closed. Since any closed subset of a compact set is again compact, we conclude that $\mathbb{V}$ is compact. This completes the proof.

□

Proof of Theorem 3.5.2. As $\xi = (X, Y)$, we can rewrite

$$\begin{aligned}&\min_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2] && (3.22\text{a}) \\ &= \min_{\beta \in \mathbb{R}^d} \sup_{\mathbb{Q} \in \mathbb{B}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}} \begin{bmatrix} \beta \\ -1 \end{bmatrix}^\top \mathbb{E}_{\mathbb{Q}}[\xi\xi^\top] \begin{bmatrix} \beta \\ -1 \end{bmatrix} && (3.22\text{b}) \\ &= \min_{\beta \in \mathbb{R}^d} \sup_{(\mu, M - \mu\mu^\top) \in \mathbb{U}_{\rho_{\mathrm{S}},\rho_{\mathrm{T}}}} \begin{bmatrix} \beta \\ -1 \end{bmatrix}^\top M \begin{bmatrix} \beta \\ -1 \end{bmatrix} \\ &= \min_{\beta \in \mathbb{R}^d} \sup_{(\mu, M) \in \mathbb{V}} \begin{bmatrix} \beta \\ -1 \end{bmatrix}^\top M \begin{bmatrix} \beta \\ -1 \end{bmatrix} \\ &= \sup_{(\mu, M) \in \mathbb{V}} \min_{\beta \in \mathbb{R}^d} \begin{bmatrix} \beta \\ -1 \end{bmatrix}^\top M \begin{bmatrix} \beta \\ -1 \end{bmatrix} && (3.22\text{c})\end{aligned}$$

$$= \sup_{(\mu,M)\in\mathbb{V}} M_{YY} - M_{XY}^\top M_{XX}^{-1} M_{XY} \tag{3.22d}$$

where (3.22c) follows from the Sion's minimax theorem, which holds because the objective function is convex in β, concave in M, and Lemma 3.6.3. Equation (3.22d) exploits the unique optimal solution in β as $\beta^\star = M_{XX}^{-1} M_{XY}$, in which the matrix inverse is well defined because $M \succ 0$ for any feasible M.

Finally, after an application of the Schur complement reformulation to (3.22d), the nonlinear semidefinite program in the theorem statement follows from representations (3.19) and (3.21). This completes the proof. □

Proof of Proposition 3.5.3. It is well-known that the space of probability measures equipped with the Wasserstein distance W_2 is a geodesic metric space (see [Vil08, Section 7] for example), meaning that for any two probability distributions $\mathcal{N}_0$ and $\mathcal{N}_1$, there exists a constant-speed geodesic curve $[0,1] \ni a \mapsto \mathcal{N}_a$ satisfying

$$W_2(\mathcal{N}_a, \mathcal{N}_{a'}) = |a - a'| W_2(\mathcal{N}_0, \mathcal{N}_1) \quad \forall a, a' \in [0,1].$$

The claim follows trivially if $W_2(\mathcal{N}_{\mathrm{S}}, \mathcal{N}_{\mathrm{T}}) \leq \sqrt{\rho_{\mathrm{S}}}$. Therefore, we assume $W_2(\mathcal{N}_{\mathrm{S}}, \mathcal{N}_{\mathrm{T}}) > \sqrt{\rho_{\mathrm{S}}}$.

Consider the the geodesic $\mathcal{N}_t$ from $\mathcal{N}_0 = \mathcal{N}_{\mathrm{S}}$ to $\mathcal{N}_1 = \mathcal{N}_{\mathrm{T}}$. Also, denote by $\mathbb{U}_k = \{(\mu, \Sigma) \in \mathbb{R}^p \times \mathbb{S}_+^p : \mathbb{D}((\mu, \Sigma) \parallel (\widehat{\mu}_k, \widehat{\Sigma}_k)) \leq \rho_k\}$ for $k \in \{\mathrm{S}, \mathrm{T}\}$. Then, $\mathbb{U}_{\mathrm{S}}$ and $\mathbb{U}_{\mathrm{T}}$ has empty intersection if and only if

$$W_2(\mathcal{N}_a, \mathcal{N}_{\mathrm{S}}) \leq \sqrt{\rho_{\mathrm{S}}} \Longrightarrow W_2(\mathcal{N}_a, \mathcal{N}_{\mathrm{T}}) > \sqrt{\rho_{\mathrm{T}}} \quad \forall a \in [0,1],$$

which is in turn equivalent to

$$a W_2(\mathcal{N}_{\mathrm{T}}, \mathcal{N}_{\mathrm{S}}) \leq \sqrt{\rho_{\mathrm{S}}} \Longrightarrow (1-a) W_2(\mathcal{N}_{\mathrm{T}}, \mathcal{N}_{\mathrm{S}}) \leq \sqrt{\rho_{\mathrm{T}}} \quad \forall a \in [0,1].$$

Picking $a = \frac{\sqrt{\rho_{\mathrm{S}}}}{W_2(\mathcal{N}_{\mathrm{T}}, \mathcal{N}_{\mathrm{S}})} \in (0,1)$, then we have

$$\left(1 - \frac{\sqrt{\rho_{\mathrm{S}}}}{W_2(\mathcal{N}_{\mathrm{T}}, \mathcal{N}_{\mathrm{S}})}\right) W_2(\mathcal{N}_{\mathrm{T}}, \mathcal{N}_{\mathrm{S}}) \leq \sqrt{\rho_{\mathrm{T}}}.$$

The above inequality can be rewritten as

$$W_2(\mathcal{N}_{\mathrm{T}}, \mathcal{N}_{\mathrm{S}}) \leq \sqrt{\rho_{\mathrm{S}}} + \sqrt{\rho_{\mathrm{T}}},$$

which contradicts with our supposition

$$\rho_{\mathrm{T}} \geq \left(\sqrt{\mathbb{W}((\widehat{\mu}_{\mathrm{S}}, \widehat{\Sigma}_{\mathrm{S}}) \parallel (\widehat{\mu}_{\mathrm{T}}, \widehat{\Sigma}_{\mathrm{T}}))} - \sqrt{\rho_{\mathrm{S}}}\right)^2.$$

Thus, $\mathbb{U}_{\mathrm{S}}$ and $\mathbb{U}_{\mathrm{T}}$ has non-empty intersection. □

Proof of Theorem 3.5.4. As $\xi = (X, Y)$, we can rewrite

$$\begin{aligned}
&\min_{\beta\in\mathbb{R}^d} \sup_{\mathbb{Q}\in\mathbb{B}_{\rho_\mathrm{S},\rho_\mathrm{T}}(\widehat{\mathbb{P}})} \mathbb{E}_{\mathbb{Q}}[(\beta^\top X - Y)^2] && (3.23\mathrm{a})\\
&= \min_{\beta\in\mathbb{R}^d} \sup_{(\mu, M-\mu\mu^\top)\in\mathbb{U}_{\rho_\mathrm{S},\rho_\mathrm{T}}} \begin{bmatrix}\beta\\-1\end{bmatrix}^\top M \begin{bmatrix}\beta\\-1\end{bmatrix} &&\\
&= \sup_{(\mu, M-\mu\mu^\top)\in\mathbb{U}_{\rho_\mathrm{S},\rho_\mathrm{T}}} \min_{\beta\in\mathbb{R}^d} \begin{bmatrix}\beta\\-1\end{bmatrix}^\top M \begin{bmatrix}\beta\\-1\end{bmatrix} && (3.23\mathrm{b})\\
&= \sup_{(\mu, M-\mu\mu^\top)\in\mathbb{U}_{\rho_\mathrm{S},\rho_\mathrm{T}}} M_{YY} - M_{XY}^\top M_{XX}^{-1} M_{XY} && (3.23\mathrm{c})
\end{aligned}$$

where (3.23b) follows from the Sion's minimax theorem, which holds because the objective function is convex in β, concave in M, and the set $\mathbb{U}_{\rho_\mathrm{S},\rho_\mathrm{T}}$ is compact [SA+18, Lemma A.6]. Equation (3.23c) exploits the unique optimal solution in β as $\beta^\star = M_{XX}^{-1} M_{XY}$, in which the matrix inverse is well defined because $M - \mu\mu^\top \succeq \varepsilon I$ for any feasible M. □

Additional Numerical Results

In the following the details of the datasets used in Section 3.6 are presented.

- **Uber&Lyft**[4] has $N_\mathrm{S} = 5000$ instances in the source domain and 5000 available samples in the target domain.
- **US Births (2018)**[5] has $N_\mathrm{S} = 5172$ samples in the source domain and 4828 available samples in the target domain.
- **Life Expectancy**[6] has $N_\mathrm{S} = 1407$ instances in the source domain and 242 available samples in the target domain.
- **House Prices in King County**[7] has $N_\mathrm{S} = 543$ instances in the source domain and 334 available samples in the target domain.
- **California Housing Prices**[8] has $N_\mathrm{S} = 9034$ instances in the source domain, and 6496 available instances in the target domain.

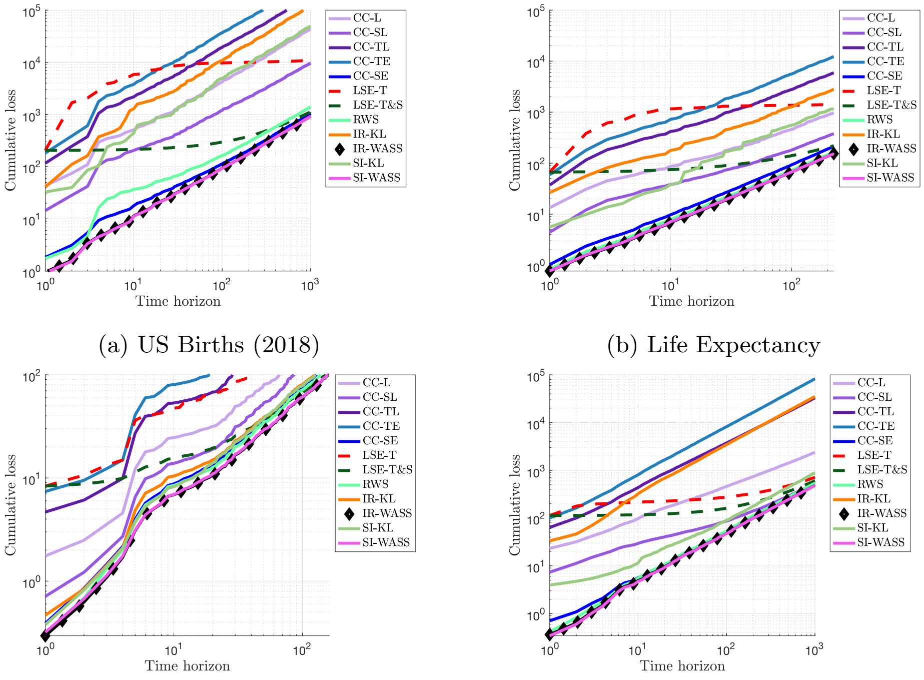
[4] Available publicly at https://www.kaggle.com/brllrb/uber-and-lyft-dataset-boston-ma

[5] Available publicly at https://www.kaggle.com/des137/us-births-2018

[6] Available publicly at https://www.kaggle.com/kumarajarshi/life-expectancy-who

[7] Available publicly at https://www.kaggle.com/c/house-prices-advanced-regression-techniques/data

[8] The modified version that we use is available publicly at https://www.kaggle.com/camnugent/california-housing-prices and the original dataset is available publicly at https://www.dcc.fc.up.pt/~ltorgo/Regression/cal_housing.html



(a) US Births (2018)

(b) Life Expectancy

(c) House Prices in KC

(d) California Housing

Figure 3.5: Cumulative loss averaged over 100 runs on logarithmic scale

Figure 3.5 demonstrates how the average cumulative loss in (3.1) grows over time for the US Births (2018), Life Expectancy, House Prices in KC and California Housing datasets. The results suggest that the IR-WASS and SI-WASS experts perform favorably over the competitors in that their cumulative loss at each time step is lower than that of most other competitors.

4. Learning Fair and Robust Models

4.1. Introduction

Machine learning models are increasingly being harnessed to aid decision-making across key sectors, influencing crucial outcomes such as employment, loan approvals, medical prescriptions, and judicial decisions on bail or parole. The potential for algorithms to surpass human decision-makers lies in their remarkable capacity to analyze extensive datasets beyond human capabilities and their efficiency in executing intricate calculations rapidly. Moreover, algorithms are perceived to provide a more objective alternative to human decision-making, which is often marred by subjectivity and vulnerability to biases. While algorithmic decision-making processes boast efficiency and comprehensive data utilization, their supposed objectivity is sometimes misleading. For instance, research has highlighted biases in algorithms within the U.S. criminal justice system, falsely suggesting that African Americans are likelier to commit crimes than white Americans [Cho17; Mul16]. Similarly, Google's ad-targeting algorithm has been shown to preferentially present higher-paying executive job ads to men over women [DTD15]. Moreover, an AI-driven hiring tool used by Amazon was found to be biased against women applying for software development and technical roles [Das18].

There are several possible explanations for biased behaviour of machine learning algorithms. First, the training data could already be corrupted by human biases due to biased device measurements or historically biased human decisions, amongst others [FN96]. Machine learning algorithms are designed to learn and preserve these biases [BG18; Man+16]. Second, minimizing the average prediction error privileges the majority populations over the minorities. Third, sensitive attributes can have an implicit detrimental effect on the decision making process even if they are not explicitly represented in the training data. Sensitive attributes are any attributes such as the race, gender or age of a person that distinguish privileged from unprivileged individuals. It is often illegal to use these sensitive attributes for decision making. Thus, a naïve approach to mitigate algorithmic biases would be to remove all sensitive information from the training data. This leads to *fairness through unawareness*. However, sensitive attributes are often correlated to other attributes that seem less problematic (such as a person's hair length or skin pigmentation), and this enables algorithms to make unfair recommendations based on predictions of the sensitive attributes. Ultimately, this results in an implicit use of the sensitive attributes under the guise of fairness [BS16; BYF20; Kle+18; LMC18].

The scientific community has spent substantial efforts to establish mathematical definitions of algorithmic fairness and to ensure that machine learning models are actually fair in the sense of these definitions. In the following, we explain some of the most popular fairness definitions in the context of binary classification and identify without loss of generality the positive outcome with the "advantaged" outcome, such as "admission to a college" or "receiving a promotion." *Demographic parity* [Dwo+12] requires the likelihood of a positive outcome (*e.g.*, a person being hired) to be the same regardless of whether the person is in the protected (*e.g.*, female) group or not. *Equalized odds* [Har+16] requires the probability of a person in the positive class being correctly classified and the probability of a person in a negative class being misclassified should both be the same for persons in the privileged and unprivileged groups. *Equal opportunities* [Har+16] can be viewed as a relaxation of the equalized odds criterion as it requires non-discrimination only within the privileged group. Hence, equal opportunities requires the true positive rates to be equal in the privileged and unprivileged groups. Other notions of fairness include the *disparate impact* [Fel+15] and *disparate mistreatment* [Zaf+17a] criteria. The central idea behind any notion of fairness is to require the decisions of a classifier to be balanced among the privileged and unprivileged groups and label sets. For a comprehensive survey and further discussions of fairness in machine learning we refer to [Ber+18; CR20; CD+17; Meh+19].

Logistic regression is one of the most popular classification methods [HJLS13]. Its objective is to establish a probabilistic relationship between a random feature vector $X \in \mathcal{X} = \mathbb{R}^p$ and a random binary explanatory variable $Y \in \mathcal{Y} = \{0, 1\}$. We assume here that there is a single sensitive attribute $A \in \mathcal{A} = \{0, 1\}$, which is also random and is *not* contained in the feature vector X, and we consider the privileged learning setting [VV09; QS17], where the sensitive information is only available at the training stage but not at the testing stage. Note that predicting Y from X ensures fairness through unawareness. In the remainder, we denote by $\{(\hat{x}_i, \hat{a}_i, \hat{y}_i)\}_{i=1}^N$ a finite set of training samples that are drawn independently from the probability distribution $\mathbb{P}$ of the joint random vector (X, A, Y). In logistic regression, the conditional probability $\mathbb{P}[Y = 1 | X = x]$ is modeled as the sigmoidal hypothesis

$$h_\beta(x) = (1 + \exp(-\beta^\top x))^{-1},$$

where the weight vector $\beta \in \mathbb{R}^p$ constitutes an unknown regression parameter. Classical logistic regression determines β by solving the tractable convex optimization problem

$$\min_{\beta} \ \frac{1}{N}\sum_{i=1}^{N} \ell_\beta(\widehat{x}_i, \widehat{y}_i), \quad \ell_\beta(x,y) = -y\log(h_\beta(x)) - (1-y)\log(1-h_\beta(x)) \tag{4.1}$$

which minimizes the empirical *log-loss*, that is, the negative log-likelihood function of the training data. To make logistic regression fair, we will include an *unfairness measure* in problem (4.1). Specifically, we will either include a fairness constraint that requires the unfairness measure to fall below a given threshold, or we will include the unfairness measure as a penalty term in the objective function. As it is not possible to satisfy multiple notions of fairness simultaneously [Ber+18; KMR16], we focus on unfairness measures related to equal opportunities. However, our method is general enough to cater for other notions of fairness.

Definition 4 (Unfairness measure). *If $f : [0,1] \to \mathbb{R}$ is measurable, then the unfairness of a hypothesis $h : \mathcal{X} \to [0,1]$ with respect to f under a distribution $\mathbb{Q}$ of (X, A, Y) is*

$$\mathbb{U}_f(\mathbb{Q}, h) = \big|\mathbb{E}_{\mathbb{Q}}[f(h(X))|A=1, Y=1] - \mathbb{E}_{\mathbb{Q}}[f(h(X))|A=0, Y=1]\big|.$$

The larger $\mathbb{U}_f(\mathbb{Q}, h)$, the more unfair is the hypothesis h, and if $\mathbb{U}_f(\mathbb{Q}, h) = 0$, then the hypothesis is maximally fair. Different choices of f induce different notions of fairness. If $f(z) = \mathbb{1}_{\{z\geq\tau\}}$, then $\mathbb{U}_f(\mathbb{Q}, h) = 0$ means that h is fair in view of the *equal opportunities* criterion [Har+16]. Here, $\tau \in [0,1]$ is the classification threshold. If $f(z) = z$, then $\mathbb{U}_f(\mathbb{Q}, h) = 0$ means that the hypothesis h is fair in view of the *probabilistic equal opportunities* criterion for probabilistic classifiers [Ple+17].

It is well known that increasing the fairness of an algorithm typically reduces its accuracy [Fri+19; LMC18; MW18]. This prompts us to introduce an ideal *fair* logistic regression model

$$\min_{\beta} \ \mathbb{E}_{\mathbb{P}}[-Y\log(h_\beta(X)) - (1-Y)\log(1-h_\beta(X))] + \eta\mathbb{U}_f(\mathbb{P}, h_\beta), \tag{4.2}$$

where $\eta \in \mathbb{R}_+$ is a tuning parameter that balances the trade-off between accuracy and fairness. Unfortunately, problem (4.2) is difficult to solve for several reasons. If $f(z) = \mathbb{1}_{\{z\geq\tau\}}$, then the unfairness measure $\mathbb{U}_f(\mathbb{P}, h_\beta)$ is discontinuous in β, and if $f(z) = z$, then $\mathbb{U}_f(\mathbb{P}, h_\beta)$—though smooth—is still non-convex in β. In both cases, it seems difficult to solve (4.2) to global optimality. In addition, the distribution $\mathbb{P}$ is unknown and only indirectly observable through the N independent training samples. Thus, an important input for problem (4.2) is unavailable in practice. The latter shortcoming could be addressed by simply replacing the unknown true distribution $\mathbb{P}$ in (4.2) with the empirical distribution $\widehat{\mathbb{P}}_N$, which is defined as the discrete uniform distribution on the N training samples. However, this naïve approach could result in over-fitting and yield classifiers with a poor out-of-sample performance (both in terms of accuracy and fairness) if N is small relative to p.

The concerns over poor out-of-sample performance prompt us to pursue a *distributionally robust* approach, whereby the objective function in (4.2) is minimized in view of the most adverse distribution $\mathbb{Q}$ within some *ambiguity set* that reflects all available distributional information. The ambiguity set could be characterized through moment and support information [DY10; GS10; WKS14], or it could be defined as a ball around $\widehat{\mathbb{P}}_N$ with respect to a distance measure for distributions such as the Prohorov metric [EI06] or the Kullback-Leibler divergence [HH13]. Due to its attractive measure concentration properties, we use here the Wasserstein metric to construct ambiguity sets [Kuh+19; MEK18; PW07]. Moreover, Wasserstein distributional robustness offers probabilistic interpretations for popular regularization techniques [BKM19; GCK17; SMK15].

The main contributions of this chapter can be summarized as follows.

1. **Log-probabilistic equal opportunities:** We propose a new unfairness measure and the corresponding fairness criterion, termed log-probabilistic equal opportunities, which approximates the probabilistic equal opportunities criterion. We then prove that the empirical (*i.e.*, $\mathbb{P} = \widehat{\mathbb{P}}_N$) fair logistic regression model (4.2) with the new unfairness measure is equivalent to a tractable convex program.

2. **Distributionally robust fair logistic regression:** We robustify the fair logistic regression model against all distributions in a Wasserstein ball centered at $\widehat{\mathbb{P}}_N$, and we prove that this model is still equivalent to a tractable convex program if unfairness is quantified under the log-probabilistic equal opportunities criterion. Experiments suggest that the resulting classifiers improve fairness at a marginal loss of accuracy.

3. **Unfairness quantification:** Using similar techniques from Wasserstein distributionally robust optimization, we develop two highly tractable linear programs whose optimal values provide confidence bounds on the unfairness of *any* fixed classifier with respect to the (classical) probabilistic equal opportunities criterion. We also devise a hypothesis test that checks whether a given classifier is fair in view of equal opportunities.

The existing literature on algorithmic fairness can be subdivided into three categories. Papers in the first category propose to pre-process the training data before solving a plain-vanilla classification problem [Cal+17; Gor+19; Fel+15; KC12; LRT11; Sam+18; Zem+13]. Papers in the second category enforce fairness during the training step by appending fairness constraints to the classification problem [Don+18; MW18; Woo+17; Zaf+17a; Zaf+17b], by including regularization terms that penalize discrimination [Bah+20; HV19; Kam+12; KAS11] or by (approximately) penalizing any mismatches between the true positive rates and the false negative rates across different groups [BL17]. Several other papers in this category propose adversarial approaches to algorithmic fairness [ES15; Gar+19; Has+18; KKG18; Mad+18; Rez+20; YBS20; ZLM18].

Papers in the third category modify a pre-trained classifier in order to increase its fairness properties while preserving its classification performance as much as possible [CD+17; Dwo+18; Har+16; MW18].

The method proposed here can be viewed as an adversarial approach pertaining to the second category. There are only few other papers that study fairness from a distributionally robust perspective. A classification model with fairness constraints embedded in the ambiguity set is proposed in [Rez+20], a repeated loss minimization model with a χ^2-divergence ambiguity set is considered in [Has+18] and robust fairness constraints based on a total variation ambiguity set that captures noisy protected group information is described in [Wan+20b]. In addition, a fair distributionally robust classification model with a Wasserstein ambiguity set is studied in [YBS20], but this model deals with individual fairness and does not admit a tractable convex reformulation. In contrast, we consider marginally constrained Wasserstein ambiguity sets to enforce a notion of group fairness and provide a tractable convex reformulation.

4.2. Fair Logistic Regression

Recall that the fair logistic regression model (4.2) is non-convex if $f(z) = \mathbb{1}_{\{z \geq \tau\}}$, which induces equal opportunities, or if $f(z) = z$, which induces probabilistic equal opportunities. In order to convexify (4.2), we thus propose a new unfairness measure corresponding to $f(z) = \log(z)$, and we refer to the fairness criterion induced by the condition $\mathbb{U}_f(\mathbb{Q}, h) = 0$ as *log-probabilistic equal opportunities*. A classifier is fair in view of this criterion if the expected log-probability of a person in the positive class being correctly classified is the same for persons in the privileged and unprivileged groups. We also note that the log-probability function $f(h_\beta(x)) = -\log(1 + \exp(-\beta^\top x))$ can be viewed as a concave approximation of the sigmoid function $h_\beta(x)$. Concave (or convex) approximations of non-convex functions are routinely used in machine learning and arise, for example, when one replaces a non-convex loss function (such as the zero-one loss) with a convex surrogate loss function (such as the hinge loss or the log-loss) or when one replaces a non-convex risk measure (such as the value-at-risk) with a convex one (such as the conditional value-at-risk).

We now denote by $\widehat{p}_{ay} = \widehat{\mathbb{P}}_N(A = a, Y = y)$ the empirical proportion of people with attribute $a \in \mathcal{A}$ in class $y \in \mathcal{Y}$, and we define $r_a = 1/\hat{p}_{a1}$ for all $a \in \mathcal{A}$. Using this notation, we can prove that the logistic regression model (4.2) with the log-probabilistic equal opportunities unfairness measure is tractable under the empirical distribution for all sufficiently small η.

Theorem 4.2.1 (Fair logistic regression)**.** *If $f(z) = \log(z)$, $\eta \leq \min\{\hat{p}_{11}, \hat{p}_{01}\}$*

and $\mathbb{P} = \widehat{\mathbb{P}}_N$, *then problem* (4.2) *is equivalent to the tractable convex program*

$$\begin{array}{cl} \min\limits_{\beta\in\mathbb{R}^p, t\in\mathbb{R}} & t \\ \text{s.t.} & \mathbb{E}_{\widehat{\mathbb{P}}_N}[\ell_\beta(X,Y) + \eta r_1 \log(h_\beta(X))\mathbb{1}_{(1,1)}(A,Y) - \\ & \qquad \eta r_0 \log(h_\beta(X))\mathbb{1}_{(0,1)}(A,Y)] \leq t \\ & \mathbb{E}_{\widehat{\mathbb{P}}_N}[\ell_\beta(X,Y) + \eta r_0 \log(h_\beta(X))\mathbb{1}_{(0,1)}(A,Y) - \\ & \qquad \eta r_1 \log(h_\beta(X))\mathbb{1}_{(1,1)}(A,Y)] \leq t, \end{array}$$

where the expectation under $\widehat{\mathbb{P}}_N$ *is a finite sum.*

4.3. Distributionally Robust Fair Logistic Regression

Approximating the unknown data-generating distribution $\mathbb{P}$ with the empirical distribution $\widehat{\mathbb{P}}_N$ may result in overfitting. Following [BKM19; GCK17; SMK15], we thus regularize the nominal classification problem under $\widehat{\mathbb{P}}_N$ by robustifying it against all distributions in a Wasserstein ball around $\widehat{\mathbb{P}}_N$ that contains the unknown true distribution $\mathbb{P}$ with high confidence.

Definition 5 (Wasserstein distance)**.** *The type-*1 *Wasserstein distance between two probability distributions* $\mathbb{Q}_1$ *and* $\mathbb{Q}_2$ *of a random vector* $\xi \in \mathbb{R}^n$ *is defined as*

$$\mathbb{W}(\mathbb{Q}_1, \mathbb{Q}_2) = \inf_{\pi\in\Pi(\mathbb{Q}_1,\mathbb{Q}_2)} \mathbb{E}_\pi[c(\xi_1, \xi_2)], \tag{4.3}$$

where $\Pi(\mathbb{Q}_1, \mathbb{Q}_2)$ *denotes the set of all joint distributions of the random vectors* $\xi_1 \in \mathbb{R}^n$ *and* $\xi_2 \in \mathbb{R}^n$ *under which* ξ_1 *and* ξ_2 *have marginal distributions* $\mathbb{Q}_1$ *and* $\mathbb{Q}_2$*, respectively, and where* $c : \mathbb{R}^n \times \mathbb{R}^n \to [0, \infty]$ *constitutes a lower semi-continuous ground metric.*

When computing Wasserstein distances between distributions on $\mathcal{X}\times\mathcal{A}\times\mathcal{Y}$, we will use

$$c\big((x, a, y), (x', a', y')\big) = \|x - x'\| + \kappa_{\mathcal{A}}|a - a'| + \kappa_{\mathcal{Y}}|y - y'| \tag{4.4}$$

as the ground metric, where $\|\cdot\|$ is a norm on $\mathbb{R}^p$ and $\kappa_{\mathcal{A}}, \kappa_{\mathcal{Y}} \in (0, \infty]$. Using the Wasserstein distance with the ground metric (4.4), we define the ambiguity set $\mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$ as the Wasserstein ball of radius $\rho \geq 0$ around the empirical distribution $\widehat{\mathbb{P}}_N$, intersected with the set of all distributions under which the marginal of (A, Y) matches the empirical marginal. Thus,

$$\mathbb{B}_\rho(\widehat{\mathbb{P}}_N) = \left\{\mathbb{Q} \in \mathcal{M} : \mathbb{W}(\mathbb{Q}, \widehat{\mathbb{P}}_N) \leq \rho,\ \mathbb{Q}(A = a, Y = y) = \hat{p}_{ay} \quad \forall a \in \mathcal{A},\ y \in \mathcal{Y}\right\},$$

where $\mathcal{M}$ stands for the set of all possible distributions on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$. Note that $\mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$ is non-empty as it contains at least $\widehat{\mathbb{P}}_N$. Note also that all distributions in $\mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$ can be obtained by reshaping $\widehat{\mathbb{P}}_N$ at a transportation cost of at most ρ. The parameter $\kappa_\mathcal{A}$ represents the transportation cost of changing the sensitive attribute from A to $1 - A$, and thus it can be viewed as our trust in A. A similar interpretation applies to $\kappa_\mathcal{Y}$. We can now formally introduce the *distributionally robust fair* logistic regression model

$$\min_{\beta} \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\hat{\mathbb{P}}_N)} \mathbb{E}_\mathbb{Q}[-Y \log(h_\beta(X)) - (1-Y)\log(1-h_\beta(X))] + \eta \mathbb{U}_f(\mathbb{Q}, h_\beta), \quad (4.5)$$

which minimizes a combination of the expected log-loss and some unfairness measure under the most adverse distribution in $\mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$. Wasserstein ambiguity sets with marginal constraints were first studied in [Fro+19], where it was found that restricting the marginals of the outputs and/or the features eliminates unrealistic data distributions from the ambiguity set and often improves the performance of the resulting classifiers while maintaining strong robustness guarantees. We are now ready to prove that (4.5) is tractable if the log-probabilistic equal opportunities unfairness measure is used and if η is sufficiently small.

Theorem 4.3.1 (Distributionally robust fair logistic regression). *If $f(z) = \log(z)$ and $\eta \leq \min\{\hat{p}_{11}, \hat{p}_{01}\}$, then problem (4.5) is equivalent to the tractable convex program*

$$\begin{array}{ll}
\min & t \\
\text{s.t.} & \beta \in \mathbb{R}^p,\ t \in \mathbb{R},\ \lambda_0, \lambda_1 \in \mathbb{R}_+,\ \mu_0, \mu_1 \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{Y}|},\ \nu_0, \nu_1 \in \mathbb{R}^N \\
& \|\beta\|_*(1 + \eta r_0) \leq \lambda_1, \quad \|\beta\|_*(1 + \eta r_1) \leq \lambda_0 \\
& \rho\lambda_{a'} + \sum\limits_{a \in \mathcal{A},\, y \in \mathcal{Y}} \hat{p}_{ay}\, \mu_{a'ay} + \frac{1}{N}\sum\limits_{i=1}^N \nu_{a'i} \leq t \quad \forall a' \in \{0, 1\} \\
& \left.\begin{array}{l}
\log(h_\beta(-\hat{x}_i)) + \kappa_\mathcal{A}|a - \hat{a}_i|\lambda_a + \kappa_\mathcal{Y}|\hat{y}_i|\lambda_a + \mu_{aa0} + \nu_{ai} \geq 0 \\
\log(h_\beta(-\hat{x}_i)) + \kappa_\mathcal{A}|a' - \hat{a}_i|\lambda_a + \kappa_\mathcal{Y}|\hat{y}_i|\lambda_a + \mu_{aa'0} + \nu_{ai} \geq 0 \\
(1 - \eta r_a)\log(h_\beta(\hat{x}_i)) + \kappa_\mathcal{A}|a - \hat{a}_i|\lambda_a + \kappa_\mathcal{Y}|1 - \hat{y}_i|\lambda_a + \mu_{aa1} + \nu_{ai} \geq 0 \\
(1 + \eta r_{a'})\log(h_\beta(\hat{x}_i)) + \kappa_\mathcal{A}|a' - \hat{a}_i|\lambda_a + \kappa_\mathcal{Y}|1 - \hat{y}_i|\lambda_a + \mu_{aa'1} + \nu_{ai} \geq 0
\end{array}\right\} \\
& \qquad\qquad\qquad\qquad\qquad\qquad \forall i \in [N], \forall a, a' \in \mathcal{A} : a' = 1 - a,
\end{array}$$

where $\|\cdot\|_*$ represents the norm dual to $\|\cdot\|$ on $\mathbb{R}^p$.

Note that the assumption on η implies that $\eta r_a = \eta/\hat{p}_{a1} \leq 1$ for all $a \in \mathcal{A}$, and thus it is easy to verify that the reformulation of Theorem 4.3.1 is indeed convex. For many commonly used norms, this reformulation can be addressed with an exponential cone solver such as MOSEK. Alternatively, one may develop customized first-order methods by adapting the algoritghm proposed in [LHS19] to account for an unfairness measure in the objective.

4.4. Unfairness Quantification

A regulator may find it difficult to quantify the degree of discrimination of each pre-trained logistic classifier because this quantity depends critically on the test data at hand. To this end, we first define the worst (highest) possible unfairness levels of the classifier h across all distributions in a Wasserstein ambiguity set of the form $\mathbb{B}_\rho(\widehat{\mathbb{P}}^N)$ as

$$\overline{\mathbb{U}}_f = \sup\nolimits_{\mathbb{Q}\in\mathbb{B}_\rho(\widehat{\mathbb{P}}^N)} \mathbb{U}_f(\mathbb{Q}, h).$$

Here, by slight abuse of notation, $\widehat{\mathbb{P}}^N$ should be interpreted as the discrete uniform distribution on N *test samples* $\{(\hat{x}_i, \hat{a}_i, \hat{y}_i)\}_{i=1}^N$ drawn independently from $\mathbb{P}$.

The first main result of this section is to show that $\overline{\mathbb{U}}_f$ can be re-expressed in terms of the optimal values of two highly scalable linear programs when $f(z) = \mathbb{1}_{\{z\geq\tau\}}$, that is, when unfairness is measured with respect to the standard equal opportunity criterion. Thus, there is no need to resort to approximations involving log-probabilities.

To see this, we define $\mathcal{X}_0 = \{x \in \mathcal{X} : h(x) < \tau\}$ and $\mathcal{X}_1 = \{x \in \mathcal{X} : h(x) \geq \tau\}$, and we set

$$\mathbb{V}(a,a') = \sup_{\mathbb{Q}\in\mathbb{B}_\rho(\hat{\mathbb{P}}_N)} \mathbb{Q}[X \in \mathcal{X}_1 | A = a, Y = 1] - \mathbb{Q}[X \in \mathcal{X}_1 | A = a', Y = 1] \ \forall a, a' \in \mathcal{A}. \tag{4.6}$$

In addition, we define $d_{yi} = \inf_{x\in\mathcal{X}_y} \|x - \widehat{x}_i\|$ for all $y \in \mathcal{Y}$ and $i \in [N]$ as the distances of the testing features $\hat{x}_i$ to the sets $\mathcal{X}_y$. Our ability to quantify the fairness of h will critically depend on whether d_{yi} can be computed efficiently. For linear classifiers the sets $\mathcal{X}_1$ and $\mathcal{X}_0$ constitute half-spaces, and therefore d_{yi} can be computed in closed form. For more complicated classifiers such as neural networks, however, one may have to resort to heuristics to estimate d_{yi}. Using this notation, we can state the following main result.

Theorem 4.4.1 (Unfairness quantification)**.** *If $f(z) = \mathbb{1}_{\{z\geq\tau\}}$, then we obtain $\overline{\mathbb{U}}_f = \max\{\mathbb{V}(1, 0), \mathbb{V}(0, 1)\}$, where $\mathbb{V}(a, a')$ can be computed for all $a, a' \in \mathcal{A}$ with $a \neq a'$ as the optimal value of a tractable linear program, that is,*

$$\mathbb{V}(a,a') = \left\{ \begin{array}{ll} \min & \rho\lambda + \widehat{p}^\top \mu + N^{-1}\mathbf{1}^\top \nu \\ \text{s.t.} & \lambda \in \mathbb{R}_+,\ \mu \in \mathbb{R}^{2\times 2},\ \nu \in \mathbb{R}^N \\ & \left. \begin{array}{l} \nu_i + \kappa_{\mathcal{A}}|a - \widehat{a}_i|\lambda + \kappa_{\mathcal{Y}}|\widehat{y}_i|\lambda + \mu_{a0} \geq 0 \\ \nu_i + \kappa_{\mathcal{A}}|a' - \widehat{a}_i|\lambda + \kappa_{\mathcal{Y}}|\widehat{y}_i|\lambda + \mu_{a'0} \geq 0 \\ \nu_i + \kappa_{\mathcal{A}}|a - \widehat{a}_i|\lambda + \kappa_{\mathcal{Y}}|1 - \widehat{y}_i|\lambda + \mu_{a1} \geq 0 \\ \nu_i + d_{1i}\lambda + \kappa_{\mathcal{A}}|a - \widehat{a}_i|\lambda + \kappa_{\mathcal{Y}}|1 - \widehat{y}_i|\lambda + \mu_{a1} \geq r_a \\ \nu_i + d_{0i}\lambda + \kappa_{\mathcal{A}}|a' - \widehat{a}_i|\lambda + \kappa_{\mathcal{Y}}|1 - \widehat{y}_i|\lambda + \mu_{a'1} \geq 0 \\ \nu_i + \kappa_{\mathcal{A}}|a' - \widehat{a}_i|\lambda + \kappa_{\mathcal{Y}}|1 - \widehat{y}_i|\lambda + \mu_{a'1} \geq -r_{a'} \end{array} \right\} \forall i \in [N]. \end{array} \right.$$

The bounds on the unfairness measure related to equal opportunity can be computed even faster if we have absolute trust in A and Y, that is, if $\kappa_{\mathcal{A}} = \kappa_{\mathcal{Y}} = \infty$. To see this, we select $\hat{x}_i^\star \in \operatorname{argmin}_{x_i \in \partial \mathcal{X}_1} \|x_i - \hat{x}_i^\star\|$ and we assume for the simplicity of exposition that $\|\hat{x}_i - \hat{x}_i^\star\| > 0$ for all $i \in [N]$. We define non-negative rewards and weights through

$$(c_{aa'i}, w_{aa'i}) = \begin{cases} (r_a, d_{1i}) & \text{if } \widehat{x}_i \in \operatorname{int}(\mathcal{X}_0),\ \widehat{a}_i = a,\ \widehat{y}_i = 1, \\ (r_{a'}, d_{0i}) & \text{if } \widehat{x}_i \in \operatorname{int}(\mathcal{X}_1),\ \widehat{a}_i = a',\ \widehat{y}_i = 1, \\ (0, +\infty) & \text{otherwise} \end{cases}$$

for all $a, a' \in \mathcal{A}$ and $i \in [N]$. In addition, we introduce the notational shorthand

$$\hat{\mathbb{V}}(a, a') = \widehat{\mathbb{P}}_N[X \in \mathcal{X}_1 | A = a, Y = 1] - \widehat{\mathbb{P}}_N[X \in \mathcal{X}_1 | A = a', Y = 1] \quad \forall a, a' \in \mathcal{A},$$

which can be evaluated by computing a finite sum. We can then prove the following theorem.

Theorem 4.4.2 (Absolute trust in A and Y). *If $f(z) = \mathbb{1}_{\{z \geq \tau\}}$ and $\kappa_{\mathcal{A}} = \kappa_{\mathcal{Y}} = \infty$, then*

$$\mathbb{V}(a, a') = \hat{\mathbb{V}}(a, a') + \max_{z \in [0,1]^N} \left\{ \frac{1}{N} \sum_{i \in [N]} c_{aa'i} z_i \ :\ \frac{1}{N} \sum_{i \in [N]} w_{aa'i} z_i \leq \rho \right\} \ \forall a, a' \in \mathcal{A}. \tag{4.7}$$

Theorem 4.4.2 asserts that evaluating $\mathbb{V}(a, a')$ is tantamount to solving a continuous knapsack problem in N variables, which can be solved by a greedy heuristics in time $\mathcal{O}(N \log N)$.

It is instructive to study the worst- and best-case distributions that determine $\overline{\mathbb{U}}_f$ and $\underline{\mathbb{U}}_f$. By Theorem 4.4.1, these extremal distributions can be constructed from the extremal distributions that determine $\mathbb{V}(1, 0)$ and $\mathbb{V}(0, 1)$. As the objective function of (4.6) represents a conditional expectation of a discontinuous integrand that fails to be upper semi-continuous, however, the supremum in (4.6) is not attained. We thus construct suboptimal distributions that attain the supremum of (4.6) asymptotically. For linear classifiers, the projections $\hat{x}_i^\star$ of the test samples to the decision boundary may be constructed analytically. For more sophisticated classifiers, however, they may have to be approximated using heuristic methods.

Proposition 4.4.3 (Extremal distributions). *If $f(z) = \mathbb{1}_{\{z \geq \tau\}}$, $\kappa_{\mathcal{A}} = \kappa_{\mathcal{Y}} = \infty$ and $z^\star$ is a maximizer of the linear program in (4.7) for some fixed $a, a' \in \mathcal{A}$, then*

$$\mathbb{Q}^\star = \frac{1}{N} \left(\textstyle\sum_{i=1}^N z_i^\star \delta_{(\widehat{x}_i^\star, \widehat{a}_i, \widehat{y}_i)} + \sum_{i=1}^N (1 - z_i^\star) \delta_{(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)} \right),$$

is feasible in (4.6), and for any $\varepsilon > 0$ there exists $\mathbb{Q}_\varepsilon^\star \in \mathbb{B}_\varepsilon(\mathbb{Q}^\star)$ that is ε-suboptimal in (4.6).

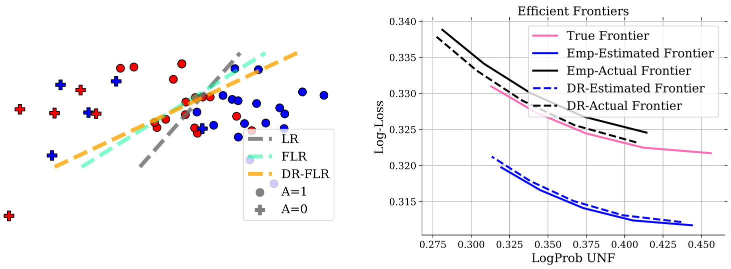


Figure 4.1: Classification boundaries (left), Pareto Frontiers (right)

Note that $\mathbb{Q}^\star$ is in general strictly suboptimal in (4.6), but every neighborhood of $\mathbb{Q}^\star$ contains ε-suboptimal distributions $\mathbb{Q}^\star_\varepsilon$ for any $\varepsilon > 0$. In principle, $\mathbb{Q}^\star_\varepsilon$ can be constructed explicitly from $\mathbb{Q}^\star$. However, the construction is cumbersome and therefore omitted.

4.5. Numerical Experiments

Below we refer $\mathbb{U}_f(\mathbb{P}, h)$ as the deterministic unfairness (Det-UNF) if $f(z) = \mathbb{1}_{\{z \geq \tau\}}$, the probabilistic unfairness (Prob-UNF) if $f(z) = z$ and the log-probabilistic unfairness (LogProb-UNF) if $f(z) = \log(z)$. Details regarding the setup of the experiments such as the data generation procedure and parameter selection etc. are relegated to Appendix 4.6.

Synthetic Experiments. To show the effects of the unfairness penalty and the robustification, we compare the classical, fair and distributionally robust fair logistic regression models (LR, FLR and DR-FLR, respectively) on a dataset with $N = 25$ training samples and $p = 2$ features. As the sensitive attribute A strongly correlates with X_1, fair classifiers assign low weight to X_1, which leads to horizontal decision boundaries. Penalizing unfairness with $\eta = 0.1$ and robustifying the model with a Wasserstein radius of $\rho = 0.05$ ostensibly increases the fairness of the classifier, see Figure 4.1 (left). Compared to the LR classifier, the DR-FLR classifier lowers Det-UNF from 0.86 to 0.58 at the expense of reducing the accuracy from 69% to 62%.

The fair logistic regression model (4.5) constitutes a bi-criteria optimization problem that simulataneously minimizes the log-loss and the log-probabilistic unfairness. It is thus reminiscent of the Markowitz mean-variance model that seeks an optimal trade-off between the risk and return of an investment portfolio. The optimal classifers for different values of η trace out a Pareto frontier in the unfairness/loss plane. Following [Bro93], we can now distinguish true, estimated and actual Pareto frontiers. The true frontier is obtained by training and evaluating the classifier under the (unknown) true distribution, while the estimated and actual frontiers are obtained by training the classifier on the training dataset and evaluating it on the training and testing datasets, respectively. It is

Dataset	Metric	LR	FLR	DOB$^+$[Don+18]	ZVRG [Zaf+17a]	DR-FLR
Drug	Accuracy	0.78±0.01	0.78 ± 0.01	0.78 ± 0.01	**0.79 ± 0.01**	0.78 ± 0.00
	Det-UNF	0.08 ± 0.06	0.08 ± 0.05	0.10 ± 0.09	0.48 ± 0.09	**0.03 ± 0.05**
	Prob-UNF	0.08 ± 0.04	0.08 ± 0.04	-	-	**0.05 ± 0.02**
	LogProb-UNF	0.23 ± 0.19	0.24 ± 0.19	-	-	**0.15 ± 0.10**
Adult	Accuracy	**0.80±0.01**	**0.80 ± 0.01**	0.78 ± 0.02	0.77 ± 0.01	0.79 ± 0.01
	Det-UNF	0.08±0.05	**0.06 ± 0.05**	0.08 ± 0.08	0.10 ± 0.06	**0.06 ± 0.04**
	Prob-UNF	0.17±0.07	0.12 ± 0.07	–	–	**0.12 ± 0.07**
	LogProb-UNF	0.98±0.55	0.64 ± 0.51	–	–	**0.56 ± 0.42**
Compas	Accuracy	**0.65±0.01**	**0.65 ± 0.02**	0.58 ± 0.04	**0.65 ± 0.01**	0.58 ± 0.04
	Det-UNF	0.25±0.03	0.24 ± 0.03	0.12 ± 0.07	0.22 ± 0.01	**0.11 ± 0.07**
	Prob-UNF	0.12±0.02	0.11 ± 0.02	–	–	**0.02 ± 0.02**
	LogProb-UNF	0.28±0.07	0.24 ± 0.07	–	–	**0.06 ± 0.04**
Arrhythmia	Accuracy	**0.63±0.03**	0.62 ± 0.03	0.61 ± 0.03	0.62 ± 0.03	0.61 ± 0.03
	Det-UNF	0.17±0.08	0.12 ± 0.07	0.08 ± 0.06	0.23 ± 0.13	**0.07 ± 0.06**
	Prob-UNF	0.10±0.05	0.06 ± 0.04	–	–	**0.03 ± 0.03**
	LogProb-UNF	0.21±0.10	0.14 ± 0.08	–	–	**0.07 ± 0.05**

Table 4.1: Testing accuracy and unfairness (average ± standard deviation) for $N = 150$.

known that the estimated frontier optimistically underestimates and the actual frontier pessimistically overestimates the true frontier on average [Bro93]. It has also been argued that robustifying a bi-criteria model tends to move the actual and estimated frontiers closer to each other as well as closer to the true frontier [MCG10], thus improving out-of-sample performance. Figure 4.1 (right) visualizes this effect for a synthetic dataset, where the sensitive attributes correlate with the labels.

Experiments with Real Data. We now benchmark the LR, FLR and DR-FLR classifiers against fair classifiers proposed in [Don+18] (DOB$^+$) and [Zaf+17a] (ZVRG) on four publicly available datasets (Adult, Drug, COMPAS, Arrhythmia[1]). While the Adult dataset comes with designated training and testing samples, in all other datasets we randomly select 2/3 of the samples for training. Ultimately, the ratio of training samples to features is of the order of 10 in all datasets.

To train the DR-FLR classifier, we draw 150 training samples and keep the others as validation samples. We then set $\eta = \min\{\hat{p}_{11}, \hat{p}_{01}\}/2$, $\kappa_{\mathcal{A}} = \kappa_{\mathcal{Y}} = 0.5$ and tune $\rho \in [10^{-5}, 10^{-1}]$[2] on a logarithmic search grid with 50 discretization points using the validation procedure from [Don+18]. Using these hyperparameters, we then re-train the DR-FLR classifier on another set of 150 randomly drawn training samples. The DOB$^+$ and ZVRG classifiers are computed using

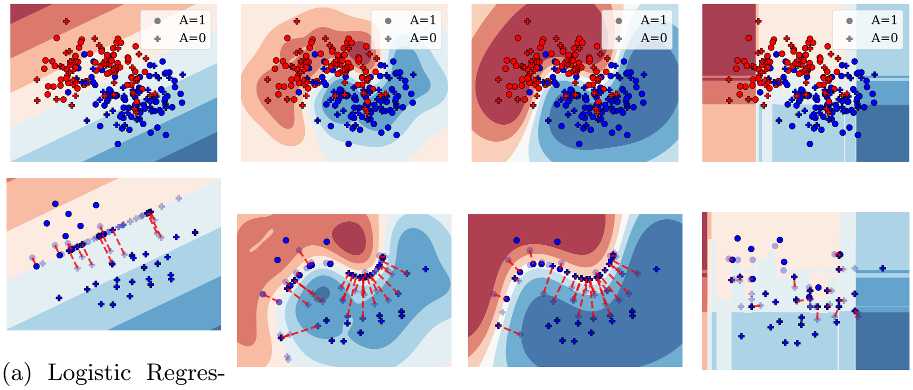
[1] We only use the first 12 out of 278 non-sensitive features of the Arrhythmia dataset so that we can use the same search grid for ρ across all datasets (in the other datasets p ranges from 5 to 12).

[2] After we obtain the logarithmic scale, we multiply the values by 5, and thus $\rho \in [5.10^{-5}, 5.10^{-1}]$ at the end.

the authors' code. The accuracy and unfairness measures of all classifiers is then evaluated on the testing data.

Table 4.1 suggests that the DR-FLR classifier performs favorably relative to its competitors in that it always decreases LogProb-UNF substantially and often yields the lowest Det-UNF with only a moderate loss in accuracy.

Worst-Case Distribution. Next, we visualize the extremal distribution $\mathbb{Q}^\star$ from Proposition 4.4.3 for 4 pre-trained classifiers (classical logistic regression, support vector machine with RBF kernel, Gaussian processes with RBF kernel, AdaBoost). Figure 4.2 illustrates which test samples are projected to the decision boundary under the adversarial distribution $\mathbb{Q}^\star$ until the transportation budget corresponding to the Wasserstein radius ρ is exhausted.



(a) Logistic Regression

(b) SVM (RBF)

(c) GP (RBF)

(d) AdaBoost

Figure 4.2: Visualization of the extremal distribution $\mathbb{Q}^\star$ for different classifiers. The red/blue background color represents the class partitions. The top row shows the test data, and the bottom row (zoomed) shows how samples with $z_i^\star > 0$ are moved to the decision boundary.

Appendix

This appendix is organized as follows. Section 4.5 contains all proofs omitted from the main text, while Section 4.6 provides detailed information on the numerical experiments and reports on additional numerical experiments.

Proofs

We first describe a strong semi-infinite duality result that forms the basis for several proofs. To this end, assume that $\phi : \mathcal{X} \times \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ is a Borel measurable loss function, and recall that $\widehat{p}_{ay} = \widehat{\mathbb{P}}_N(A = a, Y = y)$ for all $a \in \mathcal{A}$ and $y \in \mathcal{Y}$.

The semi-infinite program

$$\begin{aligned} \sup_{\mathbb{Q}\in\mathcal{M}} \quad & \mathbb{E}_{\mathbb{Q}}[\phi(X,A,Y)] \\ \text{s.t.} \quad & \mathbb{W}(\mathbb{Q},\widehat{\mathbb{P}}_N) \leq \rho \\ & \mathbb{Q}(A=a,Y=y)=\widehat{p}_{ay} \quad \forall a\in\mathcal{A},\ \forall y \in \mathcal{Y} \end{aligned} \tag{4.8}$$

thus evaluates the worst-case expected loss over all distributions in a Wasserstein ball of radius $\rho \geq 0$ around the discrete nominal distribution $\widehat{\mathbb{P}}_N$ under which the marginal distributions of A and Y coincide with their *nominal* marginal distributions. The following proposition generalizes existing strong duality results without marginal distribution information [BM19; GK22; MEK18; ZG18] and can be seen as a variant of [Fro+19, Theorem 2], which includes information on the marginal distribution of features and outputs. The proposition can also be derived from a general theory of moment problems [Sha01, Section 3]. We omit the proof for brevity.

Proposition 4.5.1 (Strong duality)**.** *If $\widehat{p}_{ay} \in (0,1)$ for all $a\in\mathcal{A}$ and $y\in\mathcal{Y}$ and if $\rho>0$, then (4.8) admits the strong semi-infinite dual*

$$\begin{aligned} \inf \quad & \rho\lambda + \sum_{a\in\mathcal{A}}\sum_{y\in\mathcal{Y}} \widehat{p}_{ay}\mu_{ay} + \frac{1}{N}\sum_{i=1}^N \nu_i \\ \text{s.t.} \quad & \lambda\in\mathbb{R}_+,\ \mu\in\mathbb{R}^{2\times 2},\ \nu\in\mathbb{R}^N \\ & \lambda\, c\big((x_i,a_i,y_i),(\widehat{x}_i,\widehat{a}_i,\widehat{y}_i)\big) + \mu_{a_i y_i} + \nu_i \geq \phi(x_i,a_i,y_i) \\ & \qquad\qquad\qquad \forall (x_i,a_i,y_i)\in\mathcal{X}\times\mathcal{A}\times\mathcal{Y},\ \forall i\in[N]. \end{aligned} \tag{4.9}$$

Also, if the supremum of (4.8) is finite, then the infimum of (4.9) is attained.

Corollary 7 (Absolute trust in A and Y)**.** *If $\widehat{p}_{ay}\in(0,1)$ for all $a\in\mathcal{A}$ and $y\in\mathcal{Y}$ and if $\rho>0$ and $\kappa_{\mathcal{A}}=\kappa_{\mathcal{Y}}=\infty$, then (4.8) admits the strong semi-infinite dual*

$$\begin{aligned} \inf \quad & \rho\lambda + \tfrac{1}{N}\textstyle\sum_{i=1}^N \nu_i \\ \text{s.t.} \quad & \lambda\in\mathbb{R}_+,\ \nu\in\mathbb{R}^N \\ & \lambda\|x_i-\hat{x}_i\| + \nu_i \geq \phi(x_i,\hat{a}_i,\hat{y}_i) \quad \forall x_i\in\mathcal{X},\ \forall i\in[N]. \end{aligned} \tag{4.10}$$

Proof of Corollary 7. When $\kappa_{\mathcal{A}}=\kappa_{\mathcal{Y}}=\infty$, the left hand side of the i-th semi-infinite constraint in (4.9) evaluates to ∞ unless $a_i=\hat{a}_i$ and $y_i=\hat{y}_i$. In this case, the constraint is trivially satisfied and can be omitted. Furthermore, by definition of $\hat{p}_{ay}$ we have

$$\sum_{a\in\mathcal{A}}\sum_{y\in\mathcal{Y}} \hat{p}_{ay}\mu_{ay} = \frac{1}{N}\sum_{i=1}^N \mu_{\hat{a}_i\hat{y}_i}.$$

Consequently, problem in (4.9) reduces to

$$\begin{array}{ll} \inf & \rho\lambda + \dfrac{1}{N}\displaystyle\sum_{i=1}^{N} \mu_{\hat{a}_i \hat{y}_i} + \dfrac{1}{N}\sum_{i=1}^{N} \nu_i \\ \text{s.t.} & \lambda \in \mathbb{R}_+,\ \mu \in \mathbb{R}^{2\times 2},\ \nu \in \mathbb{R}^N \\ & \lambda \left\| x_i - \hat{x}_i \right\| + \mu_{\hat{a}_i \hat{y}_i} + \nu_i \geq \phi(x_i, \hat{a}_i, \hat{y}_i) \quad \forall x_i \in \mathcal{X},\ \forall i \in [N]. \end{array} \tag{4.11}$$

We can further simplify problem (4.11) by applying the change of variables $\nu_i \leftarrow \mu_{\hat{a}_i \hat{y}_i} + \nu_i$, $i \in [N]$, which yields the reformulation (4.10). This observation completes the proof. □

Proofs of Section 4.2

Proof of Theorem 4.2.1. We define the log-loss function through

$$\ell_\beta(x, y) = -y \log(h_\beta(x)) - (1-y)\log(1 - h_\beta(x)) \quad \forall x \in \mathcal{X},\ \forall y \in \mathcal{Y}.$$

By introducing an auxiliary epigraphical variable, problem (4.2) can then be reformulated as

$$\begin{array}{ll} \min\limits_{\beta, t} & t \\ \text{s.t.} & \mathbb{E}_{\mathbb{P}}[\ell_\beta(X, Y)] + \eta \mathbb{U}_f(\mathbb{P}, h_\beta) \leq t. \end{array} \tag{4.12}$$

As $f(z) = \log(z)$ and $\mathbb{P} = \widehat{\mathbb{P}}_N$ by assumption, the unfairness measure simplifies to

$$\mathbb{U}_f(\widehat{\mathbb{P}}_N, h_\beta) = |\mathbb{E}_{\widehat{\mathbb{P}}_N}[\log(h_\beta(X))|A = 1, Y = 1] - \mathbb{E}_{\widehat{\mathbb{P}}_N}[\log(h_\beta(X))|A = 0, Y = 1]|.$$

By the definition of conditional expectations, we further have

$$\begin{aligned} \mathbb{E}_{\widehat{\mathbb{P}}_N}[\log h_\beta(X)|A = a, Y = 1] &= \frac{\mathbb{E}_{\widehat{\mathbb{P}}_N}[\log h_\beta(X) \mathbb{1}_{\{(a,1)\}}(A, Y)]}{\widehat{\mathbb{P}}_N(A = a, Y = 1)} \\ &= r_a\, \mathbb{E}_{\widehat{\mathbb{P}}_N}[\log h_\beta(X) \mathbb{1}_{\{(a,1)\}}(A, Y)] \end{aligned}$$

for all $a \in \mathcal{A}$, where the second equality follows from the definition of r_a. For any fixed $a, a' \in \mathcal{A}$ with $a \neq a'$ and $\beta \in \mathbb{R}^p$ we then introduce the function

$$\hat{\mathbb{T}}_\beta^{aa'} = \mathbb{E}_{\widehat{\mathbb{P}}_N}[\ell_\beta(X, Y) + \eta r_a \log(h_\beta(X)) \mathbb{1}_{\{(a,1)\}}(A, Y) - \eta r_{a'} \log(h_\beta(X)) \mathbb{1}_{\{(a',1)\}}(A, Y)].$$

By expanding the absolute value in the definition of $\mathbb{U}_f(\widehat{\mathbb{P}}_N, h_\beta)$, problem (4.12) simplifies to

$$\begin{array}{ll} \min\limits_{\beta, t} & t \\ \text{s.t.} & \hat{\mathbb{T}}_\beta^{10} \leq t,\ \hat{\mathbb{T}}_\beta^{01} \leq t, \end{array} \tag{4.13}$$

which is manifestly equivalent to the optimization problem in the theorem statement. Note that by the definition of the log-loss function, we obtain

$$\begin{aligned}\hat{\mathbb{T}}_\beta^{aa'} &= \mathbb{E}_{\widehat{\mathbb{P}}_N}[-Y\log(h_\beta(X)) - (1-Y)\log(1-h_\beta(X)) + \eta r_a \log(h_\beta(X))\mathbb{1}_{\{(a,1)\}}(A,Y) \\ &\qquad - \eta r_{a'}\log(h_\beta(X))\mathbb{1}_{\{(a',1)\}}(A,Y)] \\ &= -\frac{1}{N}\Big(\sum_{\substack{i\in[N]:\\ \hat{y}_i=1\\ \hat{a}_i=a}}(1-\eta r_a)\log(h_\beta(\hat{x}_i)) + \sum_{\substack{i\in[N]:\\ \hat{y}_i=1\\ \hat{a}_i=a'}}(\eta r_{a'}+1)\log(h_\beta(\hat{x}_i)) + \sum_{\substack{i\in[N]:\\ \hat{y}_i=0}}\log(1-h_\beta(\hat{x}_i))\Big),\end{aligned}$$

where the second equality holds because the expectation under the empirical distribution $\widehat{\mathbb{P}}_N$ can be expressed as a finite sum, and terms can be grouped by the labels and the sensitive attributes of the training samples. Thus, $\hat{\mathbb{T}}_\beta^{aa'}$ is convex in β for $\eta \leq \min\{\hat{p}_{11}, \hat{p}_{01}\}$, in which case problem (4.13) becomes a tractable convex program. This concludes the proof. □

Proofs of Section 4.3

The proof of Theorem 4.3.1 relies on the following simple corollary of [SMK15, Lemma 1].

Lemma 4.5.2. *If $\beta \in \mathbb{R}^p$ and $\gamma \in \mathbb{R}_+$, while $g_\beta(x) = \gamma\log(1+\exp(-\langle\beta, x\rangle))$ is a convex function of $x \in \mathbb{R}^p$, then we have*

$$\sup_{x\in\mathbb{R}^p} \; \gamma g_\beta(x) - \lambda\|x-\hat{x}\| = \begin{cases}\gamma g_\beta(\hat{x}) & \text{if } \gamma\|\beta\|_* \leq \lambda \\ +\infty & \text{otherwise}\end{cases}$$

for all $\lambda \in \mathbb{R}_{++}$, where $\|\cdot\|_*$ represents the dual norm of $\|\cdot\|$.

Proof of Theorem 4.3.1. To simplify notation, we define the log-loss function as usual as

$$\ell_\beta(x,y) = -y\log h_\beta(x) - (1-y)\log(1-h_\beta(x)) \quad \forall x \in \mathcal{X},\ \forall y \in \mathcal{Y}.$$

By introducing an auxiliary epigraphical variable, problem (4.5) can then be reformulated as

$$\begin{array}{ll}\min\limits_{\beta,t} & t \\ \text{s.t.} & \sup\limits_{\mathbb{Q}\in\mathbb{B}_\rho(\widehat{\mathbb{P}}_N)} \mathbb{E}_\mathbb{Q}[\ell_\beta(X,Y)] + \eta\mathbb{U}_f(\mathbb{Q}, h_\beta) \leq t.\end{array} \tag{4.14}$$

As $f(z) = \log(z)$ by assumption, the unfairness measure simplifies to

$$\mathbb{U}_f(\mathbb{Q}, h_\beta) = |\mathbb{E}_\mathbb{Q}[\log(h_\beta(X))|A=1, Y=1] - \mathbb{E}_\mathbb{Q}[\log(h_\beta(X))|A=0, Y=1]|.$$

By the definition of conditional expectations, we have for all $\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$ and $a \in \mathcal{A}$ that

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[\log h_\beta(X)|A = a, Y = 1] &= \frac{\mathbb{E}_{\mathbb{Q}}[\log h_\beta(X)\mathbb{1}_{\{(a,1)\}}(A, Y)]}{\mathbb{Q}(A = a, Y = 1)} \\ &= r_a \mathbb{E}_{\mathbb{Q}}[\log h_\beta(X)\mathbb{1}_{\{(a,1)\}}(A, Y)],\end{aligned}$$

where the second equality holds because $\mathbb{Q}(A = a, Y = 1) = \widehat{p}_{ay} = 1/r_a$ for any $\mathbb{Q} \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$. For any fixed $a, a' \in \mathcal{A}$ with $a \neq a'$ and $\beta \in \mathbb{R}^p$ we then introduce the function

$$\phi_\beta^{aa'}(\tilde{x}, \tilde{a}, \tilde{y}) = \ell_\beta(\tilde{x}, \tilde{y}) + \eta\, r_a \log(h_\beta(\tilde{x}))\mathbb{1}_{\{(a,1)\}}(\tilde{a}, \tilde{y}) - \eta\, r_{a'} \log(h_\beta(\tilde{x}))\mathbb{1}_{\{(a',1)\}}(\tilde{a}, \tilde{y})$$

of $\tilde{x} \in \mathcal{X}$, $\tilde{a} \in \mathcal{A}$ and $\tilde{y} \in \mathcal{Y}$, and we define

$$\mathbb{T}_\beta^{aa'} = \sup_{\mathbb{Q} \in \mathbb{B}_\rho(\hat{\mathbb{P}}_N)} \mathbb{E}_{\mathbb{Q}}[\phi_\beta^{aa'}(X, A, Y)].$$

The integrand $\phi_\beta^{aa'}$ satisfies the linear growth condition of [YKW20, Theorem 2.2], which guarantees that $\mathbb{T}_\beta^{aa'}$ is finite. By using the above notational conventions and introducing an auxiliary epigraphical variable as in the proof of Theorem 4.2.1, problem (4.14) is simplified to

$$\begin{array}{cl} \min\limits_{\beta, t} & t \\ \text{s.t.} & \mathbb{T}_\beta^{10} \leq t, \ \mathbb{T}_\beta^{01} \leq t. \end{array} \tag{4.15}$$

To convert problem (4.5) to a convex program, we need to simplify the constraints that involve $\mathbb{T}_\beta^{aa'}$. To this end, we may use Proposition 4.5.1 to obtain

$$\mathbb{T}_\beta^{aa'} = \left\{ \begin{array}{cl} \min & \rho\lambda + \sum\limits_{a \in \mathcal{A}} \sum\limits_{y \in \mathcal{Y}} \hat{p}_{ay}\mu_{ay} + \frac{1}{N}\sum\limits_{i=1}^N \nu_i \\ \text{s.t.} & \lambda \in \mathbb{R}_+, \ \mu \in \mathbb{R}^{2\times 2}, \ \nu \in \mathbb{R}^N \\ & \lambda\, c\big((x_i, a_i, y_i), (\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)\big) + \mu_{a_i y_i} + \nu_i \geq \phi_\beta^{aa'}(x_i, a_i, y_i) \\ & \qquad\qquad \forall (x_i, a_i, y_i) \in \mathcal{X} \times \mathcal{A} \times \mathcal{Y}, \ \forall i \in [N]. \end{array} \right. \tag{4.16}$$

As $\mathbb{T}_\beta^{aa'}$ is finite, Proposition 4.5.1 also ensures that the minimum of problem (4.16) is attained.

We now investigate the i-th semi-infinite constraint in (4.16) for a fixed a_i and y_i. Thanks to the additive separability of the transportation cost, this constraint can be reformulated as

$$\nu_i \geq \sup_{x_i \in \mathcal{X}} \left\{ \phi_\beta^{aa'}(x_i, a_i, y_i) - \lambda \|x_i - \widehat{x}_i\| \right\} - \kappa_{\mathcal{A}} |a_i - \widehat{a}_i| \lambda - \kappa_{\mathcal{Y}} |y_i - \widehat{y}_i| \lambda - \mu_{a_i y_i} \tag{4.17}$$

If $y_i = 0$ and $a_i \in \mathcal{A}$, then $\phi_\beta^{aa'}(x_i, a_i, 0) = -\log(1 - h_\beta(x_i))$, and by Lemma 4.5.2, we have

$$\begin{aligned} \sup_{x_i \in \mathcal{X}} \left\{ \phi_\beta^{aa'}(x_i, a_i, 0) - \lambda \|x_i - \widehat{x}_i\| \right\} &= \sup_{x_i \in \mathcal{X}} -\log(1 - h_\beta(x_i)) - \lambda \|x_i - \hat{x}_i\| \\ &= \begin{cases} -\log(1 - h_\beta(\hat{x}_i)) & \text{if } \|\beta\|_* \leq \lambda, \\ +\infty & \text{otherwise,} \end{cases} \end{aligned}$$

which implies that the constraint (4.17) is equivalent to the inequalities

$$\|\beta\|_* \leq \lambda \quad \text{and} \quad \nu_i \geq -\log(1 - h_\beta(\hat{x}_i)) - \kappa_{\mathcal{A}} |a_i - \widehat{a}_i| \lambda - \kappa_{\mathcal{Y}} |\widehat{y}_i| \lambda - \mu_{a_i 0}.$$

If $a_i = a$ and $y_i = 1$, then $\phi_\beta^{aa'}(x_i, a, 1) = (\eta r_a - 1) \log(h_\beta(x_i))$, and by Lemma 4.5.2, we have

$$\sup_{x_i \in \mathcal{X}} \left\{ \phi_\beta^{aa'}(x_i, a, 1) - \lambda \|x_i - \widehat{x}_i\| \right\} = \begin{cases} (\eta r_a - 1) \log(h_\beta(\widehat{x}_i)) & \text{if } (1 - \eta r_a) \|\beta\|_* \leq \lambda, \\ +\infty & \text{otherwise,} \end{cases}$$

which implies that the constraint (4.17) is equivalent to

$$(1 - \eta r_a) \|\beta\|_* \leq \lambda \quad \text{and} \quad \nu_i \geq (\eta r_a - 1) \log(h_\beta(\widehat{x}_i)) - \kappa_{\mathcal{A}} |a - \widehat{a}_i| \lambda - \kappa_{\mathcal{Y}} |1 - \widehat{y}_i| \lambda - \mu_{a1}.$$

If $a_i = a'$ and $y_i = 1$, finally, then $\phi_\beta^{aa'}(x_i, a', 1) = -(1 + \eta r_{a'}) \log(h_\beta(x_i))$, and we can use an analogous argument involving Lemma 4.5.2 to show that

$$\sup_{x_i \in \mathcal{X}} \left\{ \phi_\beta^{aa'}(x_i, a', 1) - \lambda \|x_i - \widehat{x}_i\| \right\} = \begin{cases} -(1 + \eta r_{a'}) \log(h_\beta(\widehat{x}_i)) & \text{if } (1 + \eta r_{a'}) \|\beta\|_* \leq \lambda, \\ +\infty & \text{otherwise,} \end{cases}$$

which implies that the constraint (4.17) is equivalent to

$$(1 + \eta\, r_{a'}) \|\beta\|_* \leq \lambda \text{ and } \nu_i \geq -(1 + \eta\, r_{a'}) \log(h_\beta(\widehat{x}_i)) - \kappa_{\mathcal{A}} |a' - \widehat{a}_i| \lambda - \kappa_{\mathcal{Y}} |1 - \widehat{y}_i| \lambda - \mu_{a'1}.$$

Substituting the above reformulations of constraint (4.17) corresponding to all possible combinations of a_i and y_i into (4.16) yields

$$\mathbb{T}_{\beta}^{aa'} = \left\{ \begin{array}{l} \min\ \rho\lambda + \displaystyle\sum_{a\in\mathcal{A}}\sum_{y\in\mathcal{Y}} \hat{p}_{ay}\mu_{ay} + \frac{1}{N}\sum_{i=1}^{N}\nu_i \\ \text{s.t.}\ \ \lambda\in\mathbb{R}_+,\ \mu\in\mathbb{R}^{2\times 2},\ \nu\in\mathbb{R}^N \\ \qquad \|\beta\|_*\leq\lambda,\ \|\beta\|_*(1-\eta r_a)\leq\lambda,\ \|\beta\|_*(1+\eta r_{a'})\leq\lambda \\ \qquad \left.\begin{array}{l} \nu_i \geq -\log(1-h_\beta(\hat{x}_i)) - \kappa_{\mathcal{A}}|a-\hat{a}_i|\lambda - \kappa_{\mathcal{Y}}|\hat{y}_i|\lambda - \mu_{a0} \\ \nu_i \geq -\log(1-h_\beta(\hat{x}_i)) - \kappa_{\mathcal{A}}|a'-\hat{a}_i|\lambda - \kappa_{\mathcal{Y}}|\hat{y}_i|\lambda - \mu_{a'0} \\ \nu_i \geq (\eta r_a - 1)\log(h_\beta(\hat{x}_i)) - \kappa_{\mathcal{A}}|a-\hat{a}_i|\lambda - \kappa_{\mathcal{Y}}|1-\hat{y}_i|\lambda - \mu_{a1} \\ \nu_i \geq -(1+\eta r_{a'})\log(h_\beta(\hat{x}_i)) - \kappa_{\mathcal{A}}|a'-\hat{a}_i|\lambda - \kappa_{\mathcal{Y}}|1-\hat{y}_i|\lambda - \mu_{a'1} \end{array}\right\} \\ \hfill \forall i\in[N]. \end{array} \right.$$

Note that the constraints $\|\beta\|_* \leq \lambda$ and $\|\beta\|_*(1-\eta r_a) \leq \lambda$ are redundant in view of the constraint $\|\beta\|_*(1+\eta r_{a'}) \leq \lambda$. The claim then follows by substituting the dual reformulations for $\mathbb{T}_{\beta}^{aa'}$ into (4.15) and eliminating the embedded minimization operators. □

4.5.1. Proofs of Section 4.4

Proof of Theorem 4.4.1. By the definition of $\mathbb{V}(a,a')$ for $a,a'\in\mathcal{A}$, one readily verifies that the bounds on the unfairness measure can be expressed as

$$\overline{\mathbb{U}}_f = \max\{\mathbb{V}(1,0),\mathbb{V}(0,1)\} \quad \text{and} \quad \underline{\mathbb{U}}_f = \max\{0,-\mathbb{V}(1,0),-\mathbb{V}(0,1)\}.$$

For any fixed $a,a'\in\mathcal{A}$ with $a\neq a'$ we then introduce the function

$$\phi^{aa'}(\tilde{x},\tilde{a},\tilde{y}) = r_a\mathbb{1}_{\mathcal{X}_1\times\{(a,1)\}}(\tilde{x},\tilde{a},\tilde{y}) - r_{a'}\mathbb{1}_{\mathcal{X}_1\times\{(a',1)\}}(\tilde{x},\tilde{a},\tilde{y}),$$

which depens on $\tilde{x}\in\mathcal{X}$, $\tilde{a}\in\mathcal{A}$ and $\tilde{y}\in\mathcal{Y}$, and which allows us to re-express $\mathbb{V}(a,a')$ as

$$\mathbb{V}(a,a') = \left\{ \begin{array}{ll} \sup\limits_{\mathbb{Q}\in\mathcal{M}} & \mathbb{E}_{\mathbb{Q}}[\phi^{aa'}(X,A,Y)] \\ \text{s.t.} & \mathbb{W}(\mathbb{Q},\widehat{\mathbb{P}}_N)\leq\rho \\ & \mathbb{Q}(A=a,Y=y)=\widehat{p}_{ay} \quad \forall a\in\mathcal{A},\ \forall y\in\mathcal{Y}. \end{array} \right.$$

Note that the function $\phi^{aa'}$ is piecewise constant and thus bounded, which implies that $\mathbb{V}(a,a')$ is finite. The strong duality result from Proposition 4.5.1

further implies that

$$\mathbb{V}(a,a') = \begin{cases} \min & \rho\lambda + \widehat{p}^\top \mu + \frac{1}{N}\mathbf{1}^\top \nu \\ \text{s.t.} & \lambda \in \mathbb{R}_+,\ \mu \in \mathbb{R}^4,\ \nu \in \mathbb{R}^N \\ & \lambda c\big((x_i,a_i,y_i),(\widehat{x}_i,\widehat{a}_i,\widehat{y}_i)\big) + \mu_{a_i y_i} + \nu_i \geq \phi^{aa'}(x_i,a_i,y_i) \\ & \qquad\qquad\qquad \forall (x_i,a_i,y_i) \in \mathcal{X}\times\mathcal{A}\times\mathcal{Y},\ \forall i \in [N]. \end{cases} \tag{4.18}$$

Note that the minimum of problem (4.18) is attained because $\mathbb{V}(a,a')$ is finite. By the definition of the transportation cost, the i-th semi-infinite constraint in (4.18) can be expressed more explicitly as

$$\nu_i \geq \sup_{x_i\in\mathcal{X}} \left\{\phi^{aa'}(x_i,a_i,y_i) - \lambda\|x_i - \widehat{x}_i\|\right\} - \kappa_{\mathcal{A}}|a_i - \widehat{a}_i|\lambda - \kappa_{\mathcal{Y}}|y_i - \widehat{y}_i|\lambda - \mu_{a_i y_i} \quad \forall a_i \in \mathcal{A},\ \forall y_i \in \mathcal{Y}. \tag{4.19}$$

If $y_i = 0$ and $a_i \in \mathcal{A}$, then $\phi^{aa'}(x_i,a_i,0) = 0$, and thus (4.19) simplifies to

$$\nu_i \geq -\kappa_{\mathcal{A}}|a_i - \widehat{a}_i|\lambda - \kappa_{\mathcal{Y}}|\widehat{y}_i|\lambda - \mu_{a_i 0} \quad \forall a_i \in \mathcal{A}.$$

If $a_i = a$ and $y_i = 1$, then $\phi^{aa'}(x_i,a,1) = r_a \mathbb{1}_{\mathcal{X}_1}(x_i)$, and we have

$$\begin{aligned} \sup_{x_i\in\mathcal{X}}\ r_a\mathbb{1}_{\mathcal{X}_1}(x_i) - \lambda\|x_i - \widehat{x}_i\| &= \begin{cases} r_a & \text{if } \widehat{x}_i \in \mathcal{X}_1 \\ \max\{0, r_a - \lambda d_{1i}\} & \text{if } \widehat{x}_i \notin \mathcal{X}_1 \end{cases} \\ &= \max\{0, r_a - \lambda d_{1i}\}, \end{aligned}$$

where the last equality holds because $d_{1i} = 0$ if $\widehat{x}_i \in \mathcal{X}_1$. Thus, constraint (4.19) reduces to

$$\nu_i \geq \max\{0, r_a - \lambda d_{1i}\} - \kappa_{\mathcal{A}}|a - \widehat{a}_i|\lambda - \kappa_{\mathcal{Y}}|1 - \widehat{y}_i|\lambda - \mu_{a1}.$$

If $a_i = a'$ and $y_i = 1$, finally, then we have $\phi^{aa'}(x_i,a',1) = -r_{a'}\mathbb{1}_{\mathcal{X}_1}(x_i)$, and thus

$$\begin{aligned} \sup_{x_i\in\mathcal{X}}\ - r_{a'}\mathbb{1}_{\mathcal{X}_1}(x_i) - \lambda\|x_i - \widehat{x}_i\| &= \begin{cases} \max\{-r_{a'}, -\lambda d_{0i}\} & \text{if } \widehat{x}_i \in \mathcal{X}_1 \\ 0 & \text{if } \widehat{x}_i \notin \mathcal{X}_1 \end{cases} \\ &= \max\{-r_{a'}, -\lambda d_{0i}\}, \end{aligned}$$

where the last equality holds because $d_{0i} = 0$ whenever $\widehat{x}_i \notin \mathcal{X}_1$. Because the set $\mathcal{X}_1$ is closed, the supremum in the above expression is not attained. Constraint (4.19) now becomes

$$\nu_i \geq \max\{-r_{a'}, -\lambda d_{0i}\} - \kappa_{\mathcal{A}}|a' - \widehat{a}_i|\lambda - \kappa_{\mathcal{Y}}|1 - \widehat{y}_i|\lambda - \mu_{a'1}.$$

In summary, the semi-infinite constraint (4.19) is equivalent to the six linear constraints

$$\begin{aligned}
\nu_i &\geq -\kappa_{\mathcal{A}}|a-\widehat{a}_i|\lambda - \kappa_{\mathcal{Y}}|\widehat{y}_i|\lambda - \mu_{a0}\\
\nu_i &\geq -\kappa_{\mathcal{A}}|a'-\widehat{a}_i|\lambda - \kappa_{\mathcal{Y}}|\widehat{y}_i|\lambda - \mu_{a'0}\\
\nu_i &\geq r_a - \lambda d_{1i} - \kappa_{\mathcal{A}}|a-\widehat{a}_i|\lambda - \kappa_{\mathcal{Y}}|1-\widehat{y}_i|\lambda - \mu_{a1}\\
\nu_i &\geq -\kappa_{\mathcal{A}}|a-\widehat{a}_i|\lambda - \kappa_{\mathcal{Y}}|1-\widehat{y}_i|\lambda - \mu_{a1}\\
\nu_i &\geq -r_{a'} - \kappa_{\mathcal{A}}|a'-\widehat{a}_i|\lambda - \kappa_{\mathcal{Y}}|1-\widehat{y}_i|\lambda - \mu_{a'1}\\
\nu_i &\geq -\lambda d_{0i} - \kappa_{\mathcal{A}}|a'-\widehat{a}_i|\lambda - \kappa_{\mathcal{Y}}|1-\widehat{y}_i|\lambda - \mu_{a'1}.
\end{aligned}$$

The claim now follows by substituting this reformulation into (4.18) for every $i \in [N]$. □

Proof of Theorem 4.4.2. If $\kappa_{\mathcal{A}} = \kappa_{\mathcal{Y}} = \infty$, then the linear programming reformulation derived in Theorem 4.4.1 simplifies to

$$\begin{array}{ll}
\min & \rho\lambda + \sum\limits_{a\in\mathcal{A}}\sum\limits_{y\in\mathcal{Y}} \widehat{p}_{ay}\mu_{ay} + \dfrac{1}{N}\sum\limits_{i=1}^{N}\nu_i \\
\text{s.t.} & \lambda\in\mathbb{R}_+,\ \mu\in\mathbb{R}^{2\times 2},\ \nu\in\mathbb{R}^N \\
& \left.\begin{array}{ll}
\nu_i + \mu_{\widehat{a}_i 0} \geq 0 & \text{if } \widehat{y}_i = 0\\
\nu_i + \mu_{a1} + d_{1i}\lambda \geq r_a & \text{if } \widehat{a}_i = a, \widehat{y}_i = 1\\
\nu_i + \mu_{a1} \geq 0 & \text{if } \widehat{a}_i = a, \widehat{y}_i = 1\\
\nu_i + \mu_{a'1} \geq -r_{a'} & \text{if } \widehat{a}_i = a', \widehat{y}_i = 1\\
\nu_i + \mu_{a'1} + d_{0i}\lambda \geq 0 & \text{if } \widehat{a}_i = a', \widehat{y}_i = 1
\end{array}\right\}\ \forall i\in[N].
\end{array} \tag{4.20}$$

Furthermore, the first constraint $\mu_{\hat{a}_i} \geq -\nu_i$ force $\mu_{\hat{a}_i 0} = -\nu_i$ for all $\{i \in [N] : \hat{y}_i = 0\}$. Hence, by definition of $\hat{p}_{ay}$ we have

$$\sum_{a\in\mathcal{A}} \hat{p}_{a0}\mu_{a0} = -\frac{1}{N}\sum_{i\in[N]:\hat{y}_i=0} \nu_i.$$

Consequently, by defining the sets $\bar{\mathcal{I}}_a = \{i \in [N] : \widehat{a}_i = a, \widehat{y}_i = 1\}$ and $\bar{\mathcal{I}}_{a'} = \{i \in [N] : \widehat{a}_i = a', \widehat{y}_i = 1\}$, problem in (4.20) is further simplified to

$$\begin{array}{ll}
\min & \rho\lambda + \widehat{p}_{a1}\mu_{a1} + \widehat{p}_{a'1}\mu_{a'1} + \frac{1}{N}\sum\limits_{i\in\bar{\mathcal{I}}_a\cup\bar{\mathcal{I}}_{a'}} \nu_i \\
\text{s.t.} & \lambda\in\mathbb{R}_+,\ \mu\in\mathbb{R}^{2\times 2},\ \nu\in\mathbb{R}^N \\
& \left.\begin{array}{l}
\nu_i + \mu_{a1} + d_{1i}\lambda \geq r_a\\
\nu_i + \mu_{a1} \geq 0
\end{array}\right\}\ \forall i\in\bar{\mathcal{I}}_a \\
& \left.\begin{array}{l}
\nu_i + \mu_{a'1} \geq -r_{a'}\\
\nu_i + \mu_{a'1} + d_{0i}\lambda \geq 0
\end{array}\right\}\ \forall i\in\bar{\mathcal{I}}_{a'}.
\end{array}$$

By introducing the Lagrangian multipliers $\gamma_1, \gamma_2 \in \mathbb{R}_+^{|\bar{\mathcal{I}}_a|}$ and $\gamma_3, \gamma_4 \in \mathbb{R}_+^{|\bar{\mathcal{I}}_{a'}|}$, we obtain the linear dual problem of the above problem as

$$\begin{array}{lll}
\max & r_a \sum\limits_{i \in \bar{\mathcal{I}}_a} \gamma_{1i} - r_{a'} \sum\limits_{i \in \bar{\mathcal{I}}_{a'}} \gamma_{3i} & \\
\text{s.t.} & \gamma_1 \in \mathbb{R}_+^{|\bar{\mathcal{I}}_a|},\ \gamma_2 \in \mathbb{R}_+^{|\bar{\mathcal{I}}_a|},\ \gamma_3 \in \mathbb{R}_+^{|\bar{\mathcal{I}}_{a'}|},\ \gamma_4 \in \mathbb{R}_+^{|\bar{\mathcal{I}}_{a'}|} & \\
& \rho - \sum\limits_{i \in \bar{\mathcal{I}}_a} \gamma_{1i} d_{1i} - \sum\limits_{i \in \bar{\mathcal{I}}_{a'}} \gamma_{4i} d_{0i} \geq 0 & \\
& \hat{p}_{a1} - \sum\limits_{i \in \bar{\mathcal{I}}_a} (\gamma_{2i} + \gamma_{1i}) = 0 & \\
& \hat{p}_{a'1} - \sum\limits_{i \in \bar{\mathcal{I}}_{a'}} (\gamma_{3i} + \gamma_{4i}) = 0 & \\
& 1/N - \gamma_{1i} - \gamma_{2i} = 0 & \forall i \in \bar{\mathcal{I}}_a \\
& 1/N - \gamma_{3i} - \gamma_{4i} = 0 & \forall i \in \bar{\mathcal{I}}_{a'}.
\end{array} \tag{4.21}$$

We now define the sets $\mathcal{I}_a = \{i \in \bar{\mathcal{I}}_a : \hat{x}_i \in \text{int}(\mathcal{X}_0)\}$ and $\mathcal{I}_{a'} = \{i \in \bar{\mathcal{I}}_{a'} : \hat{x}_i \in \text{int}(\mathcal{X}_1)\}$. Due to the last two constraints, $\gamma_{2i} + \gamma_{1i} = 1/N$ and $\gamma_{3i} + \gamma_{4i} = 1/N$, the third and the forth constraints of (4.21) become redundant as $\sum_{i \in \bar{\mathcal{I}}_a} 1/N = \hat{p}_{a1}$ and $\sum_{i \in \bar{\mathcal{I}}_{a'}} 1/N = \hat{p}_{a'1}$ by definition of the sets $\bar{\mathcal{I}}_a$ and $\bar{\mathcal{I}}_{a'}$. Notice that due to last constraint in (4.21), we have $\gamma_{3i} = 1/N - \gamma_{4i}$ for all $i \in \bar{\mathcal{I}}_{a'}$. Then, we can further simplify problem (4.21) to

$$\begin{array}{lll}
\max & r_a \sum\limits_{i \in \bar{\mathcal{I}}_a} \gamma_{1i} - r_{a'} \sum\limits_{i \in \bar{\mathcal{I}}_{a'}} \left(\frac{1}{N} - \gamma_{4i}\right) & \\
\text{s.t.} & \gamma_1 \in \mathbb{R}_+^{|\bar{\mathcal{I}}_a|},\ \gamma_2 \in \mathbb{R}_+^{|\bar{\mathcal{I}}_a|},\ \gamma_3 \in \mathbb{R}_+^{|\bar{\mathcal{I}}_{a'}|},\ \gamma_4 \in \mathbb{R}_+^{|\bar{\mathcal{I}}_{a'}|} & \\
& \rho - \sum\limits_{i \in \mathcal{I}_a} \gamma_{1i} d_{1i} - \sum\limits_{i \in \mathcal{I}_{a'}} \gamma_{4i} d_{0i} \geq 0 & \\
& \gamma_{1i} + \gamma_{2i} = 1/N & \forall i \in \bar{\mathcal{I}}_a \\
& \gamma_{3i} + \gamma_{4i} = 1/N & \forall i \in \bar{\mathcal{I}}_{a'}.
\end{array} \tag{4.22}$$

Because the variables $\gamma_{2i} \in \mathbb{R}_+^{|\bar{\mathcal{I}}_a|}$ and $\gamma_{3i} \in \mathbb{R}_+^{|\bar{\mathcal{I}}_{a'}|}$ do not appear in the objective of problem (4.22) and $r_a, r_{a'} > 0$, we can further simplify problem (4.22) to

$$\begin{array}{lll}
\max & r_a \sum\limits_{i \in \bar{\mathcal{I}}_a} \gamma_{1i} - r_{a'} \sum\limits_{i \in \bar{\mathcal{I}}_{a'}} \left(\frac{1}{N} - \gamma_{4i}\right) & \\
\text{s.t.} & \gamma_1 \in \mathbb{R}_+^{|\bar{\mathcal{I}}_a|}, \gamma_4 \in \mathbb{R}_+^{|\bar{\mathcal{I}}_{a'}|} & \\
& \rho - \sum\limits_{i \in \mathcal{I}_a} \gamma_{1i} d_{1i} - \sum\limits_{i \in \mathcal{I}_{a'}} \gamma_{4i} d_{0i} \geq 0 & \\
& \gamma_{1i} \leq 1/N & \forall i \in \bar{\mathcal{I}}_a \\
& \gamma_{4i} \leq 1/N & \forall i \in \bar{\mathcal{I}}_{a'}.
\end{array} \tag{4.23}$$

Note that for $\gamma_1^\star$ and $\gamma_4^\star$ that optimize problem (4.23) for all $i \notin \mathcal{I}_a$, $\gamma_{1i}^\star$ takes the value $1/N$, and similarly for all $i \notin \mathcal{I}_{a'}$, $\gamma_{4i}^\star$ takes the value $1/N$. Hence, it is sufficient to optimize over values of γ_{1i} and γ_{4i} for all $i \in \mathcal{I}_a \cup \mathcal{I}_{a'}$. By applying the variable transformations $\gamma_{1i} \leftarrow z_i/N$ for all $i \in \mathcal{I}_a$ and $\gamma_{4i} \leftarrow z_i/N$ for all $i \in \mathcal{I}_{a'}$, where $z \in \mathbb{R}_+^{|\mathcal{I}_a|+|\mathcal{I}_{a'}|}$, the problem (4.23) can be restated as

$$\begin{array}{ll} \max & r_a \frac{|\bar{\mathcal{I}}_a \backslash \mathcal{I}_a|}{N} + \frac{r_a}{N} \sum\limits_{i \in \mathcal{I}_a} z_i - r_{a'} \frac{|\mathcal{I}_{a'}|}{N} + \frac{r_{a'}}{N} \sum\limits_{i \in \mathcal{I}_{a'}} z_i \\ \text{s.t.} & z \in \mathbb{R}_+^{|\mathcal{I}_a|+|\mathcal{I}_{a'}|} \\ & \sum\limits_{i \in \mathcal{I}_a} z_i d_{1i} + \sum\limits_{i \in \mathcal{I}_{a'}} z_i d_{0i} \leq N\rho \\ & z_i \leq 1 \qquad\qquad\qquad\qquad\qquad\qquad \forall\, i \in \mathcal{I}_a \cup \mathcal{I}_{a'}. \end{array} \tag{4.24}$$

Observe that $r_a|\bar{\mathcal{I}}_a \backslash \mathcal{I}_a|/N - r_{a'}|\mathcal{I}_{a'}|/N$ is equivalent to empirical value function $\hat{\mathbb{V}}(a, a')$, which is defined as in the theorem statement. By introducing the non-negative rewards and weights through

$$(c_{aa'i}, w_{aa'i}) = \begin{cases} (r_a, d_{1i}) & \text{if } i \in \mathcal{I}_a, \\ (r_{a'}, d_{0i}) & \text{if } i \in \mathcal{I}_{a'}, \\ (0, +\infty) & \text{otherwise,} \end{cases}$$

we can re-write the optimization problem in (4.24) as

$$\hat{\mathbb{V}}(a, a') + \max_{z \in [0,1]^N} \left\{ \frac{1}{N} \sum_{i \in [N]} c_{aa'i} z_i \ : \ \frac{1}{N} \sum_{i \in [N]} w_{aa'i} z_i \leq \rho \right\} \ \forall a, a' \in \mathcal{A}, a \neq a',$$

where the equivalence of the two problems holds because $z_i^\star = 0$ for all $\{i \in [N] : w_{aa'i} = +\infty\}$. This observation concludes the proof. □

Proof of Proposition 4.4.3. For $\rho = 0$, we have $\mathbb{V}(a, a') = \hat{\mathbb{V}}(a, a')$ and $\mathbb{Q}^\star = \widehat{\mathbb{P}}_N$ is the optimal solution that attains the supremum in (4.6). For the rest of the proof, it suffices to consider when $\rho > 0$.

We define the set $\mathcal{I} = \{i \in [N] : \hat{a}_i = a, \hat{y}_i = 1, \hat{x}_i \in \text{int}(\mathcal{X}_0)\} \cup \{i \in [N] : \hat{a}_i = a', \hat{y}_i = 1, \hat{x}_i \in \text{int}(\mathcal{X}_1)\}$. First, we show that $\mathbb{Q}^\star$ defined in the statement of the Proposition 4.4.3 satisfies $\mathbb{Q}^\star \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$. Notice that $\mathbb{Q}^\star$ does not flip any label on A and Y as $\kappa_\mathcal{A} = \kappa_\mathcal{Y} = \infty$, thus it preserves the marginals

$$\mathbb{Q}^\star(A = a, Y = y) = \widehat{\mathbb{P}}_N(A = a, Y = y) \quad \forall a \in \mathcal{A}, y \in \mathcal{Y}.$$

Moreover, the distance from $\mathbb{Q}^\star$ to $\widehat{\mathbb{P}}_N$ satisfies

$$\mathbb{W}(\mathbb{Q}^\star, \widehat{\mathbb{P}}_N) \leq \frac{1}{N} \sum_{i \in [N]} z_i^\star \|\widehat{x}_i^\star - \widehat{x}_i\| = \frac{1}{N} \sum_{i \in \mathcal{I}} w_{aa'i} z_i^\star \leq \rho,$$

where the first inequality follows by definition of the Wasserstein distance, the equality is from the definition of $w_{aa'i}$, and the last inequality is from the feasibility of $z^\star$ in the linear program in (4.7).

In what follows, we will construct a distribution $\mathbb{Q}_\varepsilon^\star \in \mathbb{B}_\varepsilon(\mathbb{Q}^\star)$ that is ε-suboptimal in (4.7) for $\rho > 0$. For simplicity of exposition, we assume that $\hat{x}_i \neq \hat{x}_i^\star$ for all $i \in [N]$ and the norm on $\mathcal{X}$ used in the Wasserstein ground metric is a 2-norm. For any given ε, we choose $\theta \in [0, 1]$ that satisfies $\theta \geq 1 - N\varepsilon / \sum_{i \in \mathcal{J}} r_{a'} z_i^\star$, where $\mathcal{J} = \{i \in [N] : \widehat{x}_i \in \mathcal{X}_1, \widehat{a}_i = a', \widehat{y}_i = 1\}$. We set $\epsilon_0, \epsilon_1, \epsilon_2 \in \mathbb{R}_+$ to satisfy the following criteria

$$\begin{cases} \theta\epsilon_0 + (1-\theta)\epsilon_1 \leq \varepsilon, \\ (1-\theta)(\epsilon_1 + \epsilon_2) \geq \varepsilon \end{cases}$$

so that for all $i \in [N]$, the set

$$\{x \in \mathcal{X}_1 : \|x - \widehat{x}_i^\star\| \leq \epsilon_1, \|x - \widehat{x}_i\| \leq \|\widehat{x}_i^\star - \widehat{x}_i\| - \epsilon_2\}$$

is non-empty. When the norm on $\mathcal{X}$ is a 2-norm, the above condition is satisfied by setting $\theta\epsilon_0 = \varepsilon/2$, $(1-\theta)\epsilon_1 = \varepsilon/2$, and $\epsilon_2 = \epsilon_1$. For other norms, this requirement can be satisfied by properly scaling ϵ_0 down and scaling ϵ_1 and ϵ_2 up to meet the criteria. For each $i \in [N]$, consider the tuple $(\widehat{x}_{0i}^\varepsilon, \widehat{x}_{1i}^\varepsilon)$ defined as

$$(\widehat{x}_{0i}^\varepsilon, \widehat{x}_{1i}^\varepsilon) = \begin{cases} (\widehat{x}_{0i}, \widehat{x}_{1i}) & \text{if } i \in \mathcal{J}, \\ (\widehat{x}_i^\star, \widehat{x}_i^\star) & \text{otherwise,} \end{cases}$$

where $\widehat{x}_{0i} \in \mathcal{X}_0$ such that $\|\widehat{x}_{0i} - \widehat{x}_i^\star\| \leq \epsilon_0$, and $\widehat{x}_{1i} \in \mathcal{X}_1$ such that $\|\widehat{x}_{1i} - \widehat{x}_i^\star\| \leq \epsilon_1$, and $\|\hat{x}_{1i} - \hat{x}_i\| \leq \|\hat{x}_i^\star - \hat{x}_i\| - \epsilon_2$. Notice that the existence of $\widehat{x}_{0i}$ is guaranteed because $\widehat{x}_i^\star$ is the projection of $\widehat{x}_i$ onto $\partial\mathcal{X}_1$, or equivalently onto $\mathrm{cl}(\mathcal{X}_0)$, and hence $\mathcal{X}_0 \cap \{x_i : \|x_i - \widehat{x}_i^\star\| \leq \epsilon_0\}$ is non-empty for any $\epsilon_0 \in \mathbb{R}_{++}$. Consider now distribution $\mathbb{Q}_\varepsilon^\star$ that is constructed as

$$\mathbb{Q}_\varepsilon^\star = \frac{1}{N}\left(\sum_{i=1}^N \theta z_i^\star \delta_{(\widehat{x}_{0i}^\varepsilon, \widehat{a}_i, \widehat{y}_i)} + \sum_{i=1}^N z_i^\star (1-\theta) \delta_{(\widehat{x}_{1i}^\varepsilon, \widehat{a}_i, \widehat{y}_i)} + \sum_{i=1}^N (1 - z_i^\star) \delta_{(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)}\right).$$

We will show that $\mathbb{Q}_\varepsilon^\star \in \mathbb{B}_\rho(\mathbb{Q}^\star)$. By definition of $\mathbb{Q}_\varepsilon^\star$, we have

$$\begin{aligned} \mathbb{W}(\mathbb{Q}_\varepsilon^\star, \mathbb{Q}^\star) &\leq \frac{1}{N} \sum_{i \in \mathcal{J}} \left(\theta z_i^\star \|\widehat{x}_{0i} - \widehat{x}_i^\star\| + (1-\theta) z_i^\star \|\hat{x}_i^\star - \hat{x}_{1i}\|\right) \\ &\leq \theta\epsilon_0 + (1-\theta)\epsilon_1 \leq \varepsilon, \end{aligned}$$

where the first inequality is due to $z_i^\star \leq 1$ for all $i \in [N]$, $\mathcal{J} \subset [N]$, $\|\hat{x}_{0i} - \hat{x}_i^\star\| \leq \epsilon_0$ and $\|\hat{x}_{1i} - \hat{x}_i^\star\| \leq \epsilon_1$. The last inequality follows by assumption on ϵ_0 and ϵ_1.

Next, we show that $\mathbb{Q}_\varepsilon \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$. Similarly, by construction of $\mathbb{Q}_\varepsilon^\star$ we have

$$\begin{aligned}
\mathbb{W}(\mathbb{Q}_\varepsilon^\star, \widehat{\mathbb{P}}_N) &\leq \frac{1}{N}\left(\sum_{i\in[N]} \theta z_i^\star \|\widehat{x}_{0i}^\varepsilon - \widehat{x}_i\| + \sum_{i\in[N]} (1-\theta) z_i^\star \|\widehat{x}_{1i}^\varepsilon - \widehat{x}_i\|\right) \\
&= \frac{1}{N}\sum_{i\in[N]\setminus\mathcal{J}} \theta z_i^\star \|\hat{x}_i^\star - \hat{x}_i\| + \frac{1}{N}\sum_{i\in\mathcal{J}} \theta z_i^\star \|\hat{x}_{0i} - \hat{x}_i\| \\
&+ \frac{1}{N}\sum_{i\in[N]\setminus\mathcal{J}} (1-\theta) z_i^\star \|\hat{x}_i^\star - \hat{x}_i\| + \frac{1}{N}\sum_{i\in\mathcal{J}} (1-\theta) z_i^\star \|\hat{x}_{1i} - \hat{x}_i\| \\
&\leq \frac{1}{N}\sum_{i\in[N]\setminus\mathcal{J}} \theta z_i^\star \|\hat{x}_i^\star - \hat{x}_i\| + \frac{1}{N}\sum_{i\in\mathcal{J}} \theta z_i^\star \|\hat{x}_{0i} - \hat{x}_i\| \\
&+ \frac{1}{N}\sum_{i\in[N]\setminus\mathcal{J}} (1-\theta) z_i^\star \|\hat{x}_i^\star - \hat{x}_i\| + \frac{1}{N}\sum_{i\in\mathcal{J}} (1-\theta) z_i^\star \|\hat{x}_i^\star - \hat{x}_i\| - (1-\theta)\epsilon_2 \\
&= \frac{1}{N}\sum_{i\in[N]} z_i^\star \|\hat{x}_i^\star - \hat{x}_i\| + \frac{1}{N}\sum_{i\in\mathcal{J}} \theta z_i^\star (\|\hat{x}_{0i} - \hat{x}_i\| - \|\hat{x}_i^\star - \hat{x}_i\|) - (1-\theta)\epsilon_2 \\
&\leq \rho + \frac{1}{N}\sum_{i\in\mathcal{J}} \theta z_i^\star \|\hat{x}_{0i} - \hat{x}_i^\star\| - (1-\theta)\epsilon_2 \leq \rho + \theta\epsilon_0 - (1-\theta)\epsilon_2 \leq \rho,
\end{aligned}$$

where the first equality is due to the definition of $\hat{x}_{0i}^\varepsilon$ and $\hat{x}_{1i}^\varepsilon$. The second inequality follows by construction of $\hat{x}_{1i}$, that is, it satisfies $\|\hat{x}_{1i} - \hat{x}_i\| \leq \|\hat{x}_i^\star - \hat{x}_i\| - \epsilon_2$. The third inequality follows from triangle inequality, that is, $\|\hat{x}_{0i} - \hat{x}_i\| \leq \|\hat{x}_{0i} - \hat{x}_i^\star\| + \|\hat{x}_i - \hat{x}_i^\star\|$ and since $z_i^\star$ is feasible in (4.7). The last equality is due to the choice of ϵ_0 and ϵ_2 that satisfies $\theta\epsilon_0 + (1-\theta)\epsilon_2 \leq 0$. As a consequence, we have $\mathbb{Q}_\varepsilon^\star \in \mathbb{B}_\rho(\widehat{\mathbb{P}}_N)$.

In the last step, we verify that $\mathbb{Q}_\varepsilon^\star$ is an ε-suboptimal solution of the maximization problem that defines $\mathbb{V}(a, a')$. Notice that because $\widehat{\mathbb{P}}_N$ is an empirical distribution, we have

$$\hat{\mathbb{V}}(a, a') = \frac{1}{N}\sum_{\substack{i\in[N]:\, \hat{x}_i\in\mathcal{X}_1 \\ \hat{a}_i = a,\, \hat{y}_i = 1}} r_a - \frac{1}{N}\sum_{\substack{i\in[N]:\, \hat{x}_i\in\mathcal{X}_1 \\ \hat{a}_i = a',\, \hat{y}_i = 1}} r_{a'}.$$

By definition of $\mathbb{Q}_\varepsilon^\star$, we have the following equalities

$$\mathbb{Q}_\varepsilon^\star(X \in \mathcal{X}_1 | A = a, Y = 1) = \mathbb{Q}^\star(X \in \mathcal{X}_1 | A = a, Y = 1) \tag{4.25a}$$

$$\mathbb{Q}_\varepsilon^\star(X \in \mathcal{X}_1 | A = a', Y = 1) = \mathbb{Q}^\star(X \in \mathcal{X}_1 | A = a', Y = 1) - \frac{\theta}{N}\sum_{i\in\mathcal{J}} r_{a'} z_i^\star \tag{4.25b}$$

Similarly by definition of $\mathbb{Q}^\star$, we have the following equalities

$$\mathbb{Q}^\star[X \in \mathcal{X}_1 | A = a, Y = 1] - \mathbb{Q}^\star[X \in \mathcal{X}_1 | A = a', Y = 1]$$

$$= \frac{1}{N} \sum_{\substack{i\in[N]:\hat{a}_i=a\\ \hat{y}_i=1}} r_a z_i^\star + \frac{1}{N} \sum_{\substack{i\in[N]:\hat{x}_i\in\mathcal{X}_1\\ \hat{a}_i=a,\hat{y}_i=1}} r_a(1-z_i^\star) - \tag{4.26}$$

$$\frac{1}{N} \sum_{\substack{i\in[N]:\hat{a}_i=a'\\ \hat{y}_i=1}} r_{a'} z_i^\star - \frac{1}{N}\sum_{i\in\mathcal{J}} r_{a'}(1-z_i^\star)$$

$$= \hat{\mathbb{V}}(a,a') + \frac{1}{N} \sum_{\substack{i\in[N]:\hat{x}_i\in\mathrm{int}(\mathcal{X}_0)\\ \hat{a}_i=a,\hat{y}_i=1}} r_a z_i^\star - \frac{1}{N} \sum_{\substack{i\in[N]:\hat{x}_i\in\mathrm{int}(\mathcal{X}_0),\\ \hat{a}_i=a',\hat{y}_i=1}} r_{a'} z_i^\star$$

$$= \hat{\mathbb{V}}(a,a') + \frac{1}{N} \sum_{\substack{i\in[N]:\hat{x}_i\in\mathrm{int}(\mathcal{X}_0)\\ \hat{a}_i=a,\hat{y}_i=1}} r_a z_i^\star + \frac{1}{N} \sum_{\substack{i\in[N]:\hat{x}_i\in\mathrm{int}(\mathcal{X}_1)\\ \hat{a}_i=a',\hat{y}_i=1}} r_{a'} z_i^\star - \tag{4.27}$$

$$\frac{1}{N} \sum_{\substack{i\in[N]:\hat{x}_i\in\mathrm{int}(\mathcal{X}_1)\\ \hat{a}_i=a',\hat{y}_i=1}} r_{a'} z_i^\star$$

$$= \mathbb{V}(a,a') - \frac{1}{N} \sum_{\substack{i\in[N]:\hat{x}_i\in\mathrm{int}(\mathcal{X}_1)\\ \hat{a}_i=a',\hat{y}_i=1}} r_{a'} z_i^\star, \tag{4.28}$$

where the first equality follows by construction of $\mathbb{Q}^\star$, and the second equality follows from the definition of $\hat{\mathbb{V}}(a,a')$. The third equality follows by realizing that $z_i^\star = 0$ for all indices in the set $\{i \in [N] : \hat{x}_i \in \mathrm{int}(\mathcal{X}_0), \hat{a}_i = a', \hat{y}_i = 1\}$, and we add and subtract the same term to have a representation in terms of $\mathbb{V}(a,a')$. Moreover, the last equality is due to the definition of $\mathbb{V}(a,a')$.

Now, we will show that $\mathbb{Q}_\varepsilon^\star$ provides ε-suboptimal solution to the maximization problem that defines $\mathbb{V}(a,a')$. By taking the difference of (4.25a) and (4.25b),

$$\begin{aligned}
&\mathbb{Q}_\varepsilon^\star(X\in\mathcal{X}_1|A=a,Y=1) - \mathbb{Q}_\varepsilon^\star(X\in\mathcal{X}_1|A=a',Y=1)\\
&= \mathbb{Q}^\star(X\in\mathcal{X}_1|A=a,Y=1) - \mathbb{Q}^\star(X\in\mathcal{X}_1|A=a',Y=1) + \frac{\theta}{N}\sum_{i\in\mathcal{J}} r_{a'} z_i^\star\\
&= \mathbb{V}(a,a') - \frac{1}{N}\sum_{i\in\mathcal{J}} r_{a'} z_i^\star + \frac{\theta}{N}\sum_{i\in\mathcal{J}} r_{a'} z_i^\star \geq \mathbb{V}(a,a') - \varepsilon,
\end{aligned}$$

where the second equality is due to (4.28). The last inequality follows as $\theta \geq 1 - N\varepsilon/\sum_{i\in\mathcal{J}} r_{a'} z_i^\star$. This concludes the proof. □

Additional Theoretical Results

In the main chapter, we solve problem in (4.5) for general $\kappa_{\mathcal{A}}$ and $\kappa_{\mathcal{Y}}$. If $\kappa_{\mathcal{A}}$ and $\kappa_{\mathcal{Y}}$ ceases to be finite then the problem can be substantially simplified.

Corollary 8 (Absolute trust in A and Y)**.** *If $f(z) = \log(z)$, $\eta \leq \min\{\hat{p}_{11}, \hat{p}_{01}\}$ and $\kappa_{\mathcal{A}} = \kappa_{\mathcal{Y}} = \infty$, then problem (4.5) simplifies to the following tractable convex program*

$$
\begin{array}{ll}
\min & t \\
\text{s.t.} & \beta \in \mathbb{R}^p,\ t \in \mathbb{R},\ \lambda_0,\ \lambda_1 \in \mathbb{R}_+,\ \nu_0, \nu_1 \in \mathbb{R}^N \\
& \left.\begin{array}{l}
\|\beta\|_*(1 + \eta r_{a'}) \leq \lambda_a \\
\rho\lambda_a + \frac{1}{N}\sum_{i=1}^N \nu_{ai} \leq t \\
\left.\begin{array}{ll}
\nu_{ai} + \log(h_\beta(-\hat{x}_i)) \geq 0 & \textit{if } \hat{y}_i = 0 \\
\nu_{ai} + (1 - \eta r_a)\log(h_\beta(\hat{x}_i)) \geq 0 & \textit{if } \hat{a}_i = a,\ \ \hat{y}_i = 1 \\
\nu_{ai} + (1 + \eta r_{a'})\log(h_\beta(\hat{x}_i)) \geq 0 & \textit{if } \hat{a}_i = a',\ \ \hat{y}_i = 1
\end{array}\right\} \forall i \in [N]
\end{array}\right\} \\
& \hspace{10em} \forall a, a' \in \mathcal{A} : a' = 1 - a.
\end{array}
$$

Proof of Corollary 8. The proof follows the same steps as the proof of Theorem 4.3.1 until the reformulation of $\mathbb{T}_\beta^{aa'}$. Thanks to Corollary 7, $\mathbb{T}_\beta^{aa'}$ coincides with the optimal value of

$$
\begin{array}{ll}
\inf & \rho\lambda + \frac{1}{N}\sum_{i=1}^N \nu_i \\
\text{s.t.} & \lambda \in \mathbb{R}_+,\ \nu \in \mathbb{R}^N \\
& \lambda\|x_i - \hat{x}_i\| + \nu_i \geq \phi_\beta^{aa'}(x_i, \hat{a}_i, \hat{y}_i)\ \forall x_i \in \mathcal{X},\ \forall i \in [N],
\end{array}
\tag{4.29}
$$

where the function $\phi_\beta^{aa'}$ is as it is defined in the proof of Theorem 4.3.1. We now proceed to consider the constraint of problem (4.29), which can be written in a simplified form as

$$
\nu_i \geq \sup_{x_i \in \mathcal{X}} \left\{ \phi_\beta^{aa'}(x_i, \hat{a}_i, \hat{y}_i) - \lambda\|x_i - \widehat{x}_i\| \right\}. \tag{4.30}
$$

Suppose that $\hat{y}_i = 0$, then $\phi_\beta^{aa'}(x_i, \hat{a}_i, 0) = -\log(1 - h_\beta(x_i))$, and by Lemma 4.5.2, we have

$$
\begin{aligned}
\sup_{x_i \in \mathcal{X}} \left\{ \phi_\beta^{aa'}(x_i, a_i, 0) - \lambda\|x_i - \widehat{x}_i\| \right\} &= \sup_{x_i \in \mathcal{X}} -\log(1 - h_\beta(x_i)) - \lambda\|x_i - \hat{x}_i\| \\
&= \begin{cases} -\log(1 - h_\beta(\hat{x}_i)) & \text{if } \|\beta\|_* \leq \lambda, \\ +\infty & \text{otherwise,} \end{cases}
\end{aligned}
$$

and so the constraint (4.30) when $\hat{y}_i = 0$ becomes

$$
\begin{cases}
\nu_i \geq -\log(1 - h_\beta(\hat{x}_i)) \\
\|\beta\|_* \leq \lambda.
\end{cases}
$$

If $\hat{a}_i = a$ and $\hat{y}_i = 1$, then $\phi_\beta^{aa'}(x_i, a, 1) = (\eta r_a - 1)\log(h_\beta(x_i))$. We thus have by Lemma 4.5.2 that

$$\sup_{x_i \in \mathcal{X}} \left\{ \phi_\beta^{aa'}(x_i, a, 1) - \lambda \|x_i - \widehat{x}_i\| \right\} = \begin{cases} (\eta r_a - 1)\log(h_\beta(\widehat{x}_i)) & \text{if } (1 - \eta r_a)\|\beta\|_* \leq \lambda, \\ +\infty & \text{otherwise.} \end{cases}$$

If $\hat{a}_i = a$ and $\hat{y}_i = 1$, then the constraint (4.30) becomes

$$\begin{cases} \nu_i \geq (\eta r_a - 1)\log(h_\beta(\widehat{x}_i)) \\ (1 - \eta r_a)\|\beta\|_* \leq \lambda. \end{cases}$$

Using an analogous argument for the case where $\hat{a}_i = a'$ and $\hat{y}_i = 1$, we have $\phi_\beta^{aa'}(x_i, a', 1) = -(1 + \eta r_{a'})\log(h_\beta(x_i))$. By Lemma 4.5.2, we have

$$\sup_{x_i \in \mathcal{X}} \left\{ \phi_\beta^{aa'}(x_i, a', 1) - \lambda \|x_i - \widehat{x}_i\| \right\} = \begin{cases} -(1 + \eta r_{a'})\log(h_\beta(\widehat{x}_i)) & \text{if } (1 + \eta r_{a'})\|\beta\|_* \leq \lambda, \\ +\infty & \text{otherwise.} \end{cases}$$

If $\hat{a}_i = a'$ and $\hat{y}_i = 1$, then the constraint (4.30) is equivalent to

$$\begin{cases} \nu_i \geq -(1 + \eta \; r_{a'})\log(h_\beta(\widehat{x}_i)) \\ (1 + \eta \; r_{a'})\|\beta\|_* \leq \lambda. \end{cases}$$

Injecting all the specific cases of constraint (4.30) into problem (4.29), the value $\mathbb{T}_\beta^{aa'}$ is equal to the optimal value of the following optimization problem

$$\begin{array}{ll} \min & \rho\lambda + \frac{1}{N}\sum_{i=1}^N \nu_i \\ \text{s.t.} & \lambda \in \mathbb{R}_+, \; \nu \in \mathbb{R}^N \\ & \|\beta\|_* \leq \lambda, \; \|\beta\|_*(1 - \eta r_a) \leq \lambda, \; \|\beta\|_*(1 + \eta r_{a'}) \leq \lambda \\ & \left.\begin{array}{ll} \nu_i \geq -\log(1 - h_\beta(\hat{x}_i)) & \text{if } \hat{y}_i = 0 \\ \nu_i \geq (\eta r_a - 1)\log(h_\beta(\hat{x}_i)) & \text{if } \hat{a}_i = a, \hat{y}_i = 1 \\ \nu_i \geq -(1 + \eta r_{a'})\log(h_\beta(\hat{x}_i)) & \text{if } \hat{a}_i = a', \hat{y}_i = 1 \end{array}\right\} \forall i \in [N]. \end{array} \tag{4.31}$$

Note that the constraints $\|\beta\|_* \leq \lambda$ and $\|\beta\|_*(1 - \eta r_a) \leq \lambda$ are redundant in view of the constraint $\|\beta\|_*(1 + \eta r_{a'}) \leq \lambda$. The claim then follows by substituting the dual reformulations for $\mathbb{T}_\beta^{aa'}$ into (4.15) and eliminating the embedded minimization operators. □

4.6. Further Discussion and Details of Numerical Results

In this section, we provide further details about the experiments in the Section 4.5, including synthetic experiments, real dataset experiments and illustrations of the extremal distribution. All optimization problems are implemented in Python 3.7 and all experiments were run on an Intel i7-700K CPU (4.2 GHz).

Synthetic Experiments. To show the decision boundaries in Figure 4.1, we generate binary classification data that has 2 dimensional feature vectors with two subgroups one of them being the minority (i.e., $A = 0$). We generate 5000 and 2000 binary class labels $Y \in \{0, 1\}$ uniformly at random for majority subgroup ($A = 1$), and minority subgroup ($A = 0$) respectively. Then, we set the conditional true distributions of 2 dimensional feature vectors as following Gaussian distributions.

$$\begin{aligned}
X|A = 1, Y = 1 &\sim \mathcal{N}([6, 0], [3.5, 0; 0, 3.5]),\\
X|A = 1, Y = 0 &\sim \mathcal{N}([2, 0], [3.5, 0; 0, 3.5]),\\
X|A = 0, Y = 0 &\sim \mathcal{N}([-4, 0], [5, 0; 0, 5]),\\
X|A = 0, Y = 1 &\sim \mathcal{N}([-2, 0], [5, 0; 0, 5]).
\end{aligned}$$

Next, we use stratified sampling[3] to obtain $N = 50$ points from the generated data as a training dataset. We set the rest of the dataset the test dataset that we calculate the accuracy and the unfairness of the trained models.

To obtain the Pareto frontiers in Figure 4.1, we use the synthetic experiment from [Zaf+17b]. In this setting, we set the true distributions of the class labels $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 1/2$. Next, we set the conditional distributions of the 2 dimensional feature vectors as the following Gaussian distributions

$$X|Y = 1 \sim \mathcal{N}([2; 2], [5, 1; 1, 5]),\ X|Y = 0 \sim \mathcal{N}([-2; -2], [10, 1; 1, 3]).$$

Then, we draw sensitive attribute of each sample x from a Bernoulli distribution,

$$\mathbb{P}(A = 1|X = x') = pdf(x'|Y = 1)/(pdf(x'|Y = 1) + pdf(x'|Y = 0)),$$

where $x' = [\cos(\pi/4), \sin(\pi/4); \sin(\pi/4), \cos(\pi/4)]x$ is a rotated version of the feature vector x and $pdf(\cdot|Y = y)$ is the Gaussian probability density function of $X|Y = y$.

We sample 400 i.i.d. samples from $\mathbb{P}$ as our dataset, and we stratify sample 100 data points from this dataset and set it as training set, while we set the rest as the test dataset. The procedure to obtain the frontiers is explained as in Section 4.5. We fix ρ for DR-FLR to 0.01 and and the range of η is $[10^{-4}, \min\{\hat{p}_{11}, \hat{p}_{01}\}]$ with 5 equi-distant points.

Experiments with Real Data. We consider four publicly available datasets (Adult, Drug, COMPAS, Arrythmia). We obtain Adult dataset from UCI repository[4], it contains 14 features concerning demographic characteristics of 45222 instances (32561 for training and 12661 for test). The prediction task is to determine whether a person makes over 50000$ a year, where we consider *gender* as the sensitive attribute. The Drug dataset[5] have records for 1885 respondents.

[3] Stratified sampling is a method of sampling from a population which can be partitioned into subgroups, and requires sampling each subgroup independently.

[4] https://archive.ics.uci.edu/ml/datasets/adult

[5] https://archive.ics.uci.edu/ml/datasets/Drug+consumption+%28quantified%29

Each respondent is described by 12 features, including level of education, age, gender, country of residence and ethnicity. The task is to determine whether the user ever used heroin or not. We consider *ethnicity* as the sensitive attribute. COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)[6] is a popular algorithm used by judges and parole officers for scoring criminal defendant's likelihood of recidivism. It has been shown that the algorithm is biased in favor of white defendants based on a 2-year follow up study. This dataset contains variables used by the COMPAS algorithm in scoring defendants, along with their outcomes within 2 years of the decision for over 10000 criminal defendants. We concentrate on the one that includes only violent recidivism, where *ethnicity* is the sensitive attribute. We obtain the Arrhythmia dataset from UCI repository[7] which contains 279 attributes[8], where the aim is to distinguish between the presence and absence of cardiac arrhythmia and to classify it in one of the 16 groups. In our case, we changed the task with the binary classification between normal arrhythmia against 15 different classes of arrhythmia.

Training, Validation and Testing Procedure. In all other datasets we randomly select 2/3 of the samples for training and we set the rest of the data for testing. We repeat the training, validation and testing process for K_3 times, while the Adult dataset comes with designated training and testing samples, and thus $K_3 = 1$.

Validation. We select the hyper-parameter(s) of the classifier(s) (e.g., the radius of the Wasserstein ball for DR-FLR) using a cross-validation procedure on the training set similar to [Don+18]. First, we collect statistics of the parameters of the model by splitting the training set into sub-training set (N samples) and a validation set for K_1 times. In the first step, the value of the parameter in the grid with highest accuracy calculated over the validation set is identified. In the second step, we shortlist all the values of parameter in the grid with accuracy close (in our case $70\% - 98\%$) to the maximum accuracy in that range minus the lowest possible accuracy. Finally, from this list, we select the parameter value that provides the lowest unfairness measure with respect to the log-probabilistic equal opportunity.

Testing. We stratify sample N samples from the training set and we collect the statistics regarding the performance of the classifiers on the test dataset. We repeat this process for K_2 times.

Discussion on Table 4.1 in Section 4.5. Table 4.1 summarizes the testing accuracy and unfairness of averaged over $K_1 = 3, K_2 = 100, K_3 = 2$, where we tune the radius of Wasserstein ball $\rho \in [10^{-5}, 10^{-1}]$[9] for DR-FLR classifier on a logarithmic search grid with 50 discretization points: All methods are

[6] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

[7] https://archive.ics.uci.edu/ml/datasets/Arrhythmia

[8] We only use the first 12 out of 278 non-sensitive features of the Arrhythmia dataset so that we can use the same search grid for ρ across all datasets (in the other datasets p ranges from 5 to 12).

[9] After we obtain the logarithmic scale, we multiply the values by 5, and thus $\rho \in$

trained with $N = 150$ and we set $\eta = \min\{\hat{p}_{11}, \hat{p}_{01}\}/2$ both for FLR and DR-FLR, $\kappa_{\mathcal{A}} = \kappa_{\mathcal{Y}} = 0.5$ for DR-FLR, DOB$^+$ [Don+18] (the model parameter $\epsilon = 0$), and ZVRG [Zaf+17b] (the model parameter $\epsilon = 10^{-4}$). We use the following accuracy thresholds at the validation step to tune radius of Wasserstein distance for DR-FLR: 95% for Drug and Adult, 97% for Arrhythmia dataset and 73% for COMPAS dataset. The difference of the threshold is due to the structure of dataset. For example, the COMPAS dataset is mostly categorical (other than one attribute that is numerical) and thus to decrease the unfairness, the threshold that we use in the validation step for the accuracy should be smaller than the one would use for other datasets that consists mostly numerical attributes. Moreover, the accuracy threshold also depends on how unbalanced the dataset is, which determines the lowest possible accuracy that is attained when a classifier only predicts 1 (or 0) for all samples.

Worst Case Distribution. To illustrate the extremal distribution $\mathbb{Q}^\star$ from Proposition 4.4.3, we generate two interleaving half circles, which is a simple toy dataset for visualization. We assign the sensitive attributes of the data points uniformly at random by setting two-thirds of the data as the majority subgroup and the remaining one-third as the minority subgroup. We generate 500 samples and split them into training and test sets by 85%-15% proportion. Next, we train the classifiers with the training set and calculate the worst-case unfairness $\overline{\mathbb{U}}_f$ for a prescribed radius ρ. The illustrated extremal distribution $\mathbb{Q}^\star$ in Figure 4.2 are obtained with radius of the Wasserstein ball $0.02, 0.05, 0.05, 0.01$ for classical logistic regression, support vector machine with RBF kernel, Gaussian process wiht RBF kernel and AdaBoost, respectively.

4.6.1. Additional Numerical Experiments

In this section, we provide additional experiments to compare the performance of different classifiers.

Discussion on Table 4.2. An interesting experiment would be to compare the performance of DOB$^+$ and LR, FLR and DR-FLR, when we also tune the parameter of the classifier that is used in DOB$^+$. Because SVM is a deterministic classifier, we cannot calculate the (log-)probabilistic unfairness. Thus, in the cross-validation procedure, we choose the parameter that gives the lowest unfairness with respect to the deterministic equal opportunity both for DR-FLR and DOB$^+$ from the acceptable parameter grid which provides accuracy higher than the given threshold.

The results in Table 4.2 summarize the testing accuracy and unfairness averaged over $K_1 = 5$, $K_2 = 100$, $K_3 = 5$, where we tune the radius of Wasserstein ball $\rho \in [10^{-5}, 10^{-1}]$[10] for DR-FLR classifier and regularization parameter $C \in [10^{-1}, 10^2]$ of linear support vector machine for DOB$^+$ method on a logarithmic search grid with 50 discretization points. We keep the same training

$[5.10^{-5}, 5.10^{-1}]$ at the end.

[10] After we obtain the logarithmic scale, we multiply the values by 5, and thus $\rho \in [5 \cdot 10^{-5}, 5 \cdot 10^{-1}]$ at the end.

Dataset	Metric	LR	FLR	DOB$^+$[Don+18]	DR-FLR
Drug	Accuracy	**0.79±0.01**	**0.79 ± 0.01**	**0.79 ± 0.01**	**0.79 ± 0.01**
	Det-UNF	0.06 ± 0.05	0.06 ± 0.05	0.09 ± 0.07	**0.04 ± 0.04**
	Prob-UNF	0.06 ± 0.05	0.06 ± 0.05	-	**0.05 ± 0.04**
	LogProb-UNF	0.21 ± 0.20	0.20 ± 0.20	-	**0.16 ± 0.14**
Adult	Accuracy	**0.80±0.01**	**0.80 ± 0.01**	0.79 ± 0.01	0.79 ± 0.01
	Det-UNF	0.08 ± 0.06	**0.06 ± 0.06**	0.16 ± 0.10	**0.06 ± 0.06**
	Prob-UNF	0.17 ± 0.08	**0.12 ± 0.08**	–	**0.12 ± 0.08**
	LogProb-UNF	1.01 ± 0.77	0.68 ± 0.68	–	**0.64 ± 0.65**
Compas	Accuracy	**0.65±0.01**	**0.65 ± 0.02**	0.60 ± 0.03	0.60 ± 0.03
	Det-UNF	0.24 ± 0.04	0.23 ± 0.04	0.17 ± 0.06	**0.15 ± 0.07**
	Prob-UNF	0.12 ± 0.02	0.10 ± 0.03	–	**0.03 ± 0.02**
	LogProb-UNF	0.25 ± 0.06	0.22 ± 0.06	–	**0.07 ± 0.04**
Arrhythmia	Accuracy	0.63±0.03	0.63 ± 0.03	**0.65 ± 0.02**	0.62 ± 0.03
	Det-UNF	0.21±0.11	0.15 ± 0.10	0.11 ± 0.08	**0.09 ± 0.08**
	Prob-UNF	0.14±0.07	0.09 ± 0.06	–	**0.05 ± 0.04**
	LogProb-UNF	0.28 ± 0.17	0.19 ± 0.15	–	**0.09 ± 0.08**

Table 4.2: Testing accuracy and unfairness (average ± standard deviation). For DR-FLR ρ and for DOB$^+$ method regularization parameter C of linear SVM is tuned given the training data. LR, FLR, DOB$^+$ and DR-FLR are trained with $N = 150$ samples stratify sampled from the training split.

sample size $N = 150$ for all LR, FLR, DOB$^+$ and DR-FLR. We use the following accuracy thresholds at the validation step to tune ρ for DR-FLR and C for DOB$^+$: 95% for Drug, Adult, and Arrhythmia datasets and 70% for COMPAS dataset.

5. Auditing for Fairness

5.1. Introduction

The past decade witnessed data and algorithms becoming an integrative part of the human society. Recent technological advances are now allowing us to collect and store an astronomical amount of unstructured data, and the unprecedented computing power is enabling us to convert these data into decisional insights. Nowadays, machine learning algorithms can uncover complex patterns in the data to produce an exceptional performance that can match, or even surpass, that of humans. These algorithms, as a consequence, are proliferating in every corner of our lives, from suggesting us the next vacation destination to helping us create digital paintings and melodies. Machine learning algorithms are also gradually assisting humans in consequential decisions such as deciding whether a student is admitted to college, picking which medical treatment to be prescribed to a patient, and determining whether a person is convicted. Arguably, these decisions impact radically many people's lives, together with the future of their loved ones.

Algorithms are conceived and function following strict rules of logic and algebra; it is hence natural to expect that machine learning algorithms deliver objective predictions and recommendations. Unfortunately, in-depth investigations reveal the excruciating reality that state-of-the-art algorithmic assistance is far from being free of biases. For example, a predictive algorithm widely used in the United States criminal justice system is more likely to *mis*classify African-American offenders into the group of high recidivism risk compared to white-Americans [Cho17; Mul16]. The artificial intelligence tool developed by Amazon also learned to penalize gender-related keywords such as "women's" in the profile screening process, and thus may prefer to recommend hiring male candidates for software development and technical positions [Das18]. Further, Google's ad-targeting algorithm displayed advertisements for higher-paying executive jobs more often to men than to women [DTD15].

There are several possible explanations for why cold, soulless algorithms may trigger biased recommendations. First, the data used to train machine learning algorithms may already encrypt human biases manifested in the data collection process. These biases arise as the result of a suboptimal design of experiments, or from historically biased human decisions that accumulate over centuries. Machine-learned algorithms, which are apt to detect underlying patterns from data, will unintentionally learn and maintain these existing biases [BG18; Man+16]. For example, secretary or primary school teacher are professions which are predominantly taken by women, thus, natural language processing systems are inclined to associate female attributes to these jobs. Second, training a machine learning algorithm typically involves minimizing the prediction error which privileges the majority populations over the minority groups. Clinical trials, for instance, typically involve very few participants from the minority groups such as indigenous people, and thus medical interventions recommended by the algorithms may not align perfectly to the characteristics and interests of patients from the minority groups. Finally, even when the sensitive attributes are not used in the training phase, strong correlations between the sensitive attributes and the remaining variables in the dataset may be exploited to generate unjust actions. For example, the sensitive attribute of race can be easily inferred with high accuracy based on common non-sensitive attributes such as the travel history of passengers or the grocery shopping records of customers.

The pressing needs to redress undesirable algorithmic biases have propelled the rising field of fair machine learning[1]. A building pillar of this field involves the verification task: given a machine learning algorithm, we are interested in verifying if this algorithm satisfies a chosen criterion of fairness. This task is performed in two steps: first, we choose an appropriate notion of fairness, then the second step invokes a computational procedure, which may or may not involve data, to decide if the chosen fairness criterion is fulfilled. A plethora of criteria for fair machine learning were proposed in the literature, many of them are motivated by philosophical or sociological ideologies or legal constraints. For example, anti-discrimination laws may prohibit making decisions based on sensitive attributes such as age, gender, race or sexual orientation. Thus, a naïve strategy, called fairness through unawareness, involves removing all sensitive attributes from the training data. However, this strategy seldom guarantees any fairness due to the inter-correlation issues [GH+16; Gar+19], and thus potentially fails to generate inclusive outcomes [BS16; BYF20; Kle+18; LMC18]. Other notions of fairness aim to either promote individual fairness [Dwo+12], prevent disparate treatment [Zaf+17a] or avoid disparate mistreatment [Fel+15; Zaf+17b] of the algorithms. Towards similar goals, notions of *group* fairness focus on reducing the difference of favorable outcomes proportions among different sensitive groups. Examples of group fairness notions include disparate impact [Zaf+17a], demographic parity (statistical parity) [CV10; Dwo+12], equality of opportunity [Har+16] and equalized odds [Har+16]. The notion of counterfactual fairness [Gar+19] was also suggested as a measure of causal fairness. Despite the abundance of available notions, there is unfortunately no general consensus on the most suitable measure to serve as the industry standard. Moreover, except in trivial cases, it is not possible for a machine learning algorithm to simultaneously satisfy multiple notions of fairness [Ber+18; KMR16]. Therefore, the choice of the fairness notion is likely to remain more an art than a science.

This chapter focuses not on the normative approach to choosing an ideal notion of machine learning fairness. We endeavor in this chapter to shed more light on the computational procedure to complement the verification task. Concretely, we position ourselves in the classification setting, which is arguably the most popular task in machine learning. Moreover, we will focus on notions of group fairness, and we employ the framework of statistical hypothesis test instead of algorithmic test.

Contributions. Our work makes two concrete contributions to the problem of fairness testing of machine learning's classifiers.

[1] Comprehensive surveys on fair machine learning can be found in [Ber+18; CR20; CD+17; Meh+19].

1. We propose the Wasserstein projection framework to perform statistical hypothesis test of group fairness for classification algorithms. We derive in details the computation of the test statistic and the limiting distribution when fairness is measured using the probabilistic equality of opportunity and probabilistic equalized odds criteria.

2. We demonstrate that the Wasserstein projection hypothesis testing paradigm is asymptotically correct and can exploit additional information on the geometry of the feature space. Moreover, we also show that this paradigm promotes transparency and interpretability through the analysis of the most favorable distributions.

The remaining of the chapter is structured as follows. In Section 5.2, we introduce the general problem of statistical hypothesis test of classification fairness, and depict the current landscape of fairness testing in the literature. Section 5.3 details our Wasserstein projection approach to this problem. Sections 5.4 and 5.5 apply the proposed framework to test if a pre-trained logistic classifier satisfies the fairness notion of probabilistic equal opportunity and probabilistic equalized odds, respectively. Numerical experiments are presented in Section 5.6 to empirically validate the correctness and demonstrate the power of our proposed paradigm.

All technical proofs are relegated to the Appendix.

5.2. Statistical Testing Framework for Fairness and Literature Review

We consider throughout this chapter a generic binary classification setting. Let $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \{0, 1\}$ be the space of feature inputs and label outputs of interest. We assume that there is a single sensitive attribute corresponding to each data point and its space is denoted by $\mathcal{A} = \{0, 1\}$. A probabilistic classifier is represented by a function $h(\cdot) : \mathcal{X} \to [0, 1]$ that outputs for each given sample $x \in \mathcal{X}$ the probability that x belongs to the positive class. The deterministic classifier predicts class 1 if $h(x) \geq \tau$ and class 0 otherwise, where $\tau \in [0, 1]$ is a classification threshold. Note that the function h depends only on the feature X, but not on the sensitive attribute A, thus predicting Y using h satisfies fairness through unawareness.

The central goal of this chapter is to provide a statistical test to detect if a classifier h fails to satisfy a prescribed notion of machine learning fairness. A statistical hypothesis test can be cast with the null hypothesis being

$$\mathcal{H}_0\text{: the classifier } h \text{ is fair,}$$

against the alternative hypothesis being

$$\mathcal{H}_1\text{: the classifier } h \text{ is not fair.}$$

In this chapter, we focus on statistical notions of *group* fairness, which are usually defined using conditional probabilities. A prevalent notion of fairness in machine learning is the criterion of equality of opportunity[2], which requires that the true positive rate are equal between subgroups.

Definition 6 (Equal opportunity [Har+16])**.** *A classifier* $h(\cdot) : \mathcal{X} \to [0, 1]$ *satisfies the equal opportunity criterion relative to* $\mathbb{Q}$ *if*

$$\mathbb{Q}(h(X) \geq \tau | A = 1, Y = 1) = \mathbb{Q}(h(X) \geq \tau | A = 0, Y = 1),$$

where τ *is the classification threshold.*

Another popular criterion of machine learning fairness is the equalized odds, which is more stringent than the equality of opportunity: it requires that the positive outcome is conditionally independent of the sensitive attributes given the true label.

Definition 7 (Equalized odds [Har+16])**.** *A classifier* $h(\cdot) : \mathcal{X} \to [0, 1]$ *satisfies the equalized odds criterion relative to* $\mathbb{Q}$ *if*

$$\mathbb{Q}(h(X) \geq \tau | A = 1, Y = y) = \mathbb{Q}(h(X) \geq \tau | A = 0, Y = y) \; \forall y \in \mathcal{Y},$$

where τ *is the classification threshold.*

Notice that the criteria of fairness presented in Definitions 6 and 7 are dependent on the distribution $\mathbb{Q}$: a classifier h can be fair relative to a distribution $\mathbb{Q}_1$, but it may become unfair with respect to another distribution $\mathbb{Q}_2 \neq \mathbb{Q}_1$. If we denote by $\mathbb{P}$ the true population distribution that governs the random vector (X, A, Y), then it is imperative and reasonable to test for group fairness with respect to $\mathbb{P}$. For example, to test for the equality of opportunity, we can reformulate a two-sample equal conditional mean test of the null hypothesis

$$\mathcal{H}_0 : \mathbb{E}_{\mathbb{P}}[\mathbb{1}_{h(X) \geq \tau} | A = 1, Y = 1] = \mathbb{E}_{\mathbb{P}}[\mathbb{1}_{h(X) \geq \tau} | A = 0, Y = 1],$$

and one can potentially employ a Welch's t-test with proper adjustment for the randomness of the sample size. Unfortunately, deriving the test becomes complicated when the null hypothesis involves an equality of multi-dimensional quantities, which arises in the case of equalized odds, due to the complication of the covariance terms. Variations of the permutation tests were also proposed to detect discriminatory behaviour of machine learning algorithms following the same formulation of the one-dimensional two-sample equality of conditional mean test [DiC+20; Tra+17]. However, these permutation tests follow a blackbox mechanism and are unable to be generalized to multi-dimensional tests. Tests based on group fairness notions can also be accomplished using an algorithmic approach as in [DiC+20; Sal+18; Gor+19].

[2] We use two terms "equality of opportunity" and "equal opportunity" interchangeably.

From a broader perspective, deriving tests for fairness is an active area of research, and many testing procedures have been recently proposed to test for individual fairness [XYS20; JVS20], for counterfactual fairness [BYF20; Gar+19] and diverse other criteria [Bel+18; Wex+19; Tra+17].

Literature related to optimal transport. Optimal transport is a longstanding field that dates back to the seminal work of Gaspard Monge [Mon81b]. In the past few years, it has attracted significant attention in the machine learning and computer science communities thanks to the availability of fast approximation algorithms [Cut13; DGK18; Ben+15; BSR18; Gen+16]. Optimal transport is particularly successful in various learning tasks, notably generative mixture models [Kol+17; Ngu+13], image processing [AMJJ18; Fer+14; KR15; PR17; TPG16], computer vision and graphics [PW08; PW09; RTG00; Sol+14; Sol+15], clustering [Ho+17], dimensionality reduction [Caz+18; Fla+18; RCP16; Sch16; SC15], domain adaptation [Cou+16; Mur+18], signal processing [Tho+17] and data-driven distributionally robust optimization [Kuh+19; BKM19; GCK17; ZG18]. Recent comprehensive survey on optimal transport and its applications can be found in [PC19a; Kol+17].

In the context of fair classification, ideas from optimal transport have been used to construct fair logistic classifier [Taş+20], to detect classifiers that does not obey group fairness notions, or to ensure fairness by pre-processing [Gor+19], to learn a fair subspace embedding that promotes fair classification [YBS20], to test individual fairness [XYS20], or to construct a counterfactual test [BYF20].

5.3. Wasserstein Projection Framework for Statistical Test of Fairness

We hereby provide a fresh alternative to the testing problem of machine learning fairness. On that purpose, for a given classifier h, we define abstractly the following set of distributions

$$\mathcal{F}_h = \{\mathbb{Q} \in \mathcal{P} : \quad \text{the classifier } h \text{ is fair relative to } \mathbb{Q}\}, \tag{5.1}$$

where $\mathcal{P}$ denotes the space of all distributions on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$. Intuitively, the set $\mathcal{F}_h$ contains all probability distributions under which the classifier h satisfies the prescribed notion of fairness. It is trivial to see that if $\mathcal{F}_h$ contains the true data-generating distribution $\mathbb{P}$, then the classifier h is fair relative to $\mathbb{P}$. Thus, we can reinterpret the hypothesis test of fairness using the hypotheses

$$\mathcal{H}_0\colon \mathbb{P} \in \mathcal{F}_h, \qquad \mathcal{H}_1\colon \mathbb{P} \notin \mathcal{F}_h.$$

Testing the inclusion of $\mathbb{P}$ in $\mathcal{F}_h$ is convenient if $\mathcal{P}$ is endowed with a distance. In this chapter, we equip $\mathcal{P}$ with the Wasserstein distance.

Definition 8 (Wasserstein distance)**.** *The type-*2 *Wasserstein distance between* *two probability distributions* $\mathbb{Q}$ *and* $\mathbb{Q}'$ *supported on* Ξ *is defined as*

$$\mathbb{W}(\mathbb{Q}', \mathbb{Q}) = \min_{\pi \in \Pi(\mathbb{Q}', \mathbb{Q})} \sqrt{\mathbb{E}_\pi[c(\xi', \xi)^2]},$$

where the set $\Pi(\mathbb{Q}', \mathbb{Q})$ *contains all joint distributions of the random vectors* $\xi' \in \Xi$ *and* $\xi \in \Xi$ *under which* ξ' *and* ξ *have marginal distributions* $\mathbb{Q}'$ *and* $\mathbb{Q}$*, respectively, and* $c : \Xi \times \Xi \to [0, \infty]$ *constitutes a lower semi-continuous ground metric.*

The type-2 Wasserstein distance[3] is a special instance of the optimal transport. The squared Wasserstein distance between $\mathbb{Q}'$ and $\mathbb{Q}$ can be interpreted as the cost of moving the distribution $\mathbb{Q}'$ to $\mathbb{Q}$, where $c(\xi', \xi)$ is the cost of moving a unit mass from ξ' to ξ. Being a distance on $\mathcal{P}$, $\mathbb{W}$ is symmetric, non-negative and vanishes to zero if $\mathbb{Q}' = \mathbb{Q}$. The Wasserstein distance is hence an attractive measure to identify if $\mathbb{P}$ belongs to $\mathcal{F}_h$. Using this insight, the hypothesis test for fairness has the equivalent representation

$$\mathcal{H}_0\colon \inf_{\mathbb{Q} \in \mathcal{F}_h} \mathbb{W}(\mathbb{P}, \mathbb{Q}) = 0, \qquad \mathcal{H}_1\colon \inf_{\mathbb{Q} \in \mathcal{F}_h} \mathbb{W}(\mathbb{P}, \mathbb{Q}) > 0.$$

Even though $\mathbb{P}$ remains elusive to our knowledge, we are given access to a set of i.i.d test samples $\{(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)\}_{i=1}^N$ generated from the true distribution $\mathbb{P}$. Thus we can rely on the empirical value

$$\inf_{\mathbb{Q} \in \mathcal{F}_h} \mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q}),$$

which is the distance from the empirical distribution supported on the samples $\widehat{\mathbb{P}}^N = \sum_{i=1}^N \delta_{(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)}$ to the set $\mathcal{F}_h$. To perform the test, it is sufficient to study the limiting distribution of the test statistic using proper scaling under the null hypothesis $\mathcal{H}_0$. The outcome of the test is determined by comparing the test statistic to the quantile value of the limiting distribution at a chosen level of significant $\alpha \in (0, 1)$.

Advantages. The Wasserstein projection framework to hypothesis testing that we described above offers several advantages over the existing methods.

1. Geometric flexibility: The definition of the Wasserstein distance implies that there exists a joint ground metric c on the space of the features, the sensitive attribute and the label. If the modelers or the regulators possess any structural information on an appropriate metric on $\Xi = \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, then this information can be exploited in the testing procedure. Thus, the Wasserstein projection framework equips the users with an additional freedom to inject prior geometric information into the statistical test.

2. Mutivariate generalizability: Certain notions of fairness, such as equalized odds, are prescribed using multiple equalities of conditional expectations. The Wasserstein projection framework encapsulates these equalities simultaneously in the definition of the set $\mathcal{F}_h$, and provides a *joint* test of these

[3] From this point, we omit the term “type-2” for brevity.

equalities without the hassle of decoupling and testing individual equalities as being done in the currently literature.

3. Interpretability: If we denote by $\mathbb{Q}^\star$ the projection of the empirical distribution $\widehat{\mathbb{P}}^N$ onto the set of distributions $\mathcal{F}_h$, i.e.,

$$\mathbb{Q}^\star = \arg\min_{\mathbb{Q} \in \mathcal{F}_h} \ \mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q}),$$

then $\mathbb{Q}^\star$ encodes the minimal perturbation to the empirical samples so that the classifier h becomes fair. The distribution $\mathbb{Q}^\star$ is thus termed the most favorable distribution, and examining $\mathbb{Q}^\star$ can reveal the underlying mechanism and explain the outcome of the hypothesis test. The accessibility to $\mathbb{Q}^\star$ showcases the expressiveness of the Wasserstein projection framework.

Whilst theoretically sound and attractive, there are three potential difficulties with the Wasserstein projection approach to statistical test of fairness. First, to project $\widehat{\mathbb{P}}^N$ onto the set $\mathcal{F}_h$, we need to solve an infinite-dimensional optimization problem, which is inherently difficult. Second, for many notions of machine learning fairness such as the equality of opportunity and the equalized odds, the corresponding set $\mathcal{F}_h$ in (5.1) is usually prescribed using *non*linear constraints. For example, if we consider the equal opportunity criterion in Definition 6, then the set $\mathcal{F}_h$ can be re-expressed using a fractional function of the probability measure as

$$\mathcal{F}_h = \left\{ \mathbb{Q} \in \mathcal{P} \text{ s.t. } \frac{\mathbb{Q}(h(X) \geq \tau, A = 1, Y = 1)}{\mathbb{Q}(A = 1, Y = 1)} = \frac{\mathbb{Q}(h(X) \geq \tau, A = 0, Y = 1)}{\mathbb{Q}(A = 0, Y = 1)} \right\}.$$

Apart from involving nonlinear constraints, it is easy to verify that the set $\mathcal{F}_h$ is also non-convex, which amplifies the difficulty of computing the projection onto $\mathcal{F}_h$. Finally, the limiting distribution of the test statistic is difficult to analyze due to the discontinuity of the probability function at the set $\{x \in \mathcal{X} : h(x) = \tau\}$. The asymptotic analysis with this discontinuity is of a combinatorial nature, and is significantly more problematic than the asymptotic analysis of smooth quantities.

While these difficulties may be overcome via various ways, in this chapter we choose the following combination of remedies. First, we will use a relaxed notion of fairness termed *probabilistic fairness*, which was originally introduced in [Ple+17]. Second, when computing the Wasserstein distances between distributions on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, we use

$$c\big((x', a', y'), (x, a, y)\big) = \|x - x'\| + \infty|a - a'| + \infty|y - y'| \tag{5.2}$$

as the ground metric, where $\|\cdot\|$ is a norm on $\mathbb{R}^d$. This case corresponds to having an absolute trust in the label and in the sensitive attribute of the training samples. This absolute trust restriction is common in the literature of fair machine learning [XYS20; Taş+20].

We now briefly discuss the advantage of using the ground metric of the form (5.2). Denote by $p \in \mathbb{R}_{++}^{|\mathcal{A}|\times|\mathcal{Y}|}$ the array of the true marginals of (A, Y), in particular, $p_{ay} = \mathbb{P}(A = a, Y = y)$ for all $a \in \mathcal{A}$ and $y \in \mathcal{Y}$. Further, let $\widehat{p}^N \in \mathbb{R}_{++}^{|\mathcal{A}| \times |\mathcal{Y}|}$ be the array of the empirical marginals of (A, Y) under the empirical measure $\widehat{\mathbb{P}}^N$, that is, $\widehat{p}_{ay}^N = \widehat{\mathbb{P}}^N(A = a, Y = y)$ for all $a \in \mathcal{A}$ and $y \in \mathcal{Y}$. Throughout this chapter, we assume that the empirical marginals are proper, that is, $\widehat{p}_{ay}^N \in (0, 1)$ for any $(a, y) \in \mathcal{A} \times \mathcal{Y}$. We define temporarily the simplex set $\Delta := \{\bar{p} \in \mathbb{R}_{++}^{|\mathcal{A}| \times |\mathcal{Y}|} : \sum_{a \in \mathcal{A}, y \in \mathcal{Y}} \bar{p}_{ay} = 1\}$. Subsequently, for any marginals $\bar{p} \in \Delta$, we define the marginally-constrained set of distributions

$$\mathcal{F}_h(\bar{p}) \triangleq \left\{ \mathbb{Q} \in \mathcal{P} : \begin{array}{l} h \text{ is fair relative to } \mathbb{Q} \\ \mathbb{Q}(A = a, Y = y) = \bar{p}_{ay} \quad \forall (a, y) \in \mathcal{A} \times \mathcal{Y} \end{array} \right\}.$$

Using these notations, one can readily verify that

$$\mathcal{F}_h = \cup_{\bar{p} \in \Delta} \mathcal{F}_h(\bar{p}).$$

Moreover, the next result asserts that in order to compute the projection of $\widehat{\mathbb{P}}^N$ onto $\mathcal{F}_h$, to suffices to project onto the marginally-constrained set $\mathcal{F}_h(\widehat{p}^N)$.

Lemma 5.3.1 (Projection with marginal restrictions)**.** *Suppose that the ground metric is chosen as in* (5.2)*. If a measure $\mathbb{Q} \in \mathcal{F}_h$ satisfies $\mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q}) < \infty$, then $\mathbb{Q} \in \mathcal{F}_h(\widehat{p}^N)$.*

A useful consequence of Lemma 5.3.1 is that

$$\inf_{\mathbb{Q} \in \mathcal{F}_h} \mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q}) = \inf_{\mathbb{Q} \in \mathcal{F}_h(\widehat{p}^N)} \mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q}), \tag{5.3}$$

where the feasible set of the problem on the right-hand side is the marginally-constrained set $\mathcal{F}_h(\widehat{p}^N)$ using the empirical marginals $\widehat{p}^N$. For two notions of probabilistic fairness that we will explore in this chapter, projecting $\widehat{\mathbb{P}}^N$ onto $\mathcal{F}_h(\widehat{p}^N)$ is arguably easier than onto $\mathcal{F}_h$. Thus, this choice of ground metric improves the tractability when computing the test statistic.

Third, and finally, we will focus on the logistic regression setting, which is one of the most popular classification methods [HJLS13]. In this setting, the conditional probability $\mathbb{P}[Y = 1 | X = x]$ is modelled by the sigmoid function

$$h_\beta(x) = \frac{1}{1 + \exp(-\beta^\top x)},$$

where $\beta \in \mathbb{R}^d$ is the regression parameter. Moreover, a classifier with $\beta = 0$, is trivially fair. Thus, it suffices to consider $\beta \neq 0$.

Notations. We use $\| \cdot \|_*$ to denote the dual norm of $\| \cdot \|$. For any integer N, we define $[N] := \{1, 2, \ldots, N\}$. Given N test samples $(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)_{i=1}^N$, we use $\mathcal{I}_y \triangleq \{i \in [N] : \widehat{y}_i = y\}$ to denote the index set of observations with label y. The parameters λ_i are defined as

$$\forall i \in [N] : \quad \lambda_i = \begin{cases} (\widehat{p}_{11}^N)^{-1} & \text{if } (\widehat{a}_i, \widehat{y}_i) = (1, 1), \\ -(\widehat{p}_{01}^N)^{-1} & \text{if } (\widehat{a}_i, \widehat{y}_i) = (0, 1), \\ (\widehat{p}_{10}^N)^{-1} & \text{if } (\widehat{a}_i, \widehat{y}_i) = (1, 0), \\ -(\hat{p}_{00}^N)^{-1} & \text{if } (\widehat{a}_i, \widehat{y}_i) = (0, 0). \end{cases} \tag{5.4}$$

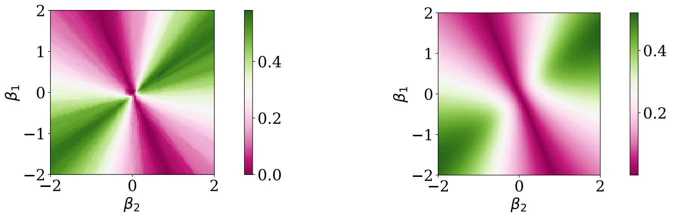
5.4. Testing Fairness for Probabilistic Equal Opportunity Criterion

In this section, we use the ingredients introduced in the previous section to concretely construct a statistical test for the fairness of a logistic classifier h_β. Specifically, we will employ the probabilistic equal opportunity criterion which was originally proposed in [Ple+17].

Definition 9 (Probabilistic equal opportunity criterion [Ple+17])**.** *A logistic classifier* $h_\beta : \mathcal{X} \to [0, 1]$ *satisfies the probabilistic equalized opportunity criteria relative to a distribution* $\mathbb{Q}$ *if*

$$\mathbb{E}_{\mathbb{Q}}[h_\beta(X)|A = 1, Y = 1] = \mathbb{E}_{\mathbb{Q}}[h_\beta(X)|A = 0, Y = 1].$$

The probabilistic equal opportunity criterion, which serves as a surrogate for the equal opportunity criterion in Definition 6, depends on the smooth and bounded sigmoid function h_β but is independent of the classification threshold τ. Motivated by [LPL20], we empirically illustrate in Figure 5.1 that the probabilistic surrogate provides a good approximation of the equal opportunity criterion. Figure 5.1a plots the absolute difference of the classification probabilities $|\mathbb{P}(h(X) \geq \frac{1}{2}|A = 1, Y = 1) - \mathbb{P}(h(X) \geq \frac{1}{2}|A = 0, Y = 1)|$, while Figure 5.1b plots the absolute difference of the sigmoid expectations $|\mathbb{E}_{\mathbb{P}}[h(X)|A = 1, Y = 1] - \mathbb{E}_{\mathbb{P}}[h(X)|A = 0, Y = 1]|$. One may observe that the regions of β so that the absolute differences fall close to zero are similar in both plots. This implies that a logistic classifier h_β which is equal opportunity fair is also likely to be *probabilistic* equal opportunity fair, and vice versa.



(a) Equal opportunity

(b) Probabilistic equal opportunity

Figure 5.1: Comparison of fairness notions for $d = 2$ and $h_\beta(x) = 1/(1+\exp(\frac{1}{3} - \beta_1 x_1 - \beta_2 x_2))$.

We use the superscript "opp" to emphasize that fairness is measured using the probabilistic equal *opp*ortunity criterion. Consequentially, the set of distributions $\mathcal{F}_{h_\beta}^{\text{opp}}$ that makes the logistic classifier h_β fair is

$$\mathcal{F}_{h_\beta}^{\text{opp}} = \left\{ \ \mathbb{Q} \in \mathcal{P} \text{ such that } \mathbb{E}_{\mathbb{Q}}[h_\beta(X)|A = 1, Y = 1] = \mathbb{E}_{\mathbb{Q}}[h_\beta(X)|A = 0, Y = 1] \ \right\}.$$

The statistical hypothesis test to verify whether the classifier h_β is fair is formulated with the null and alternative hypotheses

$$\mathcal{H}_0^{\mathrm{opp}} : \mathbb{P} \in \mathcal{F}_{h_\beta}^{\mathrm{opp}}, \quad \mathcal{H}_1^{\mathrm{opp}} : \mathbb{P} \notin \mathcal{F}_{h_\beta}^{\mathrm{opp}}.$$

The remainder of this section unfolds as follows. In Section 5.4.1, we delineate the computation of the projection of $\widehat{\mathbb{P}}^N$ onto $\mathcal{F}_{h_\beta}^{\mathrm{opp}}$. Section 5.4.2 studies the limiting distribution of the test statistic, while Section 5.4.3 examines the most favorable distribution.

5.4.1. Wasserstein Projection

Lemma 5.3.1 suggests that it is sufficient to consider the projection onto the marginally-constrained set $\mathcal{F}_{h_\beta}^{\mathrm{opp}}(\widehat{p}^N)$, where $\widehat{p}^N$ is the empirical marginals of the empirical distribution $\widehat{\mathbb{P}}^N$. In particular, $\mathcal{F}_{h_\beta}^{\mathrm{opp}}(\widehat{p}^N)$ is

$$\mathcal{F}_{h_\beta}^{\mathrm{opp}}(\widehat{p}^N) = \left\{ \mathbb{Q} \in \mathcal{P} : \begin{array}{l} (\widehat{p}_{11}^N)^{-1}\mathbb{E}_{\mathbb{Q}}[h_\beta(X)\mathbb{1}_{(1,1)}(A,Y)] = (\widehat{p}_{01}^N)^{-1}\mathbb{E}_{\mathbb{Q}}[h_\beta(X)\mathbb{1}_{(0,1)}(A,Y)] \\ \mathbb{Q}(A=a, Y=y) = \widehat{p}_{ay}^N \quad \forall (a,y) \in \mathcal{A} \times \mathcal{Y} \end{array} \right\},$$

where the equality follows from the law of conditional expectation. Notice that the set $\mathcal{F}_{h_\beta}^{\mathrm{opp}}(\widehat{p}^N)$ is prescribed using *linear* constraints of $\mathbb{Q}$, and thus it is more amenable to optimization than the set $\mathcal{F}_{h_\beta}^{\mathrm{opp}}$. It is also more convenient to work with the *squared* distance function $\mathcal{R}$ whose input is the empirical distribution $\widehat{\mathbb{P}}^N$ and its corresponding vector of empirical marginals $\widehat{p}^N$ by

$$\mathcal{R}^{\mathrm{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) := \begin{cases} \inf & \mathbb{W}(\mathbb{Q}, \widehat{\mathbb{P}}^N)^2 \\ \text{s.t.} & \mathbb{E}_{\mathbb{Q}}[h_\beta(X)((\widehat{p}_{11}^N)^{-1}\mathbb{1}_{(1,1)}(A,Y) - (\widehat{p}_{01}^N)^{-1}\mathbb{1}_{(0,1)}(A,Y))] = 0 \\ & \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{(a,y)}(A,Y)] = \widehat{p}_{ay}^N \quad \forall (a,y) \in \mathcal{A} \times \mathcal{Y}. \end{cases}$$

Notice that the constraints of the above infimum problem are linear in the measure $\mathbb{Q}$, but the functions inside the expectation operators are possibly *non*linear functions of $\widehat{p}^N$. Using the equivalent characterization (5.3), the following relation holds

$$\inf_{\mathbb{Q} \in \mathcal{F}_{h_\beta}^{\mathrm{opp}}} \mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q}) = \inf_{\mathbb{Q} \in \mathcal{F}_{h_\beta}^{\mathrm{opp}}(\widehat{p}^N)} \mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q}) = \sqrt{\mathcal{R}^{\mathrm{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)}.$$

We now proceed to show how computing the projection can be reduced to solving a finite-dimensional optimization problem.

Proposition 5.4.1 (Dual reformulation)**.** *The squared projection distance $\mathcal{R}^{\mathrm{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$ equals to the optimal value of the following finite-dimensional optimization problem*

$$\sup_{\kappa \in \mathbb{R}} \frac{1}{N} \sum_{i \in \mathcal{I}_1} \inf_{x_i \in \mathcal{X}} \left\{ \|x_i - \widehat{x}_i\|^2 + \kappa \lambda_i h_\beta(x_i) \right\}. \tag{5.5}$$

While Proposition 5.4.1 asserts that computing the *squared* projection distance $\mathcal{R}^{\mathrm{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$ is equivalent to solving a finite-dimensional problem, unfortunately, this saddle point problem is in general difficult. Indeed, because h_β is non-convex, even finding the optimal inner solution $x_i^\star$ for a fixed value of the outer variable $\gamma \in \mathbb{R}$ is generally NP-hard [MK85]. The situation can be partially alleviated if $\|\cdot\|$ is an Euclidean norm on $\mathbb{R}^d$.

Lemma 5.4.2 (Univariate reduction)**.** *Suppose that $\|\cdot\|$ is the Euclidean norm on $\mathbb{R}^d$, we have*

$$\mathcal{R}^{\mathrm{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) = \sup_{\kappa \in \mathbb{R}} \frac{1}{N} \sum_{i \in \mathcal{I}_1} \min_{k_i \in [0, \frac{1}{8}]} \; \gamma^2 \lambda_i^2 \|\beta\|_2^2 k_i^2 + \frac{\gamma \lambda_i}{1 + \exp(\gamma \lambda_i \|\beta\|_2^2 k_i - \beta^\top \widehat{x}_i)}. \tag{5.6}$$

The proof of Lemma 5.4.2 follows trivially from application of Lemma 5.8.1 to reformulate the inner infimum problems for each $i \in \mathcal{I}_1$. Lemma 5.4.2 offers a significant reduction in the computational complexity to solve the inner subproblems of (5.5). Instead of optimizing over d-dimensional vector x_i, the representation in Lemma 5.4.2 suggests that it suffices to search over a 1-dimensional space for k_i. While the objective function is still non-convex in k_i, we can perform a grid search over a compact interval to find the optimal solution for k_i to high precision. The grid search operations can also be parallelized across the index i thanks to the independent structure of the inner problems. Furthermore, the objective function of the supremum problem is a point-wise minimum of linear, thus concave, functions of γ. Hence, the outer problem is a concave maximization problem in γ, which can be solved using a golden section search algorithm.

5.4.2. Limiting Distribution

We now characterize the limit properties of $\mathcal{R}^{\mathrm{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$. The next theorem assert that the limiting distribution is of the chi-square type.

Theorem 5.4.3 (Limiting distribution – Probabilistic equal opportunity)**.** *Suppose that $(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)$ are i.i.d. samples from $\mathbb{P}$. Under the null hypothesis $\mathcal{H}_0^{\mathrm{opp}}$, we have*

$$N \times \mathcal{R}^{\mathrm{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) \xrightarrow{d.} \theta \chi_1^2,$$

where χ_1^2 is a chi-square distribution with 1 degree of freedom,

$$\theta = \left(\mathbb{E}_{\mathbb{P}} \left[\left\| \nabla h_\beta(X) \left(\frac{\mathbb{1}_{(1,1)}(A, Y)}{p_{11}} - \frac{\mathbb{1}_{(0,1)}(A, Y)}{p_{01}} \right) \right\|_*^2 \right] \right)^{-1} \frac{\sigma_1^2}{p_{01}^2 p_{11}^2}$$

with $\sigma_1^2 = \mathrm{Cov}(Z_1)$*, and* Z_1 *is the random variable*

$$\begin{aligned} Z_1 = {} & h_\beta(X)\left(p_{01}\mathbb{1}_{(1,1)}(A,Y) - p_{11}\mathbb{1}_{(0,1)}(A,Y)\right) \\ & + \mathbb{1}_{(0,1)}(A,Y)\mathbb{E}_{\mathbb{P}}[\mathbb{1}_{(1,1)}(A,Y)h_\beta(X)] \\ & - \mathbb{1}_{(1,1)}(A,Y)\mathbb{E}_{\mathbb{P}}[\mathbb{1}_{(0,1)}(A,Y)h_\beta(X)]. \end{aligned}$$

Construction of the hypothesis test. Based on the result of Theorem 5.4.3, the statistical hypothesis test proceeds as follows. Let $\eta_{1-\alpha}^{\mathrm{opp}}$ denote the $(1-\alpha)\times 100\%$ quantile of $\theta\chi_1^2$, where $\alpha \in (0,1)$ is the predetermined significance level. By Theorem 5.4.3, the statistical decision has the form

$$\text{Reject } \mathcal{H}_0^{\mathrm{opp}} \text{ if } \widehat{s}_N^{\mathrm{opp}} > \eta_{1-\alpha}^{\mathrm{opp}}$$

with $\widehat{s}_N^{\mathrm{opp}} = N \times \mathcal{R}^{\mathrm{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$. The limiting distribution $\theta\chi_1^2$ is nonpivotal because θ depends on the true distribution $\mathbb{P}$. Luckily, because the quantile function of $\theta\chi_1^2$ is continuous in θ, if $\widehat{\theta}_N$ is a consistent estimator of θ then it is also valid to use the quantile of $\widehat{\theta}_N\chi_1^2$ for the purpose of testing. We thus proceed to discuss a consistent estimator $\widehat{\theta}_N$ constructed from the available data. First, notice that $\widehat{p}_{01}^N$ and $\widehat{p}_{11}^N$ are consistent estimator for p_{01} and p_{11}. Similarly, the law of large numbers asserts that the denominator term in the definition of θ can be estimated by the sample average

$$\begin{aligned} & \mathbb{E}_{\mathbb{P}}\left[\left\|\nabla h_\beta(X)\left(\frac{\mathbb{1}_{(1,1)}(A,Y)}{p_{11}} - \frac{\mathbb{1}_{(0,1)}(A,Y)}{p_{01}}\right)\right\|_*^2\right] \\ & \approx \widehat{T}^N = \frac{\|\beta\|_*^2}{N}\sum_{i=1}^N h_\beta(\widehat{x}_i)^2(1-h_\beta(\widehat{x}_i))^2\left(\frac{\mathbb{1}_{(1,1)}(\widehat{a}_i,\widehat{y}_i)}{(\widehat{p}_{11}^N)^2} + \frac{\mathbb{1}_{(0,1)}(\widehat{a}_i,\widehat{y}_i)}{(\widehat{p}_{01}^N)^2}\right). \end{aligned}$$

Under the null hypothesis $\mathcal{H}_0^{\mathrm{opp}}$, Z_1 has mean 0. The sample average estimate of σ_1^2 is $\sigma_1^2 \approx (\widehat{\sigma}^N)^2$ with

$$\begin{aligned} (\widehat{\sigma}_1^N)^2 = \frac{1}{N}\sum_{i=1}^N \Big[& h_\beta(\widehat{x}_i)\left(p_{01}\mathbb{1}_{(1,1)}(\widehat{a}_i,\widehat{y}_i) - p_{11}\mathbb{1}_{(0,1)}(A,Y)\right) \\ & + \mathbb{1}_{(0,1)}(\widehat{a}_i,\widehat{y}_i)\big(\sum_{j=1}^N \mathbb{1}_{(1,1)}(\widehat{a}_j,\widehat{y}_j)h_\beta(\widehat{x}_j)\big) \\ & - \mathbb{1}_{(1,1)}(\widehat{a}_i,\widehat{y}_i)\big(\sum_{j=1}^N \mathbb{1}_{(0,1)}(\widehat{a}_j,\widehat{y}_j)h_\beta(\widehat{x}_j)\big)\Big]^2. \end{aligned} \tag{5.7}$$

Using a nested arguments involving the continuous mapping theorem and Slutsky's theorem, the estimator

$$\widehat{\theta}^N = \frac{(\widehat{\sigma}_1^N)^2}{\widehat{T}^N(\widehat{p}_{01}^N)^2(\widehat{p}_{11}^N)^2}$$

is consistent for θ. Let the corresponding $(1-\alpha)\times 100\%$ quantile of the random variable $\widehat{\theta}^N\chi_1^2$ be $\widehat{\eta}_{1-\alpha}^{\mathrm{opp}}$. The statistical test decision using the plug-in consistent estimate becomes

$$\text{Reject } \mathcal{H}_0^{\mathrm{opp}} \text{ if } \widehat{s}_N^{\mathrm{opp}} > \widehat{\eta}_{1-\alpha}^{\mathrm{opp}}.$$

5.4.3. Most Favorable Distributions

We now discuss the construction of the most favorable distribution $\mathbb{Q}^\star$, the projection of the empirical distribution $\widehat{\mathbb{P}}^N$ onto the set $\mathcal{F}_{h_\beta}^{\mathrm{opp}}$. Intuitively, $\mathbb{Q}^\star$ is the distribution closest to $\widehat{\mathbb{P}}^N$ that makes h_β a fair classifier under the equal opportunity criterion. If $\|\cdot\|$ is the Euclidean norm, the information about $\mathbb{Q}^\star$ can be recovered from the optimal solution of problem (5.6) by the result of the following lemma.

Lemma 5.4.4 (Most favorable distribution)**.** *Suppose that $\|\cdot\|$ is the Euclidean norm. Let $\kappa^\star$ be the optimal solution of problem* (5.6)*, and for any $i\in\mathcal{I}_1$, let $k_i^\star$ be a solution of the inner minimization of* (5.6) *with respect to $\kappa^\star$. Then the most favorable distribution $\mathbb{Q}^\star = \arg\min_{\mathbb{Q}\in\mathcal{F}_{h_\beta}^{\mathrm{opp}}} \mathbb{W}(\widehat{\mathbb{P}}^N,\mathbb{Q})$ is a discrete distribution of the form*

$$\mathbb{Q}^\star = \frac{1}{N}\Big(\sum_{i\in\mathcal{I}_0}\delta_{(\widehat{x}_i,\widehat{a}_i,\widehat{y}_i)} + \sum_{i\in\mathcal{I}_1}\delta_{(\widehat{x}_i-k_i^\star\kappa^\star\lambda_i\beta,\widehat{a}_i,\widehat{y}_i)}\Big).$$

By using the result of Lemma 5.4.2, it is easy to verify that $\mathbb{Q}^\star$ satisfies $\mathbb{W}(\mathbb{Q}^\star,\widehat{\mathbb{P}}^N)^2 = \mathcal{R}^{\mathrm{opp}}(\widehat{\mathbb{P}}^N,\widehat{p}^N)$. Moreover, one can also show that $\mathbb{Q}^\star\in\mathcal{F}_{h_\beta}^{\mathrm{opp}}$. These two observations imply that $\mathbb{Q}^\star$ is the projection of $\widehat{\mathbb{P}}^N$ onto $\mathcal{F}_{h_\beta}^{\mathrm{opp}}$. The detailed proof is omitted.

Lemma 5.4.4 suggests that in order to obtain the most favorable distribution, it suffices to perturb only the data points with positive label. This is intuitively rational because the notion of probabilistic equality of opportunity only depends on the positive label, and thus the perturbation with a minimal energy requirement should only move sample points with $\widehat{y}_i = 1$. When the underlying geometry is the Euclidean norm, the optimal perturbation of the point $\widehat{x}_i$ is to move it along a line dictated by β with a scaling factor $k_i^\star\gamma^\star\lambda_i$. Notice that λ_i defined in (5.4) are of opposite signs between samples of different sensitive attributes, which implies that it is optimal to perturb $\widehat{x}_i$ in opposite directions dependent on whether $\widehat{a}_i = 0$ or $\widehat{a}_i = 1$. This is, again, rational because moving points in opposite direction brings the clusters of points closer to the others, which reduces the discrepancy in the expected value of $h_\beta(X)$ between subgroups.

As a final remark, we note that $\mathbb{Q}^\star$ is not necessarily unique. This is because of the non-convexity of the inner problem over k_i in (5.6), which leads to the non-uniqueness of the optimal solution $k_i^\star$ (see Appendix 5.8 and Figure 5.5).

5.5. Testing Fairness for Probabilistic equalized odds Criterion

In this section, we extend the Wasserstein projection framework to the statistical test of probabilistic equalized odds for a pre-trained logistic classifier.

Definition 10 (Probabilistic equalized odds criterion [Ple+17]). *A logistic classifier $h_\beta(\cdot) : \mathcal{X} \to [0,1]$ satisfies the probabilistic equalized odds criteria relative to $\mathbb{Q}$ if*

$$\mathbb{E}_{\mathbb{Q}}[h_\beta(X)|A=1, Y=y] = \mathbb{E}_{\mathbb{Q}}[h_\beta(X)|A=0, Y=y] \quad \forall y \in \mathcal{Y}.$$

The notion of probabilistic equalized odds requires that the conditional expectation of h_β to be independent of A for any label subgroup, thus it is more stringent than the probabilistic equal opportunity studied in the previous section. We use the superscript "odd" in this section to emphasize on this specific notion of fairness. The definition of the probabilistic equalized odds prescribes the following set of distributions

$$\mathcal{F}_{h_\beta}^{\mathrm{odd}} = \left\{ \mathbb{Q} \in \mathcal{P} : \begin{array}{l} \mathbb{E}_{\mathbb{Q}}[h_\beta(X)|A=1, Y=1] = \mathbb{E}_{\mathbb{Q}}[h_\beta(X)|A=0, Y=1] \\ \mathbb{E}_{\mathbb{Q}}[h_\beta(X)|A=1, Y=0] = \mathbb{E}_{\mathbb{Q}}[h_\beta(X)|A=0, Y=0] \end{array} \right\}.$$

Correspondingly, the Wasserstein projection hypothesis test for probabilisitc equalized odds can be formulated as

$$\mathcal{H}_0^{\mathrm{odd}} : \mathbb{P} \in \mathcal{F}_{h_\beta}^{\mathrm{odd}}, \quad \mathcal{H}_1^{\mathrm{odd}} : \mathbb{P} \notin \mathcal{F}_{h_\beta}^{\mathrm{odd}}.$$

In the sequence, we study the projection onto the manifold $\mathcal{F}_{h_\beta}^{\mathrm{odd}}$ in Section 5.5.1. Section 5.5.2 examines the asymptotic behaviour of the test statistic, and we close this section by studying the most favorable distribution $\mathbb{Q}^\star$ in Section 5.5.3.

5.5.1. Wasserstein Projection

Following a similar strategy as in Section 5.4, we define the set

$$\mathcal{F}_{h_\beta}^{\mathrm{odd}}(\widehat{p}^N) = \left\{ \mathbb{Q} \in \mathcal{P} : \begin{array}{l} (\widehat{p}_{11}^N)^{-1}\mathbb{E}_{\mathbb{Q}}[h_\beta(X)\mathbb{1}_{(1,1)}(A,Y)] = (\widehat{p}_{01}^N)^{-1}\mathbb{E}_{\mathbb{Q}}[h_\beta(X)\mathbb{1}_{(0,1)}(A,Y)] \\ (\widehat{p}_{10}^N)^{-1}\mathbb{E}_{\mathbb{Q}}[h_\beta(X)\mathbb{1}_{(1,0)}(A,Y)] = (\widehat{p}_{00}^N)^{-1}\mathbb{E}_{\mathbb{Q}}[h_\beta(X)\mathbb{1}_{(0,0)}(A,Y)] \\ \mathbb{Q}(A=a, Y=y) = \widehat{p}_{ay}^N \ \ \forall (a,y) \in \mathcal{A} \times \mathcal{Y} \end{array} \right\},$$

and the squared distance function

$$\mathcal{R}^{\mathrm{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) = \left\{ \begin{array}{cl} \inf & \mathbb{W}(\mathbb{Q}, \widehat{\mathbb{P}}^N)^2 \\ \text{s.t.} & \mathbb{E}_{\mathbb{Q}}[h_\beta(X)((\widehat{p}_{11}^N)^{-1}\mathbb{1}_{(1,1)}(A,Y) - (\widehat{p}_{01}^N)^{-1}\mathbb{1}_{(0,1)}(A,Y))] = 0 \\ & \mathbb{E}_{\mathbb{Q}}[h_\beta(X)((\widehat{p}_{10}^N)^{-1}\mathbb{1}_{(1,0)}(A,Y) - (\widehat{p}_{00}^N)^{-1}\mathbb{1}_{(0,0)}(A,Y))] = 0 \\ & \mathbb{E}_{\mathbb{Q}}[\mathbb{1}_{(a,y)}(A,Y)] = \widehat{p}_{ay}^N \quad \forall (a,y) \in \mathcal{A} \times \mathcal{Y}. \end{array} \right.$$

The equivalent relation (5.3) suggests that the projection onto the set of distributions $\mathcal{F}_{h_\beta}^{\text{odd}}$ satisfies

$$\inf_{\mathbb{Q} \in \mathcal{F}_{h_\beta}^{\text{odd}}} \mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q}) = \inf_{\mathbb{Q} \in \mathcal{F}_{h_\beta}^{\text{odd}}(\widehat{p}^N)} \mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q}) = \sqrt{\mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)}.$$

The squared distance $\mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$ can be computed by solving the saddle point problem in the following proposition.

Proposition 5.5.1 (Dual reformulation)**.** *The squared projection distance* $\mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$ *equals to the optimal value of the following finite-dimensional optimization problem*

$$\sup_{\kappa \in \mathbb{R}, \zeta \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^N \inf_{x_i \in \mathcal{X}} \left\{ \|x_i - \widehat{x}_i\|^2 + (\kappa \lambda_i \mathbb{1}_1(\widehat{y}_i) + \zeta \lambda_i \mathbb{1}_0(\widehat{y}_i)) h_\beta(x_i) \right\}. \tag{5.8}$$

To complete this section, we now discuss an efficient way to compute $\mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$. The next lemma reveals that computing $\mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$ can be decomposed into two subproblems of similar structure.

Lemma 5.5.2 (Univariate reduction)**.** *We have*

$$\mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) = \mathcal{R}^{\text{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) + U_N,$$

where U_N *is computed as*

$$U_N = \sup_{\zeta \in \mathbb{R}} \frac{1}{N} \sum_{i \in \mathcal{I}_0} \inf_{x_i \in \mathcal{X}} \left\{ \|x_i - \widehat{x}_i\|^2 + \zeta \lambda_i h_\beta(x_i) \right\}.$$

Furthermore, if $\|\cdot\|$ *is the Euclidean norm on* $\mathbb{R}^d$*, then*

$$U_N = \sup_{\zeta \in \mathbb{R}} \frac{1}{N} \left\{ \sum_{i \in \mathcal{I}_0} \min_{k_i \in [0, \frac{1}{8}]} \zeta^2 \lambda_i^2 \|\beta\|_2^2 k_i^2 + \frac{\zeta \lambda_i}{1 + \exp(\zeta \lambda_i \|\beta\|_2^2 k_i - \beta^\top \widehat{x}_i)} \right\}. \tag{5.9}$$

Notice that problem (5.9) has a similar structure to problem (5.6): the mere difference is that the summation in the objective function of (5.9) runs over the index set $\mathcal{I}_0 = \{i \in [N] : \widehat{y}_i = 0\}$ instead of $\mathcal{I}_1$ in (5.6). Solving for U_N thus incurs the same computational complexity as, and can also be performed in parallel with, computing $\mathcal{R}^{\text{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$.

5.5.2. Limiting Distribution

The next result asserts that the squared projection distance $\mathcal{R}^{\text{odd}}$ has the $O(N^{-1})$ convergence rate.

Theorem 5.5.3 (Limiting distribution – Probabilistic equalized odds)**.** *Suppose that $(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)$ are i.i.d. samples from $\mathbb{P}$. Under the null hypothesis $\mathcal{H}_0^{\text{odd}}$, we have*

$$
\begin{aligned}
& N \times \mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) \xrightarrow{d.} \\
& \sup_{\gamma,\zeta} \Bigg\{ \gamma H_1 + \zeta H_0 + \\
& \mathbb{E}_{\mathbb{P}} \left[\left\| \begin{pmatrix} \gamma \\ \zeta \end{pmatrix}^\top \begin{pmatrix} p_{11}^{-1} \mathbb{1}_{(1,1)}(A,Y) - p_{01}^{-1} \mathbb{1}_{(0,1)}(A,Y) \\ p_{10}^{-1} \mathbb{1}_{(1,0)}(A,Y) - p_{00}^{-1} \mathbb{1}_{(0,0)}(A,Y) \end{pmatrix} \nabla h_\beta(X) \right\|_*^2 \right] \Bigg\},
\end{aligned}
$$

where $\nabla h_\beta(X) = h_\beta(X)(1 - h_\beta(X)\beta$, and $H_y = \mathcal{N}(0, \sigma_y^2)/(p_{1y} p_{0y})$ with $\sigma_y^2 = \operatorname{Cov}(Z_y)$, and Z_y are random variables

$$
\begin{aligned}
Z_y = {} & h_\beta(X) \left(p_{0y} \mathbb{1}_{(1,y)}(A,Y) - p_{1y} \mathbb{1}_{(0,y)}(A,Y) \right) \\
& + \mathbb{1}_{(0,y)}(A,Y) \mathbb{E}_{\mathbb{P}}[\mathbb{1}_{(1,y)}(A,Y) h_\beta(X)] \\
& - \mathbb{1}_{(1,y)}(A,Y) \mathbb{E}_{\mathbb{P}}[\mathbb{1}_{(0,y)}(A,Y) h_\beta(X)].
\end{aligned}
$$

Construction of the hypothesis test. Contrary to the explicit chi-square limiting distribution for the probabilistic equal opportunity fairness in Theorem 5.4.3, the limiting distribution for the probabilistic equalized odds fairness is not available in closed form. Nevertheless, the limiting distribution in this case can be obtained by sampling H_0 and H_1 and solving a collection of optimization problems for each sample. Notice that the objective function of the supremum problem presented in Theorem 5.5.3 is continuous in H_1 and H_0, one thus can define

$$\widehat{H}_y = \mathcal{N}(0, \widehat{\sigma}_y^2)/(\widehat{p}_{1y}^N \widehat{p}_{0y}^N),$$

where $\widehat{\sigma}_y^2$ is the sample average estimate of σ_y^2, which can be computed using an equation similar to (5.7). The limiting distribution can be computed by solving the optimization problem with plug-in values

$$
\begin{aligned}
& \sup_{\gamma,\zeta} \Bigg\{ \gamma \widehat{H}_1 + \zeta \widehat{H}_0 + \\
& \mathbb{E}_{\widehat{\mathbb{P}}^N} \left[\left\| \begin{pmatrix} \gamma \\ \zeta \end{pmatrix}^\top \begin{pmatrix} (\widehat{p}_{11}^N)^{-1} \mathbb{1}_{(1,1)}(A,Y) - (\widehat{p}_{01}^N)^{-1} \mathbb{1}_{(0,1)}(A,Y) \\ (\widehat{p}_{10}^N)^{-1} \mathbb{1}_{(1,0)}(A,Y) - (\widehat{p}_{00}^N)^{-1} \mathbb{1}_{(0,0)}(A,Y) \end{pmatrix} \nabla h_\beta(X) \right\|_*^2 \right] \Bigg\}.
\end{aligned}
$$

Notice that the expectation in taken over the empirical distribution $\widehat{\mathbb{P}}^N$, and can be written as a finite sum. The last optimization problem can be solved efficiently using quadratic programming for any realization of $\widehat{H}_1$ and $\widehat{H}_0$. The objective values can be collected to compute the $(1-\alpha)\times 100\%$-quantile estimate $\widehat{\eta}^{\text{odd}}_{1-\alpha}$ of the limiting distribution. The statistical test decision using the plug-in estimate becomes

$$\text{Reject } \mathcal{H}^{\text{odd}}_0 \text{ if } \widehat{s}^{\text{odd}}_N > \widehat{\eta}^{\text{odd}}_{1-\alpha},$$

where $\widehat{s}^{\text{odd}}_N = N \times \mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$.

5.5.3. Most Favorable Distributions

If the feature space $\mathcal{X}$ is endowed with an Euclidean norm, then the most favorable distribution $\mathbb{Q}^\star$, defined in this section as the projection of $\widehat{\mathbb{P}}^N$ onto $\mathcal{F}^{\text{odd}}_{h_\beta}$, can be constructed by exploiting Lemma 5.5.2.

Lemma 5.5.4 (Most favorable distribution)**.** *Suppose that $\|\cdot\|$ is the Euclidean norm. Let $\kappa^\star$ and $\zeta^\star$ be the optimal solution of problems* (5.6) *and* (5.9)*, respectively. For any $i \in \mathcal{I}_1$, let $k^\star_i$ be the solution of the inner minimization of* (5.6) *with respect to $\kappa^\star$, and for any $i \in \mathcal{I}_0$, let $k^\star_i$ be a solution of the inner minimization of* (5.9) *with respect to $\zeta^\star$. Then the most favorable distribution $\mathbb{Q}^\star = \arg\min_{\mathbb{Q}\in\mathcal{F}^{\text{odd}}_{h_\beta}} \mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q})$ is a discrete distribution of the form*

$$\mathbb{Q}^\star = \frac{1}{N}\Big(\sum_{i\in\mathcal{I}_0} \delta_{(\widehat{x}_i - k^\star_i\zeta^\star\lambda_i\beta, \widehat{a}_i, \widehat{y}_i)} + \sum_{i\in\mathcal{I}_1} \delta_{(\widehat{x}_i - k^\star_i\kappa^\star\lambda_i\beta, \widehat{a}_i, \widehat{y}_i)}\Big).$$

The proof of Lemma 5.5.4 follows from verifying that $\mathbb{Q}^\star \in \mathcal{F}^{\text{odd}}_{h_\beta}$ and that $\mathbb{W}(\mathbb{Q}^\star, \widehat{\mathbb{P}}^N)^2 = \mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$ using Lemma 5.5.2, the detailed proof is omitted. For probabilistic equalized odds, the most favorable distribution $\mathbb{Q}^\star$ alters the locations of both $i \in \mathcal{I}_0$ and $i \in \mathcal{I}_1$. The directions of perturbation are dependent on λ_i, which is determined using (5.4). Notice that λ_i carry opposite signs corresponding to whether $\widehat{a}_i = 0$ or $\widehat{a}_i = 1$, thus the perturbations will move $\widehat{x}_i$ in opposite directions based on the value of the sensitive attribute $\widehat{a}_i$.

5.6. Numerical Experiment

All experiments are run on an Intel Xeon based cluster composed of 287 compute nodes each with 2 Skylake processors running at 2.3 GHz with 18 cores each. We only use 2 nodes of this cluster and all optimization problems are implemented in Python version 3.7.3. In all experiments, we use the 2-norm to measure distances in the feature space. Moreover, we focus on the hypothesis test of probabilistic equal opportunity, and thus the Wasserstein projection, the limiting distribution and the most favorable distribution follow from the results presented in Section 5.4.

5.6.1. Validation of the Hypothesis Test

We now demonstrate that our proposed Wasserstein projection framework for statistical test of fairness is a valid, or asymptotically correct, test. We consider a binary classification setting in which $\mathcal{X}$ is 2-dimensional feature space. The true distribution $\mathbb{P}$ has true marginal values p_{ay} being

$$p_{11} = 0.2,\ p_{01} = 0.1,\ p_{10} = 0.3,\ p_{00} = 0.4.$$

Moreover, conditioning on (A, Y), the feature X follows a Gaussian distribution of the form

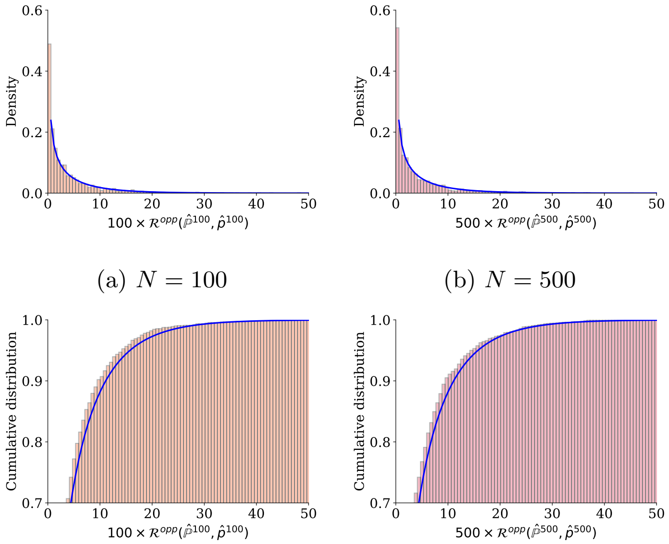
$$\begin{aligned} X|A = 1, Y = 1 &\sim \mathcal{N}([6, 0], [3.5, 0; 0, 5]),\\ X|A = 0, Y = 1 &\sim \mathcal{N}([-2, 0], [5, 0; 0, 5]),\\ X|A = 1, Y = 0 &\sim \mathcal{N}([6, 0], [3.5, 0; 0, 5]),\\ X|A = 0, Y = 0 &\sim \mathcal{N}([-4, 0], [5, 0; 0, 5]). \end{aligned}$$

The true distribution $\mathbb{P}$ is thus a mixture of Gaussian, and under this specification, a simple algebraic calculation indicates that a logistic classifier with $\beta = (0, 1)^\top$ is fair with respect to the probabilistic equal opportunity criteria in Definition 9. We thus focus on verifying fairness for this specific classifier. In the first experiment, we empirically validate Theorem 5.4.3. To this end, we generate $N \in \{100, 500\}$ i.i.d. samples from $\mathbb{P}$ to be used as the test data, and then calculate the squared projection distance $\mathcal{R}^{\text{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$ using Proposition 5.4.1. The process is repeated 2,000 times to obtain an empirical estimate of the distribution of $N \times \mathcal{R}^{\text{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$. We also generate another set of one million i.i.d. samples from $\mathbb{P}$ to estimate the limiting distribution $\theta\chi^2_1$. Figure 5.2 shows that the empirical distribution of $N \times \mathcal{R}^{\text{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$ converges to the limiting distribution $\theta\chi^2_1$ as N increases.

The second set of experiments aims to show that our proposed Wasserstein projection hypothesis test is asymptotically valid. We generate $N \in \{100, 500, 1000\}$ i.i.d. samples from $\mathbb{P}$ and calculate the test statistic $N \times \mathcal{R}^{\text{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$. The same data is used to estimate $\widehat{\theta}^N$ and compute the $(1-\alpha) \times 100\%$-quantile of $\widehat{\theta}^N \chi^2_1$ to perform the quantile based test as laid out in Section 5.4.2. We repeat this procedure for 2,000 replications to keep track of the rejection projection at different significant values of $\alpha \in \{0.5, 0.3, 0.1, 0.05, 0.01\}$. Table 5.1 summarizes the rejection probabilities of Wasserstein projection tests for equal opportunity criterion under the null hypothesis $\mathcal{H}^{\text{opp}}_0$. We can observe that at sample size $N > 100$, the rejection probability is close to the desired level α, which empirically validates our testing procedure.

5.6.2. Most Favorable Distribution Analysis

In this section, we visualize the most favorable distribution $\mathbb{Q}^\star$ from Lemma 5.4.4 for a vanilla logistic regression classifier with weight $\beta = (0.4, 0.12)^\top$. We simply generate 28 samples with equal subgroup proportions to form the empirical



(a) $N = 100$ (b) $N = 500$

(c) $N = 100$ (d) $N = 500$

Figure 5.2: Empirical distribution of $N \times \mathcal{R}^{\mathrm{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$ taken over 2,000 replications (histogram) versus the limiting distribution $\theta\chi_1^2$ (blue curve) with different sample sizes N. Fig. 5.2a-5.2b are density plots, Fig. 5.2c-5.2d are cumulative distribution plots.

distribution $\widehat{\mathbb{P}}^N$. To find the support of $\mathbb{Q}^\star$, we solve problem (5.6), whose optimizer dictates the transportation plan of each sample $\widehat{x}_i$. Figure 5.3 visualizes the original test samples that forms $\widehat{\mathbb{P}}^N$, along with the most favorable distribution $\mathbb{Q}^\star$. Green lines in the figure represent how samples are perturbed. As we are testing for the probabilistic notion of equal opportunity, only the samples with positive label $\widehat{y}_i = 1$ presented in blue are perturbed in order to obtain $\mathbb{Q}^\star$. Furthermore, we observe that the positively-labeled test samples are transported along the axis directed by β (black arrow). Moreover, the samples with different sensitive attributes, represented by different shapes, move in opposite direction so that they get closer to each other, which reduces the discrepancy in the expected value of $h_\beta(X)$ between the relevant subgroups.

5.6.3. The COMPAS Dataset

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions)[4] is a commercial tool used by judges and parole officers for scoring criminal defendant's likelihood of recidivism. The COMPAS dataset is used by the

[4] https://www.propublica.org/datastore/dataset/compas-recidivism-risk-score-data-and-analysis

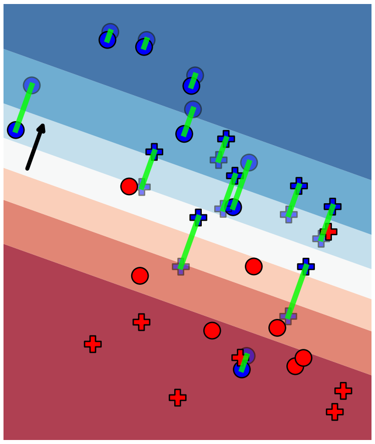


Figure 5.3: Visualization of the most favorable distribution $\mathbb{Q}^\star$ for a logistic classifier with weight $\beta = (0.4, 0.12)^\top$. The black arrow indicates the vector β. Colors represent class, while symbolic shapes encode the sensitive values. The green lines show the transport plan of the empirical test samples from their original positions (indicated with transparent colors) to their ultimate destinations (with non-transparent colors).

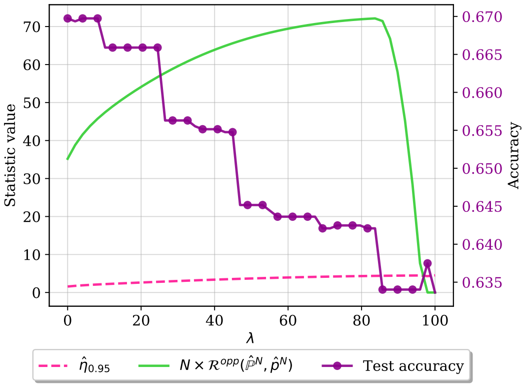


Figure 5.4: Test statistic and accuracy of Tikhonov regularized logistic regression on test data with rejection threshold $\widehat{\eta}_{0.95}$.

$N = 100$	$N = 500$	$N = 1000$	α
0.511	0.4905	0.5	0.50
0.282	0.2895	0.299	0.30
0.048	0.0895	0.093	0.10
0.007	0.0425	0.0405	0.05
0.0	0.0065	0.005	0.01

Table 5.1: Comparison of the null rejection probabilities of probabilistic equal opportunity tests with different significance levels α and test sample sizes N.

COMPAS algorithm to compute the risk score of reoffending for defendants, and also contains the criminal records within 2 years after the decision. The dataset consists of 6,172 samples with 10 attributes including gender, age category, race, etc. We concentrate on the subset of the data with violent recidivism, and we use race (African-American and Caucasian) as the sensitive attribute. We split 70% of the COMPAS data to train a Tikhonov-regularized logistic classifier, with the tuning penalty parameter λ chosen in the range from 0 to 100 with 50 equi-distant points. The remaining 30% of the data is used as the test samples for auditing.

Figure 5.4 demonstrates the relation between the accuracy and the degree of fairness with respect to the regularization parameter λ. Strong regularization penalty (high values of λ) results in small values of the test statistic, but the classifier has low test accuracy. On the contrary, weak penalization leads to undesirable fairness level but higher prediction accuracy. The pink dashed line in Figure 5.4 shows the rejection threshold of the Wasserstein projection test at significance level $\alpha = 0.05$ for varying value of the regularization parameter λ. We can observe that the Wasserstein projection test recommends a rejection of the null hypothesis $\mathcal{H}_0^{\mathrm{opp}}$ for a wide range of λ. Only at λ sufficiently large that the test fails to reject the null hypothesis.

5.7. Appendix - Proofs

5.7.1. Proofs of Section 5.2

Proof of Lemma 5.3.1. Because the fairness constraints are similar in both sets $\mathcal{F}_h$ and $\mathcal{F}_h(\widehat{p}^N)$, it thus suffice to verify that $\mathbb{Q}$ satisfies the marginal conditions $\mathbb{Q}(A = a, Y = y) = \widehat{p}_{ay}^N$ for all $(a, y) \in \mathcal{A} \times \mathcal{Y}$. By the definition of the Wasserstein distance and the ground metric c, there exists a coupling π such that

$$\mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q})^2 = \mathbb{E}_\pi[(\|X' - X\| + \infty|A' - A| + \infty|Y' - Y|)^2]$$

and the marginal distribution of π are $\widehat{\mathbb{P}}^N$ and $\mathbb{Q}$, respectively. By the law of total probability and because $\widehat{\mathbb{P}}^N$ is an empirical distribution, we can write $\pi = N^{-1}\sum_{i=1}^N \delta_{(\hat{x}_i, \hat{a}_i, \hat{y}_i)} \otimes \mathbb{Q}_i$, where $\mathbb{Q}_i$ denotes the conditional distributions of (X, A, Y) given $(X', A', Y') = (\hat{x}_i, \hat{a}_i, \hat{y}_i)$ for all $i \in [N]$.

Suppose without any loss of generality that there exists a tuple $(a, y) \in \mathcal{A} \times \mathcal{Y}$ such that $\mathbb{Q}(A = a, Y = y) > \widehat{p}_{ay}^N$. This means

$$\mathbb{Q}(A = a, Y = y) = \frac{1}{N}\sum_{i=1}^N \mathbb{Q}_i(A = a, Y = y) > \frac{1}{N}\sum_{i=1}^N \mathbb{1}_{(a,y)}(\widehat{a}_i, \widehat{y}_i).$$

This implies that there must exist an index $i^\star \in [N]$ with $(\widehat{a}_{i^\star}, \widehat{y}_{i^\star}) \neq (a, y)$, and that

$$\mathbb{Q}_{i^\star}(A = a, Y = y) > 0.$$

However, this further implies that

$$\begin{aligned}
\mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q})^2 &= \frac{1}{N}\sum_{i=1}^N \mathbb{E}_{\mathbb{Q}_i}[(\|\widehat{x}_i - X\| + \infty|\widehat{a}_i - A| + \infty|\widehat{y}_i - Y|)^2] \\
&\geq \frac{1}{N}\mathbb{E}_{\mathbb{Q}_{i^\star}}[(\|\widehat{x}_{i^\star} - X\| + \infty|\widehat{a}_{i^\star} - A| + \infty|\widehat{y}_{i^\star} - Y|)^2] \\
&\geq \frac{1}{N}\mathbb{Q}_{i^\star}(A = a, Y = y)\left(\infty(\widehat{a}_{i^\star} - a) + \infty(\widehat{y}_{i^\star} - y)\right)^2 = \infty,
\end{aligned}$$

where the equality follows from the decomposition of π using the law of total probability and the first inequality follows because the transportation cost is nonnegative. This contradicts the fact that $\mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q}) < \infty$. □

5.7.2. Proofs of Section 5.4

Before proving Proposition 5.4.1, we first prove a preparatory lemma that verifies the Slater condition of the conic optimization problem. To shorten the notation, we write $\xi = (X, A, Y)$ and denote $\Xi = \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, $\widehat{\Xi}_N = \{(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)\}_{i=1}^N$. We assume that $N \geq 2$ and $\widehat{\xi}_i = (\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)$ are distinct. We use $\mathcal{M}_+(\Xi \times \widehat{\Xi}_N)$ to denote the set of all nonnegative measures on $\Xi \times \widehat{\Xi}_N$.

Lemma 5.7.1 (Slater condition - Probabilistic equal opportunity). *Suppose that $\beta \neq 0$, $\widehat{p}_{11}^N \in (0, 1)$ and $\widehat{p}_{01}^N \in (0, 1)$. Define the function*

$$f_\beta(X, A, Y) \triangleq \frac{1}{\widehat{p}_{11}^N} h_\beta(X)\mathbb{1}_{(1,1)}(A, Y) - \frac{1}{\widehat{p}_{01}^N} h_\beta(X)\mathbb{1}_{(0,1)}(A, Y),$$

and let f be a vector-valued function $f : \Xi \times \widehat{\Xi}_N \to \mathbb{R}^{N+1}$

$$f(\xi, \xi') = \begin{pmatrix} \mathbb{1}_{\hat{\xi}_i}(\xi') \\ \vdots \\ \mathbb{1}_{\hat{\xi}_N}(\xi') \\ f_\beta(\xi) \end{pmatrix}.$$

Then we have

$$\begin{pmatrix} 1/N \\ \vdots \\ 1/N \\ 0 \end{pmatrix} \in \operatorname{int} \left\{ \mathbb{E}_\pi[f(\xi, \xi')] : \pi \in \mathcal{M}_+(\Xi \times \widehat{\Xi}_N) \right\}.$$

Proof of Lemma 5.7.1. It suffices to show that for any

$$q \in \left(\frac{1}{2N}, \frac{3}{2N} \right)^N \times \left(-\frac{1}{4}, \frac{1}{4} \right),$$

there exists a nonnegative measure $\pi \in \mathcal{M}_+(\Xi \times \widehat{\Xi}_N)$ such that $q = \mathbb{E}_\pi[f(\xi, \xi')]$. We will verify this claim by constructing π explicitly. To this end, define the following locations

$$x_{ay} \in \mathcal{X} \quad \forall (a, y) \in \mathcal{A} \times \mathcal{Y},$$

and set $\pi \in \mathcal{M}_+(\Xi \times \widehat{\Xi}_N)$ explicitly as

$$\pi(\xi = (x_{\widehat{a}_i \widehat{y}_i}, \widehat{a}_i, \widehat{y}_i), \xi' = (\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)) = q_i,$$

and π is 0 everywhere else. By construction, one can verify that $\mathbb{E}_\pi[\mathbb{1}_{\widehat{\xi}_i}(\xi')] = q_i$ for all $i \in [N]$. If we define the following index sets $\mathcal{I}_{ay} = \{i \in [N] : \widehat{a}_i = a, \widehat{y}_i = y\}$, then

$$\mathbb{E}_\pi[f_\beta(\xi)] = (\widehat{p}_{11}^N)^{-1} h_\beta(x_{11}) \sum_{i \in \mathcal{I}_{11}} q_i - (\widehat{p}_{01}^N)^{-1} h_\beta(x_{01}) \sum_{i \in \mathcal{I}_{01}} q_i.$$

It now remains to find the locations of x_{11} and x_{01} to balance the above equation. We have the following two cases.

1. Suppose that $q_{N+1} \geq 0$. In this case, choose $x_{01} \in \mathcal{X}$ such that $h_\beta(x_{01}) = \frac{1}{6}$. The condition $\mathbb{E}_\pi[f_\beta(\xi)] = q_{N+1}$ requires that

$$h_\beta(x_{11}) = \frac{q_{N+1} + \frac{1}{6}(\widehat{p}_{01}^N)^{-1} \sum_{i \in \mathcal{I}_{01}} q_i}{(\widehat{p}_{11}^N)^{-1} \sum_{i \in \mathcal{I}_{11}} q_i}.$$

Because $q_{N+1} \geq 0$ and q_i are strictly positive, the term on the right hand side is strictly positive. Moreover, we have

$$(\widehat{p}_{01}^N)^{-1} \sum_{i \in \mathcal{I}_{01}} q_i < \frac{3}{2} \quad \text{and} \quad (\widehat{p}_{11}^N)^{-1} \sum_{i \in \mathcal{I}_{11}} q_i > \frac{1}{2}$$

for any feasible value of q_i, which implies that

$$0 < \frac{q_{N+1} + \frac{1}{6}(\widehat{p}_{01}^N)^{-1} \sum_{i \in \mathcal{I}_{01}} q_i}{(\widehat{p}_{11}^N)^{-1} \sum_{i \in \mathcal{I}_{11}} q_i} < \frac{\frac{1}{4} + \frac{1}{4}}{\frac{1}{2}} = 1.$$

This implies the existence of $x_{11} \in \mathcal{X}$ so that $\mathbb{E}_\pi[f_\beta(\xi)] = q_{N+1}$.

2. Suppose that $q_{N+1} < 0$. In this case, we can choose $x_{11} \in \mathcal{X}$ such that $h_\beta(x_{11}) = \frac{1}{6}$. A similar argument as in the previous case implies the existence of $x_{01} \in \mathcal{X}$ such that $\mathbb{E}_\pi[f_\beta(\xi)] = q_{N+1}$.

Combining the two cases leads to the postulated results. □

We are now ready to prove Proposition 5.4.1.

Proof of Proposition 5.4.1. For the purpose of this proof, we define the function $\lambda : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}$ as

$$\lambda(a, y) = \frac{\mathbb{1}_{(1,1)}(a, y)}{\widehat{p}_{11}^N} - \frac{\mathbb{1}_{(0,1)}(a, y)}{\widehat{p}_{01}^N}. \tag{5.10}$$

By definition of the squared distance function $\mathcal{R}^{\text{opp}}$, we have

$$\mathcal{R}^{\text{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) = \left\{ \begin{array}{cl} \inf\limits_{\mathbb{Q} \in \mathcal{P}} & \mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q})^2 \\ \text{s.t.} & (\hat{p}_{11}^N)^{-1} \mathbb{E}_{\mathbb{Q}}[h_\beta(X) \mathbb{1}_{(1,1)}(A, Y)] = (\widehat{p}_{01}^N)^{-1} \mathbb{E}_{\mathbb{Q}}[h_\beta(X) \mathbb{1}_{(0,1)}(A, Y)] \\ & \mathbb{Q}(A = a, Y = y) = \widehat{p}_{ay}^N \quad \forall a \in \mathcal{A},\ y \in \mathcal{Y} \end{array} \right.$$

$$= \left\{ \begin{array}{cll} \inf\limits_{\pi} & \mathbb{E}_\pi[c\big((X', A', Y'), (X, A, Y)\big)^2] & \\ \text{s.t.} & \pi \in \mathcal{P}((\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{A} \times \mathcal{Y})) & \\ & \mathbb{E}_\pi[f_\beta(X, A, Y)] = 0 & \\ & \pi(A = a, Y = y) = \widehat{p}_{ay}^N & \forall a \in \mathcal{A},\ y \in \mathcal{Y} \\ & \mathbb{E}_\pi[\mathbb{1}_{(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)}(X', A', Y')] = 1/N & \forall i \in [N], \end{array} \right.$$

where the function f_β is defined as

$$f_\beta(x, a, y) \triangleq (\widehat{p}_{11}^N)^{-1} h_\beta(x) \mathbb{1}_{(1,1)}(a, y) - (\widehat{p}_{01}^N)^{-1} h_\beta(x) \mathbb{1}_{(0,1)}(a, y) h_\beta(x) \lambda(a, y), \tag{5.11}$$

and $\mathcal{P}(\mathcal{S})$ denotes the set of all joint probability measures supported on $\mathcal{S}$. Because of the infinity individual cost on $\mathcal{A}$ and $\mathcal{Y}$ by the definition of cost in (5.2), any joint measure π with finite objective value should satisfies $\pi(A = a, Y = y) = \widehat{\mathbb{P}}^N(A' = a, Y' = y) = \hat{p}^N_{ay}$ for any $a \in \mathcal{A}$ and $y \in \mathcal{Y}$. Thus, the set of constraints $\pi(A = a, Y = y) = \hat{p}^N_{ay}$ can be eliminated without alternating the optimization problem. We thus have

$$\mathcal{R}^{\text{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) = \begin{cases} \inf\limits_{\pi} & \mathbb{E}_\pi[c\big((X', A', Y'), (X, A, Y)\big)^2] \\ \text{s.t.} & \pi \in \mathcal{P}((\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{A} \times \mathcal{Y})) \\ & \mathbb{E}_\pi[f_\beta(X, A, Y)] = 0 \\ & \mathbb{E}_\pi[\mathbb{1}_{(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)}(X', A', Y')] = 1/N \quad \forall i \in [N]. \end{cases}$$

To shorten the notations, we use $\Xi = \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ and $\widehat{\Xi}_N = \{(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)\}$. Moreover, define the vector $\bar{q}$ and the vector-valued Borel measurable function on $\Xi \times \widehat{\Xi}_N$ as

$$\bar{q} = \begin{pmatrix} 0 \\ 1/N \\ \vdots \\ 1/N \end{pmatrix} \qquad f(\xi, \xi') = \begin{pmatrix} f_\beta(\xi) \\ \mathbb{1}_{\hat{\xi}_i}(\xi') \\ \vdots \\ \mathbb{1}_{\hat{\xi}_N}(\xi') \end{pmatrix}.$$

By using the introduced notation, we can reformulate the above optimization problem as

$$\inf \left\{ \mathbb{E}_\pi[c(\xi, \xi')^2] : \pi \in \mathcal{M}_+(\Xi \times \widehat{\Xi}_N), \mathbb{E}_\pi[f(\xi, \xi')] = \bar{q} \right\}$$

which is a problem of moments. By Lemma 5.7.1, the above optimization problem satisfies the Slater condition, thus the strong duality result [Smi95, Section 2.2] implies that

$$\mathcal{R}^{\text{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) = \begin{cases} \sup & \dfrac{1}{N} \sum\limits_{i=1}^N b_i \\ \text{s.t.} & b \in \mathbb{R}^N, \ \kappa \in \mathbb{R} \\ & \sum\limits_{i=1}^N b_i \mathbb{1}_{(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)}(x', a', y') - \kappa f_\beta(x, a, y) \leq \\ & \qquad c\big((x', a', y'), (x, a, y)\big)^2 \\ & \qquad\qquad \forall (x, a, y), (x', a', y') \in \mathcal{X} \times \mathcal{A} \times \mathcal{Y}. \end{cases} \tag{5.12}$$

Note that the problem in (5.12) can be equivalently represented as

$$\begin{cases} \sup & \frac{1}{N}\sum_{i=1}^{N} b_i \\ \text{s.t.} & b \in \mathbb{R}^N,\ \kappa \in \mathbb{R} \\ & b_i - \kappa f_\beta(x_i, a_i, y_i) \le c\big((\widehat{x}_i, \widehat{a}_i, \widehat{y}_i), (x_i, a_i, y_i)\big)^2 \\ & \qquad\qquad \forall (x_i, a_i, y_i) \in \mathcal{X} \times \mathcal{A} \times \mathcal{Y}, \forall i \in [N] \end{cases}$$

$$= \sup_{\kappa \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} \inf_{x_i \in \mathcal{X}} \left\{ \|x_i - \widehat{x}_i\|^2 + \kappa f_\beta(x_i, \widehat{a}_i, \widehat{y}_i) \right\}. \tag{5.13}$$

Because f_β has the form (5.11), we have the equivalent problem

$$\sup_{\kappa \in \mathbb{R}} \frac{1}{N} \sum_{i=1}^{N} \inf_{x_i \in \mathcal{X}} \left\{ \|x_i - \widehat{x}_i\|^2 + \kappa \lambda(\widehat{a}_i, \widehat{y}_i) h_\beta(x_i) \right\}.$$

For any $i \in \mathcal{I}_0$, $\lambda(\widehat{a}_i, \widehat{y}_i) = 0$, and in this case we have the optimal solution of x_i satisfies $x_i^\star = \widehat{x}_i$. As a consequence, the summation collapses to a partial sum over $\mathcal{I}_1$. This observation completes the proof. □

Proof of Theorem 5.4.3. Leveraging equation (5.13), we can express

$$\mathcal{R}^{\text{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) = \sup_{\gamma} \mathbb{E}_{\widehat{\mathbb{P}}^N} \left[\inf_{\Delta} \gamma h_\beta(X + \Delta) \left(\frac{\mathbb{1}_{(1,1)}(A, Y)}{\widehat{p}_{11}^N} - \frac{\mathbb{1}_{(0,1)}(A, Y)}{\widehat{p}_{01}^N} \right) + \|\Delta\|^2 \right].$$

We define

$$H^N \triangleq \frac{1}{\sqrt{N}} \sum_{i=1}^{N} h_\beta(\widehat{x}_i) \left(\frac{\mathbb{1}_{(1,1)}(\widehat{a}_i, \widehat{y}_i)}{\widehat{p}_{11}^N} - \frac{\mathbb{1}_{(0,1)}(\widehat{a}_i, \widehat{y}_i)}{\widehat{p}_{01}^N} \right),$$

and using this expression we can reformulate $\mathcal{R}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$ as

$$\sup_{\gamma} \left\{ \frac{1}{\sqrt{N}} \gamma H^N + \mathbb{E}_{\widehat{\mathbb{P}}^N} \Big[\qquad \inf_{\Delta} \gamma [h_\beta(X + \Delta) - h_\beta(X)] \times \right. \tag{5.14}$$
$$\left. \left(\frac{\mathbb{1}_{(1,1)}(A, Y)}{\widehat{p}_{11}^N} - \frac{\mathbb{1}_{(0,1)}(A, Y)}{\widehat{p}_{01}^N} \right) + \|\Delta\|^2 \Big] \right\}.$$

Because h_β is a sigmoid function, it is differentiable, and by the fundamental theorem of calculus, we have for any $x \in \mathcal{X}$,

$$h_\beta(x + \Delta) - h_\beta(x) = \int_0^1 \nabla h_\beta(x + t\Delta) \cdot \Delta \mathrm{d}t,$$

where $\cdot$ represents the inner product on $\mathbb{R}^d$. By applying variable transformations $\gamma \leftarrow \gamma\sqrt{N}$ and $\Delta \leftarrow \Delta\sqrt{N}$, we have

$$\begin{aligned}
&N \times \mathcal{R}^{\text{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) \\
&= \sup_{\gamma} \Bigg\{ \gamma H^N + \mathbb{E}_{\widehat{\mathbb{P}}^N}\Big[\inf_{\Delta} \gamma \int_0^1 \nabla h_\beta \left(X + t\frac{\Delta}{\sqrt{N}} \right) \cdot \Delta \mathrm{d}t \times \\
&\quad \left(\frac{\mathbb{1}_{(1,1)}(A,Y)}{\widehat{p}^N_{11}} - \frac{\mathbb{1}_{(0,1)}(A,Y)}{\widehat{p}^N_{01}} \right) + \|\Delta\|^2 \Big] \Bigg\} \\
&= \sup_{\gamma} \Bigg\{ \gamma H^N + \frac{1}{N}\sum_{i=1}^N \inf_{\Delta_i} \gamma \int_0^1 \nabla h_\beta \left(\widehat{x}_i + t\frac{\Delta_i}{\sqrt{N}} \right) \cdot \Delta_i \mathrm{d}t \times \\
&\quad \left(\frac{\mathbb{1}_{(1,1)}(\widehat{a}_i,\widehat{y}_i)}{\widehat{p}^N_{11}} - \frac{\mathbb{1}_{(0,1)}(\widehat{a}_i,\widehat{y}_i)}{\widehat{p}^N_{01}} \right) + \|\Delta_i\|^2 \Bigg\},
\end{aligned}$$

where the second equality follows by the definition of the empirical distribution $\widehat{\mathbb{P}}^N$. For any values of $\widehat{p}^N_{01} > 0$ and $\widehat{p}^N_{11} > 0$, we have for any $\gamma \neq 0$,

$$\begin{aligned}
&\mathbb{P}\left(\left\| \gamma \nabla h_\beta(X) \left(\frac{\mathbb{1}_{(1,1)}(A,Y)}{\widehat{p}^N_{11}} - \frac{\mathbb{1}_{(0,1)}(A,Y)}{\widehat{p}^N_{01}} \right) \right\|_* = 0 \right) = \\
&\mathbb{P}\big((\hat{p}^N_{11})^{-1}\mathbb{1}_{(1,1)}(A,Y) = (\hat{p}^N_{01})^{-1}\mathbb{1}_{(0,1)}(A,Y) \big) = \mathbb{P}(Y=0) < 1,
\end{aligned}$$

which implies that

$$\mathbb{P}\left(\left\| \gamma \nabla h_\beta(X) \left(\frac{\mathbb{1}_{(1,1)}(A,Y)}{\widehat{p}^N_{11}} - \frac{\mathbb{1}_{(0,1)}(A,Y)}{\widehat{p}^N_{01}} \right) \right\|_* > 0 \right) > 0.$$

This coincides with Assumption A4 in [BKM19]. Using the same argument as in the proof of [BKM19, Theorem 3], we can show that the optimal solution for γ and Δ_i belong to a compact set with high probability. Moreover, we have

$$\frac{\mathbb{1}_{(1,1)}(\widehat{a}_i,\widehat{y}_i)}{\widehat{p}^N_{11}} - \frac{\mathbb{1}_{(0,1)}(\widehat{a}_i,\widehat{y}_i)}{\widehat{p}^N_{01}} = \frac{\mathbb{1}_{(1,1)}(\widehat{a}_i,\widehat{y}_i)}{p_{11}}\left(1 - o_{\mathbb{P}}(1)\right) - \frac{\mathbb{1}_{(0,1)}(\widehat{a}_i,\widehat{y}_i)}{p_{01}}\left(1 - o_{\mathbb{P}}(1)\right),$$

and thus

$$\begin{aligned}
&N \times \mathcal{R}^{\text{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) \\
&= \sup_{\gamma} \Bigg\{ \gamma H^N + \frac{1}{N}\sum_{i=1}^N \inf_{\Delta_i} \gamma \int_0^1 \nabla h_\beta \left(\widehat{x}_i + t\frac{\Delta_i}{\sqrt{N}} \right) \cdot \Delta_i \mathrm{d}t \times \\
&\quad \left(\frac{\mathbb{1}_{(1,1)}(\widehat{a}_i,\widehat{y}_i)}{p_{11}} - \frac{\mathbb{1}_{(0,1)}(\widehat{a}_i,\widehat{y}_i)}{p_{01}} \right) + \|\Delta_i\|^2 + o_{\mathbb{P}}(1) \Bigg\}.
\end{aligned}$$

In the next step, fix any tuple $(a, y) \in \mathcal{A} \times \mathcal{Y}$, and denote the following constant

$$M_1 = |p_{11}^{-1} \mathbb{1}_{(1,1)}(a, y) - p_{01}^{-1} \mathbb{1}_{(0,1)}(a, y)|.$$

We find

$$\begin{aligned}
&\|[\nabla h_\beta(x + \Delta) - \nabla h_\beta(x)](p_{11}^{-1} \mathbb{1}_{(1,1)}(a, y) - p_{01}^{-1} \mathbb{1}_{(0,1)}(a, y))\|_* \\
=&|h_\beta(x + \Delta) - h_\beta(x) - h_\beta(x + \Delta)^2 + h_\beta(x)^2| \|\beta\|_* M_1 \\
\leq&(|h_\beta(x + \Delta) - h_\beta(x)| + |h_\beta(x + \Delta)^2 - h_\beta(x)^2|) \|\beta\|_* M_1.
\end{aligned}$$

Because the sigmoid function is slope-restricted in the interval $[0, 1]$ [FMP19, Proposition 2], we have

$$0 \leq \frac{h_\beta(x + \Delta) - h_\beta(x)}{\beta^\top \Delta} \leq 1,$$

which implies that

$$|h_\beta(x + \Delta) - h_\beta(x)| \leq |\beta^\top \Delta| \leq \|\beta\|_* \|\Delta\|,$$

where the second inequality follows from Hölder inequality. Using a similar argument, we have

$$|h_\beta(x + \Delta)^2 - h_\beta(x)^2| = \leq (h_\beta(x + \Delta) + h_\beta(x))|h_\beta(x + \Delta) - h_\beta(x)| \leq 2\|\beta\|_* \|\Delta\|.$$

Combining these inequalities, we conclude that

$$\|[\nabla h_\beta(x + \Delta) - \nabla h_\beta(x)](p_{11}^{-1} \mathbb{1}_{(1,1)}(a, y) - p_{01}^{-1} \mathbb{1}_{(0,1)}(a, y))\|_2 \leq 3\|\beta\|_*^2 M_1 \|\Delta\|,$$

and thus Assumption 6' in [BKM19] is satisfied. If $H^N \xrightarrow{d.} \tilde{Z}$ for some random variable $\tilde{Z}$, then [BKM19, Lemma 4] asserts that

$$\begin{aligned}
&N \times \mathcal{R}^{\text{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) \\
\xrightarrow{d.}& \sup_{\gamma \in \mathbb{R}} \left\{ \gamma \tilde{Z} - \frac{\gamma^2}{4} \mathbb{E}_{\mathbb{P}} \left[\left\| \nabla h_\beta(X) \left(\frac{\mathbb{1}_{(1,1)}(A, Y)}{p_{11}} - \frac{\mathbb{1}_{(0,1)}(A, Y)}{p_{01}} \right) \right\|_*^2 \right] \right\} \\
=& \left(\mathbb{E}_{\mathbb{P}} \left[\left\| \nabla h_\beta(X) \left(\frac{\mathbb{1}_{(1,1)}(A, Y)}{p_{11}} - \frac{\mathbb{1}_{(0,1)}(A, Y)}{p_{01}} \right) \right\|_*^2 \right] \right)^{-1} \tilde{Z}^2,
\end{aligned}$$

where the equality sign follows from the fact that for any realization of $\tilde{Z}$, the optimal solution of γ is

$$\gamma^\star(\tilde{Z}) = \frac{2\tilde{Z}}{\mathbb{E}_{\mathbb{P}} \left[\left\| \nabla h_\beta(X) \left(\frac{\mathbb{1}_{(1,1)}(A,Y)}{p_{11}} - \frac{\mathbb{1}_{(0,1)}(A,Y)}{p_{01}} \right) \right\|_*^2 \right]}.$$

We now study the limit distribution $\tilde{Z}$. In the next step, we study the limit of H^N.

$$\begin{aligned}
H^N &= \frac{1}{\sqrt{N}} \sum_{i=1}^{N} h_\beta(\widehat{x}_i) \left(\frac{\mathbb{1}_{(1,1)}\left(\widehat{a}_i, \widehat{y}_i\right)}{\widehat{p}_{11}^N} - \frac{\mathbb{1}_{(0,1)}\left(\widehat{a}_i, \widehat{y}_i\right)}{\hat{p}_{01}^N} \right) \\
&= \frac{1}{\hat{p}_{11}^N \hat{p}_{01}^N} \times \frac{1}{\sqrt{N}} \sum_{i=1}^{N} h_\beta(\widehat{x}_i) \left(\widehat{p}_{01}^N \mathbb{1}_{(1,1)}\left(\widehat{a}_i, \widehat{y}_i\right) - \widehat{p}_{11}^N \mathbb{1}_{(0,1)}\left(\widehat{a}_i, \widehat{y}_i\right) \right) \\
&= \frac{1}{\hat{p}_{11}^N \hat{p}_{01}^N} \times \Big(\frac{1}{\sqrt{N}} \sum_{i=1}^{N} h_\beta(\widehat{x}_i) \big(p_{01} \mathbb{1}_{(1,1)}\left(\widehat{a}_i, \widehat{y}_i\right) - p_{11} \mathbb{1}_{(0,1)}(\widehat{a}_i, \widehat{y}_i) \big) \\
&\qquad + \sqrt{N}(\hat{p}_{01}^N - p_{01}) \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{(1,1)}(\widehat{a}_i, \widehat{y}_i) h_\beta(\widehat{x}_i) \\
&\qquad - \sqrt{N}(\hat{p}_{11}^N - p_{11}) \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{(0,1)}(\widehat{a}_i, \widehat{y}_i) h_\beta(\widehat{x}_i) \Big)
\end{aligned}$$

By Slutsky's theorem, we have

$$\sqrt{N}(\hat{p}_{01}^N - p_{01}) \times \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{1}_{(1,1)}\left(\widehat{a}_i, \widehat{y}_i\right) h_\beta(\widehat{x}_i) - \mathbb{E}_{\mathbb{P}}[\mathbb{1}_{(1,1)}(A, Y) h_\beta(X)] \right) = o_{\mathbb{P}}(1),$$

$$\sqrt{N}(\hat{p}_{11}^N - p_{11}) \times \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{1}_{(0,1)}(\widehat{a}_i, \widehat{y}_i) h_\beta(\widehat{x}_i) - \mathbb{E}_{\mathbb{P}}[1_{(0,1)}\left(A, Y\right) h_\beta(X)] \right) = o_{\mathbb{P}}(1).$$

Under the null hypothesis $\mathcal{H}_0^{\text{opp}}$, we have

$$\begin{aligned}
H^N &= \frac{1}{\hat{p}_{11}^N \hat{p}_{01}^N} \times \Big[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} h_\beta(\widehat{x}_i) \left(p_{01} \mathbb{1}_{(1,1)}\left(\widehat{a}_i, \widehat{y}_i\right) - p_{11} \mathbb{1}_{(0,1)}(\widehat{a}_i, \widehat{y}_i) \right) \\
&\quad + \sqrt{N} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{(0,1)}(\widehat{a}_i, \widehat{y}_i) - p_{01} \right) \mathbb{E}_{\mathbb{P}}[\mathbb{1}_{(1,1)}(A, Y) h_\beta(X)] \\
&\quad - \sqrt{N} \left(\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{(1,1)}(\widehat{a}_i, \widehat{y}_i) - p_{11} \right) \mathbb{E}_{\mathbb{P}}[\mathbb{1}_{(0,1)}(A, Y) h_\beta(X)] \Big] + o_{\mathbb{P}}(1) \\
&= \frac{1}{\hat{p}_{11}^N \hat{p}_{01}^N} \times \Big[\frac{1}{\sqrt{N}} \sum_{i=1}^{N} h_\beta(\widehat{x}_i) \left(p_{01} \mathbb{1}_{(1,1)}\left(\widehat{a}_i, \widehat{y}_i\right) - p_{11} \mathbb{1}_{(0,1)}(\widehat{a}_i, \widehat{y}_i) \right) \\
&\quad + \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\mathbb{1}_{(0,1)}(\widehat{a}_i, \widehat{y}_i) - p_{01} \right) \mathbb{E}_{\mathbb{P}}[\mathbb{1}_{(1,1)}(A, Y) h_\beta(X)] \\
&\quad - \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \left(\mathbb{1}_{(1,1)}(\widehat{a}_i, \widehat{y}_i) - p_{11} \right) \mathbb{E}_{\mathbb{P}}[\mathbb{1}_{(0,1)}(A, Y) h_\beta(X)] \Big] + o_{\mathbb{P}}(1) \\
&\xrightarrow{d.} \tilde{Z},
\end{aligned}$$

where $\tilde{Z} \sim \frac{1}{p_{11}p_{01}}\mathcal{N}(0,\sigma^2)$, $\sigma^2 = \mathrm{Cov}(Z)$, where Z is defined as in the theorem statement. Defining θ completes the proof. □

5.7.3. Proofs of Section 5.5

The proof of Proposition 5.5.1 necessitates the following preparatory lemma. We use the same notations with Lemma 5.7.1.

Lemma 5.7.2 (Slater condition - Probabilistic equalized odds)**.** *Suppose that $\beta \neq 0$ and $\widehat{p}_{ay}^N \in (0,1)$ for all $(a,y) \in \mathcal{A} \times \mathcal{Y}$. Define the functions*

$$\begin{aligned}
f_\beta(X,A,Y) &\triangleq \frac{1}{\widehat{p}_{11}^N} h_\beta(X) \mathbb{1}_{(1,1)}(A,Y) - \frac{1}{\widehat{p}_{01}^N} h_\beta(X) \mathbb{1}_{(0,1)}(A,Y),\\
g_\beta(X,A,Y) &\triangleq \frac{1}{\widehat{p}_{10}^N} h_\beta(X) \mathbb{1}_{(1,0)}(A,Y) - \frac{1}{\widehat{p}_{00}^N} h_\beta(X) \mathbb{1}_{(0,0)}(A,Y),
\end{aligned}$$

and let f be a vector-valued function $f : \Xi \times \widehat{\Xi}_N \to \mathbb{R}^{N+2}$

$$f(\xi,\xi') = \begin{pmatrix} \mathbb{1}_{\hat{\xi}_i}(\xi') \\ \vdots \\ \mathbb{1}_{\hat{\xi}_N}(\xi') \\ f_\beta(\xi) \\ g_\beta(\xi) \end{pmatrix}$$

Then we have

$$\begin{pmatrix} 1/N \\ \vdots \\ 1/N \\ 0 \\ 0 \end{pmatrix} \in \operatorname{int}\left\{ \mathbb{E}_\pi[f(\xi,\xi')] : \pi \in \mathcal{M}_+(\Xi \times \widehat{\Xi}_N) \right\}.$$

Proof of Lemma 5.7.2. It suffices to show that for any

$$q \in \left(\frac{1}{2N}, \frac{3}{2N}\right)^N \times \left(-\frac{1}{4}, \frac{1}{4}\right)^2,$$

there exists a nonnegative measure $\pi \in \mathcal{M}_+(\Xi \times \widehat{\Xi}_N)$ such that $q = \mathbb{E}_\pi[f(\xi,\xi')]$. The proof follows a similar argument as that of Lemma 5.7.1 by noticing that

$$\mathbb{E}_\pi[g_\beta(\xi)] = (\widehat{p}_{10}^N)^{-1} h_\beta(x_{10}) \sum_{i \in \mathcal{I}_{10}} q_i - (\widehat{p}_{00}^N)^{-1} h_\beta(x_{00}) \sum_{i \in \mathcal{I}_{00}} q_i,$$

and the specification of x_{10} and x_{00} can be achieved using similar steps. □

Proof of Proposition 5.5.1. To ease the exposition, we let the function $\Lambda : \mathcal{A} \times \mathcal{Y} \to \mathbb{R}^2$ be defined as

$$\Lambda(a, y) = \begin{pmatrix} (\widehat{p}_{11}^N)^{-1}\mathbb{1}_{(1,1)}(a, y) - (\widehat{p}_{01}^N)^{-1}\mathbb{1}_{(0,1)}(a, y) \\ (\widehat{p}_{10}^N)^{-1}\mathbb{1}_{(1,0)}(a, y) - (\widehat{p}_{00}^N)^{-1}\mathbb{1}_{(0,0)}(a, y) \end{pmatrix}.$$

Moreover, we define f_β as in (5.11), and additionally define g_β as

$$g_\beta(x, a, y) = (\hat{p}_{10}^N)^{-1} h_\beta(x)\mathbb{1}_{(1,0)}(a, y) - (\hat{p}_{00}^N)^{-1} h_\beta(x)\mathbb{1}_{(0,0)}(a, y).$$

From the definition of $\mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$, we have

$$\begin{aligned}
\mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) &= \left\{ \begin{array}{ll}
\inf\limits_{\mathbb{Q}\in\mathcal{P}} & \mathbb{W}(\widehat{\mathbb{P}}^N, \mathbb{Q})^2 \\
\text{s.t.} & (\hat{p}_{11}^N)^{-1}\mathbb{E}_{\mathbb{Q}}[h_\beta(X)\mathbb{1}_{(1,1)}(A, Y)] = \\
& (\hat{p}_{01}^N)^{-1}\mathbb{E}_{\mathbb{Q}}[h_\beta(X)\mathbb{1}_{(0,1)}(A, Y)] \\
& (\hat{p}_{10}^N)^{-1}\mathbb{E}_{\mathbb{Q}}[h_\beta(X)\mathbb{1}_{(1,0)}(A, Y)] = \\
& (\hat{p}_{00}^N)^{-1}\mathbb{E}_{\mathbb{Q}}[h_\beta(X)\mathbb{1}_{(0,0)}(A, Y)] \\
& \mathbb{Q}(A = a, Y = y) = \hat{p}_{ay}^N \quad \forall a \in \mathcal{A},\ y \in \mathcal{Y}
\end{array} \right. \\
&= \left\{ \begin{array}{lll}
\inf\limits_{\pi} & \mathbb{E}_\pi[c\big((X', A', Y'), (X, A, Y)\big)^2] & \\
\text{s.t.} & \pi \in \mathcal{P}((\mathcal{X} \times \mathcal{A} \times \mathcal{Y}) \times (\mathcal{X} \times \mathcal{A} \times \mathcal{Y})) & \\
& \mathbb{E}_\pi[f_\beta(X, A, Y)] = 0 & \\
& \mathbb{E}_\pi[g_\beta(X, A, Y)] = 0 & \\
& \pi(A = a, Y = y) = \hat{p}_{ay}^N & \forall a \in \mathcal{A},\ y \in \mathcal{Y} \\
& \mathbb{E}_\pi[\mathbb{1}_{(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)}(X', A', Y')] = 1/N & \forall i \in [N].
\end{array} \right.
\end{aligned}$$

To shorten the notations, we use $\Xi = \mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ and $\widehat{\Xi}_N = \{(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)\}$. Moreover, define the vector $\bar{q}$ and the vector-valued Borel measurable function on $\Xi \times \widehat{\Xi}_N$ as

$$\bar{q} = \begin{pmatrix} 0 \\ 0 \\ 1/N \\ \vdots \\ 1/N \end{pmatrix} \qquad f(\xi, \xi') = \begin{pmatrix} f_\beta(\xi) \\ g_\beta(\xi) \\ \mathbb{1}_{\hat{\xi}_i}(\xi') \\ \vdots \\ \mathbb{1}_{\hat{\xi}_N}(\xi') \end{pmatrix}.$$

By using the introduced notation, we can reformulate the above optimization problem as

$$\inf \left\{ \mathbb{E}_\pi[c(\xi, \xi')^2] : \pi \in \mathcal{M}_+(\Xi \times \widehat{\Xi}_N), \mathbb{E}_\pi[f(\xi, \xi')] = \bar{q} \right\}$$

which is a problem of moments. By Lemma 5.7.2, the above optimization problem satisfies the Slater condition, thus the strong duality result [Smi95, Section 2.2] implies that

$$
\begin{aligned}
\mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) &= \left\{ \begin{array}{ll}
\sup & \displaystyle \frac{1}{N}\sum_{i=1}^N b_i \\
\text{s.t.} & b \in \mathbb{R}^N,\ \kappa \in \mathbb{R},\ \zeta \in \mathbb{R} \\
& \displaystyle \sum_{i=1}^N b_i \mathbb{1}_{(\widehat{x}_i, \widehat{a}_i, \widehat{y}_i)}(x', a', y') - \kappa f_\beta(x, a, y) - \\
& \zeta g_\beta(x, a, y) \le c\big((x', a', y'), (x, a, y)\big)^2 \\
& \qquad\qquad \forall (x, a, y), (x', a', y') \in \mathcal{X} \times \mathcal{A} \times \mathcal{Y}
\end{array}\right. \\
&= \left\{ \begin{array}{ll}
\sup & \displaystyle \frac{1}{N}\sum_{i=1}^N b_i \\
\text{s.t.} & b \in \mathbb{R}^N,\ \kappa \in \mathbb{R},\ \zeta \in \mathbb{R} \\
& b_i - \kappa f_\beta(x_i, a_i, y_i) - \\
& \zeta g_\beta(x_i, a_i, y_i) \le c\big((\widehat{x}_i, \widehat{a}_i, \widehat{y}_i), (x_i, a_i, y_i)\big)^2 \\
& \qquad\quad \forall (x_i, a_i, y_i) \in \mathcal{X} \times \mathcal{A} \times \mathcal{Y}, \forall i \in [N]
\end{array}\right. \\
&= \sup_{\kappa, \zeta} \ \frac{1}{N}\sum_{i=1}^N \inf_{x_i \in \mathcal{X}} \left\{ \|x_i - \widehat{x}_i\|^2 + \gamma f_\beta(x_i, \hat{a}_i, \hat{y}_i) + \zeta g_\beta(x_i, \hat{a}_i, \hat{y}_i) \right\},
\end{aligned}
$$

By definition of f_β, g_β and the parameters λ_i, we have

$$
\gamma f_\beta(x_i, \hat{a}_i, \hat{y}_i) + \zeta g_\beta(x_i, \hat{a}_i, \hat{y}_i) = (\kappa \lambda_i \mathbb{1}_1(\widehat{y}_i) + \zeta \lambda_i \mathbb{1}_0(\widehat{y}_i)) h_\beta(x_i).
$$

The proof is complete. □

Proof of Lemma 5.5.2. Because $[N] = \mathcal{I}_0 \cup \mathcal{I}_1$, we can write

$$
\begin{aligned}
\mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) = \sup_{\kappa \in \mathbb{R}} \ \frac{1}{N} \sum_{i \in \mathcal{I}_1} \inf_{x_i \in \mathcal{X}} \left\{ \|x_i - \widehat{x}_i\|^2 + \kappa \lambda_i h_\beta(x_i) \right\} + \\
\sup_{\zeta \in \mathbb{R}} \ \frac{1}{N} \sum_{i \in \mathcal{I}_0} \inf_{x_i \in \mathcal{X}} \left\{ \|x_i - \widehat{x}_i\|^2 + \zeta \lambda_i h_\beta(x_i) \right\}.
\end{aligned}
$$

Note that the first supremum coincides with $\mathcal{R}^{\text{opp}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$, and the second supremum is U_N. Under the Euclidean norm assumption, we can use Lemma 5.8.1 to reformulate the inner infimum problems for U_N, which leads to (5.9). □

Proof of Theorem 5.5.3. By applying a similar duality argument as in the proof of Theorem 5.4.3, we can reformulate $\mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N)$ as

$$
\mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) =
$$

$$
\begin{aligned}
&\sup_{\gamma,\zeta} \mathbb{E}_{\widehat{\mathbb{P}}^N}\left[\inf_{\Delta}\left\{\begin{array}{l}\gamma h_\beta(X+\Delta)\big(\frac{\mathbb{1}_{(1,1)}(A,Y)}{\widehat{p}_{11}^N} - \frac{\mathbb{1}_{(0,1)}(A,Y)}{\widehat{p}_{01}^N}\big)\\ +\zeta h_\beta(X+\Delta)\big(\frac{\mathbb{1}_{(1,0)}(A,Y)}{\widehat{p}_{10}^N} - \frac{\mathbb{1}_{(0,0)}(A,Y)}{\widehat{p}_{00}^N}\big) + \|\Delta\|^2\end{array}\right\}\right]\\
&= \sup_{\gamma,\zeta}\left\{\frac{1}{\sqrt{N}}(\zeta H_0^N + \gamma H_1^N) + \right.\\
&\left.\mathbb{E}_{\widehat{\mathbb{P}}^N}\left[\inf_{\Delta}\left(\begin{array}{l}\gamma[h_\beta(X+\Delta) - h_\beta(X)]\big(\frac{\mathbb{1}_{(1,1)}(A,Y)}{\widehat{p}_{11}^N} - \frac{\mathbb{1}_{(0,1)}(A,Y)}{\widehat{p}_{01}^N}\big)\\ +\zeta[h_\beta(X+\Delta) - h_\beta(X)]\big(\frac{\mathbb{1}_{(1,0)}(A,Y)}{\widehat{p}_{10}^N} - \frac{\mathbb{1}_{(0,0)}(A,Y)}{\widehat{p}_{00}^N}\big)\\ +\|\Delta\|^2\end{array}\right)\right]\right\}
\end{aligned}
$$

with the random variables H_0^N and H_1^N being defined as

$$
\begin{aligned}
H_0^N &\triangleq \frac{1}{\sqrt{N}}\sum_{i=1}^N h_\beta(\widehat{x}_i)\big(\frac{\mathbb{1}_{(1,0)}(\widehat{a}_i,\widehat{y}_i)}{\widehat{p}_{10}^N} - \frac{\mathbb{1}_{(0,0)}(\widehat{a}_i,\widehat{y}_i)}{\widehat{p}_{00}^N}\big),\\
H_1^N &\triangleq \frac{1}{\sqrt{N}}\sum_{i=1}^N h_\beta(\widehat{x}_i)\big(\frac{\mathbb{1}_{(1,1)}(\widehat{a}_i,\widehat{y}_i)}{\widehat{p}_{11}^N} - \frac{\mathbb{1}_{(0,1)}(\widehat{a}_i,\widehat{y}_i)}{\widehat{p}_{01}^N}\big).
\end{aligned}
$$

Notice that the condition

$$
\mathbb{P}\left(\left\|\begin{pmatrix}\gamma_1\\ \gamma_0\end{pmatrix}^\top \Lambda(A,Y)\nabla h_\beta(X)\right\|_* > 0\right) > 0
$$

is satisfied for any $(\gamma_0,\gamma_1) \neq 0$. Using the same argument as in the proof of [BKM19, Theorem 3], we can show that the optimal solution for γ, ζ and Δ_i belong to a compact set with high probability. As $\widehat{p}_{ay} - p_{ay} = o_{\mathbb{P}}(1)$ for any $(a,y) \in \mathcal{A}\times\mathcal{Y}$, we have

$$
\begin{aligned}
N \times \mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) = \sup_{\gamma,\zeta}&\left\{\gamma H_1^N + \right.\\
&\zeta H_0^N + \frac{1}{N}\sum_{i=1}^N \inf_{\Delta_i}\gamma\int_0^1 \nabla h_\beta\left(\widehat{x}_i + t\frac{\Delta_i}{\sqrt{N}}\right)\cdot\Delta_i \mathrm{d}t\ \times\\
&\left.\begin{pmatrix}\gamma\\ \zeta\end{pmatrix}^\top\begin{pmatrix}p_{11}^{-1}\mathbb{1}_{(1,1)}(\widehat{a}_i,\widehat{y}_i) - p_{01}^{-1}\mathbb{1}_{(0,1)}(\widehat{a}_i,\widehat{y}_i)\\ p_{10}^{-1}\mathbb{1}_{(1,0)}(\widehat{a}_i,\widehat{y}_i) - p_{00}^{-1}\mathbb{1}_{(0,0)}(\widehat{a}_i,\widehat{y}_i)\end{pmatrix} + \|\Delta_i\|^2 + o_{\mathbb{P}}(1)\right\}.
\end{aligned}
$$

Using a similar argument, we can bound

$$
\|[\nabla h_\beta(x+\Delta) - \nabla h_\beta(x)](p_{10}^{-1}\mathbb{1}_{(1,0)}(a,y) - p_{00}^{-1}\mathbb{1}_{(0,0)}(a,y))\|_2 \leq 3\|\beta\|_*^2 M_0\|\Delta\|
$$

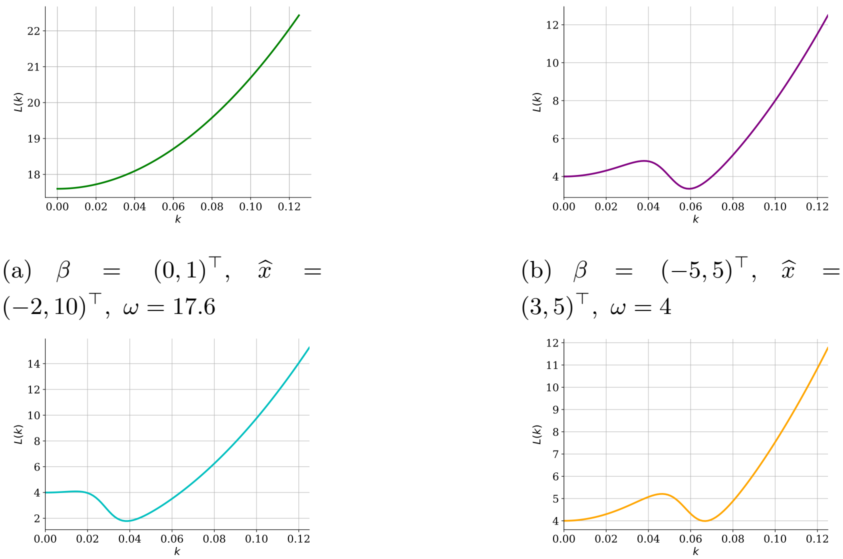
for some constant M_0, and thus Assumption 6' in [BKM19] is satisfied. If $H_0^N \xrightarrow{d.} H_0$ and $H_1^N \xrightarrow{d.} H_1$ for some random variables H_0 and H_1, then [BKM19,

Lemma 4] asserts that

$$
\begin{aligned}
&N \times \mathcal{R}^{\text{odd}}(\widehat{\mathbb{P}}^N, \widehat{p}^N) \xrightarrow{d.} \\
&\sup_{\gamma,\zeta} \Bigg\{ \gamma H_1 + \zeta H_0 + \\
&\mathbb{E}_{\mathbb{P}} \left[\left\| \begin{pmatrix} \gamma \\ \zeta \end{pmatrix}^\top \begin{pmatrix} p_{11}^{-1} \mathbb{1}_{(1,1)}(A,Y) - p_{01}^{-1} \mathbb{1}_{(0,1)}(A,Y) \\ p_{10}^{-1} \mathbb{1}_{(1,0)}(A,Y) - p_{00}^{-1} \mathbb{1}_{(0,0)}(A,Y) \end{pmatrix} \nabla h_\beta(X) \right\|_*^2 \right] \Bigg\}.
\end{aligned}
$$

Using the same limiting argument as in the proof of Theorem 5.4.3, we have the characterization of H_1 and H_0 as in the statement of the theorem. □

5.8. Appendix - Auxiliary Result



(a) $\beta = (0,1)^\top$, $\widehat{x} = (-2,10)^\top$, $\omega = 17.6$

(b) $\beta = (-5,5)^\top$, $\widehat{x} = (3,5)^\top$, $\omega = 4$

(c) $\beta = (-6,5)^\top$, $\widehat{x} = (3,5)^\top$, $\omega = 4$

(d) $\beta = (-4.7,5)^\top$, $\widehat{x} = (3,5)^\top$, $\omega = 4$

Figure 5.5: Plots of $L(k)$ with respect to k for different values of $\beta, \widehat{x}$ and ω.

The following lemma is used repeatedly to prove Lemmas 5.4.2 and 5.5.2.

Lemma 5.8.1. *For any $\omega \in \mathbb{R}$, $\widehat{x} \in \mathbb{R}^p$ and $\beta \in \mathbb{R}^p$, we have*

$$
\inf_{x \in \mathbb{R}^p} \|x - \widehat{x}\|_2^2 + \frac{\omega}{1 + \exp(-\beta^\top x)} = \min_{k \in [0, \frac{1}{8}]} \omega^2 \|\beta\|_2^2 k^2 + \frac{\omega}{1 + \exp(-\beta^\top \widehat{x} + k\omega \|\beta\|_2^2)}. \tag{5.15}
$$

Proof of Lemma 5.8.1. Any $x \in \mathbb{R}^p$ can be written using the orthogonal decomposition as $x = \widehat{x} - k\omega\beta - k'\beta^\perp$ for some $k \in \mathbb{R}$, $k' \in \mathbb{R}$ and $\beta^\perp$ perpendicular to β, that is, $\beta^\top(\beta^\perp) = 0$. Optimizing over x is equivalent to jointly optimizing over k, k' and $\beta^\perp$ as

$$\begin{array}{ll} \inf & \|k\omega\beta + k'\beta^\perp\|_2^2 + \dfrac{\omega}{1 + \exp(-\beta^\top \widehat{x} + k\omega\|\beta\|_2^2)} \\ \text{s.t.} & k \in \mathbb{R},\ k' \in \mathbb{R},\ \beta^\perp \in \mathbb{R}^p,\ \beta^\top(\beta^\perp) = 0. \end{array}$$

After extending the norm, and by noticing that the optimal solution in k' and $\beta^\perp$ should satisfy $k'\beta^\perp = 0$, the above optimization problem is equivalent to

$$\begin{array}{ll} \inf & k^2\omega^2\|\beta\|_2^2 + \dfrac{\omega}{1 + \exp(-\beta^\top \widehat{x} + k\omega\|\beta\|_2^2)} \\ \text{s.t.} & k \in \mathbb{R}. \end{array}$$

Let $L(k)$ be the objective function of the above optimization problem, we have

$$\begin{aligned} \nabla_k L(k) &= 2\omega^2\|\beta\|_2^2 k - \frac{\omega^2\|\beta\|_2^2 \exp(-\beta^\top \hat{x} + k\omega\|\beta\|_2^2)}{(1 + \exp(-\beta^\top \hat{x} + k\omega\|\beta\|_2^2))^2} \\ &= \omega^2\|\beta\|_2^2 \left(2k - \sigma(k)(1 - \sigma(k))\right), \end{aligned}$$

where for the purpose of this proof, we define $\sigma(k)$ as

$$\sigma(k) \triangleq \frac{1}{1 + \exp(-\beta^\top \hat{x} + k\omega\|\beta\|_2^2)} \in (0, 1).$$

Notice that $\sigma(k)(1 - \sigma(k)) \in (0, \frac{1}{4})$ for any value of $k \in \mathbb{R}$. Because $\nabla_k L(k)$ is continuous in k, $\nabla_k L(k) \le 0$ for any $k \le 0$, and $\nabla_k L(k) \ge 0$ for any $k \ge \frac{1}{8}$, one can conclude that there exists an optimal solution $k^\star$ that lies in the compact range $[0, \frac{1}{8}]$. This completes the proof. □

Let $L(k)$ be the objective function of the optimization problem (5.15). Figure 5.5 visualizes several instances of $L(k)$ for different values of inputs $\beta, \widehat{x}$ and ω. Note that $L(k)$ is non-convex in k, and the optimizer of $L(k)$ is not necessarily unique as indicated in Figure 5.5d.

5.9. Appendix - Numerical Results

We use the synthetic experiment from [Zaf+17b] to generate unfairness landscapes provided in Figure 5.1. We set the true distributions of the class labels $\mathbb{P}(Y = 0) = \mathbb{P}(Y = 1) = 1/2$, and conditioning on Y, the feature X is distributed as

$$X|Y = 1 \sim \mathcal{N}([2; 2], [5, 1; 1, 5]),$$

$$X|Y = 0 \sim \mathcal{N}([-2; -2], [10, 1; 1, 3]).$$

Then, we draw sensitive attribute of each sample x from a Bernoulli distribution, that is

$$\mathbb{P}(A = 1|X = x') = \mathit{pdf}(x'|Y = 1)/(\mathit{pdf}(x'|Y = 1) + \mathit{pdf}(x'|Y = 0)),$$

where $x' = [\cos(\pi/4), \sin(\pi/4); \sin(\pi/4), \cos(\pi/4)]x$ is a rotated version of the feature vector x and $\mathit{pdf}(\cdot|Y = y)$ is the Gaussian probability density function of X given $Y = y$.

Part III

Dynamic Decision Making

6. Robust Control

> Life can only be understood going backwards, but it must be lived going forwards.
>
> *Søren Kierkegard*

6.1. Introduction

The Linear Quadratic Regulator (LQR) is a classic control problem that has served as a building block for numerous applications in engineering and computer science [Aug+13; Che12], economics [HS05], or neuroscience [TJ02]. It involves controlling a system with linear dynamics and imperfect observations affected by additive noise, with the goal of minimizing a quadratic state and control cost. Under the assumption that noise terms are independent and normally distributed (a case referred to as Linear-Quadratic-Gaussian, or LQG), it is well known that the optimal control policy depends linearly on the observations and can be obtained efficiently by using the Kalman filtering procedure and dynamic programming [Ber17].

Motivated by practical settings where noise distributions may not be readily available or may not be Gaussian, this paper considers a discrete-time, finite-horizon generalization of the LQG setting where noise distributions are unknown and are chosen adversarially from ambiguity sets characterized by a Wasserstein distance and centered around nominal (Gaussian) distributions.

We show that, even under distributional ambiguity, the optimal control policy remains linear in the system's observations. Our proof is novel and does not rely on traditional recursive dynamic programming arguments. Instead, we re-parametrize the control policy in terms of the purified state observations and we derive an upper bound for the resulting minimax formulation by relaxing the ambiguity set (from a Wasserstein ball into a Gelbrich ball) while simultaneously restricting the controller to linear dependencies. We then use convex duality to prove that this upper bound matches a lower bound obtained by restricting the ambiguity set in the dual of the minimax formulation. This implies the optimality of linear output feedback controllers, thus generalizing the classic results to a distributionally robust setting.

We also find that the worst-case distribution is actually Gaussian, which leads to a very efficient algorithm for finding optimal controllers. Specifically, we propose an algorithm based on the Frank-Wolfe first-order method that at every step solves sub-problems corresponding to classic LQG control problems, using Kalman filtering and dynamic programming. We show that this algorithm enjoys a sublinear convergence rate and is susceptible to parallelization. Lastly, we implement the algorithm leveraging PyTorch's automatic differentiation module and we find that it yields uniformly lower runtimes than a direct method (based on solving semidefinite programs) across all problem horizons.

6.1.1. Literature Review

This paper is related to the ample literature in control theory and engineering aimed at designing controllers that are robust to noise. The classic LQR/LQG theory, developed in the 1960s, examined linear dynamical systems in either time or frequency domain, seeking to minimize a combination of quadratic state and control costs (in time-domain) or the $\mathcal{H}_2$ norm of the system's transfer function (in frequency domain). Motivated by findings that LQG controllers do not provide the guaranteed robust stability properties of LQR controllers [Doy78], much effort has been devoted subsequently to designing controllers that are robust to worst-case perturbations, typically evaluated in terms of the $\mathcal{H}_\infty$ norm of the system's transfer function (see, e.g., [Doy+88; ZD98] for a comprehensive review of $\mathcal{H}_\infty$ and $\mathcal{H}_2$ controllers). Because $\mathcal{H}_\infty$ controllers tend to be overly conservative [KIL22], various approaches have been proposed to balance the performance of nominal and robust controllers, e.g., by combining $\mathcal{H}_2$ and $\mathcal{H}_\infty$ approaches [BH88; DZB89]. A parallel stream of literature has considered risk-sensitive control [Whi81], which minimizes an entropic risk measure instead of the expected quadratic cost. Although risk-sensitive control has a distributionally robust flavor (as the entropic risk measure is equivalent to a distributionally robust quadratic objective penalized via Kullback-Leibler divergence), risk-sensitive control models do not admit a distributionally robust formulation because the entropic risk measure is convex, but not coherent [FS11]. In contrast, our distributionally robust model provides a direct interpretation of the exact set of noise distributions against which the controller provides safeguards, and leads to a computationally tractable framework for finding the optimal controller.

In this sense, our work is more directly related to the literature on distributionally robust control, which seeks controllers that minimize expected costs under worst-case noise distributions [BIP11; KY23; KLN21; PJD00; VP+16; Yan21]. Closest to our work are [Han23; KY23]. [KY23] proves the optimality of linear state-feedback control policies for a related minimax LQR model with a Wasserstein distance but with *perfect* state observations. With perfect observations, the optimal policies in the classic LQR formulation are independent of the noise distribution and are thus inherently already robust, so considering imperfect observations is what makes the problem significantly more challenging in our case. [Han23] studies a minimax formulation based on the Wasserstein distance with both state and observation noise but without any control policy, and focuses solely on the problem of estimating the states. Several papers have considered robust formulations with imperfect observations but for constrained systems [BTBN05; BTBN06; KLN21], which are more challenging; the common approach is to restrict attention to linear feedback policies for computational tractability, and without proving their optimality.

Also related is the recent literature stream on distributionally robust optimization using the Wasserstein distance [MEK18]. Within this stream, the closest work is [Ngu+23; SA+18], which consider the problem of minimax mean-squared-error estimation when ambiguity is modeled with a Wasserstein distance from a nominal Gaussian distribution. Our proof builds on some ideas from these papers (e.g., relying on the Gelbrich distance in the construction of the upper bound), which it combines with ideas from control theory on purified output-feedback to obtain the overall construction. Also related is [AG22], which studies multistage distributionally robust problems with ambiguity sets given by a nested Wasserstein distance for stochastic processes and identifies computationally tractable cases. For a broader overview of developments related to optimal transport and Wasserstein distance with an emphasis on computational tractability and applications in machine learning, we refer to [PC19b].

Finally, our paper is also related to literature that documents the optimality of linear/affine policies in (distributionally) robust dynamic optimization models. [BIP10; ISS13] prove optimality for one-dimensional linear systems affected by additive noise and with perfect state observations, but with general (convex) state and/or control costs, [HGK11; VPGM13] provide computationally tractable approaches to quantifying the suboptimality of affine controllers in finite or infinite-horizon settings, and [BG12; EHG21; GTW21] characterize the performance of affine policies in two-stage (distributionally) robust dynamic models.

Notation. All random objects are defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Thus, the distribution of any random vector $\xi : \Omega \to \mathbb{R}^d$ is given by the pushforward distribution $\mathbb{P}_\xi = \mathbb{P} \circ \xi^{-1}$ of $\mathbb{P}$ with respect to ξ. The expectation under $\mathbb{P}$ is denoted by $\mathbb{E}_\mathbb{P}[\cdot]$. For any $t \in \mathbb{Z}_+$, we set $[t] = \{0, \dots, t\}$.

6.2. Problem Definition

We consider a discrete-time linear dynamical system

$$x_{t+1} = A_t x_t + B_t u_t + w_t \quad \forall t \in [T-1] \tag{6.1}$$

with states $x_t \in \mathbb{R}^n$, control inputs $u_t \in \mathbb{R}^m$, process noise $w_t \in \mathbb{R}^n$ and system matrices $A_t \in \mathbb{R}^{n \times n}$ and $B_t \in \mathbb{R}^{n \times m}$. The controller only has access to imperfect state measurements

$$y_t = C_t x_t + v_t \quad \forall t \in [T-1] \tag{6.2}$$

corrupted by observation noise $v_t \in \mathbb{R}^p$, where $C_t \in \mathbb{R}^{p \times n}$ and usually $p \leq n$ (so that observing y_t would not allow reconstructing x_t even if there were no observation noise). The control inputs u_t are causal, i.e., depend on the past observations $y_0, \dots, y_t$ but not on the future observations $y_{t+1}, \dots, y_{T-1}$. More precisely, the set of feasible control inputs $\mathcal{U}_y$ is the set of random vectors $(u_0, u_1, \dots, u_{T-1})$ where for every t there exists a measurable control policy $\varphi_t : \mathbb{R}^{p(t+1)} \to \mathbb{R}^m$ such that $u_t = \varphi_t(y_0, \dots, y_t)$. Controlling the system generates costs that depend quadratically on the states and the controls:

$$J = \sum_{t=0}^{T-1} (x_t^\top Q_t x_t + u_t^\top R_t u_t) + x_T^\top Q_T x_T, \tag{6.3}$$

where $Q_t \in \mathbb{S}^n_+$ and $R_t \in \mathbb{S}^m_{++}$ represent the state and input cost matrices, respectively. The exogenous random vectors x_0, $\{w_t\}_{t=0}^{T-1}$ and $\{v_t\}_{t=0}^{T-1}$ are mutually independent and follow probability distributions given by $\mathbb{P}_{x_0}, \{\mathbb{P}_{w_t}\}_{t=0}^{T-1}$, and $\{\mathbb{P}_{v_t}\}_{t=0}^{T-1}$, respectively. As the control inputs are causal, the system equations (6.2) imply that x_t, u_t and y_t can be expressed as measurable functions of the exogenous uncertainties x_0 as well as w_s and v_s, $s \in [t]$, for every t. From now on we may thus assume without loss of generality that $\Omega = \mathbb{R}^n \times \mathbb{R}^{n\times T} \times \mathbb{R}^{p\times T}$ is the space of realizations of the exogenous uncertainties, $\mathcal{F}$ is the Borel σ-algebra on Ω and $\mathbb{P} = \mathbb{P}_{x_0} \otimes (\otimes_{t=0}^{T-1}\mathbb{P}_{w_t}) \otimes (\otimes_{t=0}^{T}\mathbb{P}_{v_t})$, where $\mathbb{P}_1 \otimes \mathbb{P}_2$ denotes the independent coupling of the distributions $\mathbb{P}_1$ and $\mathbb{P}_2$.

In this context, the classic LQG model assumes that $\mathbb{P}$ is known and Gaussian, and seeks $u \in \mathcal{U}_y$ that minimizes $\mathbb{E}_{\mathbb{P}}[J]$. Appendix §6.6 reviews the standard approach for computing optimal control inputs by estimating states through Kalman filtering techniques and using dynamic programming.

In contrast, we assume that $\mathbb{P}$ is only known to belong to an ambiguity set $\mathcal{W}$, and we formulate a distributionally robust LQG problem that seeks $u \in \mathcal{U}_y$ to minimize the worst-case expected cost:

$$\max_{\mathbb{P}\in\mathcal{W}} \mathbb{E}_{\mathbb{P}}\left[\sum_{t=0}^{T-1}(x_t^\top Q_t x_t + u_t^\top R_t u_t) + x_T^\top Q_T x_T\right]. \tag{6.4}$$

We construct the ambiguity set $\mathcal{W}$ as a ball based on the Wasserstein distance. Specifically, we assume that a *nominal* Gaussian distribution $\hat{\mathbb{P}} = \hat{\mathbb{P}}_{x_0} \otimes (\otimes_{t=0}^{T-1}\hat{\mathbb{P}}_{w_t}) \otimes (\otimes_{t=0}^{T}\hat{\mathbb{P}}_{v_t})$ is available so that $\hat{\mathbb{P}}_{x_0} = \mathcal{N}(0, \hat{X}_0)$, $\hat{\mathbb{P}}_{w_t} = \mathcal{N}(0, \hat{W}_t)$, and $\hat{\mathbb{P}}_{v_t} = \mathcal{N}(0, \hat{V}_t)$ for all $t \in [T-1]$, and $\mathcal{W}$ is given by:

$$\mathcal{W} = \mathcal{W}_{x_0} \otimes (\otimes_{t=0}^{T-1}\mathcal{W}_{w_t}) \otimes (\otimes_{t=0}^{T-1}\mathcal{W}_{v_t}),$$

where

$$\begin{aligned}
\mathcal{W}_{x_0} &= \{\mathbb{P}_{x_0} \in \mathcal{P}(\mathbb{R}^n) : \mathbb{W}(\hat{\mathbb{P}}_{x_0}, \mathbb{P}_{x_0}) \le \rho_{x_0},\ \mathbb{E}_{\mathbb{P}_{x_0}}[x_0] = 0\} \\
\mathcal{W}_{w_t} &= \{\mathbb{P}_{w_t} \in \mathcal{P}(\mathbb{R}^n) : \mathbb{W}(\hat{\mathbb{P}}_{w_t}, \mathbb{P}_{w_t}) \le \rho_{w_t},\ \mathbb{E}_{\mathbb{P}_{w_t}}[w_t] = 0\} \\
\mathcal{W}_{v_t} &= \{\mathbb{P}_{v_t} \in \mathcal{P}(\mathbb{R}^m) : \mathbb{W}(\hat{\mathbb{P}}_{v_t}, \mathbb{P}_{v_t}) \le \rho_{v_t},\ \mathbb{E}_{\mathbb{P}_{v_t}}[v_t] = 0\},
\end{aligned}$$

and $\mathbb{W}$ is the 2-Wasserstein distance. Thus, by construction, all exogenous random variables $x_0, w_0, \ldots, w_{T-1}, v_0, \ldots, v_{T-1}$ are independent under every distribution in $\mathcal{W}$.

Definition 11 (2-Wasserstein distance)**.** *The 2-Wasserstein distance between two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ on $\mathbb{R}^d$ with finite second moments is given by*

$$\mathbb{W}(\mathbb{P}_1, \mathbb{P}_2) = \left(\inf_{\pi\in\Pi(\mathbb{P}_1,\mathbb{P}_2)} \int_{\mathbb{R}^d\times\mathbb{R}^d} \|\xi_1 - \xi_2\|_2^2\, \pi(\mathrm{d}\xi_1, \mathrm{d}\xi_2)\right)^{\frac{1}{2}},$$

where $\Pi(\mathbb{P}_1, \mathbb{P}_2)$ denotes the set of all couplings, that is, all joint distributions of the random variables ξ_1 and ξ_2 with marginal distributions $\mathbb{P}_1$ and $\mathbb{P}_2$, respectively.

Our model strictly generalizes the classic LQG setting,[1] which can be recovered by choosing $\rho_{x_0} = \rho_{w_t} = \rho_{v_t} = 0$. The parameters ρ thus allow quantifying the uncertainty about the nominal model and building robustness to mis-specification. We emphasize that the Wasserstein ambiguity set $\mathcal{W}$ contains many non-Gaussian distributions and it is not readily obvious that the worst-case distribution in (6.4) is in fact Gaussian. However, the set $\mathcal{W}$ is also non-convex, as it contains only distributions under which the exogenous uncertainties are independent, which makes the distributionally robust LQG problem potentially difficult to solve.

6.3. Nash Equilibrium and Optimality of Linear Output Feedback Controllers

We henceforth view the distributionally robust LQG problem as a zero-sum game between the controller, who chooses causal control inputs, and nature, who chooses a distribution $\mathbb{P} \in \mathcal{W}$. In this section we show that this game admits a Nash equilibrium, where nature's Nash strategy is a *Gaussian* distribution $\mathbb{P}^\star \in \mathcal{W}$ and the controller's Nash strategy is a *linear* output feedback policy based on the Kalman filter evaluated under $\mathbb{P}^\star$.

Purified Observations. Before outlining our proof strategy, we first simplify the problem formulation by re-parametrizing the control inputs in a more convenient form (following [BTBN05; BTBN06; HGK11]). Note that the control inputs in the LQG formulation are subject to cyclic dependencies, as u_t depends on y_t, while y_t depends on x_t through (6.2), and x_t depends again on u_t through (6.1), etc. Because these dependencies make the problem hard to analyze, it is preferable to instead consider the controls as functions of a new set of so-called *purified* observations instead of the actual observations y_t.

Specifically, we first introduce a fictitious *noise-free* system

$$\hat{x}_{t+1} = A_t \hat{x}_t + B_t u_t \quad \forall t \in [T-1] \quad \text{and} \quad \hat{y}_t = C_t \hat{x}_t \quad \forall t \in [T-1]$$

with states $\hat{x}_t \in \mathbb{R}^n$ and outputs $\hat{y}_t \in \mathbb{R}^p$, which is initialized by $\hat{x}_0 = 0$ and controlled by the *same* inputs u_t as the original system (6.2). We then define the purified observation at time t as $\eta_t = y_t - \hat{y}_t$ and we use $\eta = (\eta_0, \ldots, \eta_{T-1})$ to denote the trajectory of *all* purified observations.

As the inputs u_t are causal, the controller can compute the fictitious state $\hat{x}_t$ and output $\hat{y}_t$ from the observations $y_0, \ldots, y_t$. Thus, η_t is representable as a function of $y_0, \ldots, y_t$. Conversely, one can show by induction that y_t can also be represented as a function of $\eta_0, \ldots, \eta_t$. Moreover, any measurable function of $y_0, \ldots, y_t$ can be expressed as a measurable function of $\eta_0, \ldots, \eta_t$ and vice-versa [HGK11, Proposition II.1]. So if we define $\mathcal{U}_\eta$ as the set of all control

[1] Our assumption that noise terms are zero-mean is consistent with the standard LQG model [Ber17]. Requiring $\mathbb{E}_{\mathbb{P}_{x_0}}[x_0] = 0$ is assumed for clarity and without loss of generality.

inputs $(u_0, u_1, \ldots, u_{T-1})$ so that $u_t = \psi_t(\eta_0, \ldots, \eta_t)$ for some measurable function $\psi_t : \mathbb{R}^{p(t+1)} \to \mathbb{R}^m$ for every $t \in [T-1]$, the above reasoning implies that $\mathcal{U}_\eta = \mathcal{U}_y$.

In view of this, we can rewrite the distributionally robust LQG problem equivalently as

$$\begin{aligned} p^\star &= \begin{cases} \min\limits_{x,u,y} & \max\limits_{\mathbb{P}\in\mathcal{W}} \ \mathbb{E}_{\mathbb{P}}\left[u^\top Ru + x^\top Qx\right] \\ \text{s.t.} & u \in \mathcal{U}_y,\ x = Hu + Gw,\ y = Cx + v \end{cases} \\ &= \begin{cases} \min\limits_{x,u} & \max\limits_{\mathbb{P}\in\mathcal{W}} \ \mathbb{E}_{\mathbb{P}}\left[u^\top Ru + x^\top Qx\right] \\ \text{s.t.} & u \in \mathcal{U}_\eta, \quad x = Hu + Gw, \end{cases} \end{aligned} \tag{6.5}$$

where $x = (x_0, \ldots, x_T)$, $u = (u_0, \ldots, u_{T-1})$, $y = (y_0, \ldots, y_{T-1})$, $w = (x_0,\ w_0,\ \ldots,\ w_{T-1})$, $v = (v_0, \ldots, v_{T-1})$, $\eta = (\eta_0, \ldots, \eta_{T-1})$, and R, Q, H, G and C are suitable block matrices (see Appendix §6.7 for their precise definitions). The latter reformulation involving the purified observations η is useful because these are *independent* of the inputs. Indeed, by recursively combining the equations of the original and the noise-free systems, one can show that $\eta = Dw + v$ for some block triangular matrix D (see Appendix §6.7 for its construction). This shows that the purified observations depend (linearly) on the exogenous uncertainties but *not* on the control inputs. Hence, the cyclic dependencies complicating the original system are eliminated in (6.5).

Subsequently, we also study the dual of (6.5), defined as

$$d^\star = \begin{cases} \max\limits_{\mathbb{P}\in\mathcal{W}} & \min\limits_{x,u} \ \mathbb{E}_{\mathbb{P}}\left[u^\top Ru + x^\top Qx\right] \\ \ \text{s.t.} & u \in \mathcal{U}_\eta, \ \ x = Hu + Gw. \end{cases} \tag{6.6}$$

The classic minimax inequality implies that $p^\star \geq d^\star$. If we can prove that $p^\star = d^\star$, that (6.5) has a solution $u^\star$ and that (6.6) has a solution $\mathbb{P}^\star$, then $(u^\star, \mathbb{P}^\star)$ must be a Nash equilibrium of the zero-sum game at hand [Roc74b, Theorem 2]. However, because $\mathcal{U}_\eta$ is an infinite-dimensional function space and $\mathcal{W}$ is an infinite-dimensional, non-convex set of non-parametric distributions, the existence of a Nash equilibrium (in pure strategies) is not at all evident. Instead, our proof strategy will rely on constructing an upper bound for $p^\star$ and a lower bound for $d^\star$, and showing that these match.

Upper Bound for $p^\star$. We obtain an upper bound for $p^\star$ by suitably *enlarging* the ambiguity set $\mathcal{W}$ and *restricting* the controllers u_t to linear dependencies. We enlarge $\mathcal{W}$ by ignoring all information about the distributions in $\mathcal{W}$ except for their covariance matrices, and by replacing the Wasserstein distance with the Gelbrich distance. To that end, we first define the Gelbrich distance on the space of covariance matrices.

Definition 12 (Gelbrich distance)**.** *The Gelbrich distance between the two co-*

variance matrices $\Sigma_1, \Sigma_2 \in \mathbb{S}_+^d$ is given by

$$\mathbb{G}(\Sigma_1, \Sigma_2) = \sqrt{\operatorname{Tr}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_2^{\frac{1}{2}}\Sigma_1\Sigma_2^{\frac{1}{2}}\right)^{\frac{1}{2}}\right)}.$$

We are interested in the Gelbrich distance because of its close connection to the 2-Wasserstein distance. Indeed, it is known that the 2-Wasserstein distance between two distributions with zero means is bounded below by the Gelbrich distance between the respective covariance matrices.

Proposition 6.3.1 (Gelbrich bound [Gel90, Theorem 2.1])**.** *For any two distributions $\mathbb{P}_1$ and $\mathbb{P}_2$ on $\mathbb{R}^d$ with zero means and covariance matrices $\Sigma_1, \Sigma_2 \in \mathbb{S}_+^d$, respectively, we have $\mathbb{W}(\mathbb{P}_1, \mathbb{P}_2) \geq \mathbb{G}(\Sigma_1, \Sigma_2)$.*

Recalling that $\hat{X}_0$, $\hat{W}_t$ and $\hat{V}_t$ respectively denote the covariance matrices for x_0, w_t and v_t under the nominal distribution $\hat{\mathbb{P}}$, we can then define the following Gelbrich ambiguity set for the exogenous uncertainties:

$$\mathcal{G} = \mathcal{G}_{x_0} \otimes (\otimes_{t=0}^{T-1}\mathcal{G}_{w_t}) \otimes (\otimes_{t=0}^{T-1}\mathcal{G}_{v_t}),$$

where

$$\begin{aligned}
\mathcal{G}_{x_0} &= \{\mathbb{P}_{x_0} \in \mathcal{P}(\mathbb{R}^n) : \mathbb{E}_{\mathbb{P}_{x_0}}[x_0] = 0,\ \mathbb{E}_{\mathbb{P}}[x_0x_0^\top] = X_0,\ \mathbb{G}(X_0, \hat{X}_0) \leq \rho_{x_0}\} \\
\mathcal{G}_{w_t} &= \{\mathbb{P}_{w_t} \in \mathcal{P}(\mathbb{R}^n) : \mathbb{E}_{\mathbb{P}_{w_t}}[w_t] = 0,\ \mathbb{E}_{\mathbb{P}}[w_tw_t^\top] = W_t,\ \mathbb{G}(W_t, \hat{W}_t) \leq \rho_{w_t}\} \\
\mathcal{G}_{v_t} &= \{\mathbb{P}_{v_t} \in \mathcal{P}(\mathbb{R}^m) : \mathbb{E}_{\mathbb{P}_{v_t}}[v_t] = 0,\ \mathbb{E}_{\mathbb{P}}[v_tv_t^\top] = V_t,\ \mathbb{G}(V_t, \hat{V}_t) \leq \rho_{v_t}\}.
\end{aligned}$$

By construction, the random vectors x_0, $\{w_t\}_{t=0}^{T-1}$ and $\{v_t\}_{t=0}^{T-1}$ are thus mutually independent under any $\mathbb{P} \in \mathcal{G}$. In addition and as a direct consequence of Proposition 6.3.1, $\mathcal{G}$ constitutes an outer approximation for the Wasserstein ambiguity set $\mathcal{W}$, as summarized in the next result.

Corollary 9 (Gelbrich hull)**.** *We have $\mathcal{W} \subseteq \mathcal{G}$.*

Because $\mathcal{G}$ covers $\mathcal{W}$, we henceforth refer to it as the *Gelbrich hull* of the Wasserstein ambiguity set $\mathcal{W}$. To finalize our construction of the upper bound on $p^\star$, we focus on linear policies[2] of the form $u = q + U\eta = q + U(Dw + v)$, where $q = (q_0, \ldots, q_{T-1})$, and U is a block lower triangular matrix

$$U = \begin{bmatrix} U_{0,0} & & & \\ U_{1,0} & U_{1,1} & & \\ \vdots & & \ddots & \\ U_{T-1,0} & \ldots & \ldots & U_{T-1,T-1} \end{bmatrix}. \tag{6.7}$$

[2] Technically, the policies are affine because they include a constant term, but we retain the more common terminology that focuses on the dependencies.

The block lower triangularity of U ensures that the corresponding controller is causal, which in turn ensures that $u \in \mathcal{U}_\eta$. In the following, we denote by $\mathcal{U}$ the set of all block lower triangular matrices of the form (6.7). An upper bound on problem (6.5) can now be obtained by *restricting* the controller's feasible set to causal controllers that are *linear* in the purified observations η and by *relaxing* nature's feasible set to the Gelbrich hull $\mathcal{G}$ of $\mathcal{W}$. The resulting bounding problem is given by

$$\overline{p}^\star = \left\{ \begin{array}{ll} \min\limits_{U,q,x,u} & \max\limits_{\mathbb{P}\in\mathcal{G}} \mathbb{E}_{\mathbb{P}}\left[u^\top R u + x^\top Q x\right] \\ \text{s.t.} & U \in \mathcal{U}, \quad u = q + U(Dw+v), \quad x = Hu + Gw. \end{array} \right. \tag{6.8}$$

As we obtained (6.8) by restricting the feasible set of the outer minimization problem and relaxing the feasible set of the inner maximization problem in (6.5), it is clear that $\overline{p}^\star \geq p^\star$. Recall also that problem (6.5) constitutes an infinite-dimensional zero-sum game, where the agents optimize over measurable policies and non-parametric distributions, respectively. In contrast, the next proposition shows that problem (6.8) is equivalent to a finite-dimensional zero-sum game.

Proposition 6.3.2. *Problem* (6.8) *is equivalent to the optimization problem*

$$\overline{p}^\star = \left\{ \begin{array}{lll} \min\limits_{\substack{q\in\mathbb{R}^{pT} \\ U\in\mathcal{U}}} & \max\limits_{\substack{W\in\mathcal{G}_W \\ V\in\mathcal{G}_V}} & \mathrm{Tr}\left(\left(D^\top U^\top (R+H^\top QH)UD + 2G^\top QHUD + G^\top QG\right)W\right) \\ & & +\mathrm{Tr}\left(\left(U^\top (R+H^\top QH)U\right)V\right) + q^\top (R+H^\top QH)q, \end{array} \right. \tag{6.9}$$

where

$$\mathcal{G}_W = \left\{ W \in \mathbb{S}_+^{n(T+1)} : \begin{array}{l} W = \mathrm{diag}(X_0, W_0, \ldots, W_{T-1}), \\ X_0 \in \mathbb{S}_+^n,\ W_t \in \mathbb{S}_+^n \ \ \forall t \in [T-1] \\ \mathbb{G}(X_0, \hat{X}_0)^2 \leq \rho_{x_0}^2, \ \ \mathbb{G}(W_t, \hat{W}_t)^2 \leq \rho_{w_t}^2 \ \ \forall t \in [T-1] \end{array} \right\}$$

$$\mathcal{G}_V = \left\{ V \in \mathbb{S}_+^{pT} : \begin{array}{l} V = \mathrm{diag}(V_0, \ldots, V_{T-1}), \\ V_t \in \mathbb{S}_+^p, \ \ \mathbb{G}(V_t, \hat{V}_t)^2 \leq \rho_{v_t}^2 \ \ \forall t \in [T-1] \end{array} \right\}.$$

We emphasize that Proposition 6.3.2 remains valid even if the nominal distribution $\hat{\mathbb{P}}$ fails to be normal. Note also that, while nature's feasible set in (6.8) is non-convex due to the independence conditions, the sets $\mathcal{G}_W$ and $\mathcal{G}_V$ are convex and even semidefinite representable thanks to the properties of the squared Gelbrich distance.[3] By dualizing the inner maximization problem, one can therefore reformulate the minimax problem (6.9) as a convex semidefinite program (SDP). Even though this SDP is computationally tractable in theory, it involves $\mathcal{O}(T(mp+n^2+p^2))$ decision variables. For practically interesting problem dimensions, it thus quickly exceeds the capabilities of existing solvers.

[3] Note that the ambiguity sets $\mathcal{G}_W$ and $\mathcal{G}_V$ appearing in (6.9) involve the squared Gelbrich distance, $\mathbb{G}(\Sigma_1, \Sigma_2)^2$. The reason is that $\mathbb{G}(\Sigma_1, \Sigma_2)^2$ is known to be jointly convex in Σ_1, Σ_2 and semidefinite representable [Ngu+23, Proposition 2.3], unlike the Gelbrich distance $\mathbb{G}(\Sigma_1, \Sigma_2)$ itself, which is generally non-convex.

Lower Bound for $d^\star$. To derive a tractable lower bound on $d^\star$, we restrict nature's feasible set to the family $\mathcal{W}_\mathcal{N}$ of all *normal* distributions in the Wasserstein ambiguity set $\mathcal{W}$. The resulting bounding problem is thus given by

$$\underline{d}^\star = \left\{ \begin{array}{lll} \max\limits_{\mathbb{P}\in\mathcal{W}_\mathcal{N}} & \min\limits_{x,u} & \mathbb{E}_\mathbb{P}\left[u^\top Ru + x^\top Qx\right] \\ & \text{s.t.} & u \in \mathcal{U}_\eta, \quad x = Hu + Gw. \end{array} \right. \tag{6.10}$$

As we obtained (6.10) by restricting the feasible set of the outer maximization problem in (6.6), it is clear that $\underline{d}^\star \leq d^\star$. Next, we show that (6.10) can be recast as a finite-dimensional zero-sum game. This result critically relies on the following known fact regarding the 2-Wasserstein distance between two *normal* distributions, which coincides with the Gelbrich distance between their covariance matrices.

Proposition 6.3.3 (Tightness for normal distributions [GS84, Proposition 7])**.** *For any two normal distributions $\mathbb{P}_1 = \mathcal{N}(0, \Sigma_1)$ and $\mathbb{P}_2 = \mathcal{N}(0, \Sigma_2)$ with zero means we have $\mathbb{W}(\mathbb{P}_1, \mathbb{P}_2) = \mathbb{G}(\Sigma_1, \Sigma_2)$.*

With this, we can provide a finite-dimensional reformulation, as summarized in the next result.

Proposition 6.3.4. *Problem (6.10) is equivalent to the optimization problem*

$$\underline{d}^\star = \left\{ \begin{array}{lll} \max\limits_{\substack{W\in\mathcal{G}_W\\ V\in\mathcal{G}_V}} & \min\limits_{\substack{q\in\mathbb{R}^{pT}\\ U\in\mathcal{U}}} & \mathrm{Tr}\left(\left(D^\top U^\top (R+H^\top QH)UD + 2G^\top QHUD + G^\top QG\right)W\right) \\ & & +\mathrm{Tr}\left(\left(U^\top (R+H^\top QH)U\right)V\right) + q^\top (R+H^\top QH)q, \end{array} \right. \tag{6.11}$$

where $\mathcal{G}_W$ and $\mathcal{G}_V$ are defined exactly as in Proposition 6.3.2.

Proposition 6.3.4 relies on Proposition 6.3.3 and thus fails to hold unless $\hat{\mathbb{P}}$ is normal. Also, one can again reformulate (6.11) as a tractable SDP by dualizing the inner minimization problem.

Conclusions. Propositions 6.3.2 and 6.3.4 reveal that problems (6.9) and (6.11) are dual to each other, that is, they can be transformed into one another by interchanging minimization and maximization. The following main theorem shows that strong duality holds irrespective of the problem data.

Theorem 6.3.5 (Strong duality of (6.9) and (6.11))**.** *We have $\overline{p}^\star = \underline{d}^\star$.*

Theorem 6.3.5 follows immediately from Sion's classic minimax theorem [Sio58], which applies because $\mathcal{G}_W$ and $\mathcal{G}_V$ are convex as well as compact thanks to [Ngu+23, Lemma A.6].

By weak duality and the construction of the bounding problems (6.9) and (6.11), we trivially have $\underline{d}^\star \leq d^\star \leq p^\star \leq \overline{p}^\star$. Theorem 6.3.5 reveals that all of these inequalities are in fact equalities, each of which gives rise to a non-trivial insight. The first key insight is that (6.5) and (6.6) are strong duals.

Corollary 10 (Strong duality of (6.5) and (6.6))**.** *We have $p^\star = d^\star$.*

We stress that, unlike Theorem 6.3.5, Corollary 10 establishes strong duality between two *infinite-dimensional* zero-sum games. The second key implication of Theorem 6.3.5 is that the distributionally robust LQG problem (6.5) is solved by a linear output-feedback controller.

Corollary 11 (The controller's Nash strategy is linear in the observations)**.** *There exist $U^\star \in \mathcal{U}$ and $q^\star \in \mathbb{R}^m$ such that the distributionally robust LQG problem* (6.5) *is solved by $u^\star = q^\star + U^\star y$.*

The identity $p^\star = \overline{p}^\star$ readily implies that (6.5) is solved by a causal controller that is linear in the *purified* observations. However, any causal controller that is linear in the purified observations η can be reformulated *exactly* as a causal controller that is linear in the original observations y and vice-versa [BTBN06, Proposition 3]. Thus, Corollary 11 follows. The third key implication of Theorem 6.3.5 is that the *dual* distributionally robust LQG problem is solved by a normal distribution.

Corollary 12 (Nature's Nash strategy is a normal distribution)**.** *The dual distributionally robust LQG problem* (6.6) *is solved by a distribution $\mathbb{P}^\star \in \mathcal{W}_{\mathcal{N}}$.*

Corollary 12 is a direct consequence of the identity $\underline{d}^\star = d^\star$. Note that the optimal normal distribution $\mathbb{P}^\star$ is uniquely determined by the covariance matrices $W^\star$ and $V^\star$ of the exogenous uncertain parameters, which can be computed by solving problem (6.11). That the worst-case distribution is actually Gaussian is not a-priori expected and is surprising given that the Wasserstein ball contains many non-Gaussian distributions.

6.4. Efficient Numerical Solution of Distributionally Robust LQG Problems

Having proven these structural results, we next turn attention to the problem of finding the optimal strategies. Our next result shows that, under a mild regularity condition, the optimal controller $u^\star$ of the distributionally robust LQG problem (6.5) can be computed efficiently from $\mathbb{P}^\star$.

Proposition 6.4.1 (Optimality of Kalman filter-based feedback controllers)**.** *If $\hat{V}_t \succ 0$ for all $t \in [T-1]$, then problem* (6.6) *is solved by a Gaussian distribution $\mathbb{P}^\star$ under which v_t has a covariance matrix $V_t^\star \succ 0$ for every $t \in [T-1]$, and* (6.5) *is solved by the optimal LQG controller corresponding to $\mathbb{P}^\star$. Additionally, the optimal value of problem* (6.9) *and its strong dual* (6.11) *does not* *change if we restrict $\mathcal{G}_W$ and $\mathcal{G}_V$ to $\mathcal{G}_W^+$ and $\mathcal{G}_V^+$, respectively, where*

$$\begin{aligned}
\mathcal{G}_W^+ &= \left\{ W \in \mathcal{G}_W : X_0 \succeq \lambda_{\min}(\hat{X}_0) I,\ W_t \succeq \lambda_{\min}(\hat{W}_t) I\ \forall t \in [T-1] \right\}, \\
\mathcal{G}_V^+ &= \left\{ V \in \mathcal{G}_V : V_t \succeq \lambda_{\min}(\hat{V}_t) I\ \forall t \in [T-1] \right\}.
\end{aligned}$$

This implies that the optimal controller can be computed by solving a classic LQG problem corresponding to nature's optimal strategy $\mathbb{P}^\star$, which can be done very efficiently through Kalman filtering and dynamic programming (see Appendix §6.6 for details). It thus suffices to design an efficient algorithm for computing $\mathbb{P}^\star$, which is uniquely determined by the covariance matrices $(W^\star, V^\star)$ that solve problem (6.11). To this end, we first reformulate (6.11) as

$$\max_{W \in \mathcal{G}_W^+, V \in \mathcal{G}_V^+} f(W, V), \tag{6.12}$$

where we restrict $\mathcal{G}_W$ and $\mathcal{G}_V$ to $\mathcal{G}_W^+$ and $\mathcal{G}_V^+$, respectively, due to Proposition 6.4.1, and where $f(W, V)$ denotes the optimal value function of the inner minimization problem in (6.11). As (6.11) is a reformulation of (6.10) and as the family of all causal purified output-feedback controllers matches the family of causal output-feedback controllers, $f(W, V)$ can also be viewed as the optimal value of the classic LQG problem corresponding to the normal distribution $\mathbb{P}$ determined by the covariance matrices W and V. These insights lead to the following structural result.

Proposition 6.4.2. *$f(W, V)$ is concave and β-smooth in $(W, V) \in \mathcal{G}_W^+ \times \mathcal{G}_V^+$ for some $\beta > 0$.*

By Proposition 6.4.2, it is possible to address problem (6.12) with a Frank-Wolfe algorithm [DR70; Dun79; Dun80; DH78; FW56; LP66]. Each iteration of this algorithm solves a direction-finding subproblem, that is, a variant of problem (6.12) that maximizes the first-order Taylor expansion of $f(W, V)$ around the current iterates.

$$\max_{L_W \in \mathcal{G}_W^+, L_V \in \mathcal{G}_V^+} \langle \nabla_W f(W, V), L_W - W \rangle + \langle \nabla_V f(W, V), L_V - V \rangle \tag{6.13}$$

The next iterates are then obtained by moving towards a maximizer $(L_W^\star, L_V^\star)$ of (6.13), i.e., we update

$$(W, V) \leftarrow (W, V) + \alpha \cdot (L_W^\star - W, L_v^\star - V),$$

where α is an appropriate step size. The proposed Frank-Wolfe algorithm enjoys a very low per-iteration complexity because problem (6.13) is separable. To see this, we reformulate (6.13) as

$$\begin{array}{cl}
\max\limits_{L_W, L_V} & \langle \nabla_{X_0} f(W, V), L_{X_0} - X_0 \rangle + \\
& \sum_{t=0}^{T-1} \langle \nabla_{W_t} f(W, V), L_{W_t} - W_t \rangle + \langle \nabla_{V_t} f(W, V), L_{V_t} - V_t \rangle \\
\text{s.t.} & \mathbb{G}(L_{X_0}, \hat{X}_0)^2 \le \rho_{x_0}^2,\ \mathbb{G}(L_{W_t}, \hat{W}_t)^2 \le \rho_{w_t}^2,\ \mathbb{G}(L_{V_t}, \hat{V}_t)^2 \le \rho_{v_t}^2\ \forall t \in [T-1] \\
& L_{X_0} \succeq \lambda_{\min}(\hat{X}_0) I,\ L_{W_t} \succeq \lambda_{\min}(\hat{W}_t) I,\ L_{V_t} \succeq \lambda_{\min}(\hat{V}_t) I\ \forall t \in [T-1].
\end{array}$$

Hence, (6.13) decomposes into $2T + 1$ separate subproblems that can be solved in parallel. That is, for any matrix $Z \in \{X_0, W_0, \dots, W_{T-1}, V_0, \dots, V_{T-1}\}$ we solve a separate subproblem of the form

$$\max_{L_Z \succeq \lambda_{\min}(\hat{Z})} \left\{ \langle \nabla_Z f(W, V), L_Z - Z \rangle : \mathbb{G}(L_Z, \hat{Z})^2 \leq \rho_z^2 \right\}. \tag{6.14}$$

These subproblems can be reformulated as tractable SDPs and are thus amenable to efficient off-the-shelf solvers. By [Ngu+23, Theorem 6.2], however, one can exploit the structure of the Gelbrich distance in order to reduce (6.14) to a univariate algebraic equation that can be solved to any desired accuracy $\delta > 0$ by a highly efficient bisection algorithm. We say that L_Z^δ is a δ-approximate solution of problem (6.14) for some $\delta \in (0, 1)$ if L_Z^δ is feasible in (6.14) and if

$$\langle \nabla_Z f(W, V), L_Z^\delta - Z \rangle \geq \delta \langle \nabla_Z f(W, V), L_Z^\star - Z \rangle,$$

where $L_Z^\star$ is an exact maximizer of (6.14). Note that, by the concavity of $f(W, V)$, the inner product on the right-hand side is nonnegative and vanishes if and only if Z maximizes $f(W, V)$ over the feasible set of (6.14). For further details we refer to Appendix §6.10 in the supplementary material.

Remark 5 (Automatic differentiation)**.** *Recall that $f(W, V)$ coincides with the optimal value of the LQG problem corresponding to the normal distribution $\mathbb{P}$ determined by the covariance matrices W and V. By using the underlying dynamic programming equations, $f(W, V)$ can thus be expressed in closed form as a serial composition of $\mathcal{O}(T)$ rational functions (see Appendix §6.6 for details). Hence, $\nabla_Z f(W, V)$ can be calculated symbolically for any $Z \in \{X_0, W_0, \dots, W_{T-1}, V_0, \dots, V_{T-1}\}$ by repeatedly applying the chain and product rules. However, the resulting formulas are lengthy and cumbersome. We thus compute the gradients numerically using backpropagation. The cost of evaluating $\nabla_Z f(W, V)$ is then of the same order of magnitude as the cost of evaluating $f(W, V)$.*

A detailed description of the proposed Frank-Wolfe method is given in Algorithm 6 below.

By [Jag13, Theorem 1 and Lemma 7], which applies thanks to the structural properties of $f(W, V)$ established in Proposition 6.4.2, Algorithm 6 attains a suboptimality gap of ϵ within $\mathcal{O}(1/\epsilon)$ iterations.

6.5. Numerical Experiments

All experiments are run on an Intel i7-8700 CPU (3.2 GHz) machine with 16GB RAM. All linear SDP problems are modeled in Python 3.8.6 using CVXPY [Agr+18; DB16] and solved with MOSEK [MOS19]. The gradients of $f(W, V)$

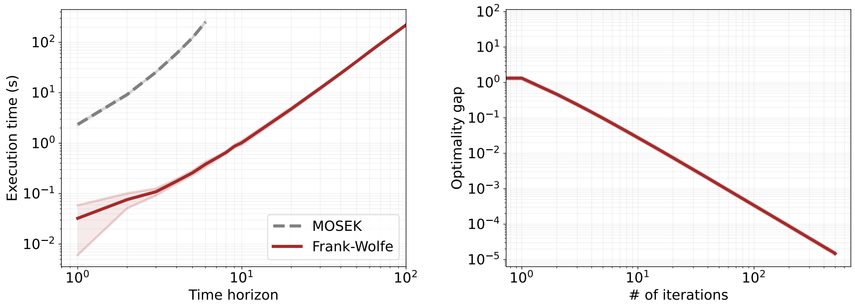
Algorithm 6 Frank-Wolfe algorithm for solving (6.12)

```
Input: initial iterates $W$, $V$, nominal covariance matrices $\hat{W}$, $\hat{V}$, oracle precision $\delta \in (0, 1)$
set initial iteration counter $k = 0$
while stopping criterion is not met do
    for $Z \in \{X_0, W_0, \dots, W_{T-1}, V_0, \dots, V_{T-1}\}$ do in parallel
        compute $\nabla_Z f(W, V)$
        find a $\delta$-approximate solution $L_Z^\delta$ of (6.14)
    end
    $g \leftarrow \langle \nabla_W f(W, V), L_W^\delta - W \rangle + \langle \nabla_V f(W, V), L_V^\delta - V \rangle$
    $(W, V) \leftarrow (W, V) + 2/(2 + k) \cdot (L_W^\delta - W, L_V^\delta - V)$
end while
Output: $W$ and $V$
```

are computed via Pymanopt [TKW16] with PyTorch's automated differentiation module [Pas+17; Pas+19].

Consider a class of distributionally robust LQG problems with $n = m = p = 10$. We set $A_t = 0.1 \times A$ to have ones on the main diagonal and the superdiagonal and zeroes everywhere else ($A_{i,j} = 1$ if $i = j$ or $i = j - 1$ and $A_{i,j} = 0$ otherwise), and the other matrices to $B_t = C_t = Q_t = R_t = I_d$. The Wasserstein radii are set to $\rho_{x_0} = \rho_{w_t} = \rho_{v_t} = 10^{-1}$. The nominal covariance matrices of the exogenous uncertainties are constructed randomly and with eigenvalues in the interval $[1, 2]$ (so as to ensure they are positive definite). The code is publicly available in the Github repository https://github.com/RAO-EPFL/DR-Control.

The optimal value of the distributionally robust LQG problem (6.5) can be computed by directly solving the SDP reformulation of (6.11) with MOSEK or by solving the nonlinear SDP (6.12) with our Frank-Wolfe method detailed in Algorithm 6. We next compare these two approaches in 10 independent simulation runs, where we set a stopping criterion corresponding to an optimality gap below 10^{-3} and we run the Frank-Wolfe method with $\delta = 0.95$. Figure 6.1a illustrates the execution time for both approaches as a function of the planning horizon T; runs where MOSEK exceeds 100s are not reported. Figure 6.1b visualizes the empirical convergence behavior of the Frank-Wolfe algorithm. The results highlight that the Frank-Wolfe algorithm achieves running times that are uniformly lower than MOSEK across all problem horizons and is able to find highly accurate solutions already after a small number of iterations (50 iterations for problem instances of time horizon $T = 10$).



(a)

(b)

Figure 6.1: (a) Execution time for MOSEK and Frank-Wolfe algorithm over 10 simulation runs as a function of the horizon T (solid lines show the mean and the shaded areas correspond to 1 standard deviation). (b) Convergence of optimality gap for Frank-Wolfe algorithm with horizon $T = 10$.

Appendix

The supplementary material is structured as follows. Appendix §6.6 presents the well-known solution to the classic LQG problem using dynamic programming and Kalman Filter estimation. Appendix §6.7 provides the definitions of the stacked system matrices utilized in the compact formulation (6.5) of the distributionally robust LQG problem. Appendix §6.8 contains the proofs of the formal statements in the main text and provides additional technical results. Appendix §6.9 derives the SDP reformulation of the dual problem (6.11). Appendix §6.10, finally, elaborates on the bisection algorithm used for solving the linearization oracle of the Frank-Wolfe algorithm.

6.6. Solution of the LQG Problem

The classic LQG problem can be solved efficiently via dynamic programming; see, e.g., [Ber17]. That is, the unique optimal control inputs satisfy $u_t^\star = K_t \hat{x}_t$ for every $t \in [T-1]$, where $K_t \in \mathbb{R}^{n \times n}$ is the optimal feedback gain matrix, and $\hat{x}_t = \mathbb{E}_{\mathbb{P}}[x_t | y_0, \ldots, y_t]$ is the minimum mean-squared-error estimator of x_t given the observation history up to time t. Thanks to the celebrated separation principle, K_t can be computed by pretending that the system is deterministic and allows for perfect state observations, and $\hat{x}_t$ can be computed while ignoring the control problem.

To compute K_t, one first solves the deterministic LQR problem corresponding to the LQG problem at hand. Its value function $x_t^\top P_t x_t$ at time t is quadratic in x_t, and P_t obeys the backward recursion

$$P_t = A_t^\top P_{t+1} A_t + Q_t - A_t^\top P_{t+1} B_t (R_t + B_t^\top P_{t+1} B_t)^{-1} B_t^\top P_{t+1} A_t \quad \forall t \in [T-1] \tag{6.15a}$$

initialized by $P_T = Q_T$. The optimal feedback gain matrix K_t can then be computed from P_{t+1} as

$$K_t = -(R_t + B_t^\top P_{t+1} B_t)^{-1} B_t^\top P_{t+1} A_t \quad \forall t \in [T-1]. \tag{6.15b}$$

Since x_t and $(y_0, \ldots, y_t)$ are jointly normally distributed, the minimum mean-squared-error estimator $\hat{x}_t$ can be calculated directly using the formula for the mean of a conditional normal distribution. Alternatively, however, one can use the Kalman filter to compute $\hat{x}_t$ recursively, which is significantly more insightful and efficient. The Kalman filter also recursively computes the covariance matrix Σ_t of x_t conditional on $y_0, \ldots, y_t$ and the covariance matrix $\Sigma_{t+1|t}$ of x_{t+1} conditional on $y_0, \ldots, y_t$ evaluated under $\mathbb{P}$. Specifically, these covariance matrices obey the forward recursion

$$\left.\begin{aligned} &\Sigma_t = \Sigma_{t|t-1} - \Sigma_{t|t-1} C_t^\top (C_t \Sigma_{t|t-1} C_t^\top + V_t)^{-1} C_t \Sigma_{t|t-1} \\ &\Sigma_{t+1|t} = A_t \Sigma_t A_t^\top + W_t \end{aligned}\right\} \quad \forall t \in [T-1] \tag{6.16}$$

initialized by $\Sigma_{0|-1} = X_0$. Using $\Sigma_{t|t-1}$, we then define the Kalman filter gain as

$$L_t = \Sigma_t C_t^\top V_t^{-1} \quad \forall t \in [T-1]$$

which allows us to compute the minimum mean-squared-error estimator via the forward recursion

$$\hat{x}_{t+1} = A_t \hat{x}_t + B_t u_t + L_{t+1} \left(y_{t+1} - C_{t+1} (A_t \hat{x}_t + B_t u_t) \right) \quad \forall t \in [T-1]$$

initialized by $\hat{x}_0 = L_0 y_0$. One can also show that the optimal value of the LQG problem amounts to

$$\sum_{t=0}^{T-1} \operatorname{Tr}((Q_t - P_t)\Sigma_t) + \sum_{t=1}^{T} \operatorname{Tr}(P_t (A_{t-1} \Sigma_{t-1} A_{t-1}^\top + W_{t-1})) + \operatorname{Tr}(P_0 X_0). \tag{6.17}$$

6.7. Definitions of Stacked System Matrices

The stacked system matrices appearing in the distributionally robust LQG problem (6.5) are defined as follows. First, the stacked state and input cost matrices $Q \in \mathbb{S}^{n(T+1)}$ and $R \in \mathbb{S}^{mT}$ are set to

$$Q = \begin{bmatrix} Q_0 & & & \\ & Q_1 & & \\ & & \ddots & \\ & & & Q_T \end{bmatrix} \quad \text{and} \quad R = \begin{bmatrix} R_0 & & & \\ & R_1 & & \\ & & \ddots & \\ & & & R_{T-1} \end{bmatrix},$$

respectively. Similarly, the stacked matrices appearing in the linear dynamics and the measurement equations $C \in \mathbb{R}^{pT \times n(T+1)}$, $G \in \mathbb{R}^{n(T+1)\times n(T+1)}$ and $H \in \mathbb{R}^{n(T+1)\times mT}$ are defined as

$$C = \begin{bmatrix} C_0 & 0 & & & \\ & C_1 & 0 & & \\ & & \ddots & \ddots & \\ & & & C_{T-1} & 0 \end{bmatrix}, \quad G = \begin{bmatrix} A_0^0 & & & \\ A_0^1 & A_1^1 & & \\ \vdots & & \ddots & \\ A_0^T & A_1^T & \dots & A_T^T \end{bmatrix}$$

and

$$H = \begin{bmatrix} 0 & & & & \\ A_1^1 B_0 & 0 & & & \\ A_1^2 B_0 & A_2^2 B_1 & 0 & & \\ \vdots & & & \ddots & \\ \vdots & & & & 0 \\ A_1^T B_0 & A_2^T B_1 & \dots & \dots & A_T^T B_{T-1} \end{bmatrix},$$

respectively, where $A_s^t = \prod_{k=s}^{t-1} A_k$ for every $s < t$ and $A_s^t = I_n$ for $s = t$.

Using the stacked system matrices, we can now express the purified observation process η as a linear function of the exogenous uncertainties w and v that is *not* impacted by u; see also [BTBN05; SB10]

Lemma 6.7.1. *We have $\eta = Dw + v$, where $D = CG$.*

Proof of Lemma 6.7.1. The purified observation process is defined as $\eta = y - \hat{y}$. Recall now that the observations of the original system satisfy $y = Cx + v$. Similarly, one readily verifies that the observations of the fictitious noise-free system satisfy $\hat{y} = C\hat{x}$. Thus, we have $\eta = C(x - \hat{x}) + v$. Next, recall that the state of the original system satisfies $x = Hu + Gw$, and note that the state of the fictitious noise-free system satisfies $\hat{x} = Hu$. Combining all of these linear equations finally shows that u cancels out and that $\eta = CGw + v = Dw + v$. □

6.8. Proofs

6.8.1. Additional Technical Results

It is well known that every causal controller that is linear in the original observations y can be reformulated as a causal controller that is linear in the purified observations η and vice versa [BTBN05; SB10]. Perhaps surprisingly, however, the one-to-one transformation between the respective coefficients of y and η is *not* linear. To keep this paper self-contained, we review these insights in the next lemma.

Lemma 6.8.1. *If $u = U\eta + q$ for some $U \in \mathcal{U}$ and $q \in \mathbb{R}^{pT}$, then $u = U'y + q'$ for $U' = (I + UCH)^{-1}U$ and $q' = (I + UCH)^{-1}q$. Conversely, if $u = U'y + q'$ for some $U' \in \mathcal{U}$ and $q' \in \mathbb{R}^{pT}$, then $u = U\eta + q$ for $U = (I - U'CH)^{-1}U'$ and $q = (I - U'CH)^{-1}q'$.*

Proof of Lemma 6.8.1. If $u = U\eta + q$ for some $U \in \mathcal{U}$ and $q \in \mathbb{R}^{pT}$, then we have

$$u = U\eta + q = U(y - \hat{y}) + q = Uy - UC\hat{x} + q = Uy - UCHu + q,$$

where the second equality follows from the definition of η, the third equality holds because $y = Cx + v$, and the last equality exploits our earlier insight that $\hat{y} = C\hat{x}$. The last expression depends only on y and u. Solving for u yields $u = U'y + q'$, where $U' = (I + UCH)^{-1}U$ and $q' = (I + UCH)^{-1}q$. Note that $(I + UCH)$ is indeed invertible because $I + UCH$ is a lower triangular matrix with all diagonal entries equal to one, ensuring a determinant of one.

Similarly, if $u = U'y + q'$ for some $U' \in \mathcal{U}$ and $q' \in \mathbb{R}^{pT}$, then we have

$$u = U'y + q' = U'(\eta + \hat{y}) + q' = U'\eta + U'C\hat{x} + q' = U'\eta + U'CHu + q'.$$

Solving for u yields $u = U\eta + q$, where $U = (I - U'CH)^{-1}U'$ and $q = (I - U'CH)^{-1}q'$. Note again that $(I - U'CH)$ is indeed invertible because $(I - U'CH)$ is a lower triangular matrix with all diagonal entries equal to one. □

6.8.2. Proofs of Section 6.3

Proof of Proposition 6.3.2. In problem (6.8), both u and x are linear in w and v, i.e., $u = q + UDw + Uv$ and $x = Hu + Gw = Hq + HUDw + HUv + Gw$. By substituting the linear representations of u and x into the objective function of problem (6.8), we obtain the following equivalent reformulation.

$$\begin{aligned}\min_{\substack{q\in\mathbb{R}^{pT}\\ U\in\mathcal{U}}} \; \max_{\mathbb{P}\in\mathcal{G}} \; & \mathbb{E}_{\mathbb{P}}\left[w^\top \left(D^\top U^\top (R + H^\top QH)UD + 2D^\top U^\top H^\top QG + G^\top QG\right) w\right] \\ & + \mathbb{E}_{\mathbb{P}}\left[v^\top \left(U^\top (R + H^\top QH)U\right) v\right] + q^\top (R + H^\top QH)q\end{aligned}$$

For any fixed $\mathbb{P} \in \mathcal{G}$, we can express the expectation in the objective function of the above problem in terms of the covariance matrices $W = \mathbb{E}_{\mathbb{P}}[ww^\top]$ and $V = \mathbb{E}_{\mathbb{P}}[vv^\top]$. Thus, the problem becomes

$$\begin{aligned}\min_{\substack{q\in\mathbb{R}^{pT}\\ U\in\mathcal{U}}} \; \max_{W,V,\mathbb{P}} \; & \operatorname{Tr}\left(\left(D^\top U^\top (R{+}H^\top QH)UD + 2G^\top QHUD + G^\top QG\right)W\right) \\ & + \operatorname{Tr}\left(\left(U^\top (R + H^\top QH)U\right)V\right) + q^\top (R{+}H^\top QH)q \\ \text{s.t.} \quad & \mathbb{P} \in \mathcal{G}, \;\; W = \mathbb{E}_{\mathbb{P}}[ww^\top], \;\; V = \mathbb{E}_{\mathbb{P}}[vv^\top].\end{aligned} \tag{6.18}$$

Recall now the definition of $\mathcal{G}$, and note that the requirements $\mathbb{G}(X_0, \hat{X}_0) \leq \rho_{x_0}$, $\mathbb{G}(W_t, \hat{W}_t) \leq \rho_{w_t}$ and $\mathbb{G}(V_t, \hat{V}_t) \leq \rho_{v_t}$ are equivalent to the convex constraints $\mathbb{G}(X_0, \hat{X}_0)^2 \leq \rho_{x_0}^2$, $\mathbb{G}(W_t, \hat{W}_t)^2 \leq \rho_{w_t}^2$ and $\mathbb{G}(V_t, \hat{V}_t)^2 \leq \rho_{v_t}^2$, respectively, for all $t \in [T-1]$. The definition of $\mathcal{G}$ also implies that

$$W = \mathbb{E}_{\mathbb{P}}[ww^\top] = \mathrm{diag}(X_0, W_0, \dots, W_{T-1})$$

and

$$V = \mathbb{E}_{\mathbb{P}}[vv^\top] = \mathrm{diag}(V_0, \dots, V_{T-1}).$$

Problem (6.18) thus constitutes a relaxation of problem (6.9). Indeed, the feasible set of the inner maximization problem in (6.18) is a subset of the feasible set of the inner maximization problem in (6.9). Moreover, for any W and V feasible in the inner maximization problem in (6.9), the distribution $\mathbb{P} = \mathbb{P}_{x_0} \otimes (\otimes_{t=0}^{T-1} \mathbb{P}_{w_t}) \otimes (\otimes_{t=0}^{T} \mathbb{P}_{v_t})$ defined through $\mathbb{P}_{x_0} = \mathcal{N}(0, X_0)$, $\mathbb{P}_{w_t} = \mathcal{N}(0, W_t)$ and $\mathbb{P}_{v_t} = \mathcal{N}(0, V_t)$, $t \in [T-1]$, is feasible in the inner maximization problem in (6.18) with the same objective value. The relaxation is thus exact, and the optimal values of (6.8), (6.9) and (6.18) coincide. □

Proof of Proposition 6.3.4. Recall that the space $\mathcal{U}_y$ of all causal output-feedback controllers coincides with the space $\mathcal{U}_\eta$ of all causal *purified* output-feedback controllers. We can thus replace the feasible set $\mathcal{U}_\eta$ of the inner minimization problem in (6.10) with $\mathcal{U}_y$. Hence, for any fixed $\mathbb{P} \in \mathcal{W}_{\mathcal{N}}$, the inner minimization problem in (6.10) constitutes a classic LQG problem. By standard LQG theory [Ber17], it is solved by a *linear* output-feedback controller of the form $u = U'y + q'$ for some $U' \in \mathcal{U}$ and $q' \in \mathbb{R}^{pT}$; see also Appendix §6.6. Lemma 6.8.1 shows, however, that any linear output-feedback controller can be equivalently expressed as a linear *purified*-output feedback controller of the form $u = U\eta + q$ for some $U \in \mathcal{U}$ and $q \in \mathbb{R}^{pT}$. In summary, the above reasoning shows that the feasible set of the inner minimization problem in (6.10) can be reduced to the family of all linear purified-output feedback controllers without sacrificing optimality. Thus, problem (6.10) is equivalent to

$$\begin{array}{lll} \max\limits_{\mathbb{P} \in \mathcal{W}_{\mathcal{N}}} & \min\limits_{q, U, x, u} & \mathbb{E}_{\mathbb{P}}\left[u^\top R u + x^\top Q x\right] \\ & \text{s.t.} & U \in \mathcal{U}, \quad u = q + U\eta, \quad x = Hu + Gw. \end{array}$$

Using a similar reasoning as in the proof of Proposition 6.3.2, we can now substitute the linear representations of u and x into the objective function and reformulate the above problem as

$$\begin{aligned}\max_{W,V,\mathbb{P}} \quad & \min_{\substack{q\in\mathbb{R}^{pT}\\ U\in\mathcal{U}}} \quad \mathrm{Tr}\left(\left(D^\top U^\top (R+H^\top QH)UD+2G^\top QHUD+G^\top QG\right)W\right)\\ & \qquad\qquad + \mathrm{Tr}\left(\left(U^\top (R+H^\top QH)U\right)V\right)+q^\top (R+H^\top QH)q\\ & \text{s.t.} \quad \mathbb{P}\in\mathcal{W}_{\mathcal{N}},\ \ W=\mathbb{E}_{\mathbb{P}}[ww^\top],\ \ V=\mathbb{E}_{\mathbb{P}}[vv^\top].\end{aligned}$$

As $\mathcal{W}_{\mathcal{N}}$ contains only *normal* distributions, Proposition 6.3.3 implies that $\mathbb{W}(\mathbb{P}_{x_0},\hat{\mathbb{P}}_{x_0})=\mathbb{G}(X_0,\hat{X}_0)$, $\mathbb{W}(\mathbb{P}_{w_t},\hat{\mathbb{P}}_{w_t})=\mathbb{G}(W_t,\hat{W}_t)$ and $\mathbb{W}(\mathbb{P}_{v_t},\hat{\mathbb{P}}_{v_t})=\mathbb{G}(V_t,\hat{V}_t)$ for all $t\in[T-1]$. We may thus replace the requirement $\mathbb{W}(\mathbb{P}_{x_0},\hat{\mathbb{P}}_{x_0})\le\rho_{x_0}$ in the definition of $\mathcal{W}_{\mathcal{N}}$ by $\mathbb{G}(X_0,\hat{X}_0)\le\rho_{x_0}$, which is equivalent to the convex constraint $\mathbb{G}(X_0,\hat{X}_0)^2\le\rho_{x_0}^2$. The conditions on the marginal distributions of w_t and v_t, $t\in[T-1]$, admit similar reformulations. The definition of $\mathcal{W}_{\mathcal{N}}$ also implies that

$$\begin{aligned}W=\mathbb{E}_{\mathbb{P}}[ww^\top]=\mathrm{diag}(X_0,W_0,\dots,W_{T-1})\quad\text{and}\\ V=\mathbb{E}_{\mathbb{P}}[vv^\top]=\mathrm{diag}(V_0,\dots,V_{T-1}).\end{aligned}$$

Thus, the feasible set of the outer maximization problem in (6.11) constitutes a relaxation of that in (6.10). One readily verifies that the relaxation is exact by using similar arguments as in the proof of Proposition 6.3.2. Thus, the claim follows. □

Proof of Theorem 6.3.5. By Proposition 6.3.2, $\bar{p}^\star$ coincides with the minimum of (6.9). Similarly, by Proposition 6.3.4 $\underline{d}^\star$ coincides with the maximum of (6.11). Note that problems (6.9) and (6.11) only differ by the order of minimization and maximization. Note also that $\mathcal{U}$ is convex and closed, $\mathcal{G}_W$ and $\mathcal{G}_V$ are convex and compact by virtue of [Ngu+23, Lemma A.6], and the (identical) trace terms in (6.9) and (6.11) are bilinear in (W,V) and (U,q). The claim thus follows from Sion's minimax theorem [Sio58]. □

6.8.3. Proofs of Section 6.4

Note that Proposition 6.4.1 is consistent with Corollary 11 because the optimal LQG controller corresponding to $\mathbb{P}^\star$ is linear in the past observations.

Proof of Proposition 6.4.1. By [Ngu+23, Lemma A.3], the inner problem in (6.9) admits a maximizer $(W^\star,V^\star)$ with $X_0^\star\succeq\lambda_{\min}(\hat{X}_0)$ as well as $W_t^\star\succeq\lambda_{\min}(\hat{W}_t)$ and $V_t^\star\succeq\lambda_{\min}(\hat{V}_t)$ for all $t\in[T-1]$. Thus, the optimal value of problem (6.9) and its strong dual (6.11) does not change if we restrict $\mathcal{G}_W$ and $\mathcal{G}_V$ to $\mathcal{G}_W^+$

and $\mathcal{G}_V^+$, respectively. We may thus conclude that problem (6.11) has a maximizer $(W^\star, V^\star)$ with $V_t^\star \succeq \lambda_{\min}(\hat{V}_t) \succ 0$ for all $t \in [T-1]$. This in turn implies that problem (6.6) is solved by a normal distribution $\mathbb{P}^\star$ under which the covariance matrix of the observation noise v_t satisfies $V_t^\star \succ 0$ for every $t \in [T-1]$. As (6.5) and (6.6) are strong duals, the optimal solution $u^\star$ of problem (6.5) forms a Nash equilibrium with $\mathbb{P}^\star$, i.e., $u^\star$ is a best response to $\mathbb{P}^\star$ and thus solves the *classic* LQG problem corresponding to $\mathbb{P}^\star$. As $R_t \succ 0$ for every $t \in [T-1]$, this best response $u^\star$ is unique, and as $V_T^\star \succ 0$ for every $t \in [T-1]$, $u^\star$ is in fact the Kalman filter-based optimal output-feedback strategy corresponding to $\mathbb{P}^\star$ (which can be obtained using the techniques highlighted in Appendix §6.6). □

Before proving Proposition 6.4.2, recall that $f(W,V)$ is called β-smooth for some $\beta > 0$ if for all $W, W' \in \mathcal{G}_W^+,\ V, V' \in \mathcal{G}_V^+$

$$|\nabla f(W,V) - \nabla f(W',V')| \le \beta \left(\|W - W'\|_F^2 + \|V - V'\|_F^2\right)^{\frac{1}{2}}$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

Proof of Proposition 6.4.2. The function $f(W,V)$ is concave because the objective function of the inner minimization problem in (6.11) is linear (and hence concave) in W and V and because concavity is preserved under minimization. To prove that $f(W,V)$ is β-smooth, we first recall from Proposition 6.3.3 that it coincides with the optimal value of the inner minimization problem in (6.10). As $\mathcal{U}_\eta = \mathcal{U}_y$, $f(W,V)$ can thus be viewed as the optimal value of the classic LQG problem corresponding to the normal distribution $\mathbb{P}$ determined by the covariance matrices W and V. Hence, $f(W,V)$ coincides with (6.17), where Σ_t, for $t \in [T-1]$, is a function of (W,V) defined recursively through the Kalman filter equations (6.16). Note that all inverse matrices in (6.16) are well-defined because any $V \in \mathcal{G}_V^+$ is strictly positive definite. Therefore, Σ_t constitutes a proper rational function (that is, a ratio of two polyonmials with the polynomial in the denominator being strictly positive) for every $t \in [T-1]$. Thus, $f(W,V)$ is infinitely often continuously differentiable on a neighborhood of $\mathcal{G}_W^+ \times \mathcal{G}_V^+$.

As $f(W,V)$ is concave and (at least) twice continuously differentiable, it is β-smooth on $\mathcal{G}_W^+ \times \mathcal{G}_V^+$ if and only if the largest eigenvalue of the Hessian matrix of $-f(W,V)$ is bounded above by β throughout $\mathcal{G}_W^+ \times \mathcal{G}_V^+$. Also, the largest eigenvalue of the positive semidefinite Hessian matrix $\nabla^2(-f(W,V))$ coincides with the spectral norm of $\nabla^2 f(W,V)$. We may thus set

$$\beta = \sup_{W \in \mathcal{G}_W^+, V \in \mathcal{G}_V^+} \|\nabla^2 f(W,V)\|_2, \tag{6.19}$$

where $\|\cdot\|_2$ denotes the spectral norm. The supremum in the above maximization problem is finite and attained thanks to Weierstrass' theorem, which applies because $f(W,V)$ is twice continuously differentiable and the spectral norm is continuous, while the sets $\mathcal{G}_W^+$ and $\mathcal{G}_V^+$ are compact by virtue of [Ngu+23, Lemma A.6]. This observation completes the proof. □

6.9. SDP Reformulation of the Lower Bounding Problem

Instead of solving the dual problem (6.11) with the customized Frank-Wolfe algorithm of Section 6.4, it can be reformulated as an SDP amenable to off-the-shelf solvers. This reformulation is obtained by dualizing the inner minimization problem and by exploiting the following preliminary lemma.

Lemma 6.9.1. *For any $\hat{Z} \in \mathbb{S}_+^d$ and $\rho_z \geq 0$, the set $\mathcal{G}_Z = \{Z \in \mathbb{S}_+^d : \mathbb{G}(Z,\hat{Z}) \leq \rho_z\}$ coincides with*

$$\left\{ Z \in \mathbb{S}_+^d : \exists E_z \in \mathbb{S}_+^d \text{ with } \operatorname{Tr}(Z + \hat{Z} - 2E_z) \leq \rho_z^2,\ \begin{bmatrix} \hat{Z}^{\frac{1}{2}} Z \hat{Z}^{\frac{1}{2}} & E_z \\ E_z & I \end{bmatrix} \succeq 0 \right\}.$$

Proof of Lemma 6.9.1. By Definition 12, we have

$$\mathcal{G}_Z = \{Z \in \mathbb{S}_+^d : \operatorname{Tr}(Z + \hat{Z} - 2(\hat{Z}^{\frac{1}{2}} Z \hat{Z}^{\frac{1}{2}})^{\frac{1}{2}}) \leq \rho_z^2\}.$$

Next, introduce an auxiliary variable $E_z \in \mathbb{S}_+^d$ subject to the matrix inequality $E_z^2 \preceq (\hat{Z}^{\frac{1}{2}} Z \hat{Z}^{\frac{1}{2}})$. By [Bel68, Theorem 1], this inequality can be recast as $E_z \preceq (\hat{Z}^{\frac{1}{2}} Z \hat{Z}^{\frac{1}{2}})^{\frac{1}{2}}$. Hence, we can reformulate the nonlinear matrix inequality in the above representation of $\mathcal{G}_Z$ as $\operatorname{Tr}(Z + \hat{Z} - 2E_z) \leq \rho_z^2$. A standard Schur complement argument reveals that the inequality $E_z^2 \preceq (\hat{Z}^{\frac{1}{2}} Z \hat{Z}^{\frac{1}{2}})$ is also equivalent to

$$\begin{bmatrix} \hat{Z}^{\frac{1}{2}} Z \hat{Z}^{\frac{1}{2}} & E_z \\ E_z & I \end{bmatrix} \succeq 0.$$

The claim then follows by combining all of these insights. □

We are now ready to derive the desired SDP reformulation of problem (6.11).

Proposition 6.9.2. *If $\hat{V} \succ 0$, then problem (6.11) is equivalent to the SDP*

$$\begin{aligned}
\max \quad & \operatorname{Tr}(G^\top QGW) - \operatorname{Tr}(F(R+H^\top QH)^{-1}) \\
\text{s.t.} \quad & W \in \mathbb{S}_+^{n(T+1)},\ V \in \mathbb{S}_+^{pT},\ M \in \mathcal{M},\ F \in \mathbb{S}_+^{Tm} \\
& E_{x_0} \in \mathbb{S}_+^n,\ E_{w_t} \in \mathbb{S}_+^n,\ E_{v_t} \in \mathbb{S}_+^p \ \ \forall t \in [T-1] \\
& \operatorname{Tr}(W_0 + \hat{X}_0 - 2E_{x_0}) \le \rho_{x_0}^2, \\
& \operatorname{Tr}(W_{t+1} + \hat{W}_t - 2E_{w_t}) \le \rho_{w_t}^2,\ \ \operatorname{Tr}(V_t + \hat{V}_t - 2E_{v_t}) \le \rho_{v_t}^2 \ \ \forall t \in [T-1] \\
& \begin{bmatrix} \hat{X}_0^{\frac{1}{2}} X_0 \hat{X}_0^{\frac{1}{2}} & E_{x_0} \\ E_{x_0} & I_n \end{bmatrix} \succeq 0, \\
& \begin{bmatrix} \hat{W}_t^{\frac{1}{2}} W_{t+1} \hat{W}_t^{\frac{1}{2}} & E_{w_t} \\ E_{w_t} & I_n \end{bmatrix} \succeq 0,\ \ \begin{bmatrix} \hat{V}_t^{\frac{1}{2}} V_t \hat{V}_t^{\frac{1}{2}} & E_{v_t} \\ E_{v_t} & I_p \end{bmatrix} \succeq 0 \ \ \forall t \in [T-1] \\
& \begin{bmatrix} F & H^\top QGWD^\top + M/2 \\ (H^\top QGWD^\top + M/2)^\top & DWD^\top + V \end{bmatrix} \succeq 0 \\
& W_0 \succeq \lambda_{\min}(\hat{X}_0)I, \quad W_{t+1} \succeq \lambda_{\min}(\hat{W}_t)I, \quad V_t \succeq \lambda_{\min}(\hat{V}_t)I \quad \forall t \in [T-1].
\end{aligned} \tag{6.20}$$

Here, $\mathcal{M}$ denotes the set of all strictly upper block triangular matrices of the form

$$\begin{bmatrix} 0 & M_{1,2} & M_{1,3} & \dots & M_{1,T} \\ & 0 & M_{2,3} & & M_{2,T} \\ & & \ddots & & \vdots \\ & & & 0 & M_{T-1,T} \\ & & & & 0 \end{bmatrix} \in \mathbb{R}^{Tm \times Tp},$$

where $M_{t,s} \in \mathbb{R}^{m \times p}$ for every $t, s \in \mathbb{Z}$ with $1 \le t < s \le T$.

Proof of Proposition 6.9.2. The proof relies on dualizing the inner minimization problem in (6.11). Note that strong duality holds because the primal problem is trivially feasible and involves only equality constraints, which implies that any feasible point is in fact a Slater point. In the following we use $M \in \mathcal{M}$ to denote the Lagrange multiplier of the constraint $U \in \mathcal{U}$, which requires all blocks of the matrix U above the main diagonal to vanish. The Lagrangian function of the inner minimization problem in (6.11) can therefore be represented as

$$\begin{aligned}
\mathcal{L}(q,U,M) = {} & \operatorname{Tr}\left(\big(D^\top U^\top (R+H^\top QH)UD + G^\top QG\big)W\right) + 2\operatorname{Tr}(G^\top QHUDW) \\
& + \operatorname{Tr}\left(\big(U^\top (R+H^\top QH)U\big)V\right) + q^\top (R+H^\top QH)q + \operatorname{Tr}(UM^\top).
\end{aligned}$$

Recall now that $R \succ 0$ and $Q \succeq 0$, and thus $R + H^\top QH \succ 0$. Consequently, $\mathcal{L}$ is minimized by $q^\star = 0$ for any fixed U and M. In addition, the partial gradient of $\mathcal{L}$ with respect U is given by

$$\frac{\partial \mathcal{L}}{\partial U} = 2(R + H^\top QH)UDWD^\top + 2(R + H^\top QH)UV + 2H^\top QGWD^\top + M.$$

Recall also that $V \in \mathcal{G}_V^+$ is strictly positive, which implies that $DWD^\top + V \succ 0$ is invertible. As we already know that $R + H^\top QH \succ 0$ is invertible, as well, $\mathcal{L}$ is minimized by

$$U^\star = -(R + H^\top QH)^{-1} \left(H^\top QGWD^\top + M/2\right)(DWD^\top + V)^{-1}$$

for any fixed M. Substituting both $q^\star$ and $U^\star$ into $\mathcal{L}$ yields the dual objective function

$$\begin{aligned}&g(M) = \mathcal{L}(q^\star, U^\star, M) = \mathrm{Tr}(G^\top QGW)\\ &- \mathrm{Tr}\left((R + H^\top QH)^{-1}(H^\top QGWD^\top + M/2)(DWD^\top + V)^{-1}(H^\top QGWD^\top + M/2)^\top\right).\end{aligned}$$

The dual of the inner minimization problem in (6.11) is thus given by $\max_{M\in\mathcal{M}} g(M)$. To linearize the dual objective function, we next introduce an auxiliary variable $F \in \mathbb{S}_+^{mT}$ subject to the matrix inequality $F \succeq (H^\top QGWD^\top + M/2)(DWD^\top + V)^{-1}(H^\top QGWD^\top + M/2)^\top$. By using a standard Schur complement reformulation, we can then rewrite the dual problem as

$$\begin{array}{ll}\max & \mathrm{Tr}(G^\top QGW) - \mathrm{Tr}((R + H^\top QH)^{-1}F)\\ \text{s.t.} & M \in \mathcal{M},\ F \in \mathbb{S}_+^{mT}\\ & \begin{bmatrix} F & H^\top QGWD^\top + M/2\\ (H^\top QGWD^\top + M/2)^\top & DWD^\top + V\end{bmatrix} \succeq 0.\end{array} \tag{6.21}$$

Next, by replacing the inner problem in (6.11) with its strong dual (6.21), we can reformulate (6.11) as

$$\begin{array}{ll}\max & \mathrm{Tr}(G^\top QGW) - \mathrm{Tr}((R + H^\top QH)^{-1}F)\\ \text{s.t.} & M \in \mathcal{M},\ F \in \mathbb{S}_+^{mT},\ W \in \mathbb{S}_+^{n(T+1)},\ V \in \mathbb{S}_+^{pT}\\ & \begin{bmatrix} F & H^\top QGWD^\top + M/2\\ (H^\top QGWD^\top + M/2)^\top & DWD^\top + V\end{bmatrix} \succeq 0\\ & \mathbb{G}(X_0, \hat{X}_0)^2 \le \rho_{x_0}^2,\ \mathbb{G}(W_t, \hat{W}_t) \le \rho_{w_t}^2,\ \mathbb{G}(V_t, \hat{V}_t) \le \rho_{v_t}^2 \quad \forall t \in [T-1].\end{array} \tag{6.22}$$

By Proposition 6.4.1, the inclusion of the constraints $X_0 \succeq \lambda_{\min}(\hat{X}_0)I$, $W_t \succeq \lambda_{\min}(\hat{W}_t)I$ and $V_t \succeq \lambda_{\min}(\hat{V}_t)I$ for all $t \in [T-1]$ has no effect on the solution to problem (6.22). In addition, by Lemma 6.9.1, each (non-linear) Gelbrich constraint in (6.22) can be reformulated as an equivalent (linear) SDP constraint. Thus, problem (6.22) reduces to (6.20), and the claim follows. □

6.10. Bisection Algorithm for the Linearization Oracle

We now show that the direction-finding subproblem (6.14) can be solved efficiently via bisection. To this end, we first establish that (6.14) can be reduced to the solution of a univariate algebraic equation.

Proposition 6.10.1 ([Ngu+23, Proposition A.4 (iii)]). *If $\hat{Z} \in \mathbb{S}^d_{++}$, $\Gamma_Z \in \mathbb{S}^d_{+}$, $\Gamma_Z \neq 0$ and $\rho_z \in \mathbb{R}_{++}$, then*

$$\begin{array}{ll} \max & \langle \Gamma_Z, L - Z \rangle \\ \text{s.t.} & \mathbb{G}(L, \hat{Z}) \leq \rho_z \\ & L \succeq \lambda_{\min}(\hat{Z}) I \end{array} \tag{6.23}$$

is uniquely solved by $L^\star = (\gamma^\star)^2(\gamma^\star I - \Gamma_Z)^{-1}\hat{Z}(\gamma^\star I - \Gamma_Z)^{-1}$, where $\gamma^\star$ is the unique solution of

$$\rho_z^2 - \langle \hat{Z}, (I - \gamma^\star(\gamma^\star I - \Gamma_Z)^{-1})^2 \rangle = 0 \tag{6.24}$$

in the interval $(\lambda_{\max}(\Gamma_Z), \infty)$.

In practice, we need to solve the algebraic equation (6.24) numerically. The numerical error in approximating $\gamma^\star$ should be contained to ensure that $L^\star$ approximates the exact maximizer of problem (6.23). The next proposition shows that, for any tolerance $\delta \in (0, 1)$, a δ-approximate solution of (6.23) can be computed with an efficient bisection algorithm.

Proposition 6.10.2 ([Ngu+23, Theorem 6.4]). *For any fixed $\rho_z \in \mathbb{R}_{++}, \hat{Z} \in \mathbb{S}^d_{++}$ and $\Gamma_Z \in \mathbb{S}^d_{+}, \Gamma_Z \neq 0$, define $\mathcal{G}_Z^+ = \{Z \in \mathbb{S}^d_+ : \mathbb{G}(Z, \hat{Z}) \leq \rho_z, Z \succeq \lambda_{\min}(\hat{Z})\}$ as the feasible set of problem (6.23), and let $Z \in \mathcal{G}_Z^+$ be any reference covariance matrix. Additionally, let $\delta \in (0, 1)$ be the desired oracle precision, and define $\varphi(\gamma) = \gamma(\rho^2 + \langle \gamma(\gamma I - \Gamma_Z)^{-1} - I, \hat{Z} \rangle) - \langle Z, \Gamma_Z \rangle$ for any $\gamma > \lambda_{\max}(\Gamma_Z)$. Then, Algorithm 7 returns in finite time a matrix $L_Z^\delta \in \mathbb{S}^d_+$ with the following properties. (i) Feasibility: $L_Z^\delta \in \mathcal{G}_Z^+$ (ii) δ-Suboptimality: $\langle L_Z^\delta - Z, \Gamma_Z \rangle \geq \delta \max_{L \in \mathcal{G}_Z^+} \langle \Gamma_Z, L - Z \rangle$.*

In summary, for any $Z \in \{X_0, W_0, \ldots, W_{T-1}, V_0, \ldots, V_{T-1}\}$, Algorithm 7 computes a δ-approximate solutions to the direction-finding subproblem (6.14) with $\Gamma_Z = \nabla_Z f(W, V)$.

6.11. Additional Information on Experiments

Generation of Nominal Covariance Matrices. The nominal covariance matrices of the exogenous uncertainties are constructed randomly using the

Algorithm 7 Bisection algorithm to compute L_Z^δ

```
Input: nominal covariance matrix $\hat{Z} \in \mathbb{S}^d_{++}$, radius $\rho \in \mathbb{R}_{++}$,
       reference covariance matrix $Z \in \mathcal{G}_Z^+$,
       gradient matrix $\Gamma_Z \in \mathbb{S}^d_+$, $\Gamma_Z \neq 0$, precision $\delta \in (0,1)$,
       dual objective function $\phi(\gamma)$ defined in Proposition 6.10.2
set $\lambda_1 \leftarrow \lambda_{\max}(\Gamma_Z)$, and let $p_1$ be an eigenvector for $\lambda_1$
set $\underline{\gamma} \leftarrow \lambda_1(1 + (p_1^\top \hat{Z} p_1)^{\frac{1}{2}}/\rho)$ and $\overline{\gamma} \leftarrow \lambda_1(1 + \mathrm{Tr}(\hat{Z})^{\frac{1}{2}}/\rho)$
repeat
    set $\tilde{\gamma} \leftarrow (\overline{\gamma} + \underline{\gamma})/2$ and $L \leftarrow (\tilde{\gamma})^2(\tilde{\gamma} I - \Gamma_Z)^{-1}\hat{Z}(\tilde{\gamma} I - \Gamma_Z)^{-1}$
    if $\frac{\mathrm{d}\phi}{\mathrm{d}\gamma}(\tilde{\gamma}) < 0$ then set $\underline{\gamma} \leftarrow \tilde{\gamma}$ else $\overline{\gamma} \leftarrow \tilde{\gamma}$ endif
until $\frac{\mathrm{d}\phi}{\mathrm{d}\gamma}(\tilde{\gamma}) > 0$ and $\langle L - Z, \Gamma_Z \rangle \geq \delta\phi(\tilde{\gamma})$
Output: $L$
```

following procedure. For each exogenous uncertainty $z \in \{x_0, w_0, \ldots, w_{T-1}, v_0, \ldots, v_{T-1}\}$, we denote the dimension of z by d and sample a matrix $M_Z \in \mathbb{R}^{d\times d}$ from the uniform distribution on the hypercube $[0,1]^{d\times d}$. Next, we define $\Xi_Z \in \mathbb{R}^{d\times d}$ as the orthogonal matrix whose columns represent the orthonormal eigenvectors of the symmetric matrix $M_Z + M_Z^\top$. Finally, we set $\hat{Z} = \Xi_Z \Lambda_Z \Xi_Z^\top$, where Λ_Z is a diagonal matrix whose main diagonal is sampled uniformly from the interval $[1,2]^d$. The rationale for adopting this cumbersome procedure is to ensure that the covariance matrix $\hat{Z}$ is positive definite.

Optimality Gap. The optimality gap of the Frank-Wolfe algorithm visualized in Figure 6.1b is calculated as the sum of the surrogate optimality gaps $\langle L_Z^\delta - Z, \nabla_Z f(W,V)\rangle$ across all $Z \in \{X_0, W_0 \ldots, W_{T-1}, V_0, \ldots, V_{T-1}\}$. For more information on the surrogate optimality gaps see [Jag13].

7. Conclusion

Now I mark a temporary pause in my exploration of reliable data-driven decision making and reflect on and summarize the progress we have made thus far.

Computation of Optimal Transport. At the beginning of this thesis, in late 2018, it was clear that optimal transport has applications across numerous domains. Surprisingly perhaps its computational complexity was not established. We fill this gap by formally establishing its hardness in Part I both for discrete and semi-discrete optimal transport problem. One might ponder the merit of investigating 'impossibility results' such as delving deep into computational complexity, which might be seen as mere markers of what we cannot achieve. However, by understanding these mathematical constraints, we not only delineate the boundaries what is feasible but also sharpen our research objectives. In essence, by acknowledging and studying our limitations, we can set clearer, more achievable goals, ultimately advancing our collective pursuit of knowledge. Indeed, at second part of each chapter we provide numerical solutions to approximately solve both problems, as we have already shown that we do not have hope for solving these problems in polynomial time.

Robust Domain Adaptation. In Chapter 3, we investigate strategies to synthesize a family of least squares estimator experts that are robust with regard to moment conditions for supervised domain adaptation problem. When these moment conditions are specified using Kullback-Leibler divergence or optimal transport, we can find robust estimators efficiently using convex optimization. The theoretical and experimental results in this chapter suggest that IR-WASS and SI-WASS are attractive schemes to generate a family of robust least squares experts. Moreover, the IR-WASS and SI-WASS experts are extremely easy to compute because it requires solving only a second-order cone or a linear semidefinite program. We observe that KL-type divergence schemes are less numerically stable due to the computation of the log-determinant and the inverse of a nearly singular covariance matrix $\widehat{\Sigma}_{\mathrm{T}}$. Setting the parameters for KL-type divergence schemes is also harder due to the asymmetry of the divergence $\mathbb{D}$. While this chapter focuses solely on *interpolating* schemes, it would also be interesting to explore *extrapolating* schemes in future research.

Learning Fair and Robust Models. The proliferation of artificial intelligence, particularly large language models, into our everyday lives, underscores the urgent need to ensure these systems are fair. As they become more intertwined with our decision-making processes in content recommendations, customer service, and even personal assistants, the ramifications of any biases within these models grow significantly. In Chapter 4, we propose distributionally robust fair classification models that prevents discrimination with respect to sensitive attributes such as gender or ethnicity. However, given the pervasive nature of these technologies, the first-order challenge has now expanded to include understanding fairness and the essential task of auditing these models for any form of prejudice. The ultimate objective of fair machine learning is to address the complex question: "*What is fair?*" Defining fairness is a challenge even in tangible, real-world situations. Thus, creating a mathematical representation for this abstract concept amplifies the difficulty. While there is no universal agreement on the exact definition, the concept of discrimination varies significantly across different domains. A prominent approach could involve collaborating with policymakers and domain experts, particularly in criminal justice, medicine, and education. By combining these insights, one could aim to develop a principled methodology to shed considerable light on this longstanding question and contribute to designing commercial auditing mechanisms for AI systems.

Auditing for Fairness. In Chapter 5, we propose a statistical hypothesis test for group fairness of classification algorithms based on the theory of optimal transport. Our test statistic relies on computing the projection distance from the empirical distribution supported on the test samples to the manifold of distributions that renders the classifier fair. When the notion of fairness is chosen to be either the probabilistic equal opportunity or the probabilistic equalized odds, we show that the projection can be computed efficiently. We provide the limiting distribution of the test statistic and show that our Wasserstein projection test is asymptotically correct. Our proposed test also offers the flexibility to incorporate the geometric information of the feature space into testing procedure. Finally, analyzing the most favorable distribution can help interpreting the reasons behind the outcome of the test.

The Wasserstein projection hypothesis test is the culmination of a benevolent motivation and effort, and it aims to furnish the developers, the regulators and the general public a quantitative method to verify certain notions of fairness in the classification setting. At the same time, we acknowledge the risks and limitations of the results presented in this chapter. First, it is essential to keep in mind that this chapter focuses on *probabilistic* notions of fairness, in particular, we provide the Wasserstein statistical test for probabilistic equality of opportunity and probabilistic equalized odds. Probabilistic notions are only approximations of the original definitions, and the employment of probabilistic notions are solely for the technical purposes. Due to the sensitivity of the test result on the choice of fairness notions, a test that is designed for probabilistic notions may not be applicable to test for original notions of fairness due to the interplay with the threshold τ and the radical difference of both the test statistic and the limiting distribution. If a logistic classifier h_β is rejected using our framework for probabilistic equal opportunity, it does *not* necessarily imply that the classifier h_β fails to satisfy the equal opportunity criterion, and vice versa. The same argument holds when we test for probabilistic equalized odds. Second, the outcome of the Wasserstein projection test is dependent on the choice of the underlying metric on the feature, the sensitive attribute and the label spaces. Indeed, the test outcome can change if we switch the metric of the feature space, for example, from the Euclidean norm to a 1-norm. In the scope of this chapter, we do not study how sensitive the test outcome is with respect to the choice of the metric, nor can we make any recommendation on the optimal choice of the metric. Nevertheless, it is reasonable to recommend that the metric should be chosen judiciously, and the action of tuning the metric in order to obtain favorable test outcome should be prohibited. Third, to simplify the computation, we have assumed absolute trust on the sensitive attributes and the label. The users of our test should be mindful if there is potential corruption to these values. Moreover, our test is constructed under the assumption that there is no missing values in the test data. This assumption, unfortunately, may not hold in real-world implementations. Constructing statistical test which is robust to adversarial attacks and missing data using the Wasserstein projection framework is an interesting research direction. Fourth, the statistical test in this chapter is for a simple null hypothesis. In practice, the regulators may be interested in a relaxed fairness test in which the difference of the conditional expectations is upper bounded by a fixed positive constant ϵ. The extension of the Wasserstein hypothesis testing framework for a composite null hypothesis is non-trivial, thus we leave this idea for future study. Finally, any auditing process for algorithmic fairness can become a dangerous tool if it falls into the hand of unqualified or vicious inspectors. The results in this chapter are developed to broaden our scientific understanding, and we recommend that the test and its outcomes should be used as an informative reference, but *not* as an absolute certification to promote any particular classifier or as a justification for any particular classification decision.

We thus sincerely recommend that the tools proposed in this chapter be exercised with utmost consideration.

Robust Control. In Chapter 6, we consider a generalization of the discrete-time, finite-horizon linear quadratic Gaussian control problem, where the noise distributions are unknown and belong to optimal transport-based ambiguity sets centered at nominal (Gaussian) distributions. The objective is to minimize a worst-case cost across all distributions in the ambiguity set, including non-Gaussian distributions. Despite the added complexity, we prove that a control policy that is linear in the observations is optimal for this problem, as in the classic LQG problem. We propose a numerical solution method that efficiently characterizes this optimal control policy using Frank-Wolfe algorithm to identify the worst-case distributions and computing the optimal control policy using Kalman filter estimation under these distributions.

In view of the popularity of LQG models, the results in this work carry important theoretical and practical implications. Despite considering a generalization of the classic LQG setting where the noise affecting the system dynamics and the observations follows unknown (and potentially non-Gaussian) distributions, our findings suggest that certain classic structural results continue to hold and that highly efficient methods can be adapted to tackle this more realistic (and more challenging) problem. Specifically, that control policies depending linearly on observations continue to be optimal and that the worst-case distribution turns out to be Gaussian is surprising from a theoretical angle and also has direct practical implications, because it allows leveraging the highly efficient Kalman filter in conjunction with dynamic programming and a Frank-Wolfe method to design an efficient computational procedure for solving the problem.

The results also raise several important questions that warrant future exploration. First, it would be highly relevant to consider extensions where the system matrices are also affected by uncertainty, as this captures many applications of practical interest in, e.g., reinforcement learning or revenue management. Second, it would be worth exploring an infinite horizon setting or relaxing the assumption that the nominal distribution is Gaussian, as both assumptions may be limiting the practical appeal of the framework. Third, one could also attempt to prove structural optimality results or design novel algorithms for generating high-quality suboptimal solutions for the more general setting involving constraints on states and/or control inputs. Lastly, one could improve the presented algorithmic proposal by exploiting topological properties of the objective so as to guarantee linear convergence rates in the Frank-Wolfe procedure.

The fact that our ambiguity set for the distributionally robust linear quadratic Gaussian control problem contains non-Gaussian distributions sheds light on the challenging problem of controlling a linear-quadratic system that does not have Gaussian noise by upper bounding its optimal cost with the optimal cost of the distributionally robust linear quadratic control problem. Nevertheless, a notable limitation of this direction stands out: identifying if an arbitrary distribution falls within an optimal transport-based ambiguity set is a computationally challenging task [TSAK23].

Academic CV

Bahar Taşkesen
Contact: bahar.taskesen@epfl.ch, www.bahartaskesen.com

Education

Doctor of Science in Risk Analytics and Optimization.
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
December 2018 - June 2024
Advisor: Prof. Daniel Kuhn

Bachelor of Science in Electrical and Electronics Engineering
Middle East Technical University (METU), Turkey
September 2014 - June 2018

Publications

Bahar Taşkesen, Dan A Iancu, Çağıl Koçyiğit, and Daniel Kuhn. Distributionally Robust Linear Quadratic Control. *NeurIPS (Spotlight)* (arXiv:2305.17037), 2023.

* **INFORMS 2023 George Nicholson Student Paper Competition Finalist**

Jose Blanchet, Daniel Kuhn, Jiajin Li, and **Bahar Taşkesen**. Unifying Distributionally Robust Optimization via Optimal Transport Theory. (arXiv:2308.-05414), 2023.

Bahar Taşkesen, Soroosh Shafieezadeh-Abadeh, and Daniel Kuhn. Semi-discrete Optimal Transport: Hardness, Regularization and Numerical Solution. *Mathematical Programming*, 199(1-2):1033–1106, 2023.

⋆ **Winner of the 2022 INFORMS Optimization Society Student Paper Prize**

Bahar Taşkesen, Soroosh Shafieezadeh-Abadeh, Daniel Kuhn, and Karthik Natarajan. Discrete Optimal Transport with Independent Marginals is #P-Hard. *SIAM Journal on Optimization*, 33(2):589– 614, 2023.

Yves Rychener, **Bahar Taşkesen**, and Daniel Kuhn. Metrizing Fairness, 2023.

Bahar Taşkesen, Man-Chung Yue, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen. Sequential Domain Adaptation by Synthesizing Distributionally Robust Experts. In *International Conference on Machine Learning*, pages 10162–10172, 2021.

⋆ **Oral Presentation - Top 3% of submissions**

Bahar Taşkesen, Jose Blanchet, Daniel Kuhn, and Viet Anh Nguyen. A Statistical Test for Probabilistic Fairness. In *ACM Conference on Fairness, Accountability, and Transparency*, pages 648–665, 2021.

Bahar Taşkesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. A Distributionally Robust Approach to Fair Classification. (arXiv:2007.09530), 2020.

Bahar Taşkesen, Alper Koz, Aydin Alatan, and Olivier Weatherbee. Change Detection for Hyperspectral Images Using Extended Mutual Information and Oversegmentation. In *Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, 2018.

Bahar Taşkesen and Anthony George Constantinides. Interactive Image Reconstruction from Irregular and Imprecise Fragments. In *Electronic Imaging and the Visual Arts Conference Florence*, 2018.

Bahar Taşkesen, Beril Beşbinar, Alper Koz, and Aydin Alatan. Unsupervised Change Detection in Satellite Images Using Oversegmentation and Mutual Information. In *Signal Processing and Communications Applications Conference*, 2017.

Teaching Experience

Teaching Assistant:

EPFL, Optimal Decision Making	*Spring 2018, Spring 2019*
EPFL, Intercultural Presentation Skills	*Fall 2019*
EPFL, Convex Optimization	*Fall 2019, Fall 2021, Fall 2022*

Awards

Finalist. INFORMS George Nicholson Student Paper Competition	*2023*
First Place. INFORMS Optimization Society Student Paper Prize	*2022*
First Place. The EDMT Research Day - Best Research Day Presentation	*2022*
Runner-up. ECSO-CMS Best Student Paper Prize	*2022*

First Place. METU Engineering Day - Poster competition *2017*
First Place. Electrical and Electronics Engineering Undergraduate Student Academic Research Program (STAR EEE) - Poster Competition *2016*

Reviewing Activities

Journals:
Computational Optimization and Applications, European Journal of Operational Research, IET Generation, Transmission & Distribution, IEEE Transactions on Signal Processing, INFORMS Journal on Computing, Journal of Optimization Theory and Applications, Management Science, Mathematical Programming, Mathematics of Operations Research, Neural Computation, Operations Research Letters, SIAM Journal on Optimization, Transportation Science.
Conferences:
NeurIPS Workshop: Algorithmic Fairness through the Lens of Causality and Interpretability (2020), IEEE CDC (2021), NeurIPS Datasets and Benchmarks Track (2021), ICLR (2022), NeurlPS (2022, 2023 [Top Reviewer]), ICML (2023), ACM FAccT (2023)

Academic Visits

Undergraduate Intern
Communications and Signal Processing Lab
Imperial College London *June 2017 - September 2017*
Advisor: Prof. A.G. Constantinides

Undergraduate researcher
OGAM (METU Center for Image Analysis) *March 2016 - June 2018*
Advisor: Prof. Aydin Alatan
CEMMETU (Computational Electromagnetics at METU) *March 2016 -*

January 2017
Advisor: Dr. Ozgur Ergul

Talks

Invited Presentations:
YOUNG Online Seminar Series Machine Learning NeEDS Mathematical Optimization, *Online*, *2024*
Northwestern University, IEMS, *Chicago*, *2024*
University of Chicago, Booth School of Business, *Chicago*, *2024*
Imperial College London, Business School, *London*, *2024*
Purdue University, School of Industrial Engineering, *West Lafayette*, *2024*
Bilkent University, Industrial Engineering Department, *Ankara*, *2024*
University of Illinois Urbana-Champaign, ISE, *Champaign*, *2023*
Georgia Technical, ISyE, *Atlanta*, *2023*
University of Southern California, ISE, *Los Angeles*, *2023*

University of British Columbia, Sauder School of Business, *Vancouver*, *2023*
Georgia Technical University, ISyE, *Atlanta*, *(Spring) 2023*
The University of Luxembourg, Centre for Logistics and Supply Chain Management, *Luxembourg*, *2023*
VinAI Research Seminar Series, *Online*, *2022*

Conference Presentations:
INFORMS Annual Meeting, *Phoenix*, *2023*
SIAM Conference on Optimization, *Seattle* *2023*
INFORMS Annual Meeting, *Indianapolis* *2022*
International Conference on Continuous Optimization (ICCOPT), *Bethlehem* *2022*
International Conference on Machine Learning (ICML), *Online* *2021*
INFORMS Annual Meeting, *Online* *2021*
ACM Conference on Fairness, Accountability, and Transparency (FAccT), *Online* *2021*
European Conference on Operations Research, *Online* *2021*
INFORMS Annual Meeting, *Online* *2020*
Electronic Imaging & Visual Arts (EVA), *Florence* *2018*

Organizer
INFORMS Annual Meeting, *Phoenix* *2023*
Session: Structuring the Ambiguous: New Perspectives on Distributionally Robust Learning
SIAM Conference on Optimization, *Seattle* *2023*
Mini-symposium: Robust Learning in Static and Dynamic Environments (co-organizer)
International Conference on Continuous Optimization (ICCOPT), *Bethlehem* *2022*
Session: Optimization under Uncertainty for Machine Learning

Work Experience
Invidyo, *Ankara* (Software Engineer) *February 2018 - August 2018*
Karel Electronics R&D, *Ankara* (Intern) *June 2016 - July 2016*

Languages
English (Fluent), Turkish (Native), Italian (Currently learning)

Programming Skills
Python, C/C++, Matlab, LabView

Bibliography

[AAN16] S. D. Ahipaşaoğlu, U. Arıkan, and K. Natarajan. “On the flexibility of using marginal distribution choice models in traffic equilibrium”. *Transportation Research Part B: Methodological* 91 (2016), pp. 130–158.

[ABA21] J. M. Altschuler and E. Boix-Adsera. “Hardness results for Multimarginal Optimal Transport problems”. *Discrete Optimization* 42 (2021), p. 100669.

[ABA22] J. M. Altschuler and E. Boix-Adsera. “Wasserstein barycenters are NP-hard to compute”. *SIAM Journal on Mathematics of Data Science* 4.1 (2022), pp. 179–203.

[AC11] M. Agueh and G. Carlier. “Barycenters in the Wasserstein space”. *SIAM Journal on Mathematical Analysis* 43.2 (2011), pp. 904–924.

[ACB17] M. Arjovsky, S. Chintala, and L. Bottou. “Wasserstein generative adversarial networks”. *International Conference on Machine Learning.* 2017, pp. 214–223.

[Adl+17] J. Adler, A. Ringh, O. Öktem, and J. Karlsson. “Learning to solve inverse problems using Wasserstein loss”. *arXiv:1710.10898* (2017).

[ADPT88] S. P. Anderson, A. De Palma, and J.-F. Thisse. “A representative consumer theory of the logit model”. *International Economic Review* 29.3 (1988), pp. 461–466.

[AG18] B. K. Abid and R. Gower. “Stochastic algorithms for entropy-regularized optimal transport problems”. *Artificial Intelligence and Statistics.* 2018, pp. 1505–1512.

[AG22] R. Arora and R. Gao. “Data-driven multistage distributionally robust optimization with nested distance: Time consistency and tractable dynamic reformulations”. *Available at Optimization Online* (2022).

[Agg+22] R. Aggarwal, K. Bibbins-Domingo, R. W. Yeh, Y. Song, N. Chiu, R. K. Wadhera, C. Shen, and D. S. Kazi. “Diabetes screening by race and ethnicity in the United States: Equivalent body mass index and age thresholds”. *Annals of Internal Medicine* 175.6 (2022), pp. 765–773.

[Agr+18] A. Agrawal, R. Verschueren, S. Diamond, and S. Boyd. “A rewriting system for convex optimization problems”. *Journal of Control and Decision* 5.1 (2018), pp. 42–60.

[AHA98] F. Aurenhammer, F. Hoffmann, and B. Aronov. “Minkowski-type theorems and least-squares clustering”. *Algorithmica* 20.1 (1998), pp. 61–76.

[ALN18] S. D. Ahipaşaoğlu, X. Li, and K. Natarajan. “A convex optimization approach for computing correlated choice probabilities with many alternatives”. *IEEE Transactions on Automatic Control* 64.1 (2018), pp. 190–205.

[Amb+18] L. Ambrogioni, U. Guclu, Y. Gucluturk, and M. van Gerven. “Wasserstein variational gradient descent: From semi-discrete optimal transport to ensemble variational inference”. *arXiv:1811.02827* (2018).

[AMJJ18] D. Alvarez-Melis, T. Jaakkola, and S. Jegelka. “Structured Optimal Transport”. *Artificial Intelligence and Statistics.* 2018, pp. 1771–1780.

[ANWS22] J. M. Altschuler, J. Niles-Weed, and A. J. Stromme. “Asymptotics for Semidiscrete Entropic Optimal Transport”. *SIAM Journal on Mathematical Analysis* 54.2 (2022), pp. 1718–1741.

[AS20] A. Ajalloeian and S. U. Stich. “Analysis of SGD with Biased Gradient Estimators”. *arXiv:2008.00051* (2020).

[Aug+13] F. Auger, M. Hilairet, J. M. Guerrero, E. Monmasson, T. Orlowska-Kowalska, and S. Katsura. “Industrial Applications of the Kalman Filter: A Review”. *IEEE Transactions on Industrial Electronics* 60.12 (2013), pp. 5458–5471.

[AWR17] J. Altschuler, J. Weed, and P. Rigollet. “Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration”. *Advances in Neural Information Processing Systems.* 2017, pp. 1964–1974.

[Azi+19] K. Azizzadenesheli, A. Liu, F. Yang, and A. Anandkumar. “Regularized learning for domain adaptation under label shifts”. *International Conference on Learning Representations.* 2019.

[Bac10] F. Bach. “Self-concordant analysis for logistic regression”. *Electronic Journal of Statistics* 4 (2010), pp. 384–414.

[Bac14] F. Bach. “Adaptivity of Averaged Stochastic Gradient Descent to Local Strong Convexity for Logistic Regression”. *Journal of Machine Learning Research* 15.19 (2014), pp. 595–627.

[Bah+20] S. Baharlouei, M. Nouiehed, A. Beirami, and M. Razaviyayn. “Rényi fair inference”. *International Conference on Learning Representations.* 2020.

[Bak+13] M. Baktashmotlagh, M. T. Harandi, B. C. Lovell, and M. Salzmann. “Unsupervised domain adaptation by domain invariant projection”. *IEEE International Conference on Computer Vision.* 2013, pp. 769–776.

[BAL85] M. E. Ben-Akiva and S. R. Lerman. *Discrete Choice Analysis: Theory and Application to Travel Demand.* MIT Press, 1985.

[Bat+13] G. Battistelli, L. Chisci, C. Fantacci, A. Farina, and A. Graziano. “Consensus CPHD Filter for Distributed Multitarget Tracking”. *IEEE Journal of Selected Topics in Signal Processing* 7.3 (2013), pp. 508–520.

[BB00] J.-D. Benamou and Y. Brenier. “A computational fluid mechanics solution to the Monge-Kantorovich mass transfer problem”. *Numerische Mathematik* 84.3 (2000), pp. 375–393.

[BD+07] S. Ben-David, J. Blitzer, K. Crammer, F. Pereira, et al. “Analysis of representations for domain adaptation”. *Advances in Neural Information Processing Systems* 19 (2007), p. 137.

[Bel+18] R. K. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilovic, et al. “AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias”. *arXiv:1810.01943* (2018).

[Bel68] R. Bellman. “Some inequalities for the square root of a positive definite matrix”. *Linear Algebra and its Applications* 1.3 (1968), pp. 321–324.

[Ben+15] J.-D. Benamou, G. Carlier, M. Cuturi, L. Nenna, and G. Peyré. “Iterative Bregman projections for regularized transportation problems”. *SIAM Journal on Scientific Computing* 37.2 (2015), A1111–A1138.

[Ber09a] D. S. Bernstein. *Matrix Mathematics: Theory, Facts, and Formulas*. Princeton University Press, 2009.

[Ber09b] D. Bertsekas. *Convex Optimization Theory*. Athena Scientific, 2009.

[Ber17] D. Bertsekas. *Dynamic Programming and Optimal Control*. Vol. I. Athena Scientific, 2017.

[Ber+18] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. “Fairness in criminal justice risk assessments: The state of the art”. *Sociological Methods & Research* (2018), p. 0049124118782533.

[Ber81] D. P. Bertsekas. “A new algorithm for the assignment problem”. *Mathematical Programming* 21.1 (1981), pp. 152–171.

[Ber92] D. P. Bertsekas. “Auction algorithms for network flow problems: A tutorial introduction”. *Computational Optimization and Applications* 1.1 (1992), pp. 7–66.

[BG12] D. Bertsimas and V. Goyal. “On the power and limitations of affine policies in two-stage adaptive optimization”. *Mathematical Programming* 134.2 (2012), pp. 491–531.

[BG18] J. Buolamwini and T. Gebru. “Gender shades: Intersectional accuracy disparities in commercial gender classification”. *Conference on Fairness, Accountability and Transparency*. 2018, pp. 77–91.

[BH88] D. S. Bernstein and W. M. Haddad. “LQG control with an $\mathcal{H}_\infty$ performance bound: A Riccati equation approach”. *American Control Conference*. 1988, pp. 796–802.

[Bil95] P. Billingsley. *Probability and Measure*. John Wiley and Sons, 1995.

[BIP10] D. Bertsimas, D. A. Iancu, and P. A. Parrilo. “Optimality of affine policies in multistage robust optimization”. *Mathematics of Operations Research* 35.2 (2010), pp. 363–394.

[BIP11] D. Bertsimas, D. A. Iancu, and P. A. Parrilo. “A Hierarchy of Near-Optimal Policies for Multistage Adaptive Optimization”. *IEEE Transactions on Automatic Control* 56.12 (2011), pp. 2809–2824.

[BKM19] J. Blanchet, Y. Kang, and K. Murthy. “Robust Wasserstein profile inference and applications to machine learning”. *Journal of Applied Probability* 56.3 (2019), pp. 830–857.

[BL17] Y. Bechavod and K. Ligett. “Penalizing unfairness in binary classification”. *arXiv:1707.00044* (2017).

[Bla+18] J. Blanchet, A. Jambulapati, C. Kent, and A. Sidford. “Towards optimal running times for optimal transport”. *arXiv:1810.07717* (2018).

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

[BM13] F. Bach and E. Moulines. “Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$”. *Advances in Neural Information Processing Systems*. 2013, pp. 773–781.

[BM19] J. Blanchet and K. Murthy. “Quantifying distributional model risk via optimal transport”. *Mathematics of Operations Research* 44.2 (2019), pp. 565–600.

[BMP06] J. Blitzer, R. McDonald, and F. Pereira. “Domain Adaptation with Structural Correspondence Learning”. *Conference on Empirical Methods in Natural Language Processing*. 2006, 120–128.

[Bon13] N. Bonnotte. “From Knothe’s rearrangement to Brenier’s optimal transport map”. *SIAM Journal on Mathematical Analysis* 45.1 (2013), pp. 64–87.

[Bre91] Y. Brenier. “Polar factorization and monotone rearrangement of vector-valued functions”. *Communications on Pure and Applied Mathematics* 44.4 (1991), pp. 375–417.

[Bro93] M. Broadie. “Computing efficient frontiers using estimated parameters”. *Annals of Operations Research* 45.1 (1993), pp. 21–58.

[BS16] S. Barocas and A. D. Selbst. “Big data’s disparate impact”. *California Law Review* 104 (2016), pp. 671–732.

[BSR18] M. Blondel, V. Seguy, and A. Rolet. “Smooth and sparse optimal transport”. *Artificial Intelligence and Statistics*. 2018, pp. 880–889.

[BT97] D. Bertsimas and J. N. Tsitsiklis. *Introduction to Linear Optimization*. Athena Scientific Belmont, 1997.

[BTBN05] A. Ben-Tal, S. Boyd, and A. Nemirovski. “Control of Uncertainty-Affected Discrete Time Linear Systems via Convex Programming”. *Available at Optimization Online* (2005).

[BTBN06] A. Ben-Tal, S. Boyd, and A. Nemirovski. “Extending scope of robust optimization: Comprehensive robust counterparts of uncertain problems”. *Mathematical Programming* 107.1 (2006), pp. 63–89.

[Bub15] S. Bubeck. “Convex optimization: Algorithms and complexity”. *Foundations and Trends in Machine Learning* 8.3-4 (2015), pp. 231–357.

[BYF20] E. Black, S. Yeom, and M. Fredrikson. “FlipTest: fairness testing via optimal transport”. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. 2020, pp. 111–121.

[Cal+17] F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. “Optimized pre-processing for discrimination prevention”. *Advances in Neural Information Processing Systems*. 2017, pp. 3992–4001.

[Caz+18] E. Cazelles, V. Seguy, J. Bigot, M. Cuturi, and N. Papadakis. “Geodesic PCA versus log-PCA of histograms in the Wasserstein space”. *SIAM Journal on Scientific Computing* 40.2 (2018), B429–B456.

[CBL06] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, 2006.

[CCO17] M. Carriere, M. Cuturi, and S. Oudot. “Sliced Wasserstein kernel for persistence diagrams”. *International Conference on Machine Learning*. 2017, pp. 664–673.

[CD+17] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. “Algorithmic decision making and the cost of fairness”. *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2017, pp. 797–806.

[CDO18] M. Cohen, J. Diakonikolas, and L. Orecchia. “On acceleration with noise-corrupted Ggradients”. *International Conference on Machine Learning*. 2018, pp. 1019–1028.

[Çel+21] T. Ö. Çelik, A. Jamneshan, G. Montúfar, B. Sturmfels, and L. Venturello. “Wasserstein distance to independence models”. *Journal of Symbolic Computation* 104 (2021), pp. 855–873.

[Che12] S.-Y. Chen. “Kalman Filter for Robot Vision: A Survey”. *IEEE Transactions on Industrial Electronics* 59.11 (2012), pp. 4409–4420.

[Che+16] X. Chen, M. Monfort, A. Liu, and B. D. Ziebart. “Robust covariate shift regression”. *Artificial Intelligence and Statistics*. 2016, pp. 1270–1279.

[Chi+18] L. Chizat, G. Peyré, B. Schmitzer, and F.-X. Vialard. “Scaling algorithms for unbalanced optimal transport problems”. *Mathematics of Computation* 87.314 (2018), pp. 2563–2609.

[Chi+20] L. Chizat, P. Roussillon, F. Léger, F.-X. Vialard, and G. Peyré. “Faster Wasserstein distance estimation with the Sinkhorn divergence”. *Advances in Neural Information Processing Systems* (2020), pp. 2257–2269.

[Cho17] A. Chouldechova. “Fair prediction with disparate impact: A study of bias in recidivism prediction instruments”. *Big Data* 5.2 (2017), pp. 153–163.

[CK20] D. Chakrabarty and S. Khanna. “Better and simpler error analysis of the Sinkhorn-Knopp algorithm for matrix scaling”. *Mathematical Programming* (2020). Forthcoming, pp. 1–13.

[Cla+21] C. Clason, D. A. Lorenz, H. Mahler, and B. Wirth. “Entropic regularization of continuous optimal transport problems”. *Journal of Mathematical Analysis and Applications* 494.1 (2021), p. 124432.

[CM14] C. Cortes and M. Mohri. “Domain adaptation and sample bias correction theory and algorithm for regression”. *Theoretical Computer Science* 519 (2014), pp. 103 –126.

[CMG21] C. Childers and M Maggard-Gibbons. “Trends in the use of robotic-assisted surgery during the COVID 19 pandemic”. *British Journal of Surgery* 108.10 (2021), e330–e331.

[CMM19] C. Cortes, M. Mohri, and A. M. Medina. “Adaptation Based on Generalized Discrepancy”. *Journal of Machine Learning Research* 20.1 (2019), pp. 1–30.

[Cor+09] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein. *Introduction to Algorithms*. MIT Press, 2009.

[Cou+16] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. “Optimal transport for domain adaptation”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.9 (2016), pp. 1853–1865.

[Cou+17] N. Courty, R. Flamary, D. Tuia, and A. Rakotomamonjy. “Optimal Transport for Domain Adaptation”. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39.9 (2017), pp. 1853–1865.

[CR20] A. Chouldechova and A. Roth. “A snapshot of the frontiers of fairness in machine learning”. *Communications of the ACM* 63.5 (2020), pp. 82–89.

[CSM94] R. Cominetti and J. San Martín. “Asymptotic Analysis of the Exponential Penalty Trajectory in Linear Programming”. *Mathematical Programming* 67.1-3 (1994), pp. 169–187.

[Csu17] G. Csurka. “A Comprehensive Survey on Domain Adaptation for Visual Applications”. *Domain Adaptation in Computer Vision Applications*. Ed. by G. Csurka. Springer International Publishing, 2017, pp. 1–35.

[CT21] G. Conforti and L. Tamanini. “A formula for the time derivative of the entropic cost and applications”. *Journal of Functional Analysis* 280.11 (2021).

[Cut13] M. Cuturi. “Sinkhorn distances: Lightspeed computation of optimal transport”. *Advances in Neural Information Processing Systems*. 2013, pp. 2292–2300.

[CV10] T. Calders and S. Verwer. “Three naive Bayes approaches for discrimination-free classification”. *Data Mining and Knowledge Discovery* 21.2 (2010), pp. 277–292.

[CW18] C. Chu and R. Wang. “A survey of domain adaptation for neural machine translation”. *International Conference on Computational Linguistics*. 2018, pp. 1304–1319.

[d’A08] A. d’Aspremont. “Smooth optimization with approximate gradient”. *SIAM Journal on Optimization* 19.3 (2008), pp. 1171–1183.

[Dag14] C. Daganzo. *Multinomial Probit: the Theory and its Application to Demand Forecasting*. Elsevier, 2014.

[Dan51] G. B. Dantzig. “Application of the simplex method to a transportation problem”. *Activity Analysis and Production and Allocation* (1951), pp. 359–373.

[Das18] J. Dastin. “Amazon scraps secret AI recruiting tool that showed bias against women”. *San Fransico, CA: Reuters. Retrieved on October* 9 (2018), p. 2018.

[DB16] S. Diamond and S. Boyd. “CVXPY: A Python-embedded modeling language for convex optimization”. *Journal of Machine Learning Research* 17.83 (2016), pp. 1–5.

[DDN21] A. Dhara, B. Das, and K. Natarajan. “Worst-Case Expected Shortfall with Univariate and Bivariate Marginals”. *INFORMS Journal on Computing* 33.1 (2021), pp. 370–389.

[Dek+12] O. Dekel, R. Gilad-Bachrach, O. Shamir, and L. Xiao. “Optimal distributed online prediction using mini-batches”. *Journal of Machine Learning Research* 13 (2012), pp. 165–202.

[Del21] A. Delalande. “Nearly Tight Convergence Bounds for Semi-discrete Entropic Optimal Transport”. *arXiv:2110.12678* (2021).

[DF88] M. E. Dyer and A. M. Frieze. “On the complexity of computing the volume of a polyhedron”. *SIAM Journal on Computing* 17.5 (1988), pp. 967–974.

[DG+12] F. De Goes, K. Breeden, V. Ostromoukhov, and M. Desbrun. “Blue noise through optimal transport”. *ACM Transactions on Graphics* 31.6 (2012), p. 171.

[DGK18] P. Dvurechensky, A. Gasnikov, and A. Kroshnin. “Computational optimal transport: Complexity by accelerated gradient descent is better than by Sinkhorn’s algorithm”. *International Conference on Machine Learning*. 2018, pp. 1367–1376.

[DH78] J. C. Dunn and S. Harshbarger. “Conditional gradient algorithms with open loop step size rules”. *Journal of Mathematical Analysis and Applications* 62.2 (1978), pp. 432–444.

[DiC+20] C. DiCiccio, S. Vasudevan, K. Basu, K. Kenthapadi, and D. Agarwal. “Evaluating fairness using permutation tests”. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 2020, pp. 1467–1477.

[DKS13] J. Dick, F. Y. Kuo, and I. H. Sloan. “High-dimensional integration: The quasi-Monte Carlo way”. *Acta Numerica* 22 (2013), pp. 133–288.

[DM84] J. A. Dubin and D. L. McFadden. “An econometric analysis of residential electric appliance holdings and consumption”. *Econometrica* 52.2 (1984), pp. 345–362.

[DN18] J. Duchi and H. Namkoong. “Learning models with uniform performance via distributionally robust optimization”. *arXiv:1810.08750* (2018).

[Don+18] M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. “Empirical risk minimization under fairness constraints”. *Advances in Neural Information Processing Systems*. 2018, pp. 2791–2801.

[Doy78] J. Doyle. “Guaranteed margins for LQG regulators”. *IEEE Transactions on Automatic Control* 23.4 (1978), pp. 756–757.

[Doy+88] J. Doyle, K. Glover, P. Khargonekar, and B. Francis. “State-space solutions to standard $\mathcal{H}_2$ and $\mathcal{H}_\infty$ control problems”. *American Control Conference*. 1988, pp. 1691–1696.

[DPR18] A. Dessein, N. Papadakis, and J.-L. Rouas. “Regularized optimal transport and the rot mover’s distance”. *Journal of Machine Learning Research* 19.1 (2018), pp. 590–642.

[DR70] V. F. Demyanov and A. M. Rubinov. *Approximate Methods in Optimization Problems*. Elsevier, 1970.

[DS06] M. Dyer and L. Stougie. “Computational complexity of stochastic programming problems”. *Mathematical Programming* 106 (2006), pp. 423–432.

[DS09] J. Duchi and Y. Singer. “Efficient online and batch learning using forward backward splitting”. *The Journal of Machine Learning Research* 10.99 (2009), pp. 2899–2934.

[DS15] M. Dyer and L. Stougie. “Erratum to: Computational complexity of stochastic programming problems”. *Mathematical Programming* 153 (2015), pp. 723–725.

[DT03] G. B. Dantzig and M. N. Thapa. *Linear Programming 2: Theory and Extensions*. Springer, 2003.

[DTD15] A. Datta, M. C. Tschantz, and A. Datta. “Automated experiments on ad privacy settings: A tale of opacity, choice, and discrimination”. *Proceedings on Privacy Enhancing Technologies* 2015.1 (2015), pp. 92–112.

[Dun79] J. C. Dunn. “Rates of convergence for conditional gradient algorithms near singular and nonsingular extremals”. *SIAM Journal on Control and Optimization* 17.2 (1979), pp. 187–211.

[Dun80] J. C. Dunn. “Convergence rates for conditional gradient sequences generated by implicit step length rules”. *SIAM Journal on Control and Optimization* 18.5 (1980), pp. 473–487.

[Dwo+12] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. “Fairness through awareness”. *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012, pp. 214–226.

[Dwo+18] C. Dwork, N. Immorlica, A. T. Kalai, and M. Leiserson. “Decoupled classifiers for group-fair and efficient machine learning”. *Conference on Fairness, Accountability and Transparency*. 2018, pp. 119–133.

[DY10] E. Delage and Y. Ye. “Distributionally robust optimization under moment uncertainty with application to data-driven problems”. *Operations Research* 58.3 (2010), pp. 595–612.

[Dye03] M. Dyer. “Approximate counting by dynamic programming”. *ACM Symposium on Theory of Computing*. 2003, pp. 693–699.

[Dye+93] M. Dyer, A. Frieze, R. Kannan, A. Kapoor, L. Perkovic, and U. Vazirani. “A mildly exponential time algorithm for approximating the number of solutions to a multidimensional knapsack problem”. *Combinatorics, Probability and Computing* 2.3 (1993), pp. 271–284.

[DZB89] J. Doyle, K. Zhou, and B. Bodenheimer. “Optimal control with mixed $\mathcal{H}_2$ and $\mathcal{H}_\infty$ performance objectives”. *American Control Conference*. 1989, pp. 2065–2070.

[EHG21] O. El Housni and V. Goyal. “On the optimality of affine policies for budgeted uncertainty sets”. *Mathematics of Operations Research* 46.2 (2021), pp. 674–711.

[EI06] E. Erdoğan and G. Iyengar. “Ambiguous chance constrained problems and robust optimization”. *Mathematical Programming* 107.1-2 (2006), pp. 37–61.

[EK18] P. M. Esfahani and D. Kuhn. “Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations”. *Mathematical Programming* 171.1-2 (2018), pp. 115–166.

[EMR15] M. Erbar, J. Maas, and M. Renger. “From large deviations to Wasserstein gradient flows in multiple dimensions”. *Electronic Communications in Probability* 20 (2015), pp. 1–12.

[ES15] H. Edwards and A. Storkey. “Censoring representations with an adversary”. *arXiv:1511.05897* (2015).

[ES18] M. Essid and J. Solomon. “Quadratically regularized optimal transport on graphs”. *SIAM Journal on Scientific Computing* 40.4 (2018), A1961–A1986.

[Eva97] L. C. Evans. “Partial differential equations and Monge-Kantorovich mass transfer”. *Current Developments in Mathematics* 1997.1 (1997), pp. 65–126.

[Fan92] S.-C. Fang. “An unconstrained convex programming view of linear programming”. *Zeitschrift für Operations Research* 36.2 (1992), pp. 149–161.

[Fel+15] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. “Certifying and removing disparate impact”. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2015, 259–268. ISBN: 9781450336642.

[Fer+14] S. Ferradans, N. Papadakis, G. Peyré, and J.-F. Aujol. “Regularized discrete optimal transport”. *SIAM Journal on Imaging Sciences* 7.3 (2014), pp. 1853–1882.

[Fey+17] J. Feydy, B. Charlier, F.-X. Vialard, and G. Peyré. “Optimal transport for diffeomorphic registration”. *International Conference on Medical Image Computing and Computer-Assisted Intervention.* 2017, pp. 291–299.

[Fla+18] R. Flamary, M. Cuturi, N. Courty, and A. Rakotomamonjy. “Wasserstein discriminant analysis”. *Machine Learning* 107.12 (2018), pp. 1923–1945.

[FLW17] G. Feng, X. Li, and Z. Wang. “On the relation between several discrete choice models”. *Operations Research* 65.6 (2017), pp. 1516–1525.

[FMP19] M. Fazlyab, M. Morari, and G. J. Pappas. “Safety Verification and Robustness Analysis of Neural Networks via Quadratic Constraints and Semidefinite Programming”. *arXiv:1903.01287* (2019).

[FN96] B. Friedman and H. Nissenbaum. “Bias in computer systems”. *ACM Transactions on information systems* 14.3 (1996), pp. 330–347.

[Fré51] M. Fréchet. “Sur les tableaux de corrélation dont les marges sont données”. *Annales de l’Université de Lyon, Sciences* 4.1/2 (1951), pp. 13–84.

[Fri+13] J. J. Friedewald, C. J. Samana, B. L. Kasiske, A. K. Israni, D. Stewart, W. Cherikh, and R. N. Formica. “The kidney allocation system”. *Surgical Clinics* 93.6 (2013), pp. 1395–1406.

[Fri+19] S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. “A comparative study of fairness-enhancing interventions in machine learning”. *Proceedings of the Conference on Fairness, Accountability, and Transparency.* 2019, pp. 329–338.

[Fro+19] C. Frogner, S. Claici, E. Chien, and J. Solomon. “Incorporating Unlabeled Data into Distributionally Robust Learning”. *arXiv:1912.07729* (2019).

[FS04] H. Föllmer and A. Schied. *Stochastic Finance: An Introduction in Discrete Time.* Walter de Gruyter, 2004.

[FS11] H. Föllmer and A. Schied. *Stochastic Finance: An Introduction in Discrete Time*. de Gruyter, 2011.

[FS12] M. P. Friedlander and M. Schmidt. “Hybrid deterministic-stochastic methods for data fitting”. *SIAM Journal on Scientific Computing* 34.3 (2012), A1380–A1405.

[Fue00] A. De la Fuente. *Mathematical Methods and Models for Economists*. Cambridge University Press, 2000.

[FW56] M. Frank and P. Wolfe. “An algorithm for quadratic programming”. *Naval Research Logistics* 3.1-2 (1956), pp. 95–110.

[Gal16] A. Galichon. *Optimal Transport Methods in Economics*. Princeton University Press, 2016.

[Gao+18] R. Gao, L. Xie, Y. Xie, and H. Xu. “Robust Hypothesis Testing Using Wasserstein Uncertainty Sets.” *Advances in Neural Information Processing Systems*. 2018, pp. 7913–7923.

[Gao20] R. Gao. “Finite-sample guarantees for Wasserstein distributionally robust optimization: Breaking the curse of dimensionality”. *arXiv:2009.04382* (2020).

[Gar+19] S. Garg, V. Perot, N. Limtiaco, A. Taly, E. H. Chi, and A. Beutel. “Counterfactual fairness in text classification through robustness”. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 219–226.

[GCK17] R. Gao, X. Chen, and A. J. Kleywegt. “Wasserstein distributional robustness and regularization in statistical learning”. *arXiv:1712.06050* (2017).

[Gel90] M. Gelbrich. “On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces”. *Mathematische Nachrichten* 147.1 (1990), pp. 185–203.

[Gen+16] A. Genevay, M. Cuturi, G. Peyré, and F. Bach. “Stochastic optimization for large-scale optimal transport”. *Advances in Neural Information Processing Systems*. 2016, pp. 3440–3448.

[GH+16] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller. “The case for process fairness in learning: Feature selection for fair decision making”. *NIPS Symposium on Machine Learning and the Law*. Vol. 1. 2016, p. 2.

[Ghi+16] M. Ghifary, W. B. Kleijn, M. Zhang, D. Balduzzi, and W. Li. “Deep reconstruction-classification networks for unsupervised domain adaptation”. *European Conference on Computer Vision*. 2016, pp. 597–613.

[GHJ20] W. Guo, N. Ho, and M. I. Jordan. “Fast algorithms for computational optimal transport and Wasserstein barycenter”. *International Conference on Artificial Intelligence and Statistics*. 2020, pp. 2088–2097.

[GHS20] U. Ghai, E. Hazan, and Y. Singer. “Exponentiated gradient meets gradient descent”. *International Conference on Algorithmic Learning Theory*. 2020, pp. 386–407.

[GJ79] M. R. Garey and D. S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman and Company, 1979.

[GK22] R. Gao and A. J. Kleywegt. “Distributionally robust stochastic optimization with Wasserstein distance”. *Mathematics of Operations Research* (2022). Forthcoming.

[GL15] Y. Ganin and V. Lempitsky. “Unsupervised domain adaptation by backpropagation”. *International Conference on Machine Learning.* 2015, pp. 1180–1189.

[GL97] L. E. Ghaoui and H. Lebret. “Robust Solutions to Least-Squares Problems with Uncertain Data”. *SIAM Journal on Matrix Analysis and Applications* 18.4 (1997), 1035–1064.

[GLS12] M. Grötschel, L. Lovász, and A. Schrijver. *Geometric Algorithms and Combinatorial Optimization.* 2012.

[Goe+15] F. de Goes, C. Wallez, J. Huang, D. Pavlov, and M. Desbrun. “Power particles: An incompressible fluid solver based on power diagrams.” *ACM Transactions on Graphics* 34.4 (2015), 50:1–50:11.

[Gop+11] P. Gopalan, A. Klivans, R. Meka, D. Štefankovic, S. Vempala, and E. Vigoda. “An FPTAS for #knapsack and related counting problems”. *Foundations of Computer Science.* 2011, pp. 817–826.

[Gor+19] P. Gordaliza, E. D. Barrio, G. Fabrice, and J.-M. Loubes. “Obtaining fairness using optimal transport theory”. *International Conference on Machine Learning.* 2019, pp. 2357–2365.

[GPC18] A. Genevay, G. Peyré, and M. Cuturi. “Learning Generative Models with Sinkhorn Divergences”. *Artificial Intelligence and Statistics.* 2018, pp. 1608–1617.

[GS10] J. Goh and M. Sim. “Distributionally robust optimization and its tractable approximations”. *Operations Research* 58.4 (2010), pp. 902–917.

[GS84] C. Givens and R. Shortt. “A class of Wasserstein metrics for probability distributions”. *The Michigan Mathematical Journal* 31.2 (1984), pp. 231–240.

[GTW21] A. Georghiou, A. Tsoukalas, and W. Wiesemann. “On the optimality of affine decision rules in robust and distributionally robust optimization”. *Available at Optimization Online* (2021).

[Gul+17] I. Gulrajani, F. Ahmed, M. Arjovsky, V. Dumoulin, and A. C. Courville. “Improved training of Wasserstein GANs”. *Advances in Neural Information Processing Systems.* 2017, pp. 5767–5777.

[GV14] J. Garcke and T. Vanck. “Importance weighted inductive transfer learning for regression”. *Joint European conference on machine learning and knowledge discovery in databases.* 2014, pp. 466–481.

[Han23] B. Han. “Distributionally robust Kalman filtering with volatility uncertainty”. *arXiv:2302.05993* (2023).

[Har+16] M. Hardt, E. Price, E. Price, and N. Srebro. “Equality of Opportunity in Supervised Learning”. *Advances in Neural Information Processing Systems 29.* 2016, pp. 3315–3323.

[Has+18] T. Hashimoto, M. Srivastava, H. Namkoong, and P. Liang. “Fairness without demographics in repeated loss minimization”. *International Conference on Machine Learning.* 2018, pp. 1929–1938.

[HGK11] M. J. Hadjiyiannis, P. J. Goulart, and D. Kuhn. "An Efficient Method to Estimate the Suboptimality of Affine Controllers". *IEEE Transactions on Automatic Control* 56.12 (2011), pp. 2841–2853.

[HH13] Z. Hu and L. J. Hong. "Kullback-Leibler divergence constrained distributionally robust optimization". *Available at Optimization Online* (2013).

[HJLS13] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant. *Applied Logistic Regression*. John Wiley & Sons, 2013.

[HKL14] E. Hazan, T. Koren, and K. Y. Levy. "Logistic regression: Tight bounds for stochastic and online optimization". *Conference on Learning Theory*. 2014, pp. 197–209.

[HKW16] G. A. Hanasusanto, D. Kuhn, and W. Wiesemann. "A comment on "computational complexity of stochastic programming problems"". *Mathematical Programming* 159.1-2 (2016), pp. 557–569.

[HM13] A. Hackbarth and R. Madlener. "Consumer preferences for alternative fuel vehicles: A discrete choice analysis". *Transportation Research Part D: Transport and Environment* 25 (2013), pp. 5–17.

[Ho+17] N. Ho, X. Nguyen, M. Yurochkin, H. H. Bui, V. Huynh, and D. Phung. "Multilevel clustering via Wasserstein means". *International Conference on Machine Learning*. 2017, pp. 1501–1509.

[Hof81] K. L. Hoffman. "A method for globally minimizing concave functions over convex sets". *Mathematical Programming* 20.1 (1981), pp. 22–32.

[HP07] R. Hochreiter and G. C. Pflug. "Financial scenario generation for stochastic multi-stage decision processes as facility location problems". *Annals of Operations Research* 152.1 (2007), pp. 257–272.

[HR07] H. Heitsch and W. Römisch. "A note on scenario reduction for two-stage stochastic programs". *Operations Research Letters* 35.6 (2007), pp. 731–738.

[HS05] L. P. Hansen and T. J. Sargent. "Robust estimation and control under commitment". *Journal of Economic Theory* 124.2 (2005), pp. 258–301.

[HSL20] B. Hu, P. Seiler, and L. Lessard. "Analysis of biased stochastic gradient descent using sequential semidefinite programs". *Mathematical Programming* (2020). Forthcoming, pp. 1–26.

[Hua+06] J. Huang, A. Gretton, K. Borgwardt, B. Schölkopf, and A. Smola. "Correcting sample selection bias by unlabeled data". *Advances in Neural Information Processing Systems* 19 (2006), pp. 601–608.

[HV19] L. Huang and N. Vishnoi. "Stable and Fair Classification". *International Conference on Machine Learning*. 2019, pp. 2879–2890.

[ISS13] D. A. Iancu, M. Sharma, and M. Sviridenko. "Supermodularity and affine policies in dynamic robust optimization". *Operations Research* 61.4 (2013), pp. 941–956.

[Jag13] M. Jaggi. "Revisiting Frank-Wolfe: Projection-free sparse convex optimization". *International Conference on Machine Learning*. 2013, pp. 427–435.

[Jer03] M. Jerrum. *Counting, sampling and integrating: Algorithms and complexity*. Springer Science & Business Media, 2003.

[JST19] A. Jambulapati, A. Sidford, and K. Tian. “A Direct $\tilde{\mathcal{O}}(1/e)$ Iteration Parallel Algorithm for Optimal Transport”. *Advances in Neural Information Processing Systems*. 2019, pp. 11359–11370.

[JVS20] P. G. John, D. Vijaykeerthy, and D. Saha. “Verifying Individual Fairness in Machine Learning Models”. *Conference on Uncertainty in Artificial Intelligence*. PMLR. 2020, pp. 749–758.

[JZ07] J. Jiang and C. Zhai. “Instance Weighting for Domain Adaptation in NLP”. *Association of Computational Linguistics*. 2007, pp. 264–271.

[Kam+12] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. “Fairness-aware classifier with prejudice remover regularizer”. *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. 2012, pp. 35–50.

[Kan42] L Kantorovich. “On the transfer of masses (in Russian)”. *Doklady Akademii Nauk* 37.2 (1942), pp. 227–229.

[Kar84] N. Karmarkar. “A new polynomial-time algorithm for linear programming”. *ACM Symposium on Theory of Computing*. 1984, pp. 302–311.

[KAS11] T. Kamishima, S. Akaho, and J. Sakuma. “Fairness-aware learning through regularization approach”. *IEEE International Conference on Data Mining Workshops*. 2011, pp. 643–650.

[Kav+19] A. Kavis, K. Y. Levy, F. Bach, and V. Cevher. “UniXGrad: A universal, adaptive algorithm with optimal guarantees for constrained optimization”. *Advances in Neural Information Processing Systems*. 2019, pp. 6257–6266.

[KC12] F. Kamiran and T. Calders. “Data preprocessing techniques for classification without discrimination”. *Knowledge and Information Systems* 33.1 (2012), pp. 1–33.

[KIL22] A. Karapetyan, A. Iannelli, and J. Lygeros. “On the regret of $\mathcal{H}_\infty$ control”. *IEEE Conference on Decision and Control*. 2022, pp. 6181–6186.

[KKG18] H. Kannan, A. Kurakin, and I. Goodfellow. “Adversarial logit pairing”. *arXiv:1803.06373* (2018).

[Kle+18] J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan. “Algorithmic fairness”. *AEA Papers and Proceedings*. Vol. 108. 2018, pp. 22–27.

[KLN21] G. Kotsalis, G. Lan, and A. S. Nemirovski. “Convex Optimization for Finite-Horizon Robust Covariance Control of Linear Stochastic Systems”. *SIAM Journal on Control and Optimization* 59.1 (2021), pp. 296–319.

[KMR16] J. Kleinberg, S. Mullainathan, and M. Raghavan. “Inherent trade-offs in the fair determination of risk scores”. *arXiv:1609.05807* (2016).

[KMT16] J. Kitagawa, Q. Mérigot, and B. Thibert. “Convergence of a Newton algorithm for semi-discrete optimal transport”. *arXiv:1603.05579* (2016).

[Kol+17] S. Kolouri, S. R. Park, M. Thorpe, D. Slepcev, and G. K. Rohde. “Optimal mass transport: Signal processing and machine-learning applications”. *IEEE Signal Processing Magazine* 34.4 (2017), pp. 43–59.

[KR15] S. Kolouri and G. K. Rohde. “Transport-based single frame super resolution of very low resolution face images”. *IEEE Conference on Computer Vision and Pattern Recognition*. 2015, pp. 4876–4884.

[KR17] J. Karlsson and A. Ringh. “Generalized Sinkhorn iterations for regularizing inverse problems using optimal mass transport”. *SIAM Journal on Imaging Sciences* 10.4 (2017), pp. 1935–1962.

[KSD10] A. Kumar, A. Saha, and H. Daume. “Co-regularization based semi-supervised domain adaptation”. *Advances in Neural Information Processing Systems* (2010), pp. 478–486.

[KSST09] S. Kakade, S. Shalev-Shwartz, and A. Tewari. “On the duality of strong convexity and strong smoothness: Learning applications and matrix regularization”. Technical report, Toyota Technological Institute. 2009.

[KTP17] P. Koniusz, Y. Tas, and F. Porikli. “Domain adaptation by mixture of alignments of second-or higher-order scatter tensors”. *IEEE Conference on Computer Vision and Pattern Recognition*. 2017, pp. 4478–4487.

[Kuh+19] D. Kuhn, P. Mohajerin Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh. “Wasserstein distributionally robust optimization: Theory and applications in machine learning”. *Operations Research & Management Science in the Age of Analytics*. 2019, pp. 130–166.

[Kuh55] H. W. Kuhn. “The Hungarian method for the assignment problem”. *Naval Research Logistics Quarterly* 2.1-2 (1955), pp. 83–97.

[Kun+18] S. Kundu, S. Kolouri, K. I. Erickson, A. F. Kramer, E. McAuley, and G. K. Rohde. “Discovery and visualization of structural biomarkers from MRI using transport-based morphometry”. *NeuroImage* 167 (2018), pp. 256–275.

[KY23] K. Kim and I. Yang. “Distributional Robustness in Minimax Linear Quadratic Control with Wasserstein Distance”. *SIAM Journal on Control and Optimization* 61.2 (2023), pp. 458–483.

[LÖ4] J. Löfberg. “YALMIP: A toolbox for modeling and optimization in MATLAB”. *IEEE International Conference on Robotics and Automation*. 2004, pp. 284–289.

[Lam19] H. Lam. “Recovering best statistical guarantees via the empirical divergence-based distributionally robust optimization”. *Operations Research* 67.4 (2019), pp. 1090–1105.

[Lan12] G. Lan. “An optimal method for stochastic composite optimization”. *Mathematical Programming* 133.1-2 (2012), pp. 365–397.

[Lév15] B. Lévy. “A numerical algorithm for L_2 semi-discrete optimal transport in 3D”. *ESAIM: Mathematical Modelling and Numerical Analysis* 49.6 (2015), pp. 1693–1715.

[LHJ19a] T. Lin, N. Ho, and M. I. Jordan. “On Efficient optimal transport: An analysis of greedy and accelerated mirror descent algorithms”. *International Conference on Machine Learning*. 2019, pp. 3982–3991.

[LHJ19b] T. Lin, N. Ho, and M. I. Jordan. “On the efficiency of the Sinkhorn and Greenkhorn algorithms for optimal transport”. *arXiv:1906.01437* (2019).

[LHS19] J. Li, S. Huang, and A. M.-C. So. “A First-Order Algorithmic Framework for Wasserstein Distributionally Robust Logistic Regression”. *Advances in Neural Information Processing Systems 32*. 2019, pp. 3937–3947.

[Li12] Q. Li. “Literature survey: Domain adaptation algorithms for natural language processing”. *Department of Computer Science The Graduate Center, The City University of New York* (2012), pp. 8–10.

[Li+16] Y. Li, L. Wang, J. Wang, J. Ye, and C. K. Reddy. “Transfer Learning for Survival Analysis via Efficient L2,1-Norm Regularized Cox Regression”. *IEEE International Conference on Data Mining*. 2016, pp. 231–240.

[Li+19] H. Li, S. Webster, N. Mason, and K. Kempf. “Product-line pricing under discrete mixed multinomial logit demand”. *Manufacturing and Service Operations Management* 21 (2019), pp. 14–28.

[LJSB12] S. Lacoste-Julien, M. Schmidt, and F. Bach. “A simpler approach to obtaining an $\mathcal{O}(1/t)$ convergence rate for the projected stochastic subgradient method”. *arXiv:1212.2002* (2012).

[LMC18] Z. Lipton, J. McAuley, and A. Chouldechova. “Does mitigating ML’s impact disparity require treatment disparity?” *Advances in Neural Information Processing Systems*. 2018, pp. 8125–8135.

[LOG16] W. Li, S. Osher, and W. Gangbo. “A fast algorithm for earth mover’s distance based on optimal transport and l_1 type Regularization”. *arXiv:1609.07092* (2016).

[Lon+16] M. Long, H. Zhu, J. Wang, and M. I. Jordan. “Unsupervised domain adaptation with residual transfer networks”. *International Conference on Neural Information Processing Systems*. 2016, 136–144.

[LP66] E. S. Levitin and B. T. Polyak. “Constrained minimization methods”. *USSR Computational Mathematics and Mathematical Physics* 6.5 (1966), pp. 1–50.

[LPHLS12] D. Lopez-Paz, J. M. Hernández-Lobato, and B. Schölkopf. “Semi-supervised domain adaptation with non-parametric copulas”. *International Conference on Neural Information Processing Systems*. 2012, 665–673.

[LPL20] M. Lohaus, M. Perrot, and U. von Luxburg. “Too relaxed to be fair”. *International Conference on Machine Learning*. 2020.

[LRT11] B. T. Luong, S. Ruggieri, and F. Turini. “k-NN as an implementation of situation testing for discrimination discovery and prevention”. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2011, pp. 502–510.

[LS14] Y. T. Lee and A. Sidford. “Path finding methods for linear programming: Solving linear programs in $\tilde{\mathcal{O}}(\sqrt{rank})$ iterations and faster algorithms for maximum flow”. *IEEE Symposium on Foundations of Computer Science*. 2014, pp. 424–433.

[LS20] T. Lattimore and C. Szepesvári. *Bandit Algorithms*. Cambridge University Press, 2020.

[LSM19] M. Leo, S. Sharma, and K. Maddulety. "Machine learning in banking risk management: A literature review". *Risks* 7.1 (2019), p. 29.

[LT19] A. Lambrecht and C. Tucker. "Algorithmic bias? An empirical study of apparent gender-based discrimination in the display of STEM career ads". *Management science* 65.7 (2019), pp. 2966–2981.

[LT93] Z.-Q. Luo and P. Tseng. "Error bounds and convergence analysis of feasible descent methods: A general approach". *Annals of Operations Research* 46.1 (1993), pp. 157–178.

[LWS18] Z. Lipton, Y.-X. Wang, and A. Smola. "Detecting and correcting for label shift with black box predictors". *International Conference on Machine Learning*. 2018, pp. 3122–3130.

[MA16] A. Martins and R. Astudillo. "From softmax to sparsemax: A sparse model of attention and multi-Label classification". *International Conference on Machine Learning*. 2016, pp. 1614–1623.

[Mad+18] D. Madras, E. Creager, T. Pitassi, and R. Zemel. "Learning Adversarially Fair and Transferable Representations". *International Conference on Machine Learning*. 2018, pp. 3384–3393.

[Man+16] A. K. Manrai, B. H. Funke, H. L. Rehm, M. S. Olesen, B. A. Maron, P. Szolovits, D. M. Margulies, J. Loscalzo, and I. S. Kohane. "Genetic misdiagnoses and the potential for health disparities". *New England Journal of Medicine* 375.7 (2016), pp. 655–665.

[Mat+20] A. de Mathelin, G. Richard, M. Mougeot, and N. Vayatis. "Adversarial Weighting for Domain Adaptation in Regression". *arXiv:2006.08251* (2020).

[MB11] E. Moulines and F. Bach. "Non-asymptotic analysis of stochastic approximation algorithms for machine learning". *Advances in Neural Information Processing Systems*. 2011, pp. 451–459.

[McC97] R. J. McCann. "A Convexity Principle for Interacting Gases". *Advances in Mathematics* 128.1 (1997), pp. 153–179.

[McF74] D. McFadden. "Conditional logit analysis of qualitative choice behavior". *Frontiers in Econometrics*. Ed. by P. Zarembka. Academic Press, 1974, pp. 105–142.

[McF78] D. McFadden. "Modeling the choice of residential location". *Transportation Research Record* 673 (1978), pp. 72–77.

[McF81] D. McFadden. "Econometric models of probabilistic choice". *Structural Analysis of Discrete Data with Econometric Application*. Ed. by C. Manski and D. McFadden. MIT Press, 1981, pp. 198–272.

[MCG10] R. D. Martin, A. Clark, and C. G. Green. "Robust portfolio construction". *Handbook of Portfolio Construction*. Springer, 2010, pp. 337–380.

[Meh+19] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan. "A survey on bias and fairness in machine learning". *arXiv:1908.09635* (2019).

[MEK18] P. Mohajerin Esfahani and D. Kuhn. “Data-driven distributionally robust optimization using the Wasserstein metric: Performance guarantees and tractable reformulations”. *Mathematical Programming* 171.1 (2018), pp. 115–166.

[Mér11] Q. Mérigot. “A Multiscale Approach to Optimal Transport”. *Computer Graphics Forum* 5.30 (2011), pp. 1583–1592.

[Mir15] J.-M. Mirebeau. “Discretization of the 3D Monge-Ampère operator, between wide stencils and power diagrams”. *Mathematical Modelling and Numerical Analysis* 49.5 (2015), pp. 1511–1523.

[Mis+12] V. K. Mishra, K. Natarajan, H. Tao, and C.-P. Teo. “Choice prediction with semidefinite optimization when utilities are correlated”. *IEEE Transactions on Automatic Control* 57.10 (2012), pp. 2450–2463.

[Mis+14] V. K. Mishra, K. Natarajan, D. Padmanabhan, C.-P. Teo, and X. Li. “On theoretical and empirical aspects of marginal distribution choice models”. *Management Science* 60.6 (2014), pp. 1511–1531.

[MK85] K. G. Murty and S. N. Kabadi. *Some NP-complete problems in quadratic and nonlinear programming*. Tech. rep. 1985.

[MM19] Y. Malitsky and K. Mishchenko. “Adaptive gradient descent without descent”. *arXiv:1910.09529* (2019).

[Mon81a] G. Monge. “Mémoire sur la théorie des déblais et des remblais”. *Histoire de l’Académie Royale des Sciences de Paris*. 1781, pp. 666–704.

[Mon81b] G. Monge. “Mémoire sur la théorie des déblais et des remblais”. *Histoire de l’Académie Royale des Sciences de Paris* (1781).

[MOS19] MOSEK ApS. *The MOSEK Optimization Toolbox. Version 9.2.* 2019.

[Mot+17a] S. Motiian, Q. Jones, S. Iranmanesh, and G. Doretto. “Few-Shot Adversarial Domain Adaptation”. *Advances in Neural Information Processing Systems*. Vol. 30. 2017, pp. 6670–6680.

[Mot+17b] S. Motiian, M. Piccirilli, D. A. Adjeroh, and G. Doretto. “Unified deep supervised domain adaptation and generalization”. *IEEE International Conference on Computer Vision*. 2017, pp. 5715–5725.

[MRZ15] H.-Y. Mak, Y. Rong, and J. Zhang. “Appointment scheduling with limited distributional information”. *Management Science* 61.2 (2015), pp. 316–334.

[MS04] B. Morris and A. Sinclair. “Random walks on truncated cubes and sampling 0-1 knapsack solutions”. *SIAM Journal on Computing* 34.1 (2004), pp. 195–226.

[Mul16] MultiMedia LLC. *Machine Bias*. Available at https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing. 2016 (accessed June 4, 2020).

[Mur+18] Z. Murez, S. Kolouri, D. Kriegman, R. Ramamoorthi, and K. Kim. “Image to image translation for domain adaptation”. *IEEE Conference on Computer Vision and Pattern Recognition*. 2018, pp. 4500–4509.

[Muz+17] B. Muzellec, R. Nock, G. Patrini, and F. Nielsen. “Tsallis regularized optimal transport and ecological inference”. *Association for the Advancement of Artificial Intelligence*. 2017, pp. 2387–2393.

[MW18] A. K. Menon and R. C. Williamson. “The cost of fairness in binary classification”. *Conference on Fairness, Accountability and Transparency.* 2018, pp. 107–118.

[Nat17] National Transportation Safety Board. *Collision Between a Car Operating With Automated Vehicle Control Systems and a Tractor-Semitrailer Truck Near Williston, Florida, May 7, 2016.* Tech. rep. NTSB/HAR-17/02. Washington, DC: National Transportation Safety Board, 2017.

[Nat20] National Transportation Safety Board. *Collision Between a Sport Utility Vehicle Operating With Partial Driving Automation and a Crash Attenuator Mountain View, California, March 23, 2018.* Tech. rep. NTSB/HAR-20/01. Washington, DC: National Transportation Safety Board, 2020.

[NB00] A. Nedić and D. Bertsekas. “Convergence rate of incremental subgradient algorithms”. *Stochastic Optimization: Algorithms and Applications.* Ed. by S. Uryasev and P. M. Pardalos. Kluwer Academic Publishers, 2000, pp. 263–304.

[ND16] H. Namkoong and J. C. Duchi. “Stochastic Gradient Methods for Distributionally Robust Optimization with f-divergences”. *Advances in Neural Information Processing Systems.* Vol. 29. 2016, pp. 2208–2216.

[Nem+09] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. “Robust stochastic approximation approach to stochastic programming”. *SIAM Journal on Optimization* 19.4 (2009), pp. 1574–1609.

[Nes83] Y. Nesterov. “A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$”. *Proceedings of the USSR Academy of Sciences* 269 (1983), pp. 543–547.

[Ngu+13] X. Nguyen et al. “Convergence of latent mixing measures in finite and infinite mixture models”. *The Annals of Statistics* 41.1 (2013), pp. 370–400.

[Ngu+19a] V. A. Nguyen, S. Shafieezadeh-Abadeh, M.-C. Yue, D. Kuhn, and W. Wiesemann. “Calculating Optimistic Likelihoods Using (Geodesically) Convex Optimization”. *Advances in Neural Information Processing Systems.* 2019.

[Ngu+19b] V. A. Nguyen, S. Shafieezadeh Abadeh, M.-C. Yue, D. Kuhn, and W. Wiesemann. “Optimistic distributionally robust optimization for nonparametric likelihood approximation”. *Advances in Neural Information Processing Systems* 32 (2019).

[Ngu+20] V. A. Nguyen, F. Zhang, J. Blanchet, E. Delage, and Y. Ye. “Distributionally Robust Local Non-parametric Conditional Estimation”. *Advances in Neural Information Processing Systems.* 2020.

[Ngu+23] V. A. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani. “Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization”. *Mathematics of Operations Research* 48.1 (2023), pp. 1–37.

[NN94] Y. Nesterov and A. Nemirovskii. *Interior-Point Polynomial Algorithms in Convex Programming.* SIAM, 1994.

[NSB20] V. A. Nguyen, N. Si, and J. Blanchet. "Robust Bayesian classification using an optimistic score ratio". *International Conference on Machine Learning*. 2020.

[NST09] K. Natarajan, M. Song, and C.-P. Teo. "Persistency model and its applications in choice modeling". *Management Science* 55.3 (2009), pp. 453–469.

[NV08] Y Nesterov and J. P. Vial. "Confidence level solutions for stochastic programming". *Automatica* 44.6 (2008), pp. 1559–1568.

[Orl97] J. B. Orlin. "A polynomial time primal network simplex algorithm for minimum cost flows". *Mathematical Programming* 78.2 (1997), pp. 109–129.

[Ort22] R. P. Ortega. *'Racially biased' devices caused delayed treatment for Black COVID-19 patients*. 2022. URL: https://www.science.org/content/article/racially-biased-devices-caused-delayed-treatment-black-covid-19-patients.

[Pal19] S. Pal. "On the difference between entropic cost and the optimal transport cost". *arXiv:1905.12206* (2019).

[Pas+17] A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer. "Automatic differentiation in PyTorch". *NIPS 2017 Autodiff Workshop*. 2017.

[Pas+19] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". *Advances in Neural Information Processing Systems*. 2019, pp. 8026–8037.

[PC19a] G. Peyré and M. Cuturi. "Computational optimal transport". *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 355–607.

[PC19b] G. Peyré and M. Cuturi. "Computational optimal transport: With applications to data science". *Foundations and Trends in Machine Learning* 11.5-6 (2019), pp. 355–607.

[PC20] F.-P. Paty and M. Cuturi. "Regularized optimal transport is ground cost adversarial". *International Conference on Machine Learning*. PMLR. 2020, pp. 7532–7542.

[Pey15] G. Peyré. "Entropic approximation of Wasserstein gradient flows". *SIAM Journal on Imaging Sciences* 8.4 (2015), pp. 2323–2351.

[Pey+17] G. Peyré, L. Chizat, F. Vialard, and J. Solomon. "Quantum entropic regularization of matrix-valued optimal transport". *European Journal of Applied Mathematics* (2017), pp. 1–24.

[Pfl01] G. C. Pflug. "Scenario tree generation for multiperiod financial optimization by optimal discretization". *Mathematical Programming* 89.2 (2001), pp. 251–271.

[Pin94] I. Pinelis. "Optimum bounds for the distributions of martingales in Banach spaces". *The Annals of Probability* 22.4 (1994), pp. 1679–1706.

[PJ92] B. T. Polyak and A. B. Juditsky. "Acceleration of stochastic approximation by averaging". *SIAM Journal on Control and Optimization* 30.4 (1992), pp. 838–855.

[PJD00] I. R. Petersen, M. R. James, and P. Dupuis. “Minimax optimal control of stochastic uncertain systems with relative entropy constraints”. *IEEE Transactions on Automatic Control* 45.3 (2000), pp. 398–412.

[PKD07] F. Pitié, A. C. Kokaram, and R. Dahyot. “Automated colour grading using colour distribution transfer”. *Computer Vision and Image Understanding* 107.1-2 (2007), pp. 123–137.

[Ple+17] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. “On fairness and calibration”. *Advances in Neural Information Processing Systems.* 2017, pp. 5680–5689.

[PPO14] N. Papadakis, G. Peyré, and E. Oudet. “Optimal transport with proximal splitting”. *SIAM Journal on Imaging Sciences* 7.1 (2014), pp. 212–238.

[PR17] N. Papadakis and J. Rabin. “Convex histogram-based joint image segmentation with regularized optimal transport cost”. *Journal of Mathematical Imaging and Vision* 59.2 (2017), pp. 161–186.

[PS10] D. Pardoe and P. Stone. “Boosting for regression transfer”. *International Conference on Machine Learning.* 2010.

[PW07] G. Pflug and D. Wozabal. “Ambiguity in portfolio selection”. *Quantitative Finance* 7.4 (2007), pp. 435–442.

[PW08] O. Pele and M. Werman. “A linear time histogram metric for improved sift matching”. *European Conference on Computer Vision.* 2008, pp. 495–508.

[PW09] O. Pele and M. Werman. “Fast and robust earth mover’s distances”. *IEEE International Conference on Computer Vision.* 2009, pp. 460–467.

[Qin+17] H. Qin, Y. Chen, J. He, and B. Chen. “Wasserstein blue noise sampling”. *ACM Transactions on Graphics* 36.4 (2017), pp. 1–14.

[QS17] N. Quadrianto and V. Sharmanska. “Recycling Privileged Learning and Distribution Matching for Fairness”. *Advances in Neural Information Processing Systems 30.* 2017, pp. 677–688.

[Qua19] K. Quanrud. “Approximating optimal transport with linear programs”. *Symposium on Simplicity in Algorithms.* 2019, 6:1–6:9.

[RCP16] A. Rolet, M. Cuturi, and G. Peyré. “Fast dictionary learning with a smoothed Wasserstein loss”. *Artificial Intelligence and Statistics.* 2016, pp. 630–638.

[Red+19] I. Redko, E. Morvant, A. Habrard, M. Sebban, and Y. Bennani. *Advances in Domain Adaptation Theory.* Elsevier, 2019.

[Rez+20] A. Rezaei, R. Fathony, O. Memarrast, and B. Ziebart. “Fairness for Robust Log Loss Classification”. *AAAI Conference on Artificial Intelligence.* 2020.

[Ric+20] G. Richard, A. de Mathelin, G. Hébrail, M. Mougeot, and N. Vayatis. “Unsupervised Multi-Source Domain Adaptation for Regression” (2020).

[RM51] H. Robbins and S. Monro. “A stochastic approximation method”. *The Annals of Mathematical Statistics* 22.3 (1951), pp. 400–407.

[Roc74a] R. T. Rockafellar. *Conjugate Duality and Optimization.* SIAM, 1974.

[Roc74b] R. T. Rockafellar. *Conjugate Duality and Optimization*. SIAM, 1974.

[RTG00] Y. Rubner, C. Tomasi, and L. J. Guibas. "The earth mover's distance as a metric for image retrieval". *International Journal of Computer Vision* 40.2 (2000), pp. 99–121.

[RU02] R. T. Rockafellar and S. Uryasev. "Conditional value-at-risk for general loss distributions". *Journal of Banking & Finance* 26.7 (2002), pp. 1443–1471.

[Rud64] W. Rudin. *Principles of Mathematical Analysis*. McGraw-Hill Education, 1964.

[Ruj+18] N. Rujeerapaiboon, K. Schindler, D. Kuhn, and W. Wiesemann. "Scenario reduction revisited: Fundamental limits and guarantees". *Mathematical Programming* (2018). Forthcoming.

[Rup88] D. Ruppert. *Efficient estimations from a slowly convergent Robbins-Monro process*. Tech. rep. School of Operations Research and Industrial Engineering, Cornell University, 1988.

[RW09] R. T. Rockafellar and R. J.-B. Wets. *Variational Analysis*. Springer Science & Business Media, 2009.

[RW18] P. Rigollet and J. Weed. "Entropic optimal transport is maximum-likelihood deconvolution". *Comptes Rendus Mathematique* 356.11-12 (2018), pp. 1228–1235.

[SA+18] S. Shafieezadeh-Abadeh, V. A. Nguyen, D. Kuhn, and P. Mohajerin Esfahani. "Wasserstein distributionally robust Kalman filtering". *Advances in Neural Information Processing Systems*. 2018, pp. 8483–8492.

[Sah+11] A. Saha, P. Rai, H. Daumé, S. Venkatasubramanian, and S. L. DuVall. "Active Supervised Domain Adaptation". *Machine Learning and Knowledge Discovery in Databases*. 2011, pp. 97–112.

[Sal+18] P. Saleiro, B. Kuester, L. Hinkson, J. London, A. Stevens, A. Anisfeld, K. T. Rodolfa, and R. Ghani. "Aequitas: A bias and fairness audit toolkit". *arXiv:1811.05577* (2018).

[Sal+19] S. M. Salaken, A. Khosravi, T. Nguyen, and S. Nahavandi. "Seeded transfer learning for regression problems with deep learning". *Expert Systems with Applications* 115 (2019), pp. 565 –577.

[Sam+18] S. Samadi, U. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. Vempala. "The price of fair PCA: One extra dimension". *Advances in Neural Information Processing Systems*. 2018, pp. 10976–10987.

[SB10] J. Skaf and S. P. Boyd. "Design of affine controllers via convex optimization". *IEEE Transactions on Automatic Control* 55.11 (2010), pp. 2476–2487.

[SC15] V. Seguy and M. Cuturi. "Principal geodesic analysis for probability measures under the optimal transport metric". *Advances in Neural Information Processing Systems*. 2015, pp. 3312–3320.

[Sch16] B. Schmitzer. "A sparse multiscale algorithm for dense optimal transport". *Journal of Mathematical Imaging and Vision* 56.2 (2016), pp. 238–259.

[Sch31] E Schrödinger. "Über die Umkehrung der Naturgesetze". *Sitzungsberichte der Preussischen Akademie der Wissenschaften. Physikalisch-Mathematische Klasse* 144.3 (1931), pp. 144–153.

[Sch98] A. Schrijver. *Theory of Linear and Integer Programming*. John Wiley & Sons, 1998.

[SDR21] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM, 2021.

[Seg+18] V. Seguy, B. B. Damodaran, R. Flamary, N. Courty, A. Rolet, and M. Blondel. "Large-scale optimal transport and mapping estimation". *International Conference on Learning Representations* (2018).

[Sha01] A. Shapiro. "On Duality Theory of Conic Linear Problems". *Semi-Infinite Programming.* Kluwer Academic Publishers, 2001, pp. 135–165.

[Sha17] A. Shapiro. "Distributionally Robust Stochastic Programming". *SIAM Journal on Optimization* 27.4 (2017), pp. 2258–2275.

[Shi00] H. Shimodaira. "Improving predictive inference under covariate shift by weighting the log-likelihood function". *Journal of Statistical Planning and Inference* 90.2 (2000), pp. 227–244.

[Sin67] R. Sinkhorn. "Diagonal equivalence to matrices with prescribed row and column sums". *The American Mathematical Monthly* 74.4 (1967), pp. 402–405.

[Sio58] M. Sion. "On general minimax theorems". *Pacific Journal of Mathematics* 8.1 (1958), pp. 171–176.

[Sjo+20] M. W. Sjoding, R. P. Dickson, T. J. Iwashyna, S. E. Gay, and T. S. Valley. "Racial bias in pulse oximetry measurement". *New England Journal of Medicine* 383.25 (2020), pp. 2477–2478.

[Smi95] J. E. Smith. "Generalized Chebychev inequalities: Theory and applications in decision analysis". *Operations Research* 43.5 (1995), pp. 807–825.

[SMK15] S. Shafieezadeh-Abadeh, P. Mohajerin Esfahani, and D. Kuhn. "Distributionally robust logistic regression". *Advances in Neural Information Processing Systems.* 2015, pp. 1576–1584.

[SNB05] V. Sindhwani, P. Niyogi, and M. Belkin. "A co-regularization approach to semi-supervised learning with multiple views". *ICML workshop on learning with multiple views.* 2005, pp. 74–79.

[Søg13] A. Søgaard. "Semi-supervised learning and domain adaptation in natural language processing". *Synthesis Lectures on Human Language Technologies* 6.2 (2013), pp. 1–103.

[Sol+14] J. Solomon, R. Rustamov, L. Guibas, and A. Butscher. "Earth mover's distances on discrete surfaces". *ACM Transactions on Graphics* 33.4 (2014), p. 67.

[Sol+15] J. Solomon, F. De Goes, G. Peyré, M. Cuturi, A. Butscher, A. Nguyen, T. Du, and L. Guibas. "Convolutional Wasserstein distances: Efficient optimal transportation on geometric domains". *ACM Transactions on Graphics* 34.4 (2015), p. 66.

[Spe+18] T. Speicher, M. Ali, G. Venkatadri, F. N. Ribeiro, G. Arvanitakis, F. Benevenuto, K. P. Gummadi, P. Loiseau, and A. Mislove. "Potential for discrimination in online targeted advertising". *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 5–19.

[SRB11] M. Schmidt, N. L. Roux, and F. Bach. "Convergence rates of inexact proximal-gradient methods for convex optimization". *Advances in Neural Information Processing Systems*. 2011, pp. 1458–1466.

[SS+09] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. "Stochastic Convex Optimization". *Conference on Learning Theory*. 2009.

[SS+11] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. "Pegasos: Primal estimated sub-gradient solver for SVM". *Mathematical programming* 127.1 (2011), pp. 3–30.

[SST10] N. Srebro, K. Sridharan, and A. Tewari. "Optimistic rates for learning with a smooth loss". *arXiv:1009.3896* (2010).

[STD19] T. Sun and Q. Tran-Dinh. "Generalized self-concordant functions: A recipe for Newton-type methods". *Mathematical Programming* 178.1-2 (2019), pp. 145–213.

[Sti18] G. Still. *Lectures on Parametric Optimization: An Introduction*. 2018.

[Sug+08] M. Sugiyama, T. Suzuki, S. Nakajima, H. Kashima, P. von Bünau, and M. Kawanabe. "Direct importance estimation for covariate shift adaptation". *Annals of the Institute of Statistical Mathematics* 60.4 (2008), pp. 699–746.

[Sun+11] Q. Sun, R. Chattopadhyay, S. Panchanathan, and J. Ye. "A Two-Stage Weighting Framework for Multi-Source Domain Adaptation". *Advances in Neural Information Processing Systems*. Vol. 24. 2011, pp. 505–513.

[ŠVV12] D. Štefankovič, S. Vempala, and E. Vigoda. "A deterministic polynomial-time approximation scheme for counting knapsack solutions". *SIAM Journal on Computing* 41.2 (2012), pp. 356–366.

[Ta23] B. Taşkesen, S. Shafieezadeh-Abadeh, D. Kuhn, and K. Natarajan. "Discrete Optimal Transport with Independent Marginals is # P-Hard". *SIAM Journal on Optimization* 33.2 (2023), pp. 589–614.

[Tal10] N. N. Taleb. *The Black Swan: The Impact of the Highly Improbable: With a new section:" On Robustness and Fragility"*. Vol. 2. Random house trade paperbacks, 2010.

[Taş+20] B. Taşkesen, V. A. Nguyen, D. Kuhn, and J. Blanchet. "A Distributionally Robust Approach to Fair Classification". *(arXiv:2007.09530)* (2020).

[Taş+21a] B. Taşkesen, J. Blanchet, D. Kuhn, and V. A. Nguyen. "A Statistical Test for Probabilistic Fairness". *ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 648–665.

[Taş+21b] B. Taşkesen, M.-C. Yue, J. Blanchet, D. Kuhn, and V. A. Nguyen. "Sequential domain adaptation by synthesizing distributionally robust experts". *International Conference on Machine Learning*. 2021, pp. 10162–10172.

[Taş+23] B. Taşkesen, D. A. Iancu, Ç. Koçyiğit, and D. Kuhn. "Distributionally Robust Linear Quadratic Control". *Conference on Neural Information Processing Systems*. 2023.

[Tho+17] M. Thorpe, S. Park, S. Kolouri, G. K. Rohde, and D. Slepčev. "A Transportation L^p Distance for Signal Analysis". *Journal of Mathematical Imaging and Vision* 59.2 (2017), pp. 187–210.

[Thu27] L. L. Thurstone. "A law of comparative judgment." *Psychological Review* 34.4 (1927), p. 273.

[TJ02] E. Todorov and M. I. Jordan. "Optimal feedback control as a theory of motor coordination". *Nature Neuroscience* 5.11 (2002), pp. 1226–1235.

[TKW16] J. Townsend, N. Koep, and S. Weichwald. "Pymanopt: A Python Toolbox for Optimization on Manifolds using Automatic Differentiation". *Journal of Machine Learning Research* 17.137 (2016), 1–5. URL: http://jmlr.org/papers/v17/16-177.html.

[TPG16] G. Tartavel, G. Peyré, and Y. Gousseau. "Wasserstein loss for image synthesis and restoration". *SIAM Journal on Imaging Sciences* 9.4 (2016), pp. 1726–1755.

[Tra09] K. E. Train. *Discrete Choice Methods with Simulation*. Cambridge University Press, 2009.

[Tra+17] F. Tramer, V. Atlidakis, R. Geambasu, D. Hsu, J.-P. Hubaux, M. Humbert, A. Juels, and H. Lin. "FairTest: Discovering unwarranted associations in data-driven applications". *2017 IEEE European Symposium on Security and Privacy (EuroS&P)*. IEEE. 2017, pp. 401–416.

[TSAK23] B. Taşkesen, S. Shafieezadeh-Abadeh, and D. Kuhn. "Semi-discrete Optimal Transport: Hardness, Regularization and Numerical Solution". *Mathematical Programming* 199.1-2 (2023), pp. 1033–1106.

[Tsy03] A. B. Tsybakov. "Optimal rates of aggregation". *Conference on Learning Theory*. 2003, pp. 303–313.

[Tze+15] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. "Simultaneous deep transfer across domains and tasks". *IEEE International Conference on Computer Vision*. 2015, pp. 4068–4076.

[Val79a] L. G. Valiant. "The complexity of computing the permanent". *Theoretical Computer Science* 8.2 (1979), pp. 189–201.

[Val79b] L. G. Valiant. "The complexity of enumeration and reliability problems". *SIAM Journal on Computing* 8.3 (1979), pp. 410–421.

[Vil03] C. Villani. *Topics in Optimal Transportation*. American Mathematical Soc., 2003.

[Vil08] C. Villani. *Optimal Transport: Old and New*. Springer, 2008.

[VL90] J. Van Leeuwen. *Handbook of Theoretical Computer Science: Algorithms and Complexity*. Elsevier, 1990.

[VP+16] B. P. G. Van Parys, D. Kuhn, P. J. Goulart, and M. Morari. "Distributionally Robust Control of Constrained Stochastic Systems". *IEEE Transactions on Automatic Control* 61.2 (2016), pp. 430–442.

[VPGM13] B. P. G. Van Parys, P. J. Goulart, and M. Morari. “Infinite Horizon Performance Bounds for Uncertain Constrained Systems”. *IEEE Transactions on Automatic Control* 58.11 (2013), pp. 2803–2817.

[VV09] V. Vapnik and A. Vashist. “A new learning paradigm: Learning using privileged information”. *Neural Networks* 22.5-6 (2009), pp. 544–557.

[Wan+10] W. Wang, J. A. Ozolek, D. Slepcev, A. B. Lee, C. Chen, and G. K. Rohde. “An optimal transportation approach for nuclear structure-based pathology”. *IEEE Transactions on Medical Imaging* 30.3 (2010), pp. 621–631.

[Wan+20a] H. Wang, A. Liu, Z. Yu, Y. Yue, and A. Anandkumar. “Distributionally Robust Learning for Unsupervised Domain Adaptation”. *arXiv:2010.05784* (2020).

[Wan+20b] S. Wang, W. Guo, H. Narasimhan, A. Cotter, M. Gupta, and M. I. Jordan. “Robust Optimization for Fairness with Noisy Protected Groups”. *arXiv:2002.09343* (2020).

[WC03] H. J. Wassenaar and W. Chen. “An approach to decision-based design with discrete choice analysis for demand modeling”. *Transactions of ASME: Journal of Mechanical Design* 125.3 (2003), pp. 490–497.

[WC20] G. Wilson and D. J. Cook. “A survey of unsupervised deep domain adaptation”. *ACM Transactions on Intelligent Systems and Technology* 11.5 (2020), pp. 1–46.

[WD18] M. Wang and W. Deng. “Deep visual domain adaptation: A survey”. *Neurocomputing* 312 (2018), pp. 135 –153.

[Wee18] J. Weed. “An explicit analysis of the entropic penalty in linear programming”. *Conference On Learning Theory.* 2018, pp. 1841–1855.

[Wex+19] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viégas, and J. Wilson. “The what-if tool: Interactive probing of machine learning models”. *IEEE transactions on visualization and computer graphics* 26.1 (2019), pp. 56–65.

[Whi81] P. Whittle. “Risk-Sensitive Linear/Quadratic/Gaussian Control”. *Advances in Applied Probability* 13.4 (1981), pp. 764–777.

[Win17] O. Wintenberger. “Optimal learning with Bernstein online aggregation”. *Machine Learning* 106.1 (2017), pp. 119–141.

[WKS14] W. Wiesemann, D. Kuhn, and M. Sim. “Distributionally Robust Convex Optimization”. *Operations Research* 62.6 (2014), pp. 1358–1376.

[WKW16] K. Weiss, T. M. Khoshgoftaar, and D. Wang. “A survey of transfer learning”. *Journal of Big Data* 3.1 (2016), pp. 1–40.

[Woo+17] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. “Learning Non-Discriminatory Predictors”. *Proceedings of the 2017 Conference on Learning Theory.* 2017, pp. 1920–1953.

[Xia09] L. Xiao. “Dual Averaging Method for Regularized Stochastic Learning and Online Optimization”. *Advances in Neural Information Processing Systems.* 2009, pp. 2116–2124.

[XYS20] S. Xue, M. Yurochkin, and Y. Sun. "Auditing ML Models for Individual Bias and Unfairness". *arXiv:2003.05048* (2020).

[Yan21] I. Yang. "Wasserstein Distributionally Robust Stochastic Control: A Data-Driven Approach". *IEEE Transactions on Automatic Control* 66.8 (2021), pp. 3863–3870.

[Yao+15] T. Yao, Y. Pan, C.-W. Ngo, H. Li, and T. Mei. "Semi-supervised domain adaptation with subspace learning for visual recognition". *IEEE conference on Computer Vision and Pattern Recognition*. 2015, pp. 2142–2150.

[YBS20] M. Yurochkin, A. Bower, and Y. Sun. "Training individually fair ML models with sensitive subspace robustness". *International Conference on Learning Representations*. 2020.

[YKW20] M.-C. Yue, D. Kuhn, and W. Wiesemann. "On Linear Optimization over Wasserstein Balls". *arXiv:2004.07162* (2020).

[Zaf+17a] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment". *International Conference on World Wide Web*. 2017, pp. 1171–1180.

[Zaf+17b] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. "Fairness constraints: Mechanisms for fair classification". *AISTATS* (2017).

[ZD98] K. Zhou and J. C. Doyle. *Essentials of Robust Control*. Prentice Hall, 1998.

[Zem+13] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. "Learning fair representations". *International Conference on Machine Learning*. 2013, pp. 325–333.

[ZG18] C. Zhao and Y. Guan. "Data-driven risk-averse stochastic optimization with Wasserstein metric". *Operations Research Letters* 46.2 (2018), pp. 262 –267.

[Zha+18] H. Zhao, S. Zhang, G. Wu, J. M. F. Moura, J. P. Costeira, and G. J. Gordon. "Adversarial Multiple Source Domain Adaptation". *Advances in Neural Information Processing Systems*. Vol. 31. 2018.

[ZLM18] B. H. Zhang, B. Lemoine, and M. Mitchell. "Mitigating unwanted biases with adversarial learning". *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 335–340.

[ZT19] Y. Zhang and P. Trubey. "Machine learning and sampling scheme: An empirical study of money laundering detection". *Computational Economics* 54 (2019), pp. 1043–1063.