Proximal Point Imitation Learning

Luca Viano
LIONS, EPFL
Lausanne, Switzerland
luca.viano@epfl.ch

Angeliki Kamoutsi
ETH Zurich
Zurich, Switzerland
kamoutsa@ethz.ch

Gergely Neu
Universitat Pompeu Fabra
Barcelona, Spain
gergely.neu@gmail.com

Igor Krawczuk
LIONS, EPFL
Lausanne, Switzerland
igor.krawczuk@epfl.ch

Volkan Cevher
LIONS, EPFL
Lausanne, Switzerland
volkan.cevher@epfl.ch

Abstract

This work develops new algorithms with rigorous efficiency guarantees for infinite horizon imitation learning (IL) with linear function approximation without restrictive coherence assumptions. We begin with the minimax formulation of the problem and then outline how to leverage classical tools from optimization, in particular, the proximal-point method (PPM) and dual smoothing, for online and offline IL, respectively. Thanks to PPM, we avoid nested policy evaluation and cost updates for online IL appearing in the prior literature. In particular, we do away with the conventional alternating updates by the optimization of a single convex and smooth objective over both cost and Q-functions. When solved inexactly, we relate the optimization errors to the suboptimality of the recovered policy. As an added bonus, by re-interpreting PPM as dual smoothing with the expert policy as a center point, we also obtain an offline IL algorithm enjoying theoretical guarantees in terms of required expert trajectories. Finally, we achieve convincing empirical performance for both linear and neural network function approximation.

1 Introduction

This work is concerned with the prototypical setting of imitation learning (IL) where

1. An expert provides demonstrations of state-action pairs in an environment. The expert could be optimal or suboptimal with respect to an unknown cost/reward function.
2. The learner chooses distance measure between its policy to be learned and the expert empirical distribution estimated from demonstrations.
3. The learner employs an algorithm, which additionally may or may not use interactions with the environment, to minimize the chosen distance.

In IL, the central goal of the learner is to recover a policy competitive with expert with respect to the underlying unknown cost function. IL is important for several real world applications like driving [62], robotics [88], and economics/finance [27] at the expense of following resources: (R1) expert demonstrations, (R2) (optional) interactions with the environment where the expert collected the demonstrations, and (R3) computational resources for solving the problem template.

Interestingly, while there is a vast amount of literature using optimization ideas on the IL problem template, i.e. Lagrangian duality [51, 38, 59, 63, 64], resource guarantees are still widely missing since the optimization literature focuses on the resource (R3) where IL literature mainly focuses on

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

the first two resources (R1) and (R2). Our work leverages deeper connections between optimization tools and IL by showing how classical optimization tools can be applied in a linear programming formulation of IL problem guaranteeing efficiency in all (R1), (R2), (R3).

Our contributions: This work aims at designing an algorithm enjoying both theoretical guarantees and convincing empirical performance. Our methodology is rooted in classical optimization tools and the LP approach to MDPs. More precisely, the method uses the recently repopularized overparameterization technique to obtain the Q-function as a Lagrangian multiplier [77, 14] and solves the associated program using a PPM update with appropriately chosen Bregman divergences. This results to an actor-critic algorithm, with the key feature that the policy evaluation step involves optimization of a single concave and smooth objective over both cost and Q-functions. In this way, we avoid instability or poor convergence due to adversarial training [51, 122, 70, 105], and can also recover an explicit cost along with Q-function. We further account for potential optimization errors, presenting an error propagation analysis that leads to rigorous guarantees for both online and offline setting. For the context of linear MDPs [14, 121, 55, 22, 116, 7, 84], we provide explicit convergence rates and error bounds for the suboptimality of the learned policy, under mild assumptions, significantly weaker than those found in the literature until now. To our knowledge, such guarantees in this setting are provided for the first time. Finally, we demonstrate that our approach achieves convincing empirical performance for both linear and neural network function approximation.

Related Literature. The first algorithm addressing the imitation learning problem is behavioral cloning [93]. Due to the covariate shift problem [98, 99], it has low efficiency in terms of expert trajectories (R1). To address this issue, [100, 87, 4, 95, 111, 85, 123, 5, 68, 69] proposed to cast the problem as inverse reinforcement learning (IRL). IRL improves the efficiency in terms of expert trajectories, at the cost of introducing the need of running reinforcement learning (RL) repetitively, which can be prohibitive in terms of environment samples (R2) and computation (R3). A successive line of work started with [112] highlights that repeated calls to an RL routine can be avoided. This work inspired generative adversarial imitation learning (GAIL) [51] and other follow-up works [38, 59, 63, 64] that leveraged optimization tools like primal-dual algorithms but did not try to deepen the optimization connections to derive efficiency guarantees in terms of all (R1),(R2),(R3). Finally, a recent line of work [40, 57] in IL bypasses the need of optimizing over cost functions and thus avoids instability due to adversarial training. Although these algorithms achieve impressive empirical performance in challenging high dimensional benchmark tasks, they are hampered by limited theoretical understanding. This is the fundamental difference from our work, which enjoys both favorable practical performance and strong theoretical guarantees.

Existing model-free IL theoretical papers with global convergence guarantees assume either a finite horizon episodic MDP setting [70], or tabular MDPs [105], or the infinite horizon case but with restrictive assumptions, such as linear quadratic regulator setting [21], continuous kernelized nonlinear regulator [26, 56], access to a generative model and coherence assumption on the choice of features [58, 14], bounded strong concentrability coefficients [122] or a linear transition law that can be completely specified by a finite-dimensional matrix [70]. On the other hand, we provide convergence guarantees and error bounds for the context of linear MDPs [14, 121, 55, 22, 116, 7, 84] under a mild *feature excitation* condition assumption. Despite being linear, the transition law can still have infinite degrees of freedom. To our knowledge, such guarantees in this setting are provided for the first time.

Our work applies the technique known as regularization in the online learning literature [6, 103] and Bregman proximal-point or smoothing in optimization literature [97, 82] to the LP formulation for MDPs [73, 35, 36, 17, 48, 49, 33, 34, 102, 91, 92, 1, 65, 30, 79, 115, 67, 13, 31, 55, 106]. From this perspective, we can see Deep Inverse Q-Learning [57] and IQ-Learn [40] that consider entropy regularization in the objective as smoothing using uniform distribution as center point. In our case, we instead use as center point the previous iteration of the algorithm (for the online case) or the expert (for the offline case).

From the technical point of view, the most important related works are the analysis of REPS/Q-REPS [90, 14, 89] and O-REPS [124] that first pointed out the connection between REPS and PPM. We build on their techniques with some important differences. In particular, while in the LP formulation of RL, PPM and mirror descent [15, 47] are equivalent, recognizing that they are *not equivalent* in IL is critical for stronger empirical performance. As an independent interest, our techniques can be used to improve upon the best rate for REPS in the tabular setting [89] and to extend the guarantees to linear MDPs. In order to discuss in more detail our research questions and situate them among prior related theoretical and practical works, we provide in Appendix A an extended literature review.

2 Background

2.1 Markov Decision Processes

The RL environment and its underlying dynamics are typically abstracted as an MDP given by a tuple $(\mathcal{S}, \mathcal{A}, P, \boldsymbol{\nu}_0, \mathbf{c}, \gamma)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition law, $\boldsymbol{\nu}_0 \in \Delta_{\mathcal{S}}$ is the initial state distribution, $\mathbf{c} \in [0, 1]^{|\mathcal{S}||\mathcal{A}|}$ is the cost, and $\gamma \in (0, 1)$ is the discount factor. For simplicity, we focus on problems where $\mathcal{S}$ and $\mathcal{A}$ are finite but too large to be enumerated. A *stationary Markov policy* $\pi \colon \mathcal{S} \to \Delta_{\mathcal{A}}$ interacts with the environment iteratively, starting with an initial state $s_0 \sim \boldsymbol{\nu}_0$. At round t, if the system is at state s_t, an action $a_t \sim \pi(\cdot|s_t)$ is sampled and applied to the environment. Then a cost $c(s, a)$ is incurred, and the system transitions to the next state $s_{t+1} \sim P(\cdot|s, a)$. The goal of RL is to solve the optimal control problem $\rho^\star_{\mathbf{c}} \triangleq \min_\pi \rho_{\mathbf{c}}(\pi)$, where $\rho_{\mathbf{c}}(\pi) \triangleq (1 - \gamma) \langle \boldsymbol{\nu}_0, \mathbf{V}^\pi_{\mathbf{c}} \rangle$ is the *normalized total discounted expected cost* of π.

The *state value function* $\mathbf{V}^\pi_{\mathbf{c}} \in \mathbb{R}^{|\mathcal{S}|}$ of π, given cost $\mathbf{c}$, is defined by $V^\pi_{\mathbf{c}}(s) \triangleq \mathbb{E}^\pi_s \left[\sum_{t=0}^\infty \gamma^t c(s_t, a_t) \right]$, where $\mathbb{E}^\pi_s$ denotes the expectation with respect to the trajectories generated by π starting from $s_0 = s$. The *optimal value function* $\mathbf{V}^\star_{\mathbf{c}} \in \mathbb{R}^{|\mathcal{S}|}$ is defined by $V^\star_{\mathbf{c}}(s) \triangleq \min_\pi V^\pi_{\mathbf{c}}(s)$. The *optimal state-action value function* $\mathbf{Q}^\star_{\mathbf{c}} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, given by $Q^\star_{\mathbf{c}}(s, a) \triangleq c(s, a) + \gamma \sum_{s'} V^\star_{\mathbf{c}}(s') P(s'|s, a)$, is known to characterize optimal behaviors. Indeed $\mathbf{V}^\star_{\mathbf{c}}$ is the unique solution to the *Bellman optimality equation* $V^\star_{\mathbf{c}}(s) = \min_a Q^\star_{\mathbf{c}}(s, a)$. In addition, any deterministic policy $\pi^\star_{\mathbf{c}}(s) = \arg\min_a Q^\star_{\mathbf{c}}(s, a)$ is known to be optimal.

For every policy π, we define the *normalized state-action occupancy measure* $\boldsymbol{\mu}_\pi \in \Delta_{\mathcal{S} \times \mathcal{A}}$, by $\mu_\pi(s, a) \triangleq (1 - \gamma) \sum_{t=0}^\infty \gamma^t \, \mathbb{P}^\pi_{\boldsymbol{\nu}_0} [s_t = s, a_t = a]$, where $\mathbb{P}^\pi_{\boldsymbol{\nu}_0}[\cdot]$ denotes the probability of an event when following π starting from $s_0 \sim \boldsymbol{\nu}_0$. The occupancy measure can be interpreted as the discounted visitation frequency of state-action pairs. This allows us to write $\rho_{\mathbf{c}}(\pi) = \langle \boldsymbol{\mu}_\pi, \mathbf{c} \rangle$.

2.2 Imitation Learning

Similarly to RL, the IL problem is posed in the MDP formalism, with the critical difference that the true cost $\mathbf{c}_{\text{true}}$ is unknown. Instead, we have access to a finite set of truncated trajectories sampled i.i.d. by executing an expert policy π_{E} in the environment. The goal is to learn a policy that performs better than π_{E} with respect to the unknown $\mathbf{c}_{\text{true}}$. To this end, we adopt the *apprenticeship learning* formalism [4, 112, 50, 51, 105], which carries the assumption that $\mathbf{c}_{\text{true}}$ belongs to a class of cost functions $\mathcal{C}$. We then seek an *apprentice policy* π_{A} that outperforms the expert across $\mathcal{C}$ by solving the following optimization problem

$$\zeta^\star \triangleq \min_\pi d_{\mathcal{C}}(\pi, \pi_{\text{E}}), \tag{1}$$

where $d_{\mathcal{C}}(\pi, \pi_{\text{E}}) \triangleq \max_{\mathbf{c} \in \mathcal{C}} \big(\rho_{\mathbf{c}}(\pi) - \rho_{\mathbf{c}}(\pi_{\text{E}}) \big)$ defines the $\mathcal{C}$-distance between π and π_{E} [51, 28, 122, 70]. Then, π_{A} satisfies the goal of IL, since it holds that $\rho_{\mathbf{c}_{\text{true}}}(\pi_{\text{A}}) - \rho_{\mathbf{c}_{\text{true}}}(\pi_{\text{E}}) \leq \zeta^\star \leq 0$. Intuitively, the cost class $\mathcal{C}$ distinguishes the expert from other policies. The maximization in (1) assigns high total cost to non-expert policies and low total cost to π_{E} [51], while the minimization aims to find the policy that matches the expert as close as possible with respect to $d_{\mathcal{C}}$.

By writing $d_{\mathcal{C}}$ in its *dual* form $\bar{d}_{\mathcal{C}}(\boldsymbol{\mu}_\pi, \boldsymbol{\mu}_{\pi_{\text{E}}}) \triangleq \max_{\mathbf{c} \in \mathcal{C}} \big(\langle \boldsymbol{\mu}_\pi, \mathbf{c} \rangle - \langle \boldsymbol{\mu}_{\pi_{\text{E}}}, \mathbf{c} \rangle \big)$, it can be interpreted as an *integral probability metric* [80, 60] between the occupancy measures $\boldsymbol{\mu}_\pi$ and $\boldsymbol{\mu}_{\pi_{\text{E}}}$. Depending on how $\mathcal{C}$ is chosen, $d_{\mathcal{C}}$ turns to a different metric of probability measures like the 1-Wasserstein distance [117, 32] for $\mathcal{C} = \text{Lip}_1(\mathcal{S} \times \mathcal{A})$, the total variation for $\mathcal{C} = \{\mathbf{c} \mid \|\mathbf{c}\|_\infty \leq 1\}$, or the maximum mean discrepancy for $\mathcal{C} = \{\mathbf{c} \mid \|\mathbf{c}\|_{\mathcal{H}} \leq 1\}$, where $\text{Lip}_1(\mathcal{S} \times \mathcal{A})$ denotes the space of 1-Lipschitz functions on $\mathcal{S} \times \mathcal{A}$, and $\|\cdot\|_{\mathcal{H}}$ denotes the norm of a reproducing kernel Hilbert space $\mathcal{H}$ [104].

In our theoretical analysis, we focus on linearly parameterized cost classes [111, 112, 51, 70, 105] of the form $\mathcal{C} \triangleq \{\mathbf{c}_{\mathbf{w}} \triangleq \sum_{i=1}^m w_i \boldsymbol{\phi}_i \mid \mathbf{w} \in \mathcal{W}\}$, where $\{\boldsymbol{\phi}_i\}_{i=1}^m \subset \mathbb{R}_+^{|\mathcal{S}||\mathcal{A}|}$ are fixed feature vectors, such that $\|\boldsymbol{\phi}_i\|_1 \leq 1$ for all $i \in [m]$, and $\mathcal{W}$ is a a convex constraint set for the cost weights $\mathbf{w}$. This assumption is not necessarily restrictive as usually in practice the true cost depends on just a few key properties, but the desirable weighting that specifies how different desiderata should be traded-off is unknown [4]. Moreover, the cost features can be complex nonlinear functions that can be obtained via unsupervised learning from raw state observations [20, 29]. The matrix $\mathbf{\Phi} \triangleq [\boldsymbol{\phi}_1 \quad \dots \quad \boldsymbol{\phi}_m]$ gives rise a *feature expectation vector* (FEV) $\boldsymbol{\rho}_{\mathbf{\Phi}}(\pi) \triangleq (\rho_{\boldsymbol{\phi}_1}(\pi_{\mathrm{E}}), \dots, \rho_{\boldsymbol{\phi}_m}(\pi_{\mathrm{E}}))^\intercal \in \mathbb{R}^m$ of a policy π. Then, by choosing $\mathcal{W}$ to be the ℓ_2 unit ball $B_1^m \triangleq \{\mathbf{w} \in \mathbb{R}^m \mid \|\mathbf{w}\|_2 \leq 1\}$ [4], we get a *feature expectation matching* objective $d_{\mathcal{C}}(\pi, \pi_{\pi_{\mathrm{E}}}) = \|\boldsymbol{\rho}_{\mathbf{\Phi}}(\pi) - \boldsymbol{\rho}_{\mathbf{\Phi}}(\pi_{\mathrm{E}})\|_2$, while for $\mathcal{W}$ being the probability simplex $\Delta_{[m]}$ [111, 112] we have a worst-case excess cost objective $d_{\mathcal{C}}(\pi, \pi_{\pi_{\mathrm{E}}}) = \max_{i\in[m]} \big(\rho_{\boldsymbol{\phi}_i}(\pi) - \rho_{\boldsymbol{\phi}_i}(\pi_{\mathrm{E}})\big)$. For clarity, we will replace $\mathbf{c}$ by $\mathbf{w}$ in the notation of the quantities defined in Section 2.1.

3 A Q-Convex-Analytic Viewpoint

Our methodology builds upon the convex-analytic approach to AL, first introduced by [112], with the key difference that we consider a different convex formulation that introduces Q-functions as slack variables. This allows to design a practical scalable model-free algorithm with theoretical guarantees.

Let $\mathfrak{F} \triangleq \{\boldsymbol{\mu} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \mid (\mathbf{B} - \gamma\mathbf{P})^\intercal\boldsymbol{\mu} = (1-\gamma)\boldsymbol{\nu}_0,\ \boldsymbol{\mu} \geq \mathbf{0}\}$ be the *state-action polytope*, where $\mathbf{P}$ is the vector form of P, i.e., $P_{(s,a),s'} \triangleq P(s'|s,a)$, and $\mathbf{B}$ is a binary matrix defined by $B_{(s,a),s'} \triangleq 1$ if $s = s'$, and $B_{(s,a),s'} \triangleq 0$ otherwise. The linear constraints that define the set $\mathfrak{F}$, also known as *Bellman flow constraints*, precisely characterize the set of state-action occupancy measures.

Proposition 1 (94). *We have that $\boldsymbol{\mu} \in \mathfrak{F}$ if and only if there exists a unique stationary Markov policy π such that $\boldsymbol{\mu} = \boldsymbol{\mu}_\pi$. If $\boldsymbol{\mu} \in \mathcal{F}$ then the policy $\pi_{\boldsymbol{\mu}}(a|x) \triangleq \frac{\boldsymbol{\mu}(x,a)}{\sum_{a'\in\mathcal{A}}\boldsymbol{\mu}(x,a')}$ has occupancy measure $\boldsymbol{\mu}$.*

Using Proposition 1 and the dual form of the $\mathcal{C}$-distance $\bar{d}_{\mathcal{C}}(\boldsymbol{\mu}, \boldsymbol{\mu}_{\pi_{\mathrm{E}}}) = \max_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}} \rangle$, it follows that (1) is equivalent to the primal convex program $\zeta^\star = \min_{\boldsymbol{\mu}}\{\bar{d}_{\mathcal{C}}(\boldsymbol{\mu}, \boldsymbol{\mu}_{\pi_{\mathrm{E}}}) \mid \boldsymbol{\mu} \in \mathfrak{F}\}$. In particular for $\mathcal{W} = \Delta_{[m]}$ and by using an epigraphic transformation, we end up with an LP program [112], while for $\mathcal{W} = B_1^m$ we get a quadratic objective with linear constraints [4].

A slight variation of the above reasoning is to introduce a mirror variable $\mathbf{d}$ and split the Bellman flow constraints in the definition of $\mathfrak{F}$. We then get the primal convex program

$$\zeta^\star = \min_{(\boldsymbol{\mu},\mathbf{d})}\{\bar{d}_{\mathcal{C}}(\boldsymbol{\mu}, \boldsymbol{\mu}_{\pi_{\mathrm{E}}}) \mid (\boldsymbol{\mu}, \mathbf{d}) \in \mathfrak{M}\}, \tag{Primal}$$

where the new polytope is given by $\mathfrak{M} \triangleq \{(\boldsymbol{\mu}, \mathbf{d}) \mid \mathbf{B}^\intercal\mathbf{d} = \gamma\mathbf{P}^\intercal\boldsymbol{\mu} + (1-\gamma)\boldsymbol{\nu}_0,\ \boldsymbol{\mu} = \mathbf{d},\ \mathbf{d} \geq \mathbf{0}\}$. This overparameterization trick has been first introduced by Mehta and Meyn [76] and has been recently revisited by [14, 84, 67, 83, 77, 71]. A salient feature of this equivalent formulation is that it introduces a Q-function as Lagrange multiplier to the equality constraint $\mathbf{d} = \boldsymbol{\mu}$, and so lends itself to data-driven algorithms. To motivate further this new formulation, in Appendix C, we shed light to its dual and provide an interpretation of the dual optimizers. In particular, when $\mathcal{W} = B_1^m$, we show that $(\mathbf{V}^\star_{\mathbf{w}_{\text{true}}}, \mathbf{Q}^\star_{\mathbf{w}_{\text{true}}}, \mathbf{w}_{\text{true}})$ is a dual optimizer.

For our theoretical analysis we focus on the linear MDP setting [55], i.e., we assume that the transition law is linear in the feature mapping. We denote by $\boldsymbol{\phi}(s,a)$ the (s,a)-th row of $\mathbf{\Phi}$.

Assumption 1 (Linear MDP). *There exists a collection of m probability measures $\boldsymbol{\omega} = (\omega_1, \dots, \omega_m)$ on $\mathcal{S}$, such that $P(\cdot|s,a) = \langle \boldsymbol{\omega}(\cdot), \phi(s,a)\rangle$, for all (s,a). Moreover $\boldsymbol{\phi}(s,a) \in \Delta_{[m]}$, for all (s,a).*

Assumption 1 essentialy says that the transition matrix $\mathbf{P}$ has rank at most m, and $\mathbf{P} = \mathbf{\Phi}\mathbf{M}$ for some matrix $\mathbf{M} \in \mathbb{R}^{m\times|\mathcal{S}|}$. It is worth noting that in the case of continuous MDPs, despite being linear, the transition law $P(\cdot|s,a)$ can still have infinite degrees of freedom. This is a substantial difference from the recent theoretical works on IL [70, 105] which consider either a linear quadratic regulator, or a transition law that can be completely specified by a finite-dimensional matrix such that the degrees of freedom are bounded.

Assumption 1 enables us to consider a relaxation of (Primal). In particular, we aggregate the constraints $\boldsymbol{\mu} = \mathbf{d}$ by imposing $\mathbf{\Phi}^\intercal\boldsymbol{\mu} = \mathbf{\Phi}^\intercal\mathbf{d}$ instead, and introduce a variable $\boldsymbol{\lambda} = \mathbf{\Phi}^\intercal\boldsymbol{\mu}$. It follows that $\boldsymbol{\lambda}$ lies in the m-dimensional simplex $\Delta_{[m]}$. Then, we get the following convex program

$$\zeta^\star = \min_{(\boldsymbol{\lambda},\mathbf{d})}\{\max_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\lambda}, \mathbf{w}\rangle - \langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}} \rangle \mid (\boldsymbol{\lambda}, \mathbf{d}) \in \mathfrak{M}_{\mathbf{\Phi}}\}, \tag{Primal$'$}$$

where $\mathfrak{M}_{\mathbf{\Phi}} \triangleq \{(\boldsymbol{\lambda}, \mathbf{d}) \mid \mathbf{B}^\intercal \mathbf{d} = \gamma \mathbf{M}^\intercal \boldsymbol{\lambda} + (1-\gamma)\boldsymbol{\nu}_0,\ \boldsymbol{\lambda} = \mathbf{\Phi}^\intercal \mathbf{d},\ \boldsymbol{\lambda} \in \Delta_{[m]},\ \mathbf{d} \in \Delta_{\mathcal{S}\times\mathcal{A}}\}$. As shown in [84, 14, 83], for linear MDPs, the set of occupancy measures $\mathfrak{F}$ can be completely characterized by the set $\mathfrak{M}_{\mathbf{\Phi}}$ (c.f., Proposition 2). While the number of constraints and variables in (Primal′) is intractable for large scale MDPs, in the next paragraph, we show how this problem can be solved using a proximal point scheme.

4 Proximal Point Imitation Learning

By using a Lagrangian decomposition, we have that (Primal′) is equivalent to the following bilinear saddle-point problem

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \langle \mathbf{y}, \mathbf{A}\mathbf{x} + \mathbf{b}\rangle, \tag{SPP}$$

where $\mathbf{A} \in \mathbb{R}^{(2m+|\mathcal{S}|)\times(m+|\mathcal{S}||\mathcal{A}|)}$, and $\mathbf{b} \in \mathbb{R}^{(m+|\mathcal{S}|+|\mathcal{S}||\mathcal{A}|)}$ are appropriately defined (see Appendix D), $\mathbf{x} \triangleq [\boldsymbol{\lambda}^\intercal, \mathbf{d}^\intercal]^\intercal$, $\mathbf{y} \triangleq [\mathbf{w}^\intercal, \mathbf{V}^\intercal, \boldsymbol{\theta}^\intercal]^\intercal$, $\mathcal{X} \triangleq \Delta_{[m]} \times \Delta_{\mathcal{S}\times\mathcal{A}}$, and $\mathcal{Y} \triangleq \mathcal{W} \times \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}^m$.

Since in practice we do not have access to the whole policy π_{E}, but instead can observe a finite set of i.i.d. sample trajectories $\mathcal{D}_{\mathrm{E}} \triangleq \{(x_0^{(l)}, a_0^{(l)}, x_1^{(l)}, a_1^{(l)}, \ldots, x_H^{(l)}, a_H^{(l)})\}_{l=1}^{n_{\mathrm{E}}} \sim \pi_{\mathrm{E}}$, we define the vector $\widehat{\mathbf{b}}$ by replacing $\boldsymbol{\rho}_{\mathbf{\Phi}}(\pi_{\mathrm{E}})$ with its empirical counterpart $\boldsymbol{\rho}_{\mathbf{\Phi}}(\widehat{\pi_{\mathrm{E}}})$ (by taking sample averages) in the definition of $\mathbf{b}$. We then consider the empirical objective $f(\mathbf{x}) \triangleq \max_{\mathbf{y}\in\mathcal{Y}} \left\langle \mathbf{y}, \mathbf{A}\mathbf{x} + \widehat{\mathbf{b}} \right\rangle$ and apply PPM on the decision variable $\mathbf{x}$. For the $\boldsymbol{\lambda}$-variable we use the relative entropy $D(\boldsymbol{\lambda}||\boldsymbol{\lambda}') \triangleq \sum_{i=1}^m \lambda(i) \log \frac{\lambda(i)}{\lambda'(i)}$, while for the occupancy measure $\mathbf{d}$ we use the conditional relative entropy $H(\mathbf{d}||\mathbf{d}') \triangleq \sum_{s,a} d(s,a) \log \frac{\pi_{\mathbf{d}}(a|s)}{\pi_{\mathbf{d}'}(a|s)}$. With this choice we can rewrite the PPM update as

$$(\boldsymbol{\lambda}_{k+1}, \mathbf{d}_{k+1}) = \underset{\boldsymbol{\lambda}\in\Delta_{[m]}, \mathbf{d}\in\Delta_{\mathcal{S}\times\mathcal{A}}}{\arg\min} \max_{\mathbf{y}\in\mathcal{Y}} \left\langle \mathbf{y}, \mathbf{A} \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{d} \end{bmatrix} + \widehat{\mathbf{b}} \right\rangle + \frac{1}{\eta} D(\boldsymbol{\lambda}||\mathbf{\Phi}^\intercal \mathbf{d}_k) + \frac{1}{\alpha} H(\mathbf{d}||\mathbf{d}_k), \tag{2}$$

where we used primal feasibility to replace $\boldsymbol{\lambda}_k$ with $\mathbf{\Phi}^\intercal \mathbf{d}_k$ as the center point of the relative entropy. PPM is implicit, meaning that it requires the evaluation of the gradient at the next iterate $\mathbf{x}_{k+1}$. Such a requirement makes it not implementable in general. However, in the following, we describe a procedure to apply proximal point to our specific $f(\mathbf{x})$. The following Proposition summarizes the result.

Proposition 2. *For a parameter $\boldsymbol{\theta} \in \mathbb{R}^m$, we define the logistic state-action value function $\mathbf{Q}_{\boldsymbol{\theta}} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ by $\mathbf{Q}_{\boldsymbol{\theta}} \triangleq \mathbf{\Phi}\boldsymbol{\theta}$, and the k-step logistic state value function $\mathbf{V}_{\boldsymbol{\theta}}^k \in \mathbb{R}^{|\mathcal{S}|}$ by*

$$V_{\boldsymbol{\theta}}^k(s) \triangleq -\frac{1}{\alpha} \log \left(\sum_a \pi_{\mathbf{d}_{k-1}}(a|s) e^{-\alpha Q_{\boldsymbol{\theta}}(s,a)} \right).$$

Moreover, we define the k-step reduced Bellman error function $\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}^k \in \mathbb{R}^m$ by $\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}^k \triangleq \mathbf{w} + \gamma \mathbf{M} \mathbf{V}_{\boldsymbol{\theta}}^k - \boldsymbol{\theta}$. Then, the PPM update $(\boldsymbol{\lambda}_k^\star, \mathbf{d}_k^\star)$ in 2 is given by

$$\lambda_k^\star(i) \propto (\mathbf{\Phi}^\intercal \mathbf{d}_{k-1})(i)\, e^{-\eta \delta_{\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star}^k(i)}, \tag{3}$$

$$\pi_{\mathbf{d}_k^\star}(a|s) \propto \pi_{\mathbf{d}_{k-1}}(a|s)\, e^{-\alpha Q_{\boldsymbol{\theta}_k^\star}(s,a)}, \tag{4}$$

where $(\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star)$ is the maximizer over $\mathcal{W} \times \mathbb{R}^m$ of the k-step logistic policy evaluation objective

$$\mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta}) \triangleq -\frac{1}{\eta} \log \sum_{i=1}^m (\mathbf{\Phi}^\intercal \mathbf{d}_{k-1})(i) e^{-\eta \delta_{\mathbf{w},\boldsymbol{\theta}}^k(i)} + (1-\gamma) \left\langle \boldsymbol{\nu}_0, \mathbf{V}_{\boldsymbol{\theta}}^k \right\rangle - \langle \boldsymbol{\rho}_{\mathbf{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w} \rangle. \tag{5}$$

Moreover, it holds that $\mathcal{G}_k(\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star) = \langle \boldsymbol{\lambda}_k^\star, \mathbf{w}_k^\star \rangle - \langle \boldsymbol{\rho}_{\mathbf{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}_k^\star \rangle + \frac{1}{\eta} D(\boldsymbol{\lambda}_k^\star||\mathbf{\Phi}^\intercal \boldsymbol{\lambda}_{k-1}) + \frac{1}{\alpha} H(\mathbf{d}_k^\star||\mathbf{d}_{k-1})$. If in addition Assumption 1 holds, then $\mathbf{d}_k^\star$ is a valid occupancy measure, i.e., $\mathbf{d}_k^\star \in \mathfrak{F}$ and so $\mathbf{d}_k^\star = \boldsymbol{\mu}_{\pi_{\mathbf{d}_k^\star}}$.

The proof of Proposition 2 is broken down into a sequence of lemmas and is presented in Appendix E. It employs an `analytical-oracle` $\mathbf{g}: \mathcal{Y} \to \mathcal{X}$ given by

$$\mathbf{g}(\mathbf{y}; \mathbf{x}_k) \triangleq \underset{\boldsymbol{\lambda}\in\Delta_{[m]}, \mathbf{d}\in\Delta_{\mathcal{S}\times\mathcal{A}}}{\arg\min} \left\langle \mathbf{y}, \mathbf{A} \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{d} \end{bmatrix} + \widehat{\mathbf{b}} \right\rangle + \frac{1}{\eta} D(\boldsymbol{\lambda}||\mathbf{\Phi}^\intercal \mathbf{d}_k) + \frac{1}{\alpha} H(\mathbf{d}||\mathbf{d}_k),$$

and a `max-oracle` $\mathbf{h} : \mathcal{X} \to \mathcal{Y}$ given by $\mathbf{h}(\mathbf{x}) \triangleq \arg\max_{\mathbf{y}\in\mathcal{Y}} \langle \mathbf{y}, \mathbf{A}\mathbf{g}(\mathbf{y};\mathbf{x})\rangle + \frac{1}{\tau} D_\Omega(\mathbf{g}(\mathbf{y};\mathbf{x})||\mathbf{x})$, where we used D_Ω to compact the two divergences. By noting that the PPM update Equation (2) can be rewritten as $\mathbf{x}_{k+1} = \mathbf{g}(\mathbf{h}(\mathbf{x}_k);\mathbf{x}_k)$, its analytical computation is reduced to the characterization of the two aforementioned oracles. In particular, the updates (3)–(4) come from the `analytical-oracle` while (5) is the objective of the `max-oracle`.

The choice of conditional entropy as Bregman divergence for the $\boldsymbol{\lambda}$ variable living in the probability simplex is standard in the optimization literature and is known to mitigate the effect of dimension. In particular, as noted in [85], the classic REPS algorithm [90] can be seen as mirror descent with relative entropy regularization. On the other hand, the choice of conditional entropy as Bregman divergence for the $\mathbf{d}$ variable is less standard and has been popularized by Q-REPS [14]. Such particular divergence leads to an actor-critic algorithm that comes with several merits. By Proposition 2, it is apparent that we get analytical softmin updates for the policy $\pi_{\mathbf{d}}$ rather than the occupancy measure $\mathbf{d}$. Moreover, these softmin updates are expressed in terms of the logistic Q-function and do not involve the unknown transition matrix $\mathbf{P}$. Consequently, we avoid the problematic occupancy measure approximation and the restrictive coherence assumption on the choice of features needed in [13, 58], as well as the biased policy updates appearing in REPS [90, 89]. In addition, the newly introduced logistic policy evaluation objective $\mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta})$ has several desired properties. It is concave and smooth in $(\mathbf{w}, \boldsymbol{\theta})$ and has bounded gradients. Therefore, it does not suffer from the pathologies of the squared Bellman error [78] and does not require heuristic gradient clipping techniques. Moreover, unlike [58] it allows a model-free implementation without the need for a generative model (see Section 4.1)

We stress the fact that the `max-oracle` of our proximal point scheme performs the cost update and policy evaluation phases jointly. This is a rather novel feature of our algorithm that differs from the separate cost update and policy evaluation step used in recent theoretical imitation learning works [122, 105, 70]. Our joint optimization over cost and Q-functions avoids instability due to adversarial training and can also recover an explicit cost along with the Q-function without requiring knowledge or additional interaction with the environment (see Section 5). It is worth noting that application of primal-dual mirror descent to (SPP) does not have this favorable property. While in the standard MDP setting, proximal point and mirror descent coincide because of the linear objective, in imitation learning proximal point optimization makes a difference. In Appendix K, we include a more detailed discussion and numerical comparison between PPM and mirror descent updates.

4.1 Practical Implementation

Exact optimization of the logistic policy evaluation objective is infeasible in practical scenarios, due to unknown dynamics and limited computation power. In this section, we design a practical algorithm that uses only sample transitions by obtaining stochastic (albeit biased) gradient estimators.

Proposition 2 gives rise to Proximal Point Imitation Learning ($\mathtt{P^2IL}$), a model-free actor-critic IRL algorithm described in Algorithm 1. The key feature of $\mathtt{P^2IL}$ is that the policy evaluation step involves optimization of a single smooth and concave objective over both cost and state-action value function parameters. In this way, we avoid instability or poor convergence in optimization due to nested policy evaluation and cost updates, as well as the undesirable properties of the widely used squared Bellman error. In particular, the kth iteration of $\mathtt{P^2IL}$ consists of the following two steps : (i) (**Critic Step**) Computation of an approximate maximizer $(\mathbf{w}_k, \boldsymbol{\theta}_k) \approx \arg\max_{\mathbf{w},\boldsymbol{\theta}} \mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta})$ of the concave logistic policy evaluation objective, by using a biased stochastic gradient ascent subroutine; (ii) (**Actor Step**) Soft-min policy update $\pi_k(a|s) \propto \pi_{k-1}(a|s)\, e^{-\alpha Q_{\boldsymbol{\theta}_k}(s,a)}$ expressed in terms of the logistic Q-function.

The domain Θ in Algorithm 1 is the ℓ_∞-ball with appropriately chosen radius D to be specified later (see Proposition 3). Moreover, $\Pi_\Theta(\mathbf{x}) \triangleq \arg\min_{\mathbf{y}\in\Theta} \|\mathbf{x}-\mathbf{y}\|_2$ (resp. $\Pi_{\mathcal{W}}(\mathbf{w})$) denotes the Euclidean projection of $\mathbf{x}$ (resp. $\mathbf{w}$) onto Θ (resp. $\mathcal{W}$).

In order to estimate the gradients $\nabla_{\boldsymbol{\theta}}\, \mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta})$ and $\nabla_{\mathbf{w}}\, \mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta})$ we invoke the Biased Stochastic Gradient Estimator subroutine (BSGE) (Algorithm 2) given in Appendix H. By using the linear MDP Assumption 1 and leveraging ridge regression and plug-in estimators, the proposed stochastic gradients can be computed via simple linear algebra with computational complexity poly$(m, n(t))$, independent of the size of the state space.

Algorithm 1 Proximal Point Imitation Learning: $\texttt{P}^2\texttt{IL}(\mathbf{\Phi}, \mathcal{D}_{\mathrm{E}}, K, \eta, \alpha)$

Input: Feature matrix $\mathbf{\Phi}$, expert demonstrations $\mathcal{D}_{\mathrm{E}}$, number of iterations K, step sizes η and α, number of SGD iterations T, SGD learning rates $\boldsymbol{\beta} = \{\beta_t\}_{t=0}^{T-1}$, number-of-samples function $n : \mathbb{N} \to \mathbb{N}$
Initialize π_0 as uniform distribution over $\mathcal{A}$
Compute the empirical FEV $\boldsymbol{\rho}_{\mathbf{\Phi}}(\widehat{\pi_{\mathrm{E}}})$ using expert demonstrations $\mathcal{D}_{\mathrm{E}}$
for $k = 1, \ldots K$ **do**
 `// Critic-step (policy evaluation)`
 Initialize $\boldsymbol{\theta}_{k,0} = \mathbf{0}$ and $\mathbf{w}_{k,0} = \mathbf{0}$
 Run π_{k-1} and collect i.i.d. samples $\mathcal{B}_k = \{(s_{k-1}^{(n)}, a_{k-1}^{(n)}, s_{k-1}^{\prime(n)})\}_{n=1}^{n(T)}$ such that $(s_{k-1}^{(n)}, a_{k-1}^{(n)}) \sim \boldsymbol{\mu}_{\pi_{k-1}}$ and $s_{k-1}^{\prime(n)} \sim \mathsf{P}(\cdot|s_{k-1}^{(n)}, a_{k-1}^{(n)})$
 for $t = 0, \ldots T-1$ **do**
  Compute biased stochastic gradient estimators

$$\big(\widehat{\nabla}_{\mathbf{w}}\mathcal{G}_k(\mathbf{w}_{k,t}, \boldsymbol{\theta}_{k,t}), \widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{G}_k(\mathbf{w}_{k,t}, \boldsymbol{\theta}_{k,t})\big) = \mathrm{BSGE}\big(k, \mathbf{w}_{k,t}, \boldsymbol{\theta}_{k,t}, n(t)\big)$$

  $\mathbf{w}_{k,t+1} = \Pi_{\mathcal{W}}\big(\mathbf{w}_{k,t} + \beta_t \widehat{\nabla}_{\mathbf{w}}\mathcal{G}_k(\mathbf{w}_{k,t}, \boldsymbol{\theta}_{k,t})\big)$
  $\boldsymbol{\theta}_{k,t+1} = \Pi_{\Theta}\big(\boldsymbol{\theta}_{k,t} + \beta_t \widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{G}_k(\mathbf{w}_{k,t}, \boldsymbol{\theta}_{k,t})\big)$
 end for
 $(\mathbf{w}_k, \boldsymbol{\theta}_k) = (\frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_{k,t}, \frac{1}{T}\sum_{t=1}^{T}\boldsymbol{\theta}_{k,t})$
 `// Actor-step (policy update)`
 Policy update: $\pi_k(a|s) \propto \pi_{k-1}(a|s)\, e^{-\alpha Q_{\boldsymbol{\theta}_k}(s,a)}$
end for
Output: Mixed policy $\widehat{\pi}_K$ of $\{\pi_k\}_{k\in[K]}$

4.2 Theoretical Analysis

The first step in our theoretical analysis is to study the propagation of optimization errors made by the algorithm on the true policy evaluation objective. In particular at each iteration step k, the ideal policy evaluation update $(\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star)$ and the ideal policy update $\pi_k^\star$ are given by $(\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star) = \arg\max_{\mathbf{w},\boldsymbol{\theta}} \mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta})$, and $\pi_k^\star(a|s) = \pi_{k-1}(a|s)e^{-\alpha(Q_{\boldsymbol{\theta}_k^\star}(s,a) - V_{\boldsymbol{\theta}_k^\star}^k(s))}$. On the other hand, consider the realised policy evaluation update $(\mathbf{w}_k, \boldsymbol{\theta}_k)$ such that $\mathcal{G}_k(\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star) - \mathcal{G}_k(\mathbf{w}_k, \boldsymbol{\theta}_k) = \epsilon_k$, the corresponding policy π_k given by $\pi_k = \pi_{k-1}(a|s)e^{-\alpha(Q_{\boldsymbol{\theta}_k}(s,a) - V_{\boldsymbol{\theta}_k}^k(s))}$, and let $\mathbf{d}_k \triangleq \boldsymbol{\mu}_{\pi_k}$. We denote by $\widehat{\pi}_K$ the extracted mixed policy of $\{\pi_k\}_{k=1}^K$. We are interested in upper-bounding the suboptimality gap $d_{\mathcal{C}}(\widehat{\pi}_K, \pi_{\mathrm{E}})$ of Algorithm 1 as a function of ε_k. To this end, we need the following assumption.

Assumption 2. *It holds that* $\lambda_{\min}(\mathbb{E}_{(s,a)\sim\mathbf{d}_k}\, \boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^\intercal) \geq \beta$, *for all* $k \in [K]$.

Assumption 2 states that every occupancy measure $\mathbf{d}_k$ induces a positive definite feature covariance matrix, and so every policy π_k explores uniformly well in the feature space. This assumption is common in the RL theory literature [2, 46, 37, 66, 3, 7]. It is also related to the condition of persistent excitation from the control literature [81].

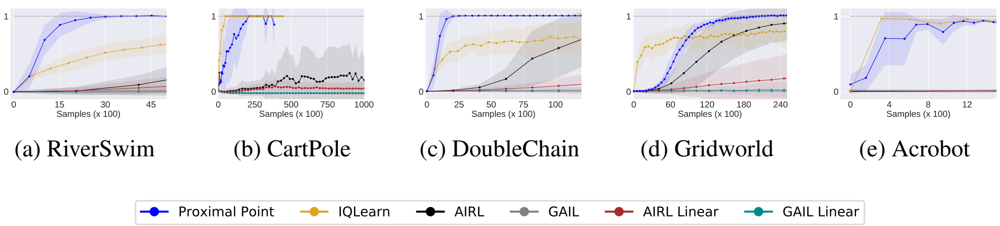
The following proposition ensures that $\max_{\mathbf{w},\boldsymbol{\theta}\in\mathcal{W}\times\mathbb{R}^m} \mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta}) = \max_{\mathbf{w},\boldsymbol{\theta}\in\mathcal{W}\times\Theta} \mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta})$. Therefore, this constraint does not change the problem optimality, but will considerably accelerate the convergence of the algorithm by considering smaller domains.

Proposition 3. *There exists a maximizer* $\boldsymbol{\theta}_k^\star$ *such that* $\|\boldsymbol{\theta}_k^\star\|_\infty \leq \frac{1+|\log\beta|}{1-\gamma} \triangleq D$.

We can now state our error propagation theorem.

Theorem 1. *Let* $\widehat{\pi}_K$ *be the output of running Algorithm 1 for* K *iterations, with* $n_{\mathrm{E}} \geq \frac{2\log(\frac{2m}{\delta})}{\varepsilon^2}$ *expert trajectories of length* $H \geq \frac{1}{1-\gamma}\log(\frac{1}{\varepsilon})$. *Let* $C \triangleq \frac{1}{\beta\eta}\Big(\sqrt{\frac{2\alpha}{1-\gamma}} + \sqrt{8\eta}\Big) + \sqrt{\frac{18\alpha}{1-\gamma}}$. *Then, with probability at least* $1-\delta$, *it holds that* $d_{\mathcal{C}}(\widehat{\pi}_K, \pi_{\mathrm{E}}) \leq \frac{1}{K}\Big(\frac{D(\boldsymbol{\lambda}^*||\mathbf{\Phi}^\intercal\mathbf{d}_0)}{\eta} + \frac{H(\mathbf{d}^*||\mathbf{d}_0)}{\alpha} + C\sum_k\sqrt{\epsilon_k} + \sum_k \epsilon_k\Big) + \varepsilon$.

By Theorem 1, whenever the policy evaluation errors ε_k, as well as the estimation error ε can be kept small, Algorithm 1 ouputs a policy $\widehat{\pi}_K$ with small suboptimality gap $\rho_{\mathbf{c}_{\mathrm{true}}}(\widehat{\pi}_K) - \rho_{\mathbf{c}_{\mathrm{true}}}(\pi_{\mathrm{E}})$.



(a) RiverSwim (b) CartPole (c) DoubleChain (d) Gridworld (e) Acrobot

Figure 1: **Online IL Experiments**. We show the total returns vs the number of env steps.

Notably, there is no direct dependence on the size of the state space or the dimension of the feature space. In the ideal case, where $\varepsilon_k = 0$ for all k, the convergence rate is $\mathcal{O}(1/K)$. The provided error propagation analysis still holds with general function approximation, i.e., in the context of deep RL. Indeed, by choosing $\mathbf{\Phi} = \mathbf{I}$, Assumption 1 is trivially satisfied and the $\boldsymbol{\theta}$ variable in the objective $\mathcal{G}_k$ is replaced by a Q-function. In practice, the estimation error ε can be made arbitrary small, by increasing the number of expert demonstrations n_{E}. Moreover, the next theorem ensures that under Assumptions 1 and 2 the biased stochastic gradient ascent (BSGA) subroutine has sublinear convergence rate.

Theorem 2. *Let $(\mathbf{w}_k, \boldsymbol{\theta}_k)$ be the output of the BSGA subroutine in Algorithm 1 for T iterations, with $n(t) \geq \max\left(\mathcal{O}\left(\frac{\gamma^2 m D t}{(\eta+\alpha)^2\beta}\log\frac{Tm}{\delta}\right), \mathcal{O}\left(\frac{mt}{(\eta+\alpha)^2\beta}\log\frac{Tm}{\delta}\right)\right)$ sample transitions, and learning rates $\beta_t = \mathcal{O}(\frac{1}{\sqrt{t}})$. Then, $\epsilon_k = \mathcal{G}_k(\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star) - \mathcal{G}_k(\mathbf{w}_k, \boldsymbol{\theta}_k) \leq \mathcal{O}(\frac{\max\{\eta,1\}mD}{\beta\sqrt{T}})$, with probability $1-\delta$.*

Corollary 1 (Resource guarantees)**.** *Choose $\eta = \alpha = 1$ and let $K = \Omega\left(\epsilon^{-1}\right)$, $T = \Omega\left(\epsilon^{-4}\right)$. Then for $\Omega\left(KT\right) = \Omega\left(\epsilon^{-5}\right)$ sample transitions, $\Omega\left(\varepsilon^{-2}\right)$ expert trajectories and approximately solving $\Omega\left(\epsilon^{-1}\right)$ concave maximization problems, we can ensure $d_{\mathcal{C}}(\widehat{\pi}, \pi_{\mathrm{E}}) \leq \mathcal{O}(\epsilon + \varepsilon)$, with high probability.*

Offline Setting. Finally, we notice that using $\mathbf{\Phi}^\intercal \boldsymbol{\mu}_{\pi_{\mathrm{E}}}$ as the reference distribution for the relative entropy we can obtain an offline algorithm that does not require environment interactions. By reinterpreting smoothing [82] as one step of proximal point, and using similar arguments as in the proof of Theorem 1, we can provide similar theoretical guarantees for the offline setting. The formal statement of the theoretical result as well as the optimization of the empirical policy evaluation objective are presented in Appendix J (see Theorems 4 and 6).

5 Experiments

In this section, we demonstrate that our approach achieves convincing empirical performance in both online and offline IL settings on several environments.[1] The precise setting is detailed in Appendix L.

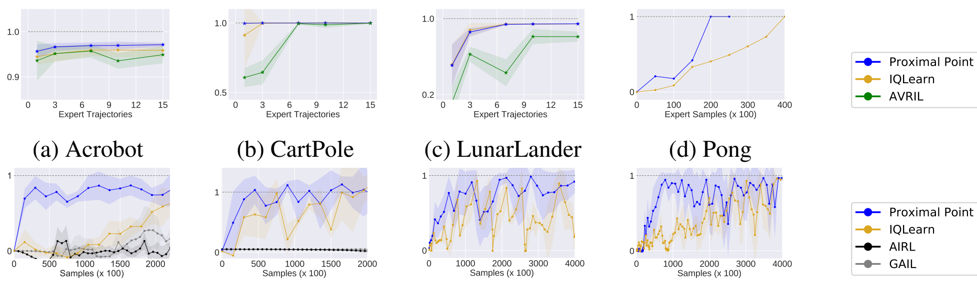
Online Setting. We first present results in various tabular environments where we can implement our algorithm without any practical relaxation outperforming GAIL [51], AIRL [38] and IQ-Learn [40]. Results are given in Figure 1. Good performance but inferior to IQ-Learn is observed also for continuous states environments (CartPole and Acrobot) where we used neural networks function approximation.

Offline Setting. Figures 2a to 2c shows that our method is competitive with the state-of-the-art offline IL methods IQLearn [40] and AVRIL [25] that recently showed performances superior to other methods like [54][64]. We also tried our algorithm in the complex image-based `Pong` task from the Atari suite. Figure 2d shows that the algorithm reaches the expert level after observing $2e5$ expert samples. We did not find AVRIL competitive in this setting, and skip it for brevity. In these settings, we verified that the algorithmic performance is convincing even for costs parameterized by neural networks.

Continuous control experiments. We attain the expert performance also in 2 MuJoCo environments: `Ant`, `HalfCheetah`, `Hopper`, and `Walker` (see Figures 2e to 2h). The additional difficulty in implementing the algorithm in continuous control experiments is that the analytical form of the policy

[1]The code is available at the following link `https://github.com/lviano/P2IL`.

improvement step is no longer computationally tractable because this would require to compute an integral over the continuous action space. Therefore, we approximated this update using the Soft Actor Critic (SAC) [44] algorithm. SAC requires environment samples making the algorithm online. The good empirical result opens the question of analyzing policy improvement errors as in [41].

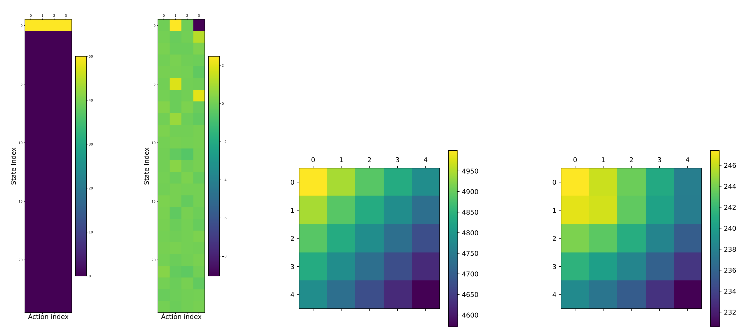


(a) Acrobot (b) CartPole (c) LunarLander (d) Pong

(e) HalfCheetah (f) Ant (g) Hopper (h) Walker2d

Figure 2: **Neural function approximation experiments.** Figures 2a to 2c show the total returns vs the number of expert trajectories. Figures 2e to 2h show the total returns vs the number of env steps. Figure 2d shows the total return vs the number of expert state-action pairs.

Recovered Costs. A unique algorithmic feature of the proposed methodology is that we can explicitly recover a cost along with the Q-function without requiring adversarial training. In Figure 3, we visualize our recovered costs in a simple 5x5 `Gridworld`. Most importantly, we verify that the recovered costs induce nearly optimal policies w.r.t. the unknown true cost function. Compared to IQ-Learn [40], we do not require knowledge or further interaction with the environment. Therefore, the recovered cost functions show promising transfer capability to new dynamics.

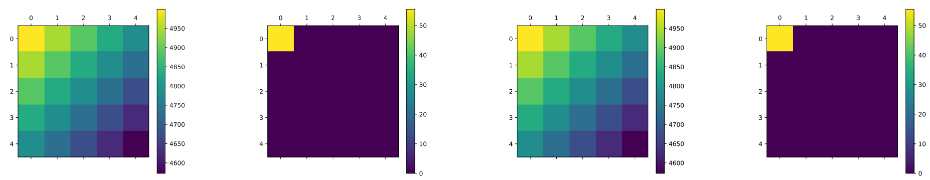


(a) $-\mathbf{c}_{\text{true}}$ (b) $-\mathbf{c}_K$ (c) -$V^\star_{\mathbf{c}_{\text{true}}}$ (d) -$V^\star_{\mathbf{c}_K}$

Figure 3: **Recovered Costs in `Gridworld`.** Comparison between the true cost $\mathbf{c}_{\text{true}}$ and the cost $\mathbf{c}_K$ recovered by $\texttt{P}^2\texttt{IL}$. We notice that the optimal value functions $V^\star_{\mathbf{c}_{\text{true}}}$ and $V^\star_{\mathbf{c}_K}$ present the same pattern. Hence, the optimal policy with respect to $\mathbf{c}_K$ is nearly optimal with respect to $\mathbf{c}_{\text{true}}$.

Cost Transfer Setting. We experimented with a transfer cost setting on a `Gridworld` (Figure 4). We consider two different Gridworld MDP environments, say M and $\widetilde{M}$, with opposite action effects. This means that action `Down` in $\widetilde{M}$ corresponds to action `Left` in M and vice versa. Similarly, the effects of `Up` and `Right` are swapped between $\widetilde{M}$ and M. We denote by $\mathbf{V}^\pi_{\widetilde{M},\mathbf{c}_{\text{true}}}$ (resp. $\mathbf{V}^\star_{\widetilde{M},\mathbf{c}_{\text{true}}}$) the value function of policy π (resp. optimal value function) in the MDP environment $\widetilde{M}$ with cost function $\mathbf{c}_{\text{true}}$. Moreover, we denote by $\pi^\star_{M,\mathbf{c}}$ the optimal policy in the MDP environment M under cost function $\mathbf{c}$. Figure (a) gives the corresponding optimal value function. Figure (b) presents the value function of the expert policy $\pi_{\text{E}} = \pi^\star_{M,\mathbf{c}_{\text{true}}}$ used as target by $\texttt{P}^2\texttt{IL}$. Figure (d) shows the value function of the learned imitating policy π_K from $\texttt{P}^2\texttt{IL}$. Finally, Figure (b) depicts the value function of the optimal policy $\pi^\star_{\widetilde{M},\mathbf{c}_K}$ for the environment $\widetilde{M}$ endowed with the recovered cost function $\mathbf{c}_K$ by

$\texttt{P}^2\texttt{IL}$ (with access to samples from M). We conclude that the policy $\pi^\star_{\widetilde{M},\mathbf{c}_K}$ is optimal in $\widetilde{M}$ with cost $\mathbf{c}_{\text{true}}$. By contrast, the expert policy $\pi_{\text{E}} = \pi^\star_{M,\mathbf{c}_{\text{true}}}$ used as target by $\texttt{P}^2\texttt{IL}$ performs poorly and as a consequence also the imitating policy π_K does so. All in all, we notice that the recovered cost induces an optimal policy for the new dynamics while the imitating policy fails. Albeit, cost transfer is successful in this experiment we do not expect this fact to be true in general because we do not tackle the issue of cost shaping [87].



(a) $-\mathbf{V}^\star_{\widetilde{M},\mathbf{c}_{\text{true}}}$ (b) $-\mathbf{V}^{\pi^\star_{M,\mathbf{c}_{\text{true}}}}_{\widetilde{M},\mathbf{c}_{\text{true}}}$ (c) $-\mathbf{V}^{\pi^\star_{\widetilde{M},\mathbf{c}_K}}_{\widetilde{M},\mathbf{c}_{\text{true}}}$ (d) $-\mathbf{V}^{\pi_K}_{\widetilde{M},\mathbf{c}_{\text{true}}}$

Figure 4: **Cost Transfer Experiment in `Gridworld`.** We compare the performance of several policies in the new MDP environment $\widetilde{M}$ with cost function $\mathbf{c}_{\text{true}}$. We notice that the recovered cost induces an optimal policy for the new dynamics while the imitating policy fails.

6 Discussion and Outlook

In this work, we studied a Proximal Point Imitation Learning ($\texttt{P}^2\texttt{IL}$) algorithm with both theoretical guarantees and convincing empirical performance. Our methodology is rooted in classical optimization tools and the LP approach to MDPs. The most significant merits of $\texttt{P}^2\texttt{IL}$ are the following: (i) It optimizes a convex and smooth logistic Bellman evaluation objective over both cost and Q-functions. In particular, it avoids instability due to adversarial training and can also recover an explicit cost along with Q function; (ii) In the context of linear MDPs, it comes with efficient resource guarantees and error bounds for the suboptimality of the learned policy (Theorem 2 and Corollary 1). In particular, given $\text{poly}(1/\varepsilon, \log(1/\delta), m)$ many samples , it recovers an ε-optimal policy, with probability $1-\delta$. Notably, the bound is independent of the size of the state-action space; (iii) Beyond the linear MDP setting, it can be implemented in a model-free manner, for both online and offline setups, with general function approximation without losing its theoretical specifications. This is justified by providing an error propagation analysis (Theorems 1 and 4), guaranteeing that small optimization errors lead to high-quality output policy; (iv) It enjoys not only strong theoretical guarantees but also favorable empirical performance. At the same time, our newly introduced methods bring challenges and open questions. One interesting question is whether one can accelerate the PPM updates and improve the convergence rate. Another direction for future work is to provide rigorous arguments for the near-optimality of the recovered cost function. On the practical side, we plan to conduct experiments in more challenging environments than MuJoCo and Atari. We hope our new techniques will be useful to future algorithm designers and lay the foundations for overcoming current limitations and challenges. In Appendix B, we point out in detail a few interesting future directions.

Acknowledgements

The authors would like to thank the anonymous reviewer for their suggestions to improve the presentation and for motivating us to inspect the recovered cost function. This work has received funding from the Enterprise for Society Center (E4S), the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme grant agreement OCAL, No. 787845, the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement n° 725594 - time-data), the Swiss National Science Foundation (SNSF) under grant number 200021_205011. Gergely Neu was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 950180). Luca Viano acknowledges travel support from ELISE (GA no 951847).

References

[1] Y. Abbasi-Yadkori, P. L. Bartlett, and A. Malek. Linear programming for large-scale Markov decision problems. In *International Conference on Machine Learning (ICML)*, 2014.

[2] Y. Abbasi-Yadkori, P. Bartlett, K. Bhatia, N. Lazic, C. Szepesvari, and G. Weisz. Politex: Regret bounds for policy iteration using expert prediction. In *International Conference on Machine Learning (ICML)*, 2019.

[3] Y. Abbasi-Yadkori, N. Lazic, C. Szepesvari, and G. Weisz. Exploration-enhanced politex. *arXiv:1908.10479*, 2019.

[4] P. Abbeel and A. Y. Ng. Apprenticeship learning via inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2004.

[5] P. Abbeel, D. Dolgov, A. Y. Ng, and S. Thrun. Apprenticeship learning for motion planning with application to parking lot navigation. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2008.

[6] J. D. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Annual Conference on Learning Theory (COLT)*, 2008.

[7] A. Agarwal, S. Kakade, A. Krishnamurthy, and W. Sun. Flambe: Structural complexity and representation learning of low rank MDPs. *Advances in neural information processing systems (NeurIPS)*, 2020.

[8] A. Ayoub, Z. Jia, C. Szepesvari, M. Wang, and L. Yang. Model-based reinforcement learning with value-targeted regression. In *International Conference on Machine Learning (ICML)*, 2020.

[9] J. A. Bagnell and J. G. Schneider. Covariant policy search. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2003.

[10] G. Banjac and J. Lygeros. A data-driven policy iteration scheme based on linear programming. In *IEEE Conference on Decision and Control (CDC)*, 2019.

[11] P. Barde, J. Roy, W. Jeon, J. Pineau, C. Pal, and D. Nowrouzezahrai. Adversarial soft advantage fitting: Imitation learning without policy optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[12] A. G. Barto, R. S. Sutton, and C. W. Anderson. Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE transactions on systems, man, and cybernetics*, pages 834–846, 1983.

[13] J. Bas-Serrano and G. Neu. Faster saddle-point optimization for solving large-scale Markov decision processes. In *Conference on Learning for Dynamics and Control (L4DC)*, 2020.

[14] J. Bas-Serrano, S. Curi, A. Krause, and G. Neu. Logistic Q-learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

[15] A. Beck and M. Teboulle. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Operations Research Letters*, 31(3):167–175, 2003.

[16] P. N. Beuchat, A. Georghiou, and J. Lygeros. Performance guarantees for model-based approximate dynamic programming in continuous spaces. *IEEE Transactions on Automatic Control*, 65(1):143–158, 2020.

[17] V. S. Borkar. A convex analytic approach to Markov decision processes. *Probability Theory and Related Fields*, 78(4):583–602, 1988.

[18] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[19] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba. OpenAI Gym. *arXiv:1606.01540*, 2016.

[20] D. S. Brown, R. Coleman, R. Srinivasan, and S. Niekum. Safe imitation learning via fast Bayesian reward inference from preferences. In *International Conference on Machine Learning (ICML)*, 2020.

[21] Q. Cai, M. Hong, Y. Chen, and Z. Wang. On the global convergence of imitation learning: a case for linear quadratic regulator. *arXiv:1901.03674*, 2019.

[22] Q. Cai, Z. Yang, C. Jin, and Z. Wang. Provably efficient exploration in policy optimization. In *International Conference on Machine Learning (ICML)*, 2020.

[23] Y. Carmon, Y. Jin, A. Sidford, and K. Tian. Variance reduction for matrix games. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[24] N. Cesa-Bianchi and G. Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[25] A. J. Chan and M. van der Schaar. Scalable Bayesian inverse reinforcement learning. *arXiv:2102.06483*, 2021.

[26] J. Chang, M. Uehara, D. Sreenivas, R. Kidambi, and W. Sun. Mitigating covariate shift in imitation learning via offline data with partial coverage. *Advances in Neural Information Processing Systems (NeuriPS)*, 2021.

[27] A. Charpentier, R. Elie, and C. Remlinger. Reinforcement learning in economics and finance. *arXiv:20031004*, 2020.

[28] M. Chen, Y. Wang, T. Liu, Z. Yang, X. Li, Z. Wang, and T. Zhao. On computation and generalization of generative adversarial imitation learning. *International Conference on Learning Representations (ICLR)*, 2020.

[29] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton. A simple framework for contrastive learning of visual representations. In *International Conference on Machine Learning (ICML)*, 2020.

[30] Y. Chen, L. Li, and M. Wang. Scalable bilinear π learning using state and action features. In *International Conference on Machine Learning (ICML)*, 2018.

[31] C.-A. Cheng, R. T. des Combes, B. Boots, and G. Gordon. A reduction from reinforcement learning to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2020.

[32] R. Dadashi, L. Hussenot, M. Geist, and O. Pietquin. Primal Wasserstein imitation learning. In *International Conference on Learning Representations (ICLR)*, 2021.

[33] D. P. De Farias and B. Van Roy. The linear programming approach to approximate dynamic programming. *Operations Research*, 51(6):850–865, 2003.

[34] D. P. De Farias and B. Van Roy. On constraint sampling in the linear programming approach to approximate dynamic programming. *Mathematics of Operations Research*, 29(3):462–478, 2004.

[35] G. T. De Ghellinck and G. D. Eppen. Linear programming solutions for separable Markovian decision problems. *Management Science*, 13(5):371–394, 1967.

[36] E. V. Denardo. On linear programming in a Markov decision problem. *Management Science*, 16(5):281–288, 1970.

[37] Y. Duan, Z. Jia, and M. Wang. Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning (ICML)*, 2020.

[38] J. Fu, K. Luo, and S. Levine. Learning robust rewards with adverserial inverse reinforcement learning. In *International Conference on Learning Representations (ICLR)*, 2018.

[39] T. Furmston and D. Barber. Variational methods for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[40] D. Garg, S. Chakraborty, C. Cundy, J. Song, and S. Ermon. IQ-learn: Inverse soft-Q learning for imitation. In *Advances in Neural Information Processing Systems (NeuRIPS)*, 2021.

[41] M. Geist, B. Scherrer, and O. Pietquin. A Theory of Regularized Markov Decision Processes. In *International Conference on Machine Learning (ICML)*, 2019.

[42] A. Geramifard, C. Dann, R. H. Klein, W. Dabney, and J. P. How. RLPy: A value-function-based reinforcement learning framework for education and research. *Journal of Machine Learning Research*, 16(46):1573–1578, 2015.

[43] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2014.

[44] T. Haarnoja, A. Zhou, P. Abbeel, and S. Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning (ICML)*, 2018.

[45] F. Hanzely, P. Richtarik, and L. Xiao. Accelerated bregman proximal gradient methods for relatively smooth convex optimization. *Computational Optimization and Applications*, 79(2): 405–440, 2021.

[46] B. Hao, T. Lattimore, C. Szepesvári, and M. Wang. Online sparse reinforcement learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2021.

[47] E. Hazan. Introduction to online convex optimization. *Foundations and Trends in Optimization*, 2(3-4):157–325, 2016.

[48] O. Hernández-Lerma and J. B. Lasserre. *Discrete-Time Markov Control Processes: Basic Optimality Criteria*. Springer-Verlag New York, 1996.

[49] O. Hernández-Lerma and J. B. Lasserre. *Further Topics on Discrete-Time Markov Control Processes*. Springer-Verlag New York, 1999.

[50] J. Ho and S. Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

[51] J. Ho, J. K. Gupta, and S. Ermon. Model-free imitation learning with policy optimization. In *International Conference on Machine Learning (ICML)*, 2016.

[52] S. R. Howard, A. Ramdas, J. McAuliffe, and J. Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2), 2021.

[53] D. Hsu, S. M. Kakade, and T. Zhang. Random design analysis of ridge regression. In *Conference on Learning Theory (COLT)*, 2012.

[54] D. Jarrett, I. Bica, and M. van der Schaar. Strictly batch imitation learning by energy-based distribution matching. *arXiv:2006.14154*, 2021.

[55] Y. Jin and A. Sidford. Efficiently solving MDPs with stochastic mirror descent. In *International Conference on Machine Learning (ICML)*, 2020.

[56] S. Kakade, A. Krishnamurthy, K. Lowrey, M. Ohnishi, and W. Sun. Information theoretic regret bounds for online nonlinear control. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[57] G. Kalweit, H. Maria, M. Werling, and J. Boedecker. Deep inverse Q-learning with constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[58] A. Kamoutsi, G. Banjac, and J. Lygeros. Efficient performance bounds for primal-dual reinforcement learning from demonstrations. In *International Conference on Machine Learning (ICML)*, 2021.

[59] L. Ke, S. Choudhury, M. Barnes, W. Sun, G. Lee, and S. Srinivasa. Imitation learning as f-divergence minimization. In *International Workshop on the Algorithmic Foundations of Robotics (WAFR)*, 2020.

[60] C. Kent, J. Li, J. Blanchet, and P. Glynn. Modified Frank Wolfe in probability space. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[61] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*, 2015.

[62] W. B. Knox, A. Allievi, H. Banzhaf, F. Schmitt, and P. Stone. Reward (mis)design for autonomous driving, 2021.

[63] I. Kostrikov, K. K. Agrawal, D. Dwibedi, S. Levine, and J. Tompson. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. In *International Conference on Learning Representations (ICLR)*, 2019.

[64] I. Kostrikov, O. Nachum, and J. Tompson. Imitation learning via off-policy distribution matching. In *International Conference on Learning Representations (ICLR)*, 2020.

[65] C. Lakshminarayanan, S. Bhatnagar, and C. Szepesvári. A linearly relaxed approximate linear program for Markov decision processes. *IEEE Transactions on Automatic Control*, 63(4): 1185–1191, 2018.

[66] N. Lazic, D. Yin, M. Farajtabar, N. Levine, D. Gorur, C. Harris, and D. Schuurmans. A maximum-entropy approach to off-policy evaluation in average-reward MDPs. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[67] D. Lee and N. He. Stochastic primal-dual Q-learning algorithm for discounted MDPs. In *American Control Conference (ACC)*, 2019.

[68] S. Levine, Z. Popović, and V. Koltun. Feature construction for inverse reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2010.

[69] S. Levine, Z. Popović, and V. Koltun. Nonlinear inverse reinforcement learning with Gaussian processes. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2011.

[70] Z. Liu, Y. Zhang, Z. Fu, Z. Yang, and Z. Wang. Learning from demonstration: Provably efficient adversarial policy imitation with linear function approximation. In *International Conference on Machine Learning (ICML)*, 2022.

[71] F. Lu, P. G. Mehta, S. P. Meyn, and G. Neu. Convex q-learning. In *2021 American Control Conference (ACC)*, pages 4749–4756, 2021. doi: 10.23919/ACC50511.2021.9483244.

[72] Y. Malitsky and M. K. Tam. A forward-backward splitting method for monotone inclusions without cocoercivity. *SIAM Journal on Optimization*, 30(2):1451–1472, 2020.

[73] A. Manne. Linear programming and sequential decisions. *Management Science*, 6(3):259–267, 1960.

[74] A. Martinelli, M. Gargiani, and J. Lygeros. Data-driven optimal control with a relaxed linear program. *arXiv:2003.08721*, 2020.

[75] C. McDiarmid. *Concentration*, pages 195–248. Springer Berlin Heidelberg, 1998.

[76] P. Mehta and S. Meyn. Q-learning and pontryagin's minimum principle. In *IEEE Conference on Decision and Control (CDC)*, 2009.

[77] P. G. Mehta and S. P. Meyn. Convex Q-learning, Part 1: Deterministic optimal control. *arXiv:2008.03559*, 2020.

[78] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

[79] P. Mohajerin Esfahani, T. Sutter, D. Kuhn, and J. Lygeros. From infinite to finite programs: explicit error bounds with applications to approximate dynamic programming. *SIAM Journal on Optimization*, 28(3):1968–1998, 2018.

[80] A. Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[81] K. S. Narendra and A. M. Annaswamy. Persistent excitation in adaptive systems. *International Journal of Control*, 45(1):127–160, 1987.

[82] Y. Nesterov. Smooth minimization of nonsmooth functions. *Math. Programming*, 103:127–152, 2005.

[83] G. Neu and J. Olkhovskaya. Online learning in mdps with linear function approximation and bandit feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[84] G. Neu and C. Pike-Burke. A unifying view of optimism in episodic reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[85] G. Neu and C. Szepesvári. Apprenticeship learning using inverse reinforcement learning and gradient methods. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, 2007.

[86] G. Neu, A. Jonsson, and V. Gómez. A unified view of entropy-regularized Markov decision processes. *arXiv:1705.07798*, 2017.

[87] A. Y. Ng and S. J. Russell. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning (ICML)*, 2000.

[88] T. Osa, J. Pajarinen, G. Neumann, J. Bagnell, P. Abbeel, and J. Peters. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 2018.

[89] A. Pacchiano, J. Lee, P. Bartlett, and O. Nachum. Near optimal policy optimization via REPS. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[90] J. Peters, K. Mülling, and Y. Altun. Relative entropy policy search. In *National Conference on Artificial Intelligence (AAAI)*, 2010.

[91] M. Petrik and S. Zilberstein. Constraint relaxation in approximate linear programs. In *International Conference on Machine Learning (ICML)*, pages 809–816, 2009.

[92] M. Petrik, G. Taylor, R. Parr, and S. Zilberstein. Feature selection using regularization in approximate linear programs for Markov decision processes. In *International Conference on International Conference on Machine Learning (ICML)*, 2010.

[93] D. A. Pomerleau. Efficient training of artificial neural networks for autonomous navigation. *Neural Computation*, 3(1):88–97, 1991.

[94] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., USA, 1st edition, 1994.

[95] N. D. Ratliff, J. A. Bagnell, and M. A. Zinkevich. Maximum margin planning. In *International Conference on Machine Learning (ICML)*, 2006.

[96] S. Reddy, A. D. Dragan, and S. Levine. SQIL: imitation learning via regularized behavioral cloning. *arXiv:1905.11108*, 2019.

[97] R. T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877–898, 1976.

[98] S. Ross and D. Bagnell. Efficient reductions for imitation learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[99] S. Ross, G. Gordon, and D. Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

[100] S. Russell. Learning agents for uncertain environments (extended abstract). In *Annual Conference on Computational Learning Theory (COLT)*, 1998.

[101] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov. Proximal policy optimization algorithms. *arXiv:1707.06347*, 2017.

[102] P. J. Schweitzer and A. Seidmann. Generalized polynomial approximations in Markovian decision processes. *Journal of Mathematical Analysis and Applications*, 110(2):568–582, 1985.

[103] S. Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012.

[104] S. Shalev-Shwartz and S. Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[105] L. Shani, T. Zahavy, and S. Mannor. Online apprenticeship learning. *arXiv:2102.06924*, 2021.

[106] R. Shariff and C. Szepesvári. Efficient planning in large MDPs with weak linear function approximation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[107] M. Sion. On general minimax theorems. *Pacific Journal of Mathematics*, 8(1):171–176, 1958.

[108] A. L. Strehl and M. L. Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

[109] T. Sutter, A. Kamoutsi, P. E. Esfahani, and J. Lygeros. Data-driven approximate dynamic programming: A linear programming approach. In *IEEE Conference on Decision and Control (CDC)*, 2017.

[110] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, second edition, 2018.

[111] U. Syed and R. E. Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2007.

[112] U. Syed, M. Bowling, and R. Schapire. Apprenticeship learning using linear programming. In *International Conference on Machine Learning (ICML)*, 2008.

[113] E. Todorov, T. Erez, and Y. Tassa. Mujoco: A physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5026–5033, 2012.

[114] N. Vieillard, T. Kozuno, B. Scherrer, O. Pietquin, R. Munos, and M. Geist. Leverage the average: an analysis of kl regularization in reinforcement learning. *Advances in Neural Information Processing Systems*, 33:12163–12174, 2020.

[115] M. Wang. Randomized linear programming solves the Markov decision problem in nearly linear (sometimes sublinear) time. *Mathematics of Operations Research*, 45(2):517–546, 2020.

[116] R. Wang, S. S. Du, L. Yang, and R. R. Salakhutdinov. On reward-free reinforcement learning with linear function approximation. *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[117] H. Xiao, M. Herman, J. Wagner, S. Ziesche, J. Etesami, and T. H. Linh. Wasserstein adversarial imitation learning. *arXiv:1906.08113*, 2019.

[118] T. Xu, Z. Li, and Y. Yu. Error bounds of imitating policies and environments. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

[119] S. Yan and N. He. Bregman augmented lagrangian and its acceleration. *arXiv preprint arXiv:2002.06315*, 2020.

[120] L. Yang and K.-C. Toh. Bregman proximal point algorithm revisited: a new inexact version and its variant. *arXiv preprint arXiv:2105.10370*, 2021.

[121] L. Yang and M. Wang. Sample-optimal parametric Q-learning using linearly additive features. In *International Conference on Machine Learning (ICML)*, 2019.

[122] Y. Zhang, Q. Cai, Z. Yang, and Z. Wang. Generative adversarial imitation learning with neural network parameterization: global optimality and convergence rate. In *International Conference on Machine Learning (ICML)*, 2020.

[123] B. D. Ziebart, A. Maas, J. A. Bagnell, and A. K. Dey. Maximum entropy inverse reinforcement learning. In *National Conference on Artificial Intelligence (AAAI)*, 2008.

[124] A. Zimin and G. Neu. Online learning in episodic Markovian decision processes by relative entropy policy search. *Advances in neural information processing systems (NeurIPS)*, 2013.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] We reflect the contribution described in the introduction with the theoretical results in Section 4, Section 4.2
   (b) Did you describe the limitations of your work? [Yes] We state the assumptions needed for the analysis in Sections 3 and 4.2
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] The work is mainly theoretical, we do not foresee potential negative impacts on the society.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes] We acknowledge habing read the review guidelines.
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We state the assumptions needed for the analysis in Section 3 Section 4.2
   (b) Did you include complete proofs of all theoretical results? [Yes] The main results are stated in Section 4.2 and the proofs are included as supplementary material.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] The code is included in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] Training details are provided in the Appendix.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] We averaged multiple seeds whenever possible computationally. We ran a single seed for the computational expensive Pong environment.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] We specified the resource in the Supplementary material.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes] We cite [40] for using their codebase and their expert data.
   (b) Did you mention the license of the assets? [Yes] We mention the license in the Supplementary.
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes] We attach the code to the supplementary.
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] We discuss this in the supplemnetary material
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A] We do not use personal data.
5. If you used crowdsourcing or conducted research with human subjects...

(a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A] We did not involve human partecipants.
(b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A] We did not involve human partecipants.
(c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A] We did not involve human partecipants.

A Related Literature (Extended)

In order to state our research questions and situate them among prior related theoretical and practical works, we provide an extended literature review.

Theoretical Imitation Learning. Our work is related to recent actor-critic IL schemes with theoretical guarantees for different MDP models, and different policy evaluation objectives (e.g., minimizing the squared Bellman error) [21, 122, 26, 70, 105]. Contrary to these actor-critic schemes, in our proximal-point imitation learning algorithm, the policy evaluation step involves optimization of a single objective over both cost and Q-functions. In this way, we avoid instability or poor convergence due to nested policy evaluation and cost update steps [40] as well as the undesirable properties of the widely used squared Bellman error [78]. Moreover, for the context of linear MDPs [14, 121, 55, 22, 115, 7, 84], we provide guarantees and convergence rates for the suboptimality of the learned policy, under mild assumptions, significantly weaker than those found in the literature until now. To our knowledge, such guarantees in this setting are provided for the first time. It is worth noting that in the case of continuous MDPs, despite being linear, the transition law can still have infinite degrees of freedom. This is a substantial difference from the recent theoretical works on IL [21, 122, 26, 70, 105] which consider either tabular MDPs [105], or a linear quadratic regulator [21], or a linear transition law that can be completely specified by a finite-dimensional matrix [70]. In the last case, the degrees of freedom are bounded, and thus mitigate the challenges in estimating the transition model. Indeed, the linear MDP setting studied in [70] reduces the unknown dynamics problem to estimating an unknown finite-dimensional matrix, which differs from our nonparametric approach. We also note that [122, 118] require the restrictive assumption of bounded concentrability coefficients, while this is not the case for the analysis in this paper. The convergence and generalization of actor-critic IL schemes for general MDPs has been studied in [28]. However, the authors in [28] only provide local optimality convergence guarantees, i.e., convergence to a stationary point. On the contrary, our algorithm provides global convergence guarantees for the linear MDP setting. Moreover, we account for potential policy evaluation errors , presenting an error propagation analysis that leads to rigorous guarantees for both online and offline setting, beyond the linear MDP assumption. Indeed, it is worth noting that the provided error propagation analysis justifies using our derived actor-critic scheme with general function approximation. A scalable deep reinforcement learning implementation is possible, without losing the theoretical guarantees of Theorem 1. The work [26] studies offline IL for the continuous kernelized nonlinear regulator and Gaussian process setting [56]. We notice that this setting is different from the linear MDP model studied in this paper, and each one does not imply the other. Finally, a recent theoretical IL work that is rooted in the LP approach to MDPs is [58]. The authors consider a Lagrangian reformulation of the problem and design a stochastic primal-dual algorithm with explicit performance bounds on the quality of the extracted policy. The most important limitations of the primal-dual algorithm [58] are (i) the need of a generative oracle, (ii) restricted coherence assumptions on the choice of features, as well as (iii) the problematic occupancy measure approximation. These limitations lead to poor practical performance for challenging high-dimensional and model-free IL setups. On the other hand, our algorithm overcomes these difficulties by applying a proximal point update to an alternative Q-LP formulation [77, 83]. This results to a model-free actor-critic scheme with explicit tractable softmax policy updates. Compared with the setting in [58], where access to a generative-model oracle is assumed, we only have the ability to execute learned policies in the underlying MDP to generate trajectories. This assumption is considerably weaker that having a simulator-based MDP, however it is stronger than having ”irreversible experience”, where the learner must follow a single trajectory without having access to a *reset action*, that obtains a new trajectory from the initial state distribution. Most importantly our algorithm enjoys not only strong theoretical guarantees, but also favorable practical performance.

Approximate Linear Programming. There is an emerging body of literature [33, 1, 30, 109, 65, 79, 115, 67, 10, 16, 74, 31, 55, 106, 14] that studies ALP for the forward RL. While this approach dates back to 1960s [73], it has recently witnessed an interesting renaissance for its potential to provide a solid formal framework for newly derived methods, as well as a deeper understanding of existing empirically successful algorithms. In this paper, we present scalable imitation learning algorithms with theoretical guarantees rooted in the LP approach, highlighting how historical key limitations have been eliminated. Prior approximate linear programming (ALP) approaches developed algorithms for solving large-scale and/or continuous MDPs on a low-dimensional subspace by reducing the number of constraints (e.g., by constraint sampling) [33, 34]. However, these prior works either scale badly with the size of the state-action spaces or require access to samples from a distribution that depends on the optimal policy. Moreover, they focus mainly on the approximation of the optimal value but not so much on extracting a near optimal policy. On the other hand, a recent line of works [30, 67, 115, 55, 106] solve the problem for large-scale MDPs by employing stochastic primal-dual methods, in light of Lagrangian duality. Although this approach achieves state-of-the-art sample complexity guarantees, it shows poor performance in practice. First, current primal-dual algorithms need access to a simulator, mitigating implicitly the problem of exploration, Second, when dealing with linear relaxations of MDPs [14, 58] one needs to impose a restrictive coherence assumptions to ensure that small duality gap for the linearly relaxed LP implies small suboptimality gap for the extracted policy. Finally, while their is enough intuition behind the use of linear function approximation for value functions, this is not the case for occupancy measure approximation. A new breed of algorithms that seem to overcome these difficulties is based on an alternative Q-LP formulation of RL. This approach has been first introduced by Mehta and Meyn [76] and has been recently revisited by [14, 84, 67, 83, 77, 71]. A salient feature of this equivalent formulation is that it introduces a Q-function as slack variables, and so lends itself to data-driven algorithms. Our work is inspired by these line of works. The most related works are the analysis of REPS/Q-REPS [90, 14, 89] and O-REPS [124] that first pointed out the connection between REPS and PPM. We build on their techniques with some important differences. In particular, while in the LP formulation of RL, PPM and mirror descent [15, 47] are equivalent, recognizing that they are *not equivalent* in IL is critical for stronger empirical performance. Moreover, our techniques can be used to improve upon the best rate for REPS in the tabular setting [89] and to extend their guarantees to Linear MDPs.

State-of-the-art Imitation Learning. Generative adversarial imitation learning (GAIL) [51] and other follow-up works [38, 59, 63, 64] formulate the IL as a minimax adversarial problem similar to a GAN [43] and leverage primal-dual optimization tools. In particular, GAIL solves IL with alternating updates of both policy and cost functions. On the other hand, a recent line of work [40, 11, 96] bypasses the need of optimizing over cost functions and thus avoids instability due to adversarial training. Although these algorithms achieve impressive empirical performance in challenging high dimensional benchmark tasks, they are hampered by limited theoretical understanding. This is the fundamental difference from our work, which enjoys both favorable practical performance and strong theoretical guarantees. Moreover, a unique algorithmic feature of our proposed methodology is a convex and smooth logistic policy evaluation objective that optimizes jointly cost and Q-functions. As a result, our algorithm has the additional practical benefit that can also recover an explicit cost along with the Q-function without requiring knowledge or further interaction with the environment (as in [40, 11, 96]). Therefore, the recovered cost functions show promising transfer capability to new dynamics. In addition, unlike IQ-Learn [40], in our online IL algorithm, instead of regularizing the IL objective, the key idea is to penalize the divergence between the current policy and the policy obtained at the previous iteration. We do so by employing a Bregman proximal point update. Most importantly, as we have already highlighted, the convergence properties of [40, 11, 96]) remain largely elusive in the function approximation and model-free regime. It is unclear whether the sampling-based variants of their algorithms converge to a global optimum or if they converge at all, even for the simple tabular setting.

B Future directions

In this work, we studied a proximal point imitation learning algorithm with both theoretical guarantees and convincing empirical performance in challenging benchmark tasks. Our methodology is rooted in classical stochastic optimization tools and in the LP approach to MDPs. We hope that our new techniques will be useful for future algorithm designers and lay foundations for overcoming current limitations and challenges. We point out a few interesting directions.

Accelerated proximal point. An appealing possibility is to study an accelerated proximal point scheme with inexact updates and achieve faster convergence rates. While there has been an effort in this direction [119, 120], the acceleration relies on the triangle/quadrangle scaling property assumption [45] that does not hold for KL divergence over the simplex. Understanding if it is possible to accelerate PPM without such an assumption is an open question, whose solution has direct application to the LP formulation of RL and imitation learning.

Primal-dual methods with conditional relative entropy. Recent primal-dual RL algorithms rooted in the LP approach to MDPs achieve state-of-the-art sample complexity guarantees. See, for example, [13] for exact gradients, [23, 55] for stochastic gradients, and [58] for the imitation learning problem. The most important disadvantages of primal-dual RL algorithms are (i) the need of a generative oracle, (ii) restricted coherence assumptions on the choice of features, as well as (iii) the problematic occupancy measure approximation. Unfortunately, these limitations lead to poor practical performance for challenging high-dimensional and model-free RL and IL setups. On the other hand, our algorithm overcomes these difficulties but requires to approximately solve a small-dimensional convex program repetitively. It is also challenging to account for the biased gradient estimates beyond the linear MDP setting. It is promising to investigate if by combining the alternative Q-LP formulation and the conditional relative entropy as Bregman divergence in a primal-dual mirror descent scheme, one can avoid the current practical limitations of primal-dual RL methods. It is also interesting that in this case, the action-value parameters will be updated by taking one gradient step each time, instead of solving a small-dimensional convex program.

Inexact policy improvement update. The error propagation presented in this work accommodates for errors only in the policy evaluation phase, while it assumes that the policy improvement step can be implemented exactly. This happens in other related works like [14, 114]. In contrast, the error propagation analysis in [41] takes into account an error in the policy improvement step but unfortunately it does not provide a way to ensure that such an error is small. Future research effort will aim to include in our error propagation analysis a term given by inexact policy improvement steps, ensure that such errors are small and characterizing the deterioration in the sample complexity under policy improvement errors. This kind of analysis would be important for continuous actions environment where the softmax policy update can not be computed in closed-form.

C Dual Program Interpretation

To motivate further the Primal formulation, we shed light to its dual and provide an interpretation of the dual optimizers. For brevity, we focus on the case $\mathcal{W} = B_1^m$. The proof can be found in Appendix C.1 and is based on strong duality between the two convex programs.

Proposition 4. *The dual convex program is given by*

$$\zeta^\star = \max_{(\mathbf{w},\mathbf{V},\mathbf{Q})} \Big\{ (1-\gamma)\,\langle \boldsymbol{\nu}_0, \mathbf{V}\rangle - \big\langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}} \big\rangle \mid \mathbf{Q} \geq \mathbf{B}\mathbf{V},\ \mathbf{Q} = \mathbf{c}_{\mathbf{w}} + \gamma \mathbf{P}\mathbf{V},\ \mathbf{V} \in \mathbb{R}^{|\mathcal{S}|},\ \mathbf{Q} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|},\ \mathbf{w} \in \mathcal{W} \Big\}. \quad \text{(Dual)}$$

Moreover, for $\mathcal{W} = B_1^m$, a triple $(\mathbf{V}_{\mathrm{A}}, \mathbf{Q}_{\mathrm{A}}, \mathbf{w}_{\mathrm{A}})$ is dual optimal if and only if (i) π_{E} is optimal for the RL problem with cost $\mathbf{c} = \mathbf{c}_{\mathbf{w}_{\mathrm{A}}}$, (ii) $\mathbf{V}_{\mathrm{A}} = \mathbf{V}^\star_{\mathbf{w}_{\mathrm{A}}}$, (iii) $\mathbf{Q}_{\mathrm{A}} = \mathbf{Q}^\star_{\mathbf{w}_{\mathrm{A}}}$, and (iv) $\mathbf{w}_{\mathrm{A}} \in \mathcal{W}$. In particular, $(\mathbf{V}^\star_{\mathbf{w}_{\mathrm{true}}}, \mathbf{Q}^\star_{\mathbf{w}_{\mathrm{true}}}, \mathbf{w}_{\mathrm{true}})$ is a dual optimizer.

Proposition 4 states that the set of dual optimal costs $\mathbf{c}_{\mathbf{w}_{\mathrm{A}}}$ is the set of costs in $\mathcal{C}$ for which the expert is optimal. In this case, the optimal $\mathbf{V}_{\mathrm{A}}$ coincides with the corresponding optimal value function[2], while the optimal $\mathbf{Q}_{\mathrm{A}}$ coincides with the corresponding optimal state-action value function. In particular, the true weights $\mathbf{w}_{\mathrm{true}}$, the true optimal value function $\mathbf{V}^\star_{\mathbf{w}_{\mathrm{true}}}$ and the true optimal state-action value function $\mathbf{Q}^\star_{\mathbf{w}_{\mathrm{true}}}$ are dual optimizers. Therefore, the presented Q-convex approach allows to recover an optimal solution to the original problem (1) from both the (Primal) and (Dual) formulations: it can be obtained either as the induced policy of a primal optimal occupancy measure or as a greedy policy associated to a dual optimal Q-function. In Section 4, we generalize the later observation to implement PPM using softmin updates in terms of Q-functions.

C.1 Proof of Proposition 4

We recall the alternative Q-LP approach to MDPs [76, 77, 84, 83, 14]. Let $\mathbf{c} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be a cost function. The forward RL problem is equivalent to the following linear programs[3]

$$\rho^\star_{\mathbf{c}} = \min_{(\boldsymbol{\mu},\mathbf{d}) \in \mathbb{R}^{2|\mathcal{S}||\mathcal{A}|}} \{ \langle \boldsymbol{\mu}, \mathbf{c}\rangle \mid \mathbf{B}^\intercal \mathbf{d} = \gamma \mathbf{P}^\intercal \boldsymbol{\mu} + (1-\gamma)\boldsymbol{\nu}_0,\ \mathbf{d} = \boldsymbol{\mu},\ \mathbf{d} \geq \mathbf{0} \} \quad \text{(Primal Q-LP)}$$

$$= \max_{(\mathbf{V},\mathbf{Q}) \in \mathbb{R}^{|\mathcal{S}|+|\mathcal{S}||\mathcal{A}|}} \Big\{ (1-\gamma)\,\langle \boldsymbol{\nu}_0, \mathbf{u}\rangle \mid \mathbf{Q} \geq \mathbf{B}\mathbf{V},\ \mathbf{Q} = \mathbf{c} + \gamma\mathbf{P}\mathbf{V},\ \mathbf{V} \in \mathbb{R}^{|\mathcal{S}|}, \Big\}, \quad \text{(Dual Q-LP)}$$

[2]To be precise, this is the case if $\boldsymbol{\nu}_0 \in \mathbb{R}^{|\mathcal{S}|}_{++}$, otherwise they coincide $\boldsymbol{\nu}_0$-almost surely.

[3]Note that usually in the literature the primal LP is (Dual Q-LP).

We have that if $\pi^\star$ is an optimal policy for the forward RL problem with cost $\mathbf{c}$, then $(\boldsymbol{\mu}_{\pi^\star}, \boldsymbol{\mu}_{\pi^\star})$ is optimal for (Primal Q-LP) and conversely if $(\boldsymbol{\mu}^\star, \mathbf{d}^\star)$ is optimal for (Primal Q-LP), then $\pi_{\boldsymbol{\mu}^\star}$ is an optimal policy for the forward RL problem with cost $\mathbf{c}$. Moreover, $(\mathbf{V}^\star_{\mathbf{c}}, \mathbf{Q}^\star_{\mathbf{c}})$ is an optimal solution to (Dual Q-LP) and it is the unique optimizer when $\boldsymbol{\nu}_0 \in \mathbb{R}^{|\mathcal{S}|}_{++}$. For the following results, we will assume without loss of generality that $\boldsymbol{\nu}_0 \in \mathbb{R}^{|\mathcal{S}|}_{++}$.

Proof of Proposition 4. We first derive the dual convex program. We have,

$$\begin{aligned}
\zeta^\star &= \min_{(\boldsymbol{\mu},\mathbf{d})\in\mathfrak{M}} \max_{\mathbf{w}\in\mathcal{W}} \left\langle \boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}} \right\rangle \\
&= \max_{\mathbf{w}\in\mathcal{W}} \min_{(\boldsymbol{\mu},\mathbf{d})\in\mathfrak{M}} \left\langle \boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}} \right\rangle \\
&= \max_{\mathbf{w}\in\mathcal{W}} \min_{\boldsymbol{\mu},\mathbf{d}\geq\mathbf{0}} \max_{\mathbf{V},\mathbf{Q}} \{ \left\langle \boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}} \right\rangle + \langle \gamma\mathbf{P}^\intercal\boldsymbol{\mu} + (1-\gamma)\boldsymbol{\nu}_0 - \mathbf{B}^\intercal\mathbf{d}, \mathbf{V}\rangle + \langle \mathbf{d} - \boldsymbol{\mu}, \mathbf{Q}\rangle \} \\
&= \max_{\mathbf{w}\in\mathcal{W}} \min_{\boldsymbol{\mu},\mathbf{d}\geq\mathbf{0}} \max_{\mathbf{V},\mathbf{Q}} \{ (1-\gamma)\langle \boldsymbol{\nu}_0, \mathbf{V}\rangle - \left\langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}} \right\rangle + \langle \boldsymbol{\mu}, \mathbf{c}_{\mathbf{w}} + \gamma\mathbf{P}\mathbf{V} - \mathbf{Q}\rangle + \langle \mathbf{d}, \mathbf{Q} - \mathbf{B}\mathbf{V}\rangle \} \\
&= \max_{\mathbf{w}\in\mathcal{W}} \max_{\mathbf{V},\mathbf{Q}} \min_{\boldsymbol{\mu},\mathbf{d}\geq\mathbf{0}} \{ (1-\gamma)\langle \boldsymbol{\nu}_0, \mathbf{V}\rangle - \left\langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}} \right\rangle + \langle \boldsymbol{\mu}, \mathbf{c}_{\mathbf{w}} + \gamma\mathbf{P}\mathbf{V} - \mathbf{Q}\rangle + \langle \mathbf{d}, \mathbf{Q} - \mathbf{B}\mathbf{V}\rangle \} \\
&= \max_{(\mathbf{w},\mathbf{V},\mathbf{Q})} \Big\{ (1-\gamma)\langle \boldsymbol{\nu}_0, \mathbf{V}\rangle - \left\langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}} \right\rangle \mid \mathbf{Q} \geq \mathbf{B}\mathbf{V},\ \mathbf{Q} = \mathbf{c}_{\mathbf{w}} + \gamma\mathbf{P}\mathbf{V}, \\
&\qquad\qquad \mathbf{V}\in\mathbb{R}^{|\mathcal{S}|},\ \mathbf{Q}\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|},\ \mathbf{w}\in\mathcal{W} \Big\}, \qquad \text{(Dual)}
\end{aligned}$$

where the second equality follows by Sion's minimax theorem [107], since $\mathcal{M}$ is convex and compact, $\mathcal{W}$ is convex and the objective is bilinear, the third equality follows by introducing Lagrange multipliers $\mathbf{V}$ and $\mathbf{Q}$, and the fifth equality follows by linear duality. Note that the derivations hold for any convex set $\mathcal{W}$.

From now on we consider the case $\mathcal{W} = B^m_1 = \{\mathbf{w} \in \mathbb{R}^m \mid \|\mathbf{w}\|_2 \leq 1\}$. Then, the (Primal) program can be written in the form

$$\begin{aligned}
\zeta^\star &= \min_{(\boldsymbol{\mu},\mathbf{d})} \{ \bar{d}_{\mathcal{C}}(\boldsymbol{\mu}, \boldsymbol{\mu}_{\pi_{\mathrm{E}}}) \mid (\boldsymbol{\mu},\mathbf{d}) \in \mathfrak{M} \} \\
&= \min_{(\boldsymbol{\mu},\mathbf{d})} \{ \max_{\mathbf{w}\in\mathcal{W}} \left\langle \boldsymbol{\mu} - \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}} \right\rangle \mid (\boldsymbol{\mu},\mathbf{d}) \in \mathfrak{M} \} \\
&= \min_{(\boldsymbol{\mu},\mathbf{d})} \{ \max_{\mathbf{w}\in\mathcal{W}} \left\langle \boldsymbol{\Phi}^\intercal\boldsymbol{\mu} - \boldsymbol{\Phi}^\intercal\boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{w} \right\rangle \mid (\boldsymbol{\mu},\mathbf{d}) \in \mathfrak{M} \} \\
&= \min_{(\boldsymbol{\mu},\mathbf{d})} \{ \left\| \boldsymbol{\Phi}^\intercal\boldsymbol{\mu} - \boldsymbol{\Phi}^\intercal\boldsymbol{\mu}_{\pi_{\mathrm{E}}} \right\|_2 \mid (\boldsymbol{\mu},\mathbf{d}) \in \mathfrak{M} \}, \qquad \text{(Primal)}
\end{aligned}$$

where in the last equality we used that the ℓ_2-norm is self-dual, that is, the dual norm of the ℓ_2-norm is still the ℓ_2-norm. Therefore, when $\mathcal{W} = B^m_1$, we get a quadratic objective with linear constraints [4].

Assume first that $(\mathbf{V}_{\mathrm{A}}, \mathbf{Q}_{\mathrm{A}}, \mathbf{w}_{\mathrm{A}})$ is optimal for (Dual). Then,

$$\mathbf{Q}_{\mathrm{A}} \geq \mathbf{B}\mathbf{V}_{\mathrm{A}}, \quad \mathbf{Q}_{\mathrm{A}} = \mathbf{c}_{\mathbf{w}_{\mathrm{A}}} + \gamma\mathbf{P}\mathbf{V}_{\mathrm{A}}, \quad \mathbf{w}_{\mathrm{A}} \in \mathcal{W}, \tag{6}$$

$$(1-\gamma)\langle \boldsymbol{\nu}_0, \mathbf{V}_{\mathrm{A}}\rangle - \left\langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}_{\mathrm{A}}} \right\rangle = \zeta^\star = 0, \tag{7}$$

where (6) holds because $(\mathbf{V}_{\mathrm{A}}, \mathbf{Q}_{\mathrm{A}}, \mathbf{w}_{\mathrm{A}})$ is feasible to (Dual), and (7) holds by optimality. Therefore, $(\mathbf{V}_{\mathrm{A}}, \mathbf{Q}_{\mathrm{A}})$ is feasible for (Dual Q-LP) with cost $\mathbf{c} = \mathbf{c}_{\mathbf{w}_{\mathrm{A}}}$. Moreover, $(\boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \boldsymbol{\mu}_{\pi_{\mathrm{E}}})$ is feasible for (Primal Q-LP) with cost $\mathbf{c} = \mathbf{c}_{\mathbf{w}_{\mathrm{A}}}$. Therefore,

$$(1-\gamma)\langle \boldsymbol{\nu}_0, \mathbf{V}_{\mathrm{A}}\rangle \leq \rho^\star_{\mathbf{w}_{\mathrm{A}}} \leq \left\langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}_{\mathrm{A}}} \right\rangle. \tag{8}$$

However, by (7) we get that $(1-\gamma)\langle \boldsymbol{\nu}_0, \mathbf{V}_{\mathrm{A}}\rangle = \left\langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}_{\mathrm{A}}} \right\rangle$. Thus, $(\boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \boldsymbol{\mu}_{\pi_{\mathrm{E}}})$ is optimal for (Primal Q-LP) with cost $\mathbf{c} = \mathbf{c}_{\mathbf{w}_{\mathrm{A}}}$ and $(\mathbf{V}_{\mathrm{A}}, \mathbf{Q}_{\mathrm{A}})$ is optimal for (Dual Q-LP) with cost $\mathbf{c} = \mathbf{c}_{\mathbf{w}_{\mathrm{A}}}$. Thus π_{E} is optimal for the forward RL problem with cost $\mathbf{c}_{\mathbf{w}_{\mathrm{A}}}$, $\mathbf{V}_{\mathrm{A}} = \mathbf{V}^\star_{\mathbf{c}_{\mathbf{w}_{\mathrm{A}}}}$, and $\mathbf{Q}_{\mathrm{A}} = \mathbf{Q}^\star_{\mathbf{c}_{\mathbf{w}_{\mathrm{A}}}}$

Conversely, assume that $\mathbf{w}_{\mathrm{A}} \in \mathcal{W}$, π_{E} is optimal for $\mathbf{c}_{\mathbf{w}_{\mathrm{A}}}$, $\mathbf{V}_{\mathrm{A}} = \mathbf{V}^\star_{\mathbf{w}_{\mathrm{A}}}$, and $\mathbf{Q}_{\mathrm{A}} = \mathbf{Q}^\star_{\mathbf{w}_{\mathrm{A}}}$. Then, we have that $(\boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \boldsymbol{\mu}_{\pi_{\mathrm{E}}})$ is optimal for (Primal Q-LP) with cost $\mathbf{c}_{\mathbf{w}_{\mathrm{A}}}$, and $(\mathbf{V}_{\mathrm{A}}, \mathbf{Q}_{\mathrm{A}})$ is optimal for (Dual Q-LP) with cost $\mathbf{c}_{\mathbf{w}_{\mathrm{A}}}$. By dual feasibility, we get

$$\mathbf{Q}_{\mathrm{A}} \geq \mathbf{B}\mathbf{V}_{\mathrm{A}}, \quad \mathbf{Q}_{\mathrm{A}} = \mathbf{c}_{\mathbf{w}_{\mathrm{A}}} + \gamma\mathbf{P}\mathbf{V}_{\mathrm{A}}. \tag{9}$$

Moreover, by primal-dual optimality, we have

$$(1-\gamma)\langle \boldsymbol{\nu}_0, \mathbf{V}_{\mathrm{A}}\rangle = \left\langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}_{\mathrm{A}}}\right\rangle. \tag{10}$$

From (9), we get that $(\mathbf{V}_{\mathrm{A}}, \mathbf{Q}_{\mathrm{A}}, \mathbf{w}_{\mathrm{A}})$ is feasible to (Dual). Since $\zeta^\star = 0$, by (10), we conclude that $(\mathbf{V}_{\mathrm{A}}, \mathbf{Q}_{\mathrm{A}}, \mathbf{w}_{\mathrm{A}})$ is optimal for (Dual). ☐

D Saddle-Point Formulation

By using a compact notation, we have that Primal′ is equivalent to the following bilinear saddle-point problem

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{y}\in\mathcal{Y}} \langle \mathbf{y}, \mathbf{A}\mathbf{x} + \mathbf{b}\rangle, \tag{SPP}$$

where

$$\mathbf{A} \triangleq \begin{bmatrix} \mathbf{I}_m & 0 \\ -\gamma\mathbf{M}^\intercal & \mathbf{B}^\intercal \\ \mathbf{I}_m & -\boldsymbol{\Phi}^\intercal \end{bmatrix}, \qquad \mathbf{b} \triangleq \begin{bmatrix} -\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{\mathrm{E}}) \\ (1-\gamma)\boldsymbol{\nu}_0 \\ \mathbf{0} \end{bmatrix},$$

$\mathbf{x} \triangleq [\boldsymbol{\lambda}^\intercal, \mathbf{d}^\intercal]^\intercal$, $\mathbf{y} \triangleq [\mathbf{w}^\intercal, \mathbf{V}^\intercal, \boldsymbol{\theta}^\intercal]^\intercal$, $\mathcal{X} \triangleq \Delta_{[m]} \times \Delta_{\mathcal{S}\times\mathcal{A}}$, and $\mathcal{Y} \triangleq \mathcal{W} \times \mathbb{R}^{|\mathcal{S}|} \times \mathbb{R}^m$.

E Proof of Proposition 2

Proof of Proposition 2. We break the proof in three parts. In the first two parts, we introduce and compute the explicit forms of the oracles, while in the third part we derive the proximal point updates.

Analytical oracle. We characterize the `analytical-oracle` by employing the first-order optimality conditions for $\boldsymbol{\lambda}$ and $\mathbf{d}$. In particular, at each iteration step k, for any $[\mathbf{w}^\intercal, \mathbf{V}^\intercal, \boldsymbol{\theta}^\intercal]^\intercal$, we have that the Lagrangian of the optimization problem in the definition of the `analytical-oracle` has the form

$$\begin{aligned}\langle \boldsymbol{\lambda}, \mathbf{w}\rangle - \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}\rangle + \left\langle \mathbf{V}, \gamma\mathbf{M}^\intercal\boldsymbol{\lambda} + (1-\gamma)\boldsymbol{\nu}_0 - \mathbf{B}^\intercal\mathbf{d}\right\rangle \\ + \left\langle \boldsymbol{\theta}, \boldsymbol{\Phi}^\intercal\mathbf{d} - \boldsymbol{\lambda}\right\rangle + \frac{1}{\eta}D(\boldsymbol{\lambda}||\boldsymbol{\lambda}_k) + \frac{1}{\alpha}H(\mathbf{d}||\mathbf{d}_k) + \langle \boldsymbol{\lambda}, \tau\mathbf{1}\rangle - \tau,\end{aligned}$$

where we considered a Lagrangian multiplier τ for the simplex constraint $\sum_i \lambda(i) = 1$. Now taking the derivatives with respect to to $\boldsymbol{\lambda}$ and $\mathbf{d}$, we obtain the following first order optimality conditions:

$$\begin{aligned}\Big(\mathbf{w} + \gamma\mathbf{M}\mathbf{V} - \boldsymbol{\theta}\Big)(i) + \tau + \frac{1}{\eta}\log\frac{\boldsymbol{\lambda}(i)}{\boldsymbol{\lambda}_k(i)} + \frac{1}{\eta} = 0,\ \text{for all } i \in [m], \\ \Big(\mathbf{B}\mathbf{V} + \boldsymbol{\Phi}\boldsymbol{\theta}\Big)(s,a) + \frac{1}{\alpha}\log\frac{\pi_{\mathbf{d}}(a|s)}{\pi_{\mathbf{d}_k}(a|s)} = 0,\ \text{for all}(s,a) \in \mathcal{S}\times\mathcal{A}.\end{aligned}$$

Therefore, we obtain

$$\lambda(i) = \lambda_k(i)\, e^{-\eta\boldsymbol{\delta}^k_{\mathbf{w},\boldsymbol{\theta}}(i) + 1 - \eta\tau}, \tag{11}$$

where $\boldsymbol{\delta}^k_{\mathbf{w},\boldsymbol{\theta}} \triangleq \mathbf{w} + \gamma\mathbf{M}\mathbf{V}^k_{\boldsymbol{\theta}} - \boldsymbol{\theta}$. In addition, the simplex constraint $\sum_i \lambda(i) = 1$ is satisfied by choosing $\tau = \tau^k_{\mathbf{w},\boldsymbol{\theta}}$, where

$$\tau^k_{\mathbf{w},\boldsymbol{\theta}} \triangleq \frac{1}{\eta}\log\left(\sum_{i=1}^m (\boldsymbol{\Phi}^\intercal\mathbf{d}_k)(i) e^{-\eta\boldsymbol{\delta}^k_{\mathbf{w},\boldsymbol{\theta}}(i)}\right). \tag{12}$$

Moreover, by setting $\mathbf{Q}_{\boldsymbol{\theta}} = \boldsymbol{\Phi}\boldsymbol{\theta}$, we get

$$\pi_{\mathbf{d}}(a|s) = \pi_{\mathbf{d}_k}(a|s)\, e^{-\alpha(Q_{\boldsymbol{\theta}}(s,a) - V(s))}. \tag{13}$$

Equation (4) follows by noting that the constraint $\sum_a \pi_{\mathbf{d}}(a|x)$ implies that $\mathbf{V}$ equals the logistic value function $\mathbf{V}^k_{\boldsymbol{\theta}}$ given in Proposition 2. Finally, since $(\boldsymbol{\lambda}_k, \mathbf{d}_k)$ are ideal updates, they are primal feasible. Hence, we can use the constraint $\boldsymbol{\lambda}_k = \boldsymbol{\Phi}^\intercal\mathbf{d}_k$ in Equation (11) to obtain Equation (3).

All in all, for any $\mathbf{y} = [\mathbf{w}^\intercal, \mathbf{V}^\intercal, \boldsymbol{\theta}^\intercal]^\intercal$ the `analytical-oracle` outputs $\mathbf{g}(\mathbf{y}; \mathbf{x}_k) = [\boldsymbol{\lambda}^\intercal, \mathbf{d}^\intercal]^\intercal$ with

$$\lambda(i) \propto (\boldsymbol{\Phi}^\intercal\mathbf{d}_k)(i)\, e^{-\eta\boldsymbol{\delta}^k_{\mathbf{w},\boldsymbol{\theta}}(i)}, \tag{14}$$

$$\pi_{\mathbf{d}}(a|s) = \pi_{\mathbf{d}_k}(a|s)\, e^{-\alpha\left(Q_{\boldsymbol{\theta}}(s,a) - V^k_{\boldsymbol{\theta}}(s)\right)}. \tag{15}$$

Note that the derivatives with respect to $\boldsymbol{\lambda}$ and $\mathbf{d}$ differ from the ones in Logistic Q-Learning [14]. In our case, $\boldsymbol{\delta}^k_{\mathbf{w},\boldsymbol{\theta}}$ depends on both cost weights $\mathbf{w}$ and logistic action-value parameters $\boldsymbol{\theta}$. In addition, $\boldsymbol{\delta}^k_{\mathbf{w},\boldsymbol{\theta}}$ is the reduced Bellman error in the feature space rather than in the high dimensional state-action space.

Max oracle. Since the objective in (2) is convex in $\mathbf{x}$ and linear in $\mathbf{y}$, $\mathcal{X}$ is convex and compact, and $\mathcal{Y}$ is convex, by virtue of Sion's minimax theorem [107], we can exchange the $\min$ and $\max$ in Equation (2). We then have

$$\min_{\mathbf{x}\in\mathcal{X}}\max_{\mathbf{y}\in\mathcal{Y}}\left\langle \mathbf{y}, \mathbf{A}\mathbf{x}+\widehat{\mathbf{b}}\right\rangle + \frac{1}{\tau}D_\Omega(\mathbf{x}||\mathbf{x}_k) = \max_{\mathbf{y}\in\mathcal{Y}}\min_{\mathbf{x}\in\mathcal{X}}\left\langle \mathbf{y}, \mathbf{A}\mathbf{x}+\widehat{\mathbf{b}}\right\rangle + \frac{1}{\tau}D_\Omega(\mathbf{x}||\mathbf{x}_k).$$

Therefore, we get

$$\begin{aligned}
\mathbf{y}^\star &= \arg\max_{\mathbf{y}\in\mathcal{Y}}\min_{\mathbf{x}\in\mathcal{X}}\left\langle \mathbf{y}, \mathbf{A}\mathbf{x}+\widehat{\mathbf{b}}\right\rangle + \frac{1}{\tau}D_\Omega(\mathbf{x}||\mathbf{x}_k)\\
&= \arg\max_{\mathbf{y}\in\mathcal{Y}}\left\langle \mathbf{y}, \mathbf{A}\mathbf{g}(\mathbf{y};\mathbf{x}_k)+\widehat{\mathbf{b}}\right\rangle + \frac{1}{\tau}D_\Omega(\mathbf{g}(\mathbf{y};\mathbf{x}_k)||\mathbf{x}_k)\\
&= \mathbf{h}(\mathbf{x}_k).
\end{aligned}$$

Proximal point updates via max and analytical oracles. It remains to prove the closed-form expressions for $\pi_{\mathbf{d}^\star}$ and $\boldsymbol{\lambda}^\star$ given in Equation (3) and Equation (4), respectively. We start rewriting the objective of the `max-oracle` as a function of $\boldsymbol{\lambda}$ and $\mathbf{d}$. In particular, we have

$$\begin{aligned}
&\left\langle \mathbf{y}, \mathbf{A}\mathbf{g}(\mathbf{y};\mathbf{x}_k)+\widehat{\mathbf{b}}\right\rangle + \frac{1}{\tau}D_\Omega(\mathbf{g}(\mathbf{y};\mathbf{x}_k)||\mathbf{x}_k)\\
&= \min_{\mathbf{d}\in\Delta_{\mathcal{S}\times\mathcal{A}},\boldsymbol{\lambda}\in\Delta_{[m]}}\left\langle \mathbf{y}, \mathbf{A}\begin{bmatrix}\boldsymbol{\lambda}\\ \mathbf{d}\end{bmatrix}+\widehat{\mathbf{b}}\right\rangle + \frac{1}{\alpha}H(\mathbf{d}||\mathbf{d}_k)+\frac{1}{\eta}D(\boldsymbol{\lambda}||\boldsymbol{\lambda}_k).
\end{aligned}$$

The minimizers of the previous expression are characterized via the `analytical-oracle`. In particular, plugging in the analytical forms for $\boldsymbol{\lambda}, \mathbf{d}$ and $\mathbf{V}$, we obtain

$$\begin{aligned}
&\min_{\mathbf{d}\in\Delta_{\mathcal{S}\times\mathcal{A}},\boldsymbol{\lambda}\in\Delta_{[m]}}\left\langle \mathbf{y}, \mathbf{A}\begin{bmatrix}\boldsymbol{\lambda}\\ \mathbf{d}\end{bmatrix}+\widehat{\mathbf{b}}\right\rangle + \frac{1}{\alpha}H(\mathbf{d}||\mathbf{d}_k)+\frac{1}{\eta}D(\boldsymbol{\lambda}||\boldsymbol{\lambda}_k)\\
&= \langle\boldsymbol{\lambda},\mathbf{w}\rangle - \langle\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}),\mathbf{w}\rangle + \frac{1}{\eta}\left\langle\boldsymbol{\lambda}, -\eta\boldsymbol{\delta}^k_{\mathbf{w},\boldsymbol{\theta}} - \eta\tau^k_{\mathbf{w},\boldsymbol{\theta}}\right\rangle\\
&\quad + \frac{1}{\alpha}\left\langle\mathbf{d}, -\alpha(\boldsymbol{\Phi}\boldsymbol{\theta}-\mathbf{B}\mathbf{V}^k_{\boldsymbol{\theta}})\right\rangle + \left\langle\boldsymbol{\lambda},\gamma\mathbf{M}^\intercal\mathbf{V}^k_{\boldsymbol{\theta}}\right\rangle\\
&\quad - \left\langle\mathbf{d},\mathbf{B}\mathbf{V}^k_{\boldsymbol{\theta}}\right\rangle + (1-\gamma)\left\langle\boldsymbol{\nu}_0,\mathbf{V}^k_{\boldsymbol{\theta}}\right\rangle + \langle\mathbf{d},\boldsymbol{\Phi}\boldsymbol{\theta}\rangle - \langle\boldsymbol{\lambda},\boldsymbol{\theta}\rangle\\
&= -\langle\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}),\mathbf{w}\rangle + (1-\gamma)\left\langle\boldsymbol{\nu}_0,\mathbf{V}^k_{\boldsymbol{\theta}}\right\rangle - \tau^k_{\mathbf{w},\boldsymbol{\theta}}\\
&= -\langle\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}),\mathbf{w}\rangle + (1-\gamma)\left\langle\boldsymbol{\nu}_0,\mathbf{V}^k_{\boldsymbol{\theta}}\right\rangle - \frac{1}{\eta}\log\left(\sum_{i=1}^m(\boldsymbol{\Phi}^\intercal\mathbf{d}_k)(i)e^{-\eta\boldsymbol{\delta}^k_{\mathbf{w},\boldsymbol{\theta}}(i)}\right)\\
&\triangleq \mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}).
\end{aligned}$$

This is the objective of the `max-oracle` in Proposition 2. Given that the `max-oracle` returns $(\mathbf{w}^\star_k,\boldsymbol{\theta}^\star_k)$, the corresponding primal variables $(\mathbf{d}^\star_k,\boldsymbol{\lambda}^\star_k)$ satisfy $(\mathbf{d}^\star_k,\boldsymbol{\lambda}^\star_k)=\mathbf{g}([\mathbf{w}^\star_k,\mathbf{V}_{\boldsymbol{\theta}^\star_k},\boldsymbol{\theta}^\star_k];\mathbf{x}_{k-1})$. This completes the proof of the first part of Proposition 2.

It remains to show the dual form of the max-oracle objective $\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})$. In particular, we will show that

$$\max_{\mathbf{w},\boldsymbol{\theta}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) = \max_{\mathbf{w}}\langle\boldsymbol{\lambda}_{k+1},\mathbf{w}\rangle - \langle\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{\mathrm{E}}),\mathbf{w}\rangle + \frac{1}{\eta}D(\boldsymbol{\lambda}_{k+1}||\boldsymbol{\lambda}_k) + \frac{1}{\alpha}H(\mathbf{d}_{k+1}||\mathbf{d}_k). \tag{16}$$

We first recall that

$$\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) = \min_{\mathbf{d}\in\Delta_{\mathcal{S}\times\mathcal{A}},\boldsymbol{\lambda}\in\Delta_{[m]}}\left\langle \mathbf{y}, \mathbf{A}\begin{bmatrix}\boldsymbol{\lambda}\\ \mathbf{d}\end{bmatrix}+\widehat{\mathbf{b}}\right\rangle + \frac{1}{\alpha}H(\mathbf{d}||\mathbf{d}_k)+\frac{1}{\eta}D(\boldsymbol{\lambda}||\boldsymbol{\lambda}_k).$$

Then, by taking the maximum over $\mathbf{y} = [\mathbf{w}, \mathbf{V}, \boldsymbol{\theta}]$ on both sides and using Sion's minimax theorem, we get

$$\begin{aligned}
\max_{\mathbf{w},\boldsymbol{\theta}} \mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta}) &= \max_{\mathbf{y}\in\mathcal{Y}} \min_{\mathbf{d}\in\Delta_{\mathcal{S}\times\mathcal{A}}, \boldsymbol{\lambda}\in\Delta_{[m]}} \left\langle \mathbf{y}, \mathbf{A} \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{d} \end{bmatrix} + \widehat{\mathbf{b}} \right\rangle + \frac{1}{\alpha} H(\mathbf{d}||\mathbf{d}_k) + \frac{1}{\eta} D(\boldsymbol{\lambda}||\boldsymbol{\lambda}_k) \\
&= \min_{\mathbf{d}\in\Delta_{\mathcal{S}\times\mathcal{A}}, \boldsymbol{\lambda}\in\Delta_{[m]}} \max_{\mathbf{y}\in\mathcal{Y}} \left\langle \mathbf{y}, \mathbf{A} \begin{bmatrix} \boldsymbol{\lambda} \\ \mathbf{d} \end{bmatrix} + \widehat{\mathbf{b}} \right\rangle + \frac{1}{\alpha} H(\mathbf{d}||\mathbf{d}_k) + \frac{1}{\eta} D(\boldsymbol{\lambda}||\boldsymbol{\lambda}_k) \\
&= \max_{\mathbf{y}\in\mathcal{Y}} \left\langle \mathbf{y}, \mathbf{A} \begin{bmatrix} \boldsymbol{\lambda}_{k+1} \\ \mathbf{d}_{k+1} \end{bmatrix} + \widehat{\mathbf{b}} \right\rangle + \frac{1}{\alpha} H(\mathbf{d}_{k+1}||\mathbf{d}_k) + \frac{1}{\eta} D(\boldsymbol{\lambda}_{k+1}||\boldsymbol{\lambda}_k),
\end{aligned}$$

where in the last equality we used the definition of proximal point update in Equation (2). Finally, by LP strong duality, we have that $\max_{\mathbf{w}} \langle \boldsymbol{\lambda}_{k+1}, \mathbf{w} \rangle - \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w} \rangle = \max_{\mathbf{y}\in\mathcal{Y}} \left\langle \mathbf{y}, \mathbf{A} \left[\boldsymbol{\lambda}_{k+1}^{\intercal}, \mathbf{d}_{k+1}^{\intercal} \right]^{\intercal} + \widehat{\mathbf{b}} \right\rangle$. Hence, we conclude that (16) holds.

□

F Proof of Proposition 3

Proof of Proposition 3. From first order optimality conditions for $\boldsymbol{\lambda}_k^\star$, we get

$$\left(\mathbf{w}_k^\star + \gamma \mathbf{M} \mathbf{V}_{\boldsymbol{\theta}_k^\star}^k - \boldsymbol{\theta}_k^\star\right)(i) + \tau_{\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star}^k + \frac{1}{\eta} \log \frac{\boldsymbol{\lambda}_k^\star(i)}{\boldsymbol{\lambda}_{k-1}(i)} - \frac{1}{\eta} = 0, \text{ for all } i \in [m]. \tag{17}$$

We define the regularized cost weights by $\widetilde{\mathbf{w}}_k^\star \triangleq \mathbf{w}_k^\star + \frac{1}{\eta} \log \frac{\boldsymbol{\lambda}_k^\star(i)}{\boldsymbol{\lambda}_{k-1}(i)}$, and the costant (wrt the vector index i) $c \triangleq -\tau_{\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star}^k + \frac{1}{\eta}$. This gives for all $i \in [m]$

$$\left(\widetilde{\mathbf{w}}_k^\star + \gamma \mathbf{M} \mathbf{V}_{\boldsymbol{\theta}_k^\star}^k - \boldsymbol{\theta}_k^\star\right)(i) = c.$$

We define the *span norm* as $\|\mathbf{x}\|_{\mathrm{sp}} = \inf_{c\in\mathbb{R}} \|\mathbf{x} - c\mathbf{1}\|_\infty$. Then multiplying by $\boldsymbol{\Phi}$ from the left, we have that $\boldsymbol{\Phi}\widetilde{\mathbf{w}}_k^\star + \gamma \mathbf{P} \mathbf{V}_{\boldsymbol{\theta}_k^\star}^k - \boldsymbol{\Phi}\boldsymbol{\theta}_k^\star = c\mathbf{1}$. Moreover, we can write

$$\begin{aligned}
\mathbf{V}_{\boldsymbol{\theta}_k^\star}^k(s) &= -\frac{1}{\alpha} \log \left(\sum_a \pi_{\mathbf{d}_{k-1}}(a|s) e^{-\alpha (\boldsymbol{\theta}_k^\star)^\intercal \boldsymbol{\phi}(s,a)} \right) \\
&= -\frac{1}{\alpha} \log \left(\sum_a \pi_{\mathbf{d}_{k-1}}(a|s) e^{-\alpha (\boldsymbol{\Phi}\widetilde{\mathbf{w}}_k^\star + \gamma \mathbf{P} \mathbf{V}_{\boldsymbol{\theta}_k^\star}^k)(s,a) + \alpha c} \right) \\
&= -\frac{1}{\alpha} \log \left(\sum_a \pi_{\mathbf{d}_{k-1}}(a|s) e^{-\alpha (\boldsymbol{\Phi}\widetilde{\mathbf{w}}_k^\star + \gamma \mathbf{P} \mathbf{V}_{\boldsymbol{\theta}_k^\star}^k)(s,a)} \right) + c
\end{aligned}$$

We set $(\mathcal{T}\mathbf{V}_{\boldsymbol{\theta}_k^\star}^k)(s) \triangleq -\frac{1}{\alpha} \log \left(\sum_a \pi_{\mathbf{d}_{k-1}}(a|s) e^{-\alpha (\boldsymbol{\Phi}\widetilde{\mathbf{w}}_k^\star + \gamma \mathbf{P} \mathbf{V}_{\boldsymbol{\theta}_k^\star}^k)(s,a)} \right)$. Note that $\mathcal{T}$ is the soft-Bellman operator [86, 41] that is a γ-contraction with respect to $\|\cdot\|_\infty$-norm. It follows that

$$\left\|\mathbf{V}_{\boldsymbol{\theta}_k^\star}^k\right\|_{\mathrm{sp}} = \left\|\mathcal{T}\mathbf{V}_{\boldsymbol{\theta}_k^\star}^k + c\right\|_{\mathrm{sp}} = \left\|\mathcal{T}\mathbf{V}_{\boldsymbol{\theta}_k^\star}^k\right\|_{\mathrm{sp}} \leq \left\|\mathcal{T}\mathbf{V}_{\boldsymbol{\theta}_k^\star}^k - \mathcal{T}\mathbf{0}\right\|_{\mathrm{sp}} + \|\mathcal{T}\mathbf{0}\|_{\mathrm{sp}} \leq \gamma \left\|\mathbf{V}_{\boldsymbol{\theta}_k^\star}^k\right\|_{\mathrm{sp}} + \|\boldsymbol{\Phi}\widetilde{\mathbf{w}}_k^\star\|_{\mathrm{sp}}.$$

Therefore, $\left\|\mathbf{V}^k_{\boldsymbol{\theta}^\star_k}\right\|_{\mathrm{sp}} \leq \frac{\|\boldsymbol{\Phi}\widetilde{\mathbf{w}}^\star_k\|_{\mathrm{sp}}}{1-\gamma} \leq \frac{1+\log\frac{1}{\beta}}{1-\gamma}$. Moreover, using the relation $\widetilde{\mathbf{w}}^\star_k + \gamma\mathbf{M}\mathbf{V}^k_{\boldsymbol{\theta}^\star_k} - \boldsymbol{\theta}^\star_k = c\mathbf{1}$, we have that

$$\begin{aligned}
\|\boldsymbol{\theta}^\star_k\|_{\mathrm{sp}} &\leq \left\|\widetilde{\mathbf{w}}^\star_k + \gamma\mathbf{M}\mathbf{V}^k_{\boldsymbol{\theta}^\star_k} - c\mathbf{1}\right\|_{\mathrm{sp}} \\
&= \left\|\widetilde{\mathbf{w}}^\star_k + \gamma\mathbf{M}\mathbf{V}^k_{\boldsymbol{\theta}^\star_k}\right\|_{\mathrm{sp}} \\
&\leq \|\widetilde{\mathbf{w}}^\star_k\|_{\mathrm{sp}} + \gamma\left\|\mathbf{M}\mathbf{V}^k_{\boldsymbol{\theta}^\star_k}\right\|_{\mathrm{sp}} \\
&\leq 1 + \log\left(\frac{1}{\beta}\right) + \gamma\frac{1+\log\left(\frac{1}{\beta}\right)}{1-\gamma} \\
&= \frac{1+\log\left(\frac{1}{\beta}\right)}{1-\gamma}.
\end{aligned}$$

This proves that for every maximizer the span norm is bounded. Finally, for showing that there exists a maximizer with bounded infinity norm, we want to prove that the negative logistic Bellman error is shift invariant in $\boldsymbol{\theta}$. That is, $\mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta} + c\mathbf{1}) = \mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta})$. Towards this goal, we start proving that $\mathbf{V}^k_{\boldsymbol{\theta}+c\mathbf{1}} = \mathbf{V}^k_{\boldsymbol{\theta}} + c$ for any constant $c \in \mathbb{R}$. Indeed,

$$\begin{aligned}
\mathbf{V}^k_{\boldsymbol{\theta}+c\mathbf{1}}(s) &= -\frac{1}{\alpha}\log\left(\sum_a \pi_{\mathbf{d}_{k-1}}(a|s)e^{-\alpha\boldsymbol{\theta}^\intercal\boldsymbol{\phi}(s,a)+-\alpha c\mathbf{1}^\intercal\boldsymbol{\phi}(s,a)}\right) \\
&= -\frac{1}{\alpha}\log\left(\sum_a \pi_{\mathbf{d}_{k-1}}(a|s)e^{-\alpha\boldsymbol{\theta}^\intercal\boldsymbol{\phi}(s,a)-\alpha c}\right) \\
&= -\frac{1}{\alpha}\log\left(\sum_a \pi_{\mathbf{d}_{k-1}}(a|s)e^{-\alpha\boldsymbol{\theta}^\intercal\boldsymbol{\phi}(s,a)}\right) - \frac{1}{\alpha}\log\left(e^{-\alpha c}\right) \\
&= \mathbf{V}^k_{\boldsymbol{\theta}}(s) + c
\end{aligned}$$

At this point, we can show the shift invariance of $\mathcal{G}_k$.

$$\begin{aligned}
\mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta} + c\mathbf{1}) &= -\frac{1}{\eta}\log\sum_{i=1}^m(\boldsymbol{\Phi}^\intercal\mathbf{d}_{k-1})(i)e^{-\eta\left(\mathbf{w}(i)+\gamma(\mathbf{M}\mathbf{V}^k_{\boldsymbol{\theta}+c\mathbf{1}})(i)-\boldsymbol{\theta}(i)-c\right)} \\
&\quad + (1-\gamma)\left\langle\boldsymbol{\nu}_0, \mathbf{V}^k_{\boldsymbol{\theta}+c\mathbf{1}}\right\rangle - \langle\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}\rangle \\
&= -\frac{1}{\eta}\log\sum_{i=1}^m(\boldsymbol{\Phi}^\intercal\mathbf{d}_{k-1})(i)e^{-\eta\left(\mathbf{w}(i)+\gamma(\mathbf{M}\mathbf{V}^k_{\boldsymbol{\theta}})(i)+\gamma c-\boldsymbol{\theta}(i)-c\right)} \\
&\quad + (1-\gamma)\left\langle\boldsymbol{\nu}_0, \mathbf{V}^k_{\boldsymbol{\theta}}\right\rangle + (1-\gamma)c - \langle\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}\rangle \\
&= -\frac{1}{\eta}\log\underbrace{\sum_{i=1}^m(\boldsymbol{\Phi}^\intercal\mathbf{d}_{k-1})(i)}_{=1}\,e^{-\eta\gamma c+\eta c} + (1-\gamma)c + \mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta}) \\
&= -(1-\gamma)c + (1-\gamma)c + \mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta}) \\
&= \mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta})
\end{aligned}$$

It follows that there exists a maximizer $\boldsymbol{\theta}^\star_k$ for which $\|\boldsymbol{\theta}^\star_k\|_{\mathrm{sp}} = \|\boldsymbol{\theta}^\star_k\|_\infty$. To see this, we show that we can find a value of c for which the span seminorm equals the ℓ_∞-norm, that is $\|\boldsymbol{\theta}^\star_k + c\mathbf{1}\|_\infty = \|\boldsymbol{\theta}^\star_k\|_{\mathrm{sp}}$. By definition of the span norm (and assuming that the infimum is attained), the equality is attained for $c = \arg\min_c \|\boldsymbol{\theta}^\star_k + c\mathbf{1}\|_\infty = \frac{\max_{i\in[m]}\boldsymbol{\theta}^\star_k(i)+\min_{i\in[m]}\boldsymbol{\theta}^\star_k(i)}{2}$. Then, choosing the shift for which $\max_{i\in[m]}\boldsymbol{\theta}^\star_k(i) = -\min_{i\in[m]}\boldsymbol{\theta}^\star_k(i)$, gives the maximizer for which $\|\boldsymbol{\theta}^\star_k\|_{\mathrm{sp}} = \|\boldsymbol{\theta}^\star_k\|_\infty$. This concludes the proof for the bound on the ℓ_∞-norm. $\square$

G Proof of Theorem 1

We will analyze the proximal point method applied to SPP. We use a similar error propagation analysis as in [14].

By Proposition 2, the ideal updates $(\boldsymbol{\theta}_k^\star, \mathbf{w}_k^\star, \pi_k^\star, \boldsymbol{\lambda}_k^\star, \mathbf{d}_k^\star)$ are given by

$$(\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star) = \arg\max_{\mathbf{w},\boldsymbol{\theta}} \mathcal{G}_k(\mathbf{w}, \boldsymbol{\theta}), \qquad \lambda_k^\star(i) = (\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}^\star)(i)\, e^{-\eta(\boldsymbol{\delta}^k_{\boldsymbol{\theta}_k^\star,\mathbf{w}_k^\star}(i) + \tau^k_{\boldsymbol{\theta}_k^\star,\mathbf{w}_k^\star})},$$

$$\mathbf{d}_k^\star = \boldsymbol{\mu}_{\pi_k^\star}, \qquad \pi_k^\star(a|s) = \pi_{\mathbf{d}_{k-1}^\star}(a|s)\, e^{-\alpha(Q_{\boldsymbol{\theta}_k^\star}(s,a) - V^k_{\boldsymbol{\theta}_k^\star}(s))},$$

where $\tau^k_{\boldsymbol{\theta}_k^\star,\mathbf{w}_k^\star}$ is a normalization constant. By feasibility of the ideal updates we also have $\boldsymbol{\lambda}_k^\star = \boldsymbol{\Phi}^\intercal \mathbf{d}_k^\star$ On the other hand, the realized updates $(\boldsymbol{\theta}_k, \mathbf{w}_k, \pi_k, \boldsymbol{\lambda}_k, \mathbf{d}_k)$ are given by

$$(\mathbf{w}_k, \boldsymbol{\theta}_k) = \arg\max_{\mathbf{w},\boldsymbol{\theta}} \mathcal{G}_k^{\epsilon_k}(\mathbf{w}, \boldsymbol{\theta}), \qquad \lambda_k(i) = (\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1})(i)\, e^{-\eta(\boldsymbol{\delta}^k_{\boldsymbol{\theta}_k,\mathbf{w}_k}(i) + \tau^k_{\boldsymbol{\theta}_k,\mathbf{w}_k})},$$

$$\mathbf{d}_k = \boldsymbol{\mu}_{\pi_k}, \qquad \pi_k(a|s) = \pi_{\mathbf{d}_{k-1}}(a|s)\, e^{-\alpha(Q_{\boldsymbol{\theta}_k}(s,a) - V^k_{\boldsymbol{\theta}_k}(s))},$$

where $\tau^k_{\boldsymbol{\theta}_k,\mathbf{w}_k}$ is a normalization constant, and the notation $(\mathbf{w}_k, \boldsymbol{\theta}_k) = \arg\max_{\mathbf{w},\boldsymbol{\theta}} \mathcal{G}_k^{\epsilon_k}(\mathbf{w}, \boldsymbol{\theta})$ means that $\mathcal{G}_k(\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star) - \mathcal{G}_k(\mathbf{w}_k, \boldsymbol{\theta}_k) = \epsilon_k$. We start by introducing some auxiliary results

Lemma 1. *For any occupancy measures $\mathbf{d}_1, \mathbf{d}_2 \in \mathfrak{F}$, and for any cost vectors $\mathbf{c}, \mathbf{c}' \in \mathcal{C}$, we have:*

$$\langle \boldsymbol{\mu}_{\pi_E} - \mathbf{d}_1, \mathbf{c}\rangle - \min_{\mathbf{c}'\in\mathcal{C}} \langle \boldsymbol{\mu}_{\pi_E} - \mathbf{d}_2, \mathbf{c}'\rangle \geq d_{\mathcal{C}}(\pi_E, \pi_{\mathbf{d}_2}) - d_{\mathcal{C}}(\pi_E, \pi_{\mathbf{d}_1}).$$

Proof. We have that

$$\begin{aligned}
\langle \boldsymbol{\mu}_{\pi_E} - \mathbf{d}_1, \mathbf{c}\rangle - \min_{\mathbf{c}'\in\mathcal{C}} \langle \boldsymbol{\mu}_{\pi_E} - \mathbf{d}_2, \mathbf{c}'\rangle &\geq \min_{\mathbf{c}\in\mathcal{C}} \langle \boldsymbol{\mu}_{\pi_E} - \mathbf{d}_1, \mathbf{c}\rangle - \min_{\mathbf{c}'\in\mathcal{C}} \langle \boldsymbol{\mu}_{\pi_E} - \mathbf{d}_2, \mathbf{c}'\rangle \\
&= \max_{\mathbf{c}'} \langle \mathbf{d}_2 - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}'\rangle - \max_{\mathbf{c}} \langle \mathbf{d}_1 - \boldsymbol{\mu}_{\pi_E}, \mathbf{c}\rangle \\
&= d_{\mathcal{C}}(\pi_E, \pi_{\mathbf{d}_2}) - d_{\mathcal{C}}(\pi_E, \pi_{\mathbf{d}_1}).
\end{aligned}$$

□

Corollary 2. *Let $\mathbf{d}^\star = \operatorname{argmin}_{\mathbf{d}\in\mathfrak{F}} \max_{\mathbf{c}\in\mathcal{C}} \langle \mathbf{d}, \mathbf{c}\rangle - \langle \boldsymbol{\mu}_{\pi_E}, \mathbf{c}\rangle$. Setting $\mathbf{c} = \boldsymbol{\Phi}\mathbf{w}_k$, $\mathbf{d}_1 = \mathbf{d}^\star$, $\mathbf{d}_2 = \mathbf{d}_k$, we ge that $\langle \boldsymbol{\mu}_{\pi_E} - \mathbf{d}^\star, \boldsymbol{\Phi}\mathbf{w}_k\rangle - \min_{\mathbf{c}\in\mathcal{C}} \langle \boldsymbol{\mu}_{\pi_E} - \mathbf{d}_k, \mathbf{c}\rangle \geq d_{\mathcal{C}}(\pi_E, \pi_{\mathbf{d}_k}) - d_{\mathcal{C}}(\pi_E, \pi_{\mathbf{d}^\star})$.*

Lemma 2. *It holds that $\frac{D(\boldsymbol{\lambda}_k^\star||\boldsymbol{\lambda}_k)}{\eta} + \frac{H(\mathbf{d}_k^\star||\mathbf{d}_k)}{\alpha} = \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\lambda}_k^\star, (\mathbf{w}_k^\star - \mathbf{w}_k)\rangle + \epsilon_k$.*

Proof. The proof is analogous to Lemma 1 in [14]. □

Lemma 3 (First order optimality conditions for $\mathcal{G}_k$)**.** *For all $k \in [K]$, it holds that*

$$\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\lambda}_k^\star, \mathbf{w}_k^\star - \mathbf{w}_k\rangle \leq 0.$$

Proof. We start by taking the gradient of $\mathcal{G}_k(\mathbf{w}_k, \boldsymbol{\theta}_k)$ with respect to $\mathbf{w}$. In particular, the partial derivative with respect to the i^{th} component is given by

$$\begin{aligned}
\frac{\partial \mathcal{G}_k(\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star)}{\partial \mathbf{w}(i)} &= -\left(\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}})\right)(i) + \frac{(\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1})(i) e^{-\eta \boldsymbol{\delta}^k_{\boldsymbol{\theta}_k^\star,\mathbf{w}_k^\star}(i)}}{\sum_{i=1}^m (\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1})(i) e^{-\eta \boldsymbol{\delta}^k_{\boldsymbol{\theta}_k^\star,\mathbf{w}_k^\star}(i)}} \\
&= -\left(\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}})\right)(i) + \boldsymbol{\lambda}_k^\star(i).
\end{aligned}$$

Therefore,

$$\nabla_{\mathbf{w}} \mathcal{G}_k(\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star) = -\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) + \boldsymbol{\lambda}_k^\star.$$

Then, by using the first-order optimality conditions for a concave function, we have

$$\langle \nabla_{\mathbf{w}} \mathcal{G}_k(\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star), \mathbf{w}_k - \mathbf{w}_k^\star\rangle \leq 0, \quad \forall k.$$

By replacing the expression for $\nabla_{\mathbf{w}} \mathcal{G}_k(\mathbf{w}_k^\star, \boldsymbol{\theta}_k^\star)$, we obtain

$$\langle -\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) + \boldsymbol{\lambda}_k^\star, \mathbf{w}_k - \mathbf{w}_k^\star\rangle \leq 0 \quad \forall k \iff \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\lambda}_k^\star, \mathbf{w}_k^\star - \mathbf{w}_k\rangle, \leq 0 \quad \forall k. \tag{18}$$

□

We also need the following auxiliary result.

Lemma 4. *For all $k \in [K]$, it holds that*

$$\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^{\intercal}\mathbf{d}_k, \mathbf{w}_k^{\star}\rangle \leq \min_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^{\intercal}\mathbf{d}_k, \mathbf{w}\rangle + 2\left\|\mathbf{d}_k - \mathbf{d}_k^{\star}\right\|_1. \tag{19}$$

Proof. By introducing $\bar{\mathbf{w}}_k^{\star} = \arg\min_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^{\intercal}\mathbf{d}_k, \mathbf{w}\rangle$, and applying triangular inequality, we obtain

$$\begin{aligned}
\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^{\intercal}\mathbf{d}_k, \mathbf{w}_k^{\star}\rangle &= \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^{\intercal}\mathbf{d}_k, \bar{\mathbf{w}}_k^{\star}\rangle + \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^{\intercal}\mathbf{d}_k, \mathbf{w}_k^{\star} - \bar{\mathbf{w}}_k^{\star}\rangle \\
&= \min_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^{\intercal}\mathbf{d}_k, \mathbf{w}\rangle + \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^{\intercal}\mathbf{d}_k, \mathbf{w}_k^{\star} - \bar{\mathbf{w}}_k^{\star}\rangle.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^{\intercal}\mathbf{d}_k, \mathbf{w}_k^{\star} - \bar{\mathbf{w}}_k^{\star}\rangle &= \max_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^{\intercal}\mathbf{d}_k, \mathbf{w}_k^{\star} - \mathbf{w}\rangle \\
&= \max_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) + \boldsymbol{\Phi}^{\intercal}\mathbf{d}_k^{\star} - \boldsymbol{\Phi}^{\intercal}\mathbf{d}_k^{\star} - \boldsymbol{\Phi}^{\intercal}\mathbf{d}_k, \mathbf{w}_k^{\star} - \mathbf{w}\rangle \\
&= \max_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^{\intercal}\mathbf{d}_k^{\star}, \mathbf{w}_k^{\star} - \mathbf{w}\rangle + \langle \mathbf{d}_k^{\star} - \mathbf{d}_k, \boldsymbol{\Phi}(\mathbf{w}_k^{\star} - \bar{\mathbf{w}}_k^{\star})\rangle \\
&\leq \underbrace{\max_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\lambda}_k^{\star}, \mathbf{w}_k^{\star} - \mathbf{w}\rangle}_{:=(A)} + \left\|\mathbf{d}_k^{\star} - \mathbf{d}_k\right\|_1 \left\|\boldsymbol{\Phi}(\mathbf{w}_k^{\star} - \bar{\mathbf{w}}_k^{\star})\right\|_{\infty} \\
&\leq 2\left\|\mathbf{d}_k^{\star} - \mathbf{d}_k\right\|_1.
\end{aligned}$$

The first equality holds because the term in $\mathbf{w}_k^{\star}$ is a constant wrt $\mathbf{w}$, the variable of the max. In the last inequality follows from (A) being zero as we show next:

$$\begin{aligned}
\max_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\lambda}_k^{\star}, \mathbf{w}_k^{\star} - \mathbf{w}\rangle &= \max_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\lambda}_k^{\star}, \mathbf{w}_k^{\star} - \mathbf{w}\rangle + \frac{1}{\eta} D(\boldsymbol{\lambda}_k^{\star} || \boldsymbol{\Phi}^{\intercal}\mathbf{d}_{k-1}) \\
&\quad - \frac{1}{\eta} D(\boldsymbol{\lambda}_k^{\star} || \boldsymbol{\Phi}^{\intercal}\mathbf{d}_{k-1}) + \frac{1}{\alpha} H(\mathbf{d}_k^{\star} || \mathbf{d}_{k-1}) - \frac{1}{\alpha} H(\mathbf{d}_k^{\star} || \mathbf{d}_{k-1}) \\
&= \max_{\mathbf{w}\in\mathcal{W}} \Bigg(\langle \boldsymbol{\lambda}_k^{\star} - \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}\rangle + \frac{1}{\eta} D(\boldsymbol{\lambda}_k^{\star} || \boldsymbol{\Phi}^{\intercal}\mathbf{d}_{k-1}) \\
&\quad + \frac{1}{\alpha} H(\mathbf{d}_k^{\star} || \mathbf{d}_{k-1}) \Bigg) - \langle \boldsymbol{\lambda}_k^{\star} - \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}_k^{\star}\rangle \\
&\quad - \frac{1}{\eta} D(\boldsymbol{\lambda}_k^{\star} || \boldsymbol{\Phi}^{\intercal}\mathbf{d}_{k-1}) - \frac{1}{\alpha} H(\mathbf{d}_k^{\star} || \mathbf{d}_{k-1}) \\
&= \max_{\mathbf{w}\in\mathcal{W}} \Bigg(\langle \boldsymbol{\lambda}_k^{\star} - \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}\rangle + \frac{1}{\eta} D(\boldsymbol{\lambda}_k^{\star} || \boldsymbol{\Phi}^{\intercal}\mathbf{d}_{k-1}) \\
&\quad + \frac{1}{\alpha} H(\mathbf{d}_k^{\star} || \mathbf{d}_{k-1}) \Bigg) - \max_{\mathbf{w}\in\mathcal{W}} \min_{\boldsymbol{\lambda}, \mathbf{d}\in\mathfrak{M}_{\boldsymbol{\Phi}}} \Bigg(\langle \boldsymbol{\lambda} - \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}\rangle \\
&\quad + \frac{1}{\eta} D(\boldsymbol{\lambda} || \boldsymbol{\Phi}^{\intercal}\mathbf{d}_{k-1}) + \frac{1}{\alpha} H(\mathbf{d} || \mathbf{d}_{k-1}) \Bigg) \\
&= 0.
\end{aligned}$$

□

Lemma 5 (Lower Bound on feature expectation vectors)**.** *Let Assumption 2 hold. We then have $(\boldsymbol{\Phi}^{\intercal}\mathbf{d}_k)(j) \geq \beta$ for all $j \in [m]$.*

Proof. Let $\mathbf{e}_j \in \mathbb{R}^m$ the vector with zeros everywhere but in position j where it takes the value of 1. Then, we observe that

$$\begin{aligned}
\beta &\leq \lambda_{\min}\big(\mathop{\mathbb{E}}_{s,a\sim\mathbf{d}_k} [\boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^{\intercal}] \big) \\
&= \min_{\{\mathbf{x}\in\mathbb{R}^m : \|\mathbf{x}\|_2 = 1\}} \mathbf{x}^{\intercal} \mathop{\mathbb{E}}_{s,a\sim\mathbf{d}_k} [\boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^{\intercal}]\, \mathbf{x} \\
&\leq \mathbf{e}_j^{\intercal} \mathop{\mathbb{E}}_{s,a\sim\mathbf{d}_k} [\boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^{\intercal}]\, \mathbf{e}_j \\
&= \mathop{\mathbb{E}}_{s,a\sim\mathbf{d}_k} \left[\boldsymbol{\phi}_j^2(s,a)\right] \leq \mathop{\mathbb{E}}_{s,a\sim\mathbf{d}_k} \left[\boldsymbol{\phi}_j(s,a)\right] = (\boldsymbol{\Phi}^{\intercal}\mathbf{d}_k)(j).
\end{aligned}$$

$\square$

Theorem 3 (Error propagation with empirical expert feature expectation vector)**.** *Let* $\mathbf{d}^\star = \mathrm{argmin}_{\mathbf{d}\in\mathfrak{F}} \max_{\mathbf{c}\in\mathcal{C}} \langle \mathbf{d}, \mathbf{c}\rangle - \langle \boldsymbol{\mu}_{\pi_E}, \mathbf{c}\rangle$*, and let* $\boldsymbol{\lambda}^\star$ *be any state-action occupancy measure such that* $(\boldsymbol{\lambda}^\star, \mathbf{d}^\star) \in \mathfrak{M}_{\boldsymbol{\Phi}}$*. Moreover, let* $C \triangleq \frac{1}{\beta\eta}\big(\sqrt{\frac{2\alpha}{1-\gamma}} + \sqrt{8\eta}\big) + \sqrt{\frac{18\alpha}{1-\gamma}}$*. Then, we have that*

$$
\begin{aligned}
&\frac{1}{K}\sum_k \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal \mathbf{d}^\star, \mathbf{w}_k\rangle - \min_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal \mathbf{d}_k, \mathbf{w}\rangle \\
&\leq \frac{D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_0)}{K\eta} + \frac{H(\mathbf{d}^\star||\mathbf{d}_0)}{K\alpha} + \frac{C}{K}\sum_k \sqrt{\epsilon_k} + \frac{\sum_k \epsilon_k}{K}.
\end{aligned}
$$

Proof. We have that

$$
\begin{aligned}
D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_k) =& D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) + \eta\langle \boldsymbol{\lambda}^\star, \mathbf{w}_k + \gamma \mathbf{M}\mathbf{V}^k_{\boldsymbol{\theta}_k} - \boldsymbol{\theta}_k\rangle + \eta\tau^k_{\boldsymbol{\theta}_k,\mathbf{w}_k} \\
=& D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) + \eta\langle \boldsymbol{\lambda}^\star, \mathbf{w}_k - \boldsymbol{\theta}_k\rangle + \eta\langle \gamma \mathbf{M}^T\boldsymbol{\lambda}^\star, \mathbf{V}^k_{\boldsymbol{\theta}_k}\rangle + \eta\tau^k_{\boldsymbol{\theta}_k,\mathbf{w}_k} \\
=& D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) + \eta\langle \boldsymbol{\lambda}^\star, \mathbf{w}_k - \boldsymbol{\theta}_k\rangle + \eta\langle \mathbf{B}^\intercal\mathbf{d}^\star, \mathbf{V}^k_{\boldsymbol{\theta}_k}\rangle - \eta(1-\gamma)\langle \boldsymbol{\nu}_0, \mathbf{V}^k_{\boldsymbol{\theta}_k}\rangle \\
&+ \eta\tau^k_{\boldsymbol{\theta}_k,\mathbf{w}_k} \\
=& D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) + \eta\langle \boldsymbol{\lambda}^\star, \mathbf{w}_k - \boldsymbol{\theta}_k\rangle + \eta\langle \mathbf{B}^\intercal\mathbf{d}^\star, \mathbf{V}^k_{\boldsymbol{\theta}_k}\rangle - \eta(1-\gamma)\langle \boldsymbol{\nu}_0, \mathbf{V}^k_{\boldsymbol{\theta}_k}\rangle \\
&+ \eta\tau^k_{\boldsymbol{\theta}_k,\mathbf{w}_k} \\
=& D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) + \eta\langle \boldsymbol{\lambda}^\star, \mathbf{w}_k - \boldsymbol{\theta}_k\rangle + \eta\langle \mathbf{B}^\intercal\mathbf{d}^\star, \mathbf{V}^k_{\boldsymbol{\theta}_k}\rangle - \eta\mathcal{G}_k(\boldsymbol{\theta}_k, \mathbf{w}_k) \\
&- \eta\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}_k\rangle \\
\leq& D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) + \eta\langle \boldsymbol{\lambda}^\star, \mathbf{w}_k - \boldsymbol{\theta}_k\rangle + \eta\langle \mathbf{B}^\intercal\mathbf{d}^\star, \mathbf{V}^k_{\boldsymbol{\theta}_k}\rangle - \eta\mathcal{G}_k(\boldsymbol{\theta}^\star_k, \mathbf{w}^\star_k) \\
&+ \eta\epsilon_k - \eta\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}_k\rangle \\
\leq& D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) + \eta\langle \boldsymbol{\lambda}^\star, \mathbf{w}_k - \boldsymbol{\theta}_k\rangle + \eta\langle \mathbf{B}^\intercal\mathbf{d}^\star, \mathbf{V}^k_{\boldsymbol{\theta}_k}\rangle + \eta\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\lambda}^\star_k, \mathbf{w}^\star_k\rangle \\
&- D(\boldsymbol{\lambda}^\star_k||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) - \eta\frac{H(\mathbf{d}^\star_k||\mathbf{d}_{k-1})}{\alpha} + \eta\epsilon_k - \eta\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}_k\rangle \\
\leq& D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) + \eta\langle \boldsymbol{\lambda}^\star, \mathbf{w}_k\rangle + \eta\langle \mathbf{d}^\star, \mathbf{B}\mathbf{V}^k_{\boldsymbol{\theta}_k} - \boldsymbol{\Phi}\boldsymbol{\theta}_k\rangle + \eta\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\lambda}^\star_k, \mathbf{w}^\star_k\rangle \\
&+ \eta\epsilon_k - \eta\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}_k\rangle \\
\leq& D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) + \eta\langle \mathbf{d}^\star, \boldsymbol{\Phi}\mathbf{w}_k\rangle + \eta\langle \mathbf{d}^\star, \mathbf{B}\mathbf{V}^k_{\boldsymbol{\theta}_k} - \boldsymbol{\Phi}\boldsymbol{\theta}_k\rangle + \eta\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal\mathbf{d}_k, \mathbf{w}^\star_k\rangle \\
&+ \eta\langle \mathbf{d}_k - \mathbf{d}^\star_k, \boldsymbol{\Phi}\mathbf{w}^\star_k\rangle + \eta\epsilon_k - \eta\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}_k\rangle \\
\leq& D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) + \eta\langle \mathbf{d}^\star, \boldsymbol{\Phi}\mathbf{w}_k\rangle + \eta\langle \mathbf{d}^\star, \mathbf{B}\mathbf{V}^k_{\boldsymbol{\theta}_k} - \boldsymbol{\Phi}\boldsymbol{\theta}_k\rangle + \eta\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal\mathbf{d}_k, \mathbf{w}^\star_k\rangle \\
&+ \eta\left\|\mathbf{d}_k - \mathbf{d}^\star_k\right\|_1 + \eta\epsilon_k - \eta\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}_k\rangle \quad \text{Using Lemma 4} \\
\leq& D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) + \eta\langle \mathbf{d}^\star, \boldsymbol{\Phi}\mathbf{w}_k\rangle + \eta\langle \mathbf{d}^\star, \mathbf{B}\mathbf{V}^k_{\boldsymbol{\theta}_k} - \boldsymbol{\Phi}\boldsymbol{\theta}_k\rangle \\
&+ \min_{\mathbf{w}\in\mathcal{W}} \eta\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal\mathbf{d}_k, \mathbf{w}\rangle + 3\eta\left\|\mathbf{d}_k - \mathbf{d}^\star_k\right\|_1 + \eta\epsilon_k - \eta\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w}_k\rangle.
\end{aligned}
$$

Therefore, it follows that

$$
\begin{aligned}
\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal\mathbf{d}^\star, \mathbf{w}_k\rangle - \min_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal\mathbf{d}_k, \mathbf{w}\rangle \leq& \frac{D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) - D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_k)}{\eta} \\
&+ \langle \mathbf{d}^\star, \mathbf{B}\mathbf{V}^k_{\boldsymbol{\theta}_k} - \boldsymbol{\Phi}\boldsymbol{\theta}_k\rangle + 3\left\|\mathbf{d}_k - \mathbf{d}^\star_k\right\|_1 + \epsilon_k. \quad (20)
\end{aligned}
$$

Then, by using $H(\mathbf{d}^\star||\mathbf{d}_k) = H(\mathbf{d}^\star||\mathbf{d}_{k-1}) - \alpha\langle \mathbf{d}^\star, \boldsymbol{\Phi}\boldsymbol{\theta}_k - \mathbf{B}\mathbf{V}^k_{\boldsymbol{\theta}_k}\rangle$, we obtain

$$
\begin{aligned}
\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal\mathbf{d}^\star, \mathbf{w}_k\rangle - \min_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal\mathbf{d}_k, \mathbf{w}\rangle \leq& \frac{D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) - D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_k)}{\eta} \\
&+ \frac{H(\mathbf{d}^\star||\mathbf{d}_{k-1}) - H(\mathbf{d}^\star||\mathbf{d}_k)}{\alpha} \\
&+ 3\left\|\mathbf{d}_k - \mathbf{d}^\star_k\right\|_1 + \epsilon_k.
\end{aligned}
$$

Summing over iteration indices k and dividing by the total number of iterations K, we obtain

$$\frac{1}{K}\sum_k \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal \mathbf{d}^\star, \mathbf{w}_k\rangle - \min_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal \mathbf{d}_k, \mathbf{w}\rangle \leq \frac{1}{K}\sum_k \Bigg(\frac{D(\boldsymbol{\lambda}^\star || \boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1})}{\eta} \\ - \frac{D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_k)}{\eta} + \frac{H(\mathbf{d}^\star||\mathbf{d}_{k-1}) - H(\mathbf{d}^\star||\mathbf{d}_k)}{\alpha} + 3\left\|\mathbf{d}_k - \mathbf{d}_k^\star\right\|_1 \Bigg) + \frac{\sum_k \epsilon_k}{K}. \quad (21)$$

Moreover, by a telescoping sum, we get

$$\begin{aligned}
&\sum_k \left(\frac{D(\boldsymbol{\lambda}^\star || \boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) - D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_k)}{\eta} + \frac{H(\mathbf{d}^\star||\mathbf{d}_{k-1}) - H(\mathbf{d}^\star||\mathbf{d}_k)}{\alpha} \right) \\
&\quad = \sum_k \Bigg(\frac{D(\boldsymbol{\lambda}^\star || \boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) - D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_k)}{\eta} + \frac{D(\boldsymbol{\lambda}^\star || \boldsymbol{\Phi}^\intercal \mathbf{d}_k) - D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_k)}{\eta} \\
&\qquad + \frac{H(\mathbf{d}^\star||\mathbf{d}_{k-1}) - H(\mathbf{d}^\star||\mathbf{d}_k)}{\alpha} \Bigg) \\
&\quad = \frac{D(\boldsymbol{\lambda}^\star || \boldsymbol{\Phi}^\intercal \mathbf{d}_0) - D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_K)}{\eta} + \frac{H(\mathbf{d}^\star||\mathbf{d}_0) - H(\mathbf{d}^\star||\mathbf{d}_K)}{\alpha} \\
&\qquad + \sum_k \frac{D(\boldsymbol{\lambda}^\star || \boldsymbol{\Phi}^\intercal \mathbf{d}_k) - D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_k)}{\eta} \\
&\quad \leq \frac{D(\boldsymbol{\lambda}^\star || \boldsymbol{\Phi}^\intercal \mathbf{d}_0)}{\eta} + \frac{H(\mathbf{d}^\star||\mathbf{d}_0)}{\alpha} + \sum_k \frac{D(\boldsymbol{\lambda}^\star || \boldsymbol{\Phi}^\intercal \mathbf{d}_k) - D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_k)}{\eta}
\end{aligned}$$

Combining this derivation with (21), we get

$$\frac{1}{K}\sum_k \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal \mathbf{d}^\star, \mathbf{w}_k\rangle - \min_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal \mathbf{d}_k, \mathbf{w}\rangle \leq \frac{D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_0)}{K\eta} + \frac{H(\mathbf{d}^\star||\mathbf{d}_0)}{K\alpha} \\ + \frac{1}{K}\sum_k \left(\frac{D(\boldsymbol{\lambda}^\star || \boldsymbol{\Phi}^\intercal \mathbf{d}_k) - D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_k)}{\eta} + 3\left\|\mathbf{d}_k - \mathbf{d}_k^\star\right\|_1 \right) + \frac{\sum_k \epsilon_k}{K}. \quad (22)$$

In order to bound the term $D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_k) - D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_k)$, we introduce the Bregman projection to the space of feature expectation vectors induced by valid occupancy measures $\tilde{\boldsymbol{\lambda}}_k = \arg\min_{\{\lambda=\boldsymbol{\Phi}\mathbf{d}|\mathbf{d}\in\mathcal{F}\}} D(\boldsymbol{\lambda}||\boldsymbol{\lambda}_k)$. We then have

$$\begin{aligned}
D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_k) - D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_k) =& D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_k) - D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_k) + D(\boldsymbol{\lambda}^\star||\tilde{\boldsymbol{\lambda}}_k) - D(\boldsymbol{\lambda}^\star||\tilde{\boldsymbol{\lambda}}_k) \\
\leq& D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_k) - D(\boldsymbol{\lambda}^\star||\tilde{\boldsymbol{\lambda}}_k) - D(\tilde{\boldsymbol{\lambda}}_k||\boldsymbol{\lambda}_k) \\
\leq& D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_k) - D(\boldsymbol{\lambda}^\star||\tilde{\boldsymbol{\lambda}}_k),
\end{aligned}$$

where in the second inequality, we used Lemma 11.3 in [24]. Furthermore,

$$
\begin{aligned}
D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal \mathbf{d}_k) - D(\boldsymbol{\lambda}^\star||\tilde{\boldsymbol{\lambda}}_k) &= \sum_{i=1}^m \boldsymbol{\lambda}^\star(i) \log \frac{\tilde{\boldsymbol{\lambda}}_k(i)}{\boldsymbol{\Phi}^\intercal \mathbf{d}_k(i)} \\
&\leq \sum_{i=1}^m \boldsymbol{\lambda}^\star(i) \left(\frac{\tilde{\boldsymbol{\lambda}}_k(i)}{\boldsymbol{\Phi}^\intercal \mathbf{d}_k(i)} - 1 \right) \\
&\leq \sum_{i=1}^m \frac{\boldsymbol{\lambda}^\star(i)}{\boldsymbol{\Phi}^\intercal \mathbf{d}_k(i)} \left| \tilde{\boldsymbol{\lambda}}_k(i) - \boldsymbol{\Phi}^\intercal \mathbf{d}_k(i) \right| \\
&\leq \max_i \frac{\boldsymbol{\lambda}^\star(i)}{\boldsymbol{\Phi}^\intercal \mathbf{d}_k(i)} \left\| \tilde{\boldsymbol{\lambda}}_k - \boldsymbol{\Phi}^\intercal \mathbf{d}_k \right\|_1 \\
&\leq \frac{1}{\beta} \left\| \tilde{\boldsymbol{\lambda}}_k - \boldsymbol{\Phi}^\intercal \mathbf{d}_k \right\|_1 \\
&\leq \frac{1}{\beta} (\left\| \tilde{\boldsymbol{\lambda}}_k - \boldsymbol{\lambda}_k \right\|_1 + \|\boldsymbol{\lambda}^\star_k - \boldsymbol{\lambda}_k\|_1 + \|\boldsymbol{\lambda}^\star_k - \boldsymbol{\Phi}^\intercal \mathbf{d}_k\|_1) \\
&\leq \frac{1}{\beta} (\sqrt{2D(\tilde{\boldsymbol{\lambda}}_k||\boldsymbol{\lambda}_k)} + \sqrt{2D(\boldsymbol{\lambda}^\star_k||\boldsymbol{\lambda}_k)} + \|\boldsymbol{\lambda}^\star_k - \boldsymbol{\Phi}^\intercal \mathbf{d}_k\|_1) \\
&\leq \frac{1}{\beta} (2\sqrt{2D(\boldsymbol{\lambda}^\star_k||\boldsymbol{\lambda}_k)} + \|\boldsymbol{\Phi}^\intercal \mathbf{d}^\star_k - \boldsymbol{\Phi}^\intercal \mathbf{d}_k\|_1) \\
&\leq \frac{1}{\beta} \Bigg(\sqrt{8\eta(\epsilon_k + \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal \mathbf{d}^\star_k, \mathbf{w}^\star_k - \mathbf{w}_k \rangle)} \\
&\qquad + \|\boldsymbol{\Phi}\|_\infty \|\mathbf{d}^\star_k - \mathbf{d}_k\|_1 \Bigg),
\end{aligned}
$$

where we used $\max_i \frac{\boldsymbol{\lambda}^\star(i)}{\boldsymbol{\Phi}^\intercal \mathbf{d}_k(i)} \leq \frac{1}{\beta}$ thanks to Lemma 5 while in the last line we use the fact that $H(\mathbf{d}^\star_k||\mathbf{d}_k)$ is positive and the equality in Lemma 2. To bound the ℓ_1-norm, we apply Pinkser's inequality and Lemma 2 in [14] to get that

$$
\|\mathbf{d}_k - \mathbf{d}^\star_k\| \leq \sqrt{2D(\mathbf{d}_k||\mathbf{d}^\star_k)} \leq \sqrt{2\frac{H(\mathbf{d}_k||\mathbf{d}^\star_k)}{1-\gamma}} \leq \sqrt{\frac{2\alpha}{1-\gamma}(\epsilon_k + \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal \mathbf{d}^\star_k, \mathbf{w}^\star_k - \mathbf{w}_k \rangle)}.
$$

Plugging the last derivation in Equation (22) gives

$$
\begin{aligned}
\frac{1}{K}\sum_k \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \mathbf{d}^\star, \boldsymbol{\Phi}\mathbf{w}_k \rangle - \min_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \mathbf{d}_k, \boldsymbol{\Phi}\mathbf{w} \rangle &\leq \frac{D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_0)}{K\eta} + \frac{H(\mathbf{d}^\star||\mathbf{d}_0)}{K\alpha} \\
&+ \frac{C}{K}\sum_k \left(\sqrt{\epsilon_k + \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\lambda}^\star_k, \mathbf{w}^\star_k - \mathbf{w}_k \rangle} \right) + \frac{\sum_k \epsilon_k}{K}. \quad (23)
\end{aligned}
$$

Finally, using Lemma 3 we have that the term $\langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\lambda}^\star_k, \mathbf{w}^\star_k - \mathbf{w}_k \rangle$ is non positive. Therefore,

$$
\begin{aligned}
\frac{1}{K}\sum_k \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal \mathbf{d}^\star, \mathbf{w}_k \rangle - \min_{\mathbf{w}\in\mathcal{W}} \langle \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}) - \boldsymbol{\Phi}^\intercal \mathbf{d}_k, \mathbf{w} \rangle &\leq \frac{D(\boldsymbol{\lambda}^\star||\boldsymbol{\lambda}_0)}{K\eta} + \frac{H(\mathbf{d}^\star||\mathbf{d}_0)}{K\alpha} \\
&+ \frac{C}{K}\sum_k \sqrt{\epsilon_k} + \frac{\sum_k \epsilon_k}{K},
\end{aligned}
$$

where $C = \frac{1}{\beta\eta}(\sqrt{\frac{2\alpha}{1-\gamma}} + \sqrt{8\eta}) + 3\sqrt{\frac{2\alpha}{1-\gamma}}$. □

Finally, we need a Lemma that provides a concentration for the estimated expert feature expectation vector.

Lemma 6 ([111]). *Let $\mathcal{D}_{\pi_E} \triangleq \{(s_0^\ell, a_0^\ell, s_1^\ell, a_1^\ell, \dots, s_H^\ell, a_H^\ell)\}_{\ell=1}^{n_{\mathrm{E}}} \sim \pi_E$ be a finite set of i.i.d. truncated sample trajectories. We consider the empirical expert feature expectation vector $\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_E})$*

by taking sample averages, i.e.,

$$\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_E}) \triangleq (1-\gamma)\frac{1}{n_{\mathrm{E}}}\sum_{t=0}^{H}\sum_{\ell=1}^{N}\gamma^t\phi_i(s_t^\ell, a_t^\ell),\ \forall\, i\in[m].$$

Suppose the trajectory length is $H \geq \frac{1}{1-\gamma}\log(\frac{1}{\varepsilon})$, and the number of of expert trajectories is $n_{\mathrm{E}} \geq \frac{2\log(\frac{2m}{\delta})}{\varepsilon^2}$. Then, with probability at least $1-\delta$, it holds that $\|\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_E) - \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_E})\|_\infty \leq \varepsilon$.

At this point, Theorem 1 is proven from the results of Theorem 3,Lemma 6 and Lemma 1.

H Biased Stochastic Gradients and their Properties

In order to estimate the gradient $\nabla_{\boldsymbol{\theta}}\, G(\mathbf{w},\boldsymbol{\theta})$, we define the policy $\pi_{k,\boldsymbol{\theta}}(a|s) \propto \pi_k(a|s)e^{-\alpha Q_{\boldsymbol{\theta}}(s,a)}$, for all $k\in\mathbb{N}$, and for all $\boldsymbol{\theta}\in\mathbb{R}^m$. Then, by standard computations we get that for all $(\mathbf{w},\boldsymbol{\theta})$, and for all $j\in[m]$,

$$\begin{aligned}
&\nabla_{\boldsymbol{\theta},j}\, G(\mathbf{w},\boldsymbol{\theta}) \\
&= \sum_{i=1}^{m}(\boldsymbol{\Phi}^\intercal\mathbf{d}_{k-1})(i)\mathbf{B}_{\mathbf{w},\boldsymbol{\theta}}^k(i)\big[\gamma\boldsymbol{\Gamma}_k(i,j) - \mathbb{1}\{i=j\}\big] + (1-\gamma)\sum_s \boldsymbol{\nu}_0(s)\sum_a \pi_{k-1,\boldsymbol{\theta}}(a|s)\boldsymbol{\phi}_i(s,a) \\
&= \mathbb{E}_{(s,a)\sim\mathbf{d}_{k-1},i\sim\boldsymbol{\phi}(s,a)}\Big[\mathbf{B}_{\mathbf{w},\boldsymbol{\theta}}^k(i)\big[\gamma\boldsymbol{\Gamma}_k(i,j)-\mathbb{1}\{i=j\}\big]\Big] + (1-\gamma)\,\mathbb{E}_{s_0\sim\boldsymbol{\nu}_0,a_0\sim\pi_{k-1,\boldsymbol{\theta}}(\cdot|s_0)}\Big[\boldsymbol{\phi}_i(s_0,a_0)\Big],
\end{aligned}$$

where $\mathbf{B}_{\mathbf{w},\boldsymbol{\theta}}^k(i) \triangleq \frac{\exp\left(-\eta\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}^{\,k}(i)\right)}{Z_k}$, $Z_k \triangleq \sum_{i=1}^m \exp\big(-\eta\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}^{\,k}(i)\big)\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{k-1})(i)$, and $\boldsymbol{\Gamma}_k(i,j) \triangleq \sum_{s',a'}\mathbf{M}_{i,s'}\pi_{k-1,\boldsymbol{\theta}}(a'|s')\boldsymbol{\phi}_j(s',a')$, . Similarly, for the gradient $\nabla_{\mathbf{w}}\, G(\mathbf{w},\boldsymbol{\theta})$, we can write

$$\begin{aligned}
\nabla_{\mathbf{w},j}\, G(\mathbf{w},\boldsymbol{\theta}) &= -\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}})(j) + \sum_{i=1}^{m}(\boldsymbol{\Phi}^\intercal\mathbf{d}_{k-1})(i)\mathbf{B}_{\mathbf{w},\boldsymbol{\theta}}^k(i)\mathbb{1}\{i=j\} \\
&= -\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}})(j) + \mathbb{E}_{(s,a)\sim\mathbf{d}_{k-1},i\sim\boldsymbol{\phi}(s,a)}\Big[\mathbf{B}_{\mathbf{w},\boldsymbol{\theta}}^k(i)\mathbb{1}\{i=j\}\Big]
\end{aligned}$$

Note that the following estimators of $\nabla_{\boldsymbol{\theta}}\, G_k(\mathbf{w},\boldsymbol{\theta})$ and $\nabla_{\mathbf{w}}\, G_k(\mathbf{w},\boldsymbol{\theta})$ are unbiased: Sample $(s',a')\sim\mathbf{d}_{k-1}$, $i'\sim\boldsymbol{\phi}(s',a')$, $s_0\sim\boldsymbol{\nu}_0$, and $a_0\sim\pi_{k-1,\boldsymbol{\theta}}(\cdot|s_0)$, then define

$$\widetilde{\nabla}_{\mathbf{w},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) = -\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}})(j) + \mathbf{B}_{\mathbf{w},\boldsymbol{\theta}}^k(i')\mathbb{1}\{i'=j\}, \tag{24}$$

$$\widetilde{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) = \mathbf{B}_{\mathbf{w},\boldsymbol{\theta}}^k(i')\Big[\gamma\boldsymbol{\Gamma}_k(i',j) - \mathbb{1}\{i'=j\}\Big] + (1-\gamma)\boldsymbol{\phi}_j(s_0,a_0). \tag{25}$$

These expressions give rise to the Biased Stochastic Gradient Estimator subroutine (BSGE) given in Algorithm 2, where we plug-in estimators $\widehat{\mathbf{B}}_{\mathbf{w},\boldsymbol{\theta}}^k \in \mathbb{R}^m$ and $\widehat{\boldsymbol{\Gamma}}_k \in \mathbb{R}^{m\times m}$ to Equations (24) and (25). It remains to show how to maintain good estimators $\widehat{\mathbf{B}}_{\mathbf{w},\boldsymbol{\theta}}^k$ and $\widehat{\boldsymbol{\Gamma}}_k$ by using the linear MDP Assumption 1. While the estimator $\widehat{\boldsymbol{\Gamma}}_k \in \mathbb{R}^m$ is a standard ridge regression estimator, the construction of $\widehat{\mathbf{B}}_{\mathbf{w},\boldsymbol{\theta}}^k$ is more involved. In particular, we first need to build an estimator for the product $\mathbf{M}\mathbf{V}_\theta^k$ via ridge regression. Then, the estimator for $\widehat{\mathbf{B}}_{\mathbf{w},\boldsymbol{\theta}}^k$ is derived by plugging-in the estimator of $\mathbf{M}\mathbf{V}_\theta^k$, and the estimator for the feature expectation vector $\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{k-1})$ to equation $\mathbf{B}_{\mathbf{w},\boldsymbol{\theta}}^k(i) \triangleq \frac{\exp\left(-\eta\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}^{\,k}(i)\right)}{Z_k}$. The reasoning and analysis is inspired by [52, 89].

H.1 Ridge estimators

This section leverages ridge regression [53] to build estimators $\widehat{\mathbf{B}}_{\mathbf{w},\boldsymbol{\theta}}^k$ and $\widehat{\boldsymbol{\Gamma}}_k \in \mathbb{R}^{m\times m}$. We work under the Assumption 2 which ensures that every iterate covers the features space. We recall that by Lemma 5, Assumption 2 implies that $\boldsymbol{\Phi}^\intercal\mathbf{d}_k(s,a)\geq\beta$, for all $k\in[K]$.

Algorithm 2 Biased Stochastic Gradient Estimator: BSGE$(k, \mathbf{w}, \boldsymbol{\theta}, N)$

Input: Policy evaluation step k, reference points $(\mathbf{w}, \boldsymbol{\theta})$, number of samples N

Compute empirical estimators $\widehat{\boldsymbol{\delta}}^{\,k}_{\mathbf{w},\boldsymbol{\theta}} \in \mathbb{R}^m$, $\widehat{\boldsymbol{\Gamma}}_k \in \mathbb{R}^{m\times m}$, $\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{k-1}}) \in \mathbb{R}^m$ using the first N samples $\{(s^{(n)}_{k-1}, a^{(n)}_{k-1}, s'^{(n)}_{k-1})\}_{n=1}^N$ from the buffer $\mathcal{B}_k$

for $i = 1, \dots, m$ **do**

Compute $\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i) = \frac{\exp\left(-\eta\widehat{\boldsymbol{\delta}}^{\,k}_{\mathbf{w},\boldsymbol{\theta}}(i)\right)}{\widehat{Z}_k}$, Where $\widehat{Z}_k = \sum_{i=1}^m \exp\big(-\eta\widehat{\boldsymbol{\delta}}^{\,k}_{\mathbf{w},\boldsymbol{\theta}}(i)\big)\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{k-1}})(i)$

end for

Sample $(s^{(N+1)}_{k-1}, a^{(N+1)}_{k-1}) \sim \boldsymbol{\mu}_{\pi_{k-1}}$, $i^{(N+1)}_{k-1} \sim \boldsymbol{\phi}(s^{(N+1)}_{k-1}, a^{(N+1)}_{k-1})$

Sample $s^{(0)}_{k-1} \sim \boldsymbol{\nu}_0$, and $a^{(0)}_{k-1} \sim \pi_{k-1,\boldsymbol{\theta}}(\cdot|s_0)$

Compute

$$\widehat{\nabla}_{\mathbf{w},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) = -\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}})(j) + \widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i^{(N+1)}_{k-1})\mathbb{1}\{i^{(N+1)}_{k-1} = j\}$$
$$\widehat{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) = \widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i^{(N+1)}_{k-1})\Big[\gamma\widehat{\boldsymbol{\Gamma}}_k(i^{(N+1)}_{k-1}, j) - \mathbb{1}\{i^{(N+1)}_{k-1} = j\}\Big] + (1-\gamma)\boldsymbol{\phi}_j(s^{(0)}_{k-1}, a^0_{k-1})$$

Output: $(\widehat{\nabla}_{\mathbf{w}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}), \widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}))$

H.1.1 Estimator for $\mathbf{MV}^k_\theta$

We first construct an estimator for $\mathbf{M}_k\mathbf{V}^k_\theta$. We can start noticing that we can rewrite $\mathbf{M}_k\mathbf{V}^k_\theta$ using the feature covariance matrix $\bar{\boldsymbol{\Lambda}}_k \triangleq \mathop{\mathbb{E}}_{(s,a)\sim\mathbf{d}_{k-1}}\left[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^\intercal\right]$ as showed by the next lemma.

Lemma 7. *It holds that* $\mathbf{MV}^k_\theta = \bar{\boldsymbol{\Lambda}}_k^{-1}\mathop{\mathbb{E}}_{(s,a)\sim\mathbf{d}_{k-1}, s'\sim P(\cdot|s,a)}\left[\boldsymbol{\phi}(s,a)V^k_\theta(s')\right]$.

Proof.

$$\begin{aligned}
\mathbf{MV}^k_\theta =& \bar{\boldsymbol{\Lambda}}_k^{-1}\bar{\boldsymbol{\Lambda}}_k\mathbf{MV}^k_\theta \\
=& \bar{\boldsymbol{\Lambda}}_k^{-1}\mathop{\mathbb{E}}_{(s,a)\sim\mathbf{d}_{k-1}}\left[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^\intercal\mathbf{MV}^k_\theta\right] \\
=& \bar{\boldsymbol{\Lambda}}_k^{-1}\mathop{\mathbb{E}}_{(s,a)\sim\mathbf{d}_{k-1}}\left[\boldsymbol{\phi}(s,a)\boldsymbol{\phi}(s,a)^\intercal\sum_{s'}\mathbf{M}_{:s'}V^k_\theta(s')\right] \\
=& \bar{\boldsymbol{\Lambda}}_k^{-1}\mathop{\mathbb{E}}_{(s,a)\sim\mathbf{d}_{k-1}}\left[\boldsymbol{\phi}(s,a)\sum_{s'}\boldsymbol{\phi}(s,a)^\intercal\mathbf{M}_{:s'}V^k_\theta(s')\right] \\
=& \bar{\boldsymbol{\Lambda}}_k^{-1}\mathop{\mathbb{E}}_{(s,a)\sim\mathbf{d}_{k-1}}\left[\boldsymbol{\phi}(s,a)\sum_{s'}P(s'|s,a)V^k_\theta(s')\right] \\
=& \bar{\boldsymbol{\Lambda}}_k^{-1}\mathop{\mathbb{E}}_{(s,a)\sim\mathbf{d}_{k-1}, s'\sim P(\cdot|s,a)}\left[\boldsymbol{\phi}(s,a)V^k_\theta(s')\right].
\end{aligned}$$

□

It follows that $\mathbf{MV}^k_\theta = \arg\min_{\mathbf{z}}\mathop{\mathbb{E}}_{(s,a)\sim\mathbf{d}_{k-1}, s'\sim P(\cdot|s,a)}\left[\left(\boldsymbol{\phi}(s,a)^\intercal\mathbf{z} - V^k_\theta(s')\right)^2\right]$.

Now, we move to the problem of estimating $\widehat{\mathbf{MV}^k_\theta}$ with a finite amount of environment interactions sampled i.i.d from $\mathbf{d}_{k-1}$. We define

$$\widehat{\mathbf{MV}^k_\theta} \triangleq \arg\min_{\mathbf{z}}\frac{1}{N}\sum_{n=1}^N\left(\boldsymbol{\phi}(s^{(n)}_k, a^{(n)}_k)^\intercal\mathbf{z} - V^k_\theta(s'^{(n)}_k)\right)^2 + \chi\left\|\mathbf{z}\right\|_2^2.$$

By optimality conditions, we can obtain a closed-form expression for $\widehat{\mathbf{MV}^k_\theta}$.

Lemma 8. *It holds that*

$$\widehat{\mathbf{MV}_\theta^k} = \frac{1}{N}\left(\mathbf{\Lambda}_{k,N} + \chi\mathbf{I}\right)^{-1}\sum_{n=1}^N \boldsymbol{\phi}(s_{k-1}^{(n)}, a_{k-1}^{(n)})V_\theta^k(s_{k-1}^{\prime(n)}),$$

where $\mathbf{\Lambda}_{k,N} \triangleq \frac{1}{N}\sum_{n=1}^N \boldsymbol{\phi}(s_{k-1}^{(n)}, a_{k-1}^{(n)})\boldsymbol{\phi}(s_{k-1}^{(n)}, a_{k-1}^{(n)})^\mathsf{T}$ *is the empirical covariance matrix.*

Proof. Let $\mathcal{L}(\mathbf{z}) \triangleq \frac{1}{N}\sum_{n=1}^N \left(\boldsymbol{\phi}(s_{k-1}^{(n)}, a_{k-1}^{(n)})^\mathsf{T}\mathbf{z} - V_\theta^k(s_{k-1}^{\prime(n)})\right)^2 + \chi\left\|\mathbf{z}\right\|_2^2$. The first derivative is given by

$$\frac{1}{2}\nabla_{\mathbf{z}}\mathcal{L}(\mathbf{z}) = \frac{1}{N}\sum_{n=1}^N \boldsymbol{\phi}(s_{k-1}^{(n)}, a_{k-1}^{(n)})\left(\boldsymbol{\phi}(s_{k-1}^{(n)}, a_{k-1}^{(n)})^\mathsf{T}\mathbf{z} - V_\theta^k(s_{k-1}^{\prime(n)})\right) + \chi\mathbf{z}. \tag{26}$$

Since $\mathcal{L}(\cdot)$ is convex in $\mathbf{z}$, by first-order optimality conditions, we get

$$\frac{1}{N}\sum_{n=1}^N \boldsymbol{\phi}(s_{k-1}^{(n)}, a_{k-1}^{(n)})\left(\boldsymbol{\phi}(s_{k-1}^{(n)}, a_{k-1}^{(n)})^\mathsf{T}\widehat{\mathbf{MV}_\theta^k} - V_\theta^k(s_{k-1}^{\prime(n)})\right) + \chi\widehat{\mathbf{MV}_\theta^k} = 0$$

The statement follows from rearranging the terms. $\square$

Remark 1. *Note that when* $\chi = 0$*, and* $\boldsymbol{\phi}(s,a)$ *is one-hot vector for every* (s,a)*, then we obtain the tabular estimators* $\mathbf{W}_v$ *proposed in [89].*

We invoke Theorem 2 in [53] to derive an upper bound for $\left\|\mathbf{MV}_\theta^k - \widehat{\mathbf{MV}_\theta^k}\right\|_{\bar{\mathbf{\Lambda}}_k}^2$.

Lemma 9. *Fix some* $\chi > 0$ *and take* $N \geq \mathcal{O}(\frac{\log(\frac{m}{\delta})}{\chi\beta})$*. Then, with probability at least* $1-\delta$*, we have*

$$\left\|\mathbf{MV}_\theta^k - \widehat{\mathbf{MV}_\theta^k}\right\|_{\bar{\mathbf{\Lambda}}_k}^2 \leq \mathcal{O}\left(\frac{m\chi^2}{\beta^3}D^2 + \frac{1}{N}\frac{m\chi}{\beta^4}D^2\log\left(\frac{1}{\delta}\right) + \frac{D^2 m}{N}\log\left(\frac{1}{\delta}\right)\right),$$

where $D \triangleq \frac{1+\log\left(\frac{1}{\beta}\right)}{1-\gamma} \geq 1$ *is the upper bound of* $\left\|\mathbf{V}_\theta^k\right\|_\infty$ *derived in Proposition 3.*

Proof. We introduce the following auxiliary quantities:

$$\begin{aligned}\mathbf{M}_\chi\mathbf{V}_\theta^k &= \arg\min_{\mathbf{z}} \mathop{\mathbb{E}}_{(s,a)\sim\mathbf{d}_{k-1}, s'\sim P(\cdot|s,a)}\left[\boldsymbol{\phi}(s,a)^\mathsf{T}\mathbf{z} - V_\theta^k(s')\right] + \chi\left\|\mathbf{z}\right\|_2^2 \\ &= \left(\bar{\mathbf{\Lambda}}_k + \chi\mathbf{I}\right)^{-1}\mathop{\mathbb{E}}_{(s,a)\sim\mathbf{d}_{k-1}, s'\sim P(\cdot|s,a)}\left[\boldsymbol{\phi}(s,a)V_\theta^k(s')\right],\end{aligned}$$

and the conditional expectation

$$\bar{\mathbf{M}}\mathbf{V}_\theta^k = \mathbb{E}\left[\widehat{\mathbf{MV}_\theta^k}|\mathcal{F}_n\right] = \frac{1}{N}\left(\mathbf{\Lambda}_{k,N} + \chi\mathbf{I}\right)^{-1}\sum_{n=1}^N \boldsymbol{\phi}(s_{k-1}^{(n)}, a_{k-1}^{(n)})\mathop{\mathbb{E}}_{s'\sim P(\cdot|s_{k-1}^{(n)}, a_{k-1}^{(n)})}\left[V_\theta^k(s')\right]$$

with $\mathcal{F}_n$ being the filtration $\mathcal{F}_n = \{s_{k-1}^{(i)}, a_{k-1}^{(i)}\}_{i=0}^n$. Then applying the general random design decomposition in ([53], Proposition 3) we obtain:

$$\left\|\mathbf{MV}_\theta^k - \widehat{\mathbf{MV}_\theta^k}\right\|_{\bar{\mathbf{\Lambda}}_k}^2 \leq 3\underbrace{\left\|\mathbf{MV}_\theta^k - \mathbf{M}_\chi\mathbf{V}_\theta^k\right\|_{\bar{\mathbf{\Lambda}}_k}^2}_{\triangleq\epsilon_{\mathrm{rg}}} + 3\underbrace{\left\|\mathbf{M}_\chi\mathbf{V}_\theta^k - \bar{\mathbf{M}}\mathbf{V}_\theta^k\right\|_{\bar{\mathbf{\Lambda}}_k}^2}_{\triangleq\epsilon_{\mathrm{bs}}} + 3\underbrace{\left\|\bar{\mathbf{M}}\mathbf{V}_\theta^k - \widehat{\mathbf{MV}_\theta^k}\right\|_{\bar{\mathbf{\Lambda}}_k}^2}_{\triangleq\epsilon_{\mathrm{vr}}}, \tag{27}$$

where similarly to [53], we define ϵ_{rg} as the ridge error, ϵ_{bs} the ridge estimator bias and with ϵ_{vr} the ridge estimator variance. By choosing $N \geq \mathcal{O}(6\rho_\chi^2 d_{1,\chi}(\log\max(1, d_{1,\chi}) + \log\frac{1}{\delta})) = \mathcal{O}(\frac{1}{\beta\chi}\log\frac{m}{\delta})$, we ensure that the conditions in Theorem 2 in [53] are satisfied. We next bound each term separately.

Ridge error. In [53], the bound derived for the ridge error is a function of the regularization parameter χ, the eigenvalues of the covariance matrix $\bar{\mathbf{\Lambda}}_k$ denoted as $\{\sigma_j\}_{j=1}^m$ and the corresponding eigenvectors $\{\mathbf{v}_j\}_{j=1}^m$. In particular, we have

$$\begin{aligned}
\epsilon_{\mathrm{rg}} \leq & \sum_{j=1}^m \frac{\sigma_j}{(\frac{\sigma_j}{\chi}+1)^2}(\mathbf{v}_j^\intercal \mathbf{M}\mathbf{V}_\theta^k)^2 \\
= & \sum_{j=1}^m \frac{\sigma_j}{(\frac{\sigma_j}{\chi}+1)^2}\left(\underset{(s,a)\sim\mathbf{d}_{k-1}, s'\sim P(\cdot|s,a)}{\mathbb{E}}\left[\boldsymbol{\phi}(s,a)V_\theta^k(s')\right]^\intercal \bar{\mathbf{\Lambda}}_k^{-1}\mathbf{v}_j\right)^2 \\
= & \sum_{j=1}^m \frac{1}{(\frac{\sigma_j}{\chi}+1)^2\sigma_j}\left(\underset{(s,a)\sim\mathbf{d}_{k-1}, s'\sim P(\cdot|s,a)}{\mathbb{E}}\left[\boldsymbol{\phi}(s,a)V_\theta^k(s')\right]^\intercal \mathbf{v}_j\right)^2 \\
\leq & \sum_{j=1}^m \frac{1}{(\frac{\sigma_j}{\chi}+1)^2\sigma_j}\left\|\mathbf{V}_\theta^k\right\|_\infty^2 \\
\leq & \sum_{j=1}^m \frac{1}{(\frac{\beta}{\chi}+1)^2\beta}D^2 \\
= & \frac{m\chi^2}{(\beta+\chi)^2\beta}D^2 \\
\leq & \frac{m\chi^2}{\beta^3}D^2,
\end{aligned}$$

where in the first inequality we used bullet (3) of Theorem 2 in [53].

Bias. It holds that

$$\epsilon_{\mathrm{bs}} \leq \mathcal{O}\left(\frac{\rho_\chi^2 d_{1,\chi}\underset{(s,a)\sim\mathbf{d}_{k-1}}{\mathbb{E}}[\mathrm{approx}(s,a)] + (1+\rho_\chi^2 d_{1,\chi})\epsilon_{\mathrm{rg}}}{N}\log\left(\frac{1}{\delta}\right)\right),$$

where we used the notation

$$\begin{aligned}
\underset{(s,a)\sim\mathbf{d}_{k-1}}{\mathbb{E}}[\mathrm{approx}(s,a)] \triangleq & \underset{(s,a)\sim\mathbf{d}_{k-1}}{\mathbb{E}}\left[\underset{s'\sim P(\cdot|s,a)}{\mathbb{E}}\left[V_\theta^k(s')\right] - \phi(s,a)^\intercal \mathbf{M}\mathbf{V}_\theta^k\right] \\
= & \underset{(s,a)\sim\mathbf{d}_{k-1}}{\mathbb{E}}\left[\underset{s'\sim P(\cdot|s,a)}{\mathbb{E}}\left[V_\theta^k(s')\right] - \mathbf{P}\mathbf{V}_\theta^k(s,a)\right] \\
= & \underset{(s,a)\sim\mathbf{d}_{k-1}}{\mathbb{E}}\left[\underset{s'\sim P(\cdot|s,a)}{\mathbb{E}}\left[V_\theta^k(s')\right] - \underset{s'\sim P(\cdot|s,a)}{\mathbb{E}}\left[V_\theta^k(s')\right]\right] = 0.
\end{aligned}$$

Moreover,

$$d_{1,\chi} \triangleq \sum_{j=1}^m \frac{\sigma_j}{\sigma_j+\chi} \leq m$$

Finally, according to Remark 2 in [53], we have that ρ_χ is bounded as follows

$$\rho_\chi^2 \leq \frac{\|\boldsymbol{\phi}(s,a)\|_2^2}{\chi d_{1,\chi}} \leq \frac{1+\chi}{\chi\beta m} \leq \frac{2}{\chi\beta m},$$

where the last inequality follows from noticing that $d_{1,\chi} \geq \frac{\beta m}{1+\chi}$. Therefore, we can conclude that:

$$\begin{aligned}
\epsilon_{\mathrm{bs}} \leq & \mathcal{O}\left(\frac{(1+\frac{2}{\chi\beta})\epsilon_{\mathrm{rg}}}{N}\log\left(\frac{1}{\delta}\right)\right) \\
= & \mathcal{O}\left(\frac{\epsilon_{\mathrm{rg}}}{\chi\beta N}\log\left(\frac{1}{\delta}\right)\right) \\
= & \mathcal{O}\left(\frac{1}{N}\frac{m\chi}{\beta^4}D^2\log\left(\frac{1}{\delta}\right)\right),
\end{aligned}$$

Variance. From the bullet (5) in [53] it follows that

$$\epsilon_{\mathrm{vr}} = \mathcal{O}\left(\frac{\mathrm{Var}\left[\mathbf{V}_\theta^k(s') \mid s,a\right] d_{2,\chi}}{N}\log\left(\frac{1}{\delta}\right)\right).$$

We have $\mathrm{Var}\left[\mathbf{V}_\theta^k(s') \mid s,a\right] \le \left\|\mathbf{V}_\theta^k\right\|_\infty^2 \le D^2$. Finally, bounding $d_{2,\chi}$ we obtain that

$$d_{2,\chi} = \sum_{j=1}^m \left(\frac{\sigma_j}{\sigma_j+\chi}\right)^2 \le m.$$

Hence we can conclude

$$\epsilon_{\mathrm{vr}} = \mathcal{O}\left(\frac{D^2 m}{N}\log\left(\frac{1}{\delta}\right)\right).$$

Final bound. By combining the above bounds with Equation (27), we get the final bound

$$\left\|\mathbf{M}\mathbf{V}_\theta^k - \widehat{\mathbf{M}\mathbf{V}_\theta^k}\right\|_{\bar{\mathbf{\Lambda}}_k}^2 \le \mathcal{O}\left(\frac{m\chi^2}{\beta^3}D^2 + \frac{1}{N}\frac{m\chi}{\beta^4}D^2\log\left(\frac{1}{\delta}\right) + \frac{D^2 m}{N}\log\left(\frac{1}{\delta}\right)\right).$$

□

The bound above is minimized by choosing χ as small as allowed. This is made precise in the next corollary.

Corollary 3. *Let $\chi = \mathcal{O}(\frac{\log\frac{m}{\delta}}{\beta N})$. With probability at least $1-\delta$, it holds that*

$$\left\|\mathbf{M}\mathbf{V}_\theta^k - \widehat{\mathbf{M}\mathbf{V}_\theta^k}\right\|_{\bar{\mathbf{\Lambda}}_k}^2 \le \mathcal{O}\left(\frac{D^2 m}{\beta^5 N^2}\left(\log\left(\frac{m}{\delta}\right)\right)^2 + \frac{mD^2}{N}\log\left(\frac{1}{\delta}\right)\right).$$

In order to upper bound $\left\|\mathbf{M}\mathbf{V}_\theta^k - \widehat{\mathbf{M}\mathbf{V}_\theta^k}\right\|_2^2$ we need the next lemma. Hence, to bound $\left\|\mathbf{M}\mathbf{V}_\theta^k - \widehat{\mathbf{M}\mathbf{V}_\theta^k}\right\|_2^2$, we can directly apply Theorem 2 in [53] that leads to the following lemma.

Lemma 10. *Given a matrix $\mathbf{A} \in \mathbb{R}^{m\times m}$ and a vector $\mathbf{x} \in \mathbb{R}^m$, we have that $\|\mathbf{x}\|_{\mathbf{A}} \ge \lambda_{\min}(\mathbf{A})\|\mathbf{x}\|_2$.*

Proof. We have that $\mathbf{A} - \lambda_{\min}(\mathbf{A})\mathbf{I} \ge 0$ that implies $\mathbf{x}^\intercal\mathbf{A}\mathbf{x} \ge \lambda_{\min}(\mathbf{A})\mathbf{x}^\intercal\mathbf{x}$. □

Corollary 4. *Let $\chi = \mathcal{O}(\frac{\log\frac{m}{\delta}}{\beta N})$. With probability at least $1-\delta$, it holds that*

$$\left\|\mathbf{M}\mathbf{V}_\theta^k - \widehat{\mathbf{M}\mathbf{V}_\theta^k}\right\|_2 \le \mathcal{O}\left(\frac{D\sqrt{m}}{\beta^3 N}\log\left(\frac{m}{\delta}\right) + \frac{D\sqrt{m}}{\sqrt{N\beta}}\sqrt{\log\left(\frac{1}{\delta}\right)}\right). \tag{28}$$

Corollary 5. *Let $\chi = \mathcal{O}\left(\frac{\log\frac{m}{\delta}}{\beta N}\right)$, and $N \ge \max\left(\frac{\gamma^2 m D^2}{\beta\epsilon^2}\log(1/\delta), \frac{\gamma\sqrt{m}D}{\beta^3\epsilon}\log(m/\delta)\right)$. Then, with probability at least $1-\delta$, it holds that $\left\|\mathbf{M}\mathbf{V}_\theta^k - \widehat{\mathbf{M}\mathbf{V}_\theta^k}\right\|_2 \le \frac{\epsilon}{\gamma}$.*

H.1.2 Estimators for $\mathbf{\Gamma}_k$

Recall that we introduced $\mathbf{\Gamma}_k(i,j) \triangleq \sum_{s',a'}\mathbf{M}_{i,s'}\pi_{k-1,\boldsymbol{\theta}}(a'|s')\boldsymbol{\phi}_j(s',a')$. We can equivalently rewrite it as

$$\begin{aligned}\mathbf{\Gamma}_k(\cdot,j) =& \mathbf{M}\underbrace{\sum_{a'}\pi_{k-1,\boldsymbol{\theta}}(a'|s')\boldsymbol{\phi}_j(s',a')}_{h_{k,j}(s')}\\ =& \bar{\mathbf{\Lambda}}_k^{-1}\mathop{\mathbb{E}}_{(s,a)\sim\mathbf{d}_{k-1},s'\sim P(\cdot|s,a)}\left[\boldsymbol{\phi}(s,a)h_{k,j}(s')\right],\end{aligned}$$

where the last equality is obtained with manipulations analogous to Lemma 7.

Similarly, We can estimate $\mathbf{\Gamma}_k(i,j)$ with a finite amount of environment interactions sampled i.i.d. from $\mathbf{d}_{k-1}$, by solving the following ridge regression problem:

$$\widehat{\mathbf{\Gamma}}_k(\cdot,j) = \arg\min_{\mathbf{z}} \frac{1}{N}\sum_{n=1}^{N}\left(\boldsymbol{\phi}(s_{k-1}^{(n)}, a_{k-1}^{(n)})^\intercal \mathbf{z} - h_{k,j}(s_{k-1}^{\prime(n)})\right)^2 + \chi \|\mathbf{z}\|_2^2$$

Lemma 11. *By optimality conditions, we can obtain a closed form for $\widehat{\mathbf{\Gamma}}_k$ as*

$$\widehat{\mathbf{\Gamma}}_k(\cdot,j) = \frac{1}{N}\left(\mathbf{\Lambda}_{k,N} + \chi\mathbf{I}\right)^{-1}\sum_{n=1}^{N}\boldsymbol{\phi}(s_{k-1}^{(n)}, a_{k-1}^{(n)}) h_{k,j}(s_{k-1}^{\prime(n)}).$$

By noting that $\|\mathbf{h}_{k-1,j}\|_\infty \leq 1$ for any k, it follows that

Corollary 6. *For $\chi = \mathcal{O}(\frac{\log\frac{m}{\delta}}{\beta N})$, with probability at least $1-\delta$, it holds that*

$$\left\|\mathbf{\Gamma}_k(\cdot,j) - \widehat{\mathbf{\Gamma}}_k(\cdot,j)\right\|_2 \leq \mathcal{O}\left(\frac{\sqrt{m}}{\sqrt{N\beta}}\sqrt{\log\left(\frac{1}{\delta}\right)} + \frac{\sqrt{m}}{\beta^3 N}\log\left(\frac{m}{\delta}\right)\right). \tag{29}$$

Corollary 7. *For $\chi = \mathcal{O}(\frac{\log\frac{m}{\delta}}{\beta N})$, and $N \geq \max\left(\mathcal{O}\left(\frac{m}{\beta\epsilon^2}\log(1/\delta)\right), \mathcal{O}\left(\frac{\sqrt{m}}{\beta^3\epsilon}\log(m/\delta)\right)\right)$, with probability at least $1-\delta$, it holds that $\left\|\mathbf{\Gamma}_k(\cdot,j) - \widehat{\mathbf{\Gamma}}_k(\cdot,j)\right\|_2 \leq \epsilon$.*

H.1.3 Estimator for feature expectation vector $\boldsymbol{\rho}_{\mathbf{\Phi}}(\pi_{k-1})$

The goal is to estimate $\boldsymbol{\rho}_{\mathbf{\Phi}}(\pi_{k-1})$. Consider the sample transitions $\{s_{k-1}^{(n)}, a_{k-1}^{(n)}\}_{n=1}^N \sim \mathbf{d}_{k-1}^N$. Then we estimate $\boldsymbol{\rho}_{\mathbf{\Phi}}(\pi_{k-1}) = \mathbf{\Phi}^\intercal \mathbf{d}_{k-1}$ by $\boldsymbol{\rho}_{\mathbf{\Phi}}(\widehat{\pi_{k-1}}) \triangleq \frac{1}{N}\sum_{n=1}^N \boldsymbol{\phi}(s_{k-1}^{(n)}, a_{k-1}^{(n)})$.

In the next lemma, we provide a useful concentration result.

Lemma 12. *With probability at least $1-\delta$, for all $N \geq \frac{1.4\log\log(2N) + \log\frac{10.4m}{\delta}}{\beta\epsilon^2}$, and for all $i \in [m]$ simultaneously, it holds that*

$$\left|\boldsymbol{\rho}_{\mathbf{\Phi}}(\widehat{\pi_{k-1}})(i) - \boldsymbol{\rho}_{\boldsymbol{\phi}}(\pi_{k-1})(i)\right| \leq 2.26\epsilon\boldsymbol{\rho}_{\boldsymbol{\phi}}(\pi_{k-1})(i) \tag{30}$$

Proof. Consider the martingale difference sequence $Z_i(n) = \boldsymbol{\phi}_i(s_{k-1}^{(n)}, a_{k-1}^{(n)}) - \boldsymbol{\rho}_{\mathbf{\Phi}}(\pi_{k-1})(i)$ with the variance process $V_i(n) = \sum_{j=1}^n \mathbb{E}\left[Z_i^2(j)|\mathcal{F}_{j-1}\right]$, where $\mathcal{F}_{j-1}$ being the filtration up to the state action pair $(s_{k-1}^{(j)}, a_{k-1}^{(j)})$. We have,

$$\begin{aligned}
V_i(n) &= \sum_{j=1}^n \mathbb{E}\left[Z_i^2(j)|\mathcal{F}_{j-1}\right] \\
&= \sum_{j=1}^n \mathop{\mathbb{E}}_{(s,a)\sim\mathbf{d}_{k-1}}\left[\left(\boldsymbol{\phi}_i(s,a) - \boldsymbol{\rho}_{\mathbf{\Phi}}(\pi_{k-1})(i)\right)^2|\mathcal{F}_{j-1}\right] \\
&= \sum_{j=1}^n \mathop{\mathbb{E}}_{(s,a)\sim\mathbf{d}_{k-1}}\left[\boldsymbol{\phi}_i^2(s,a) - 2\boldsymbol{\phi}_i(s,a)\boldsymbol{\rho}_{\mathbf{\Phi}}(\pi_{k-1})(i) + \boldsymbol{\rho}_{\mathbf{\Phi}}(\pi_{k-1})(i)^2|\mathcal{F}_{j-1}\right] \\
&\leq \sum_{j=1}^n \mathop{\mathbb{E}}_{(s,a)\sim\mathbf{d}_{k-1}}\left[\boldsymbol{\phi}_i(s,a)|\mathcal{F}_{j-1}\right] - n\boldsymbol{\rho}_{\mathbf{\Phi}}(\pi_{k-1})(i)^2 \\
&= n\left(\boldsymbol{\rho}_{\mathbf{\Phi}}(\pi_{k-1})(i) - \boldsymbol{\rho}_{\mathbf{\Phi}}(\pi_{k-1})(i)^2\right) \leq n\boldsymbol{\rho}_{\mathbf{\Phi}}(\pi_{k-1})(i).
\end{aligned}$$

The martingale difference sequence $Z_i(j)$ satisfies the sub-ψ_P condition of [52] (see Bennet case in their Table 3) with constant $c = 2$. Therefore, by Lemma 13 in [89] with $m = \boldsymbol{\rho}_{\mathbf{\Phi}}(\pi_{k-1})(i)$, with probability at least $1-\frac{\delta}{2m}$, for all $N \geq \frac{1.4\log\log(2N)+\log\frac{10.4m}{\delta}}{\beta\epsilon^2}$ simultaneously, it holds that

$$\begin{aligned}
N\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{k-1}})(i) &\geq N\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{k-1})(i) - 1.44\sqrt{\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{k-1})(i)N\left(\log\log 2N + \frac{10.4m}{\delta}\right)} \\
&\quad - 0.82\left(1.4\log\log 2N + \frac{10.4m}{\delta}\right) \\
&\geq N\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{k-1})(i) - 1.44\sqrt{\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{k-1})(i)^2N^2\epsilon^2} - 0.82N\beta\epsilon^2 \\
&\geq N\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{k-1})(i) - 2.26\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{k-1})(i)N\epsilon.
\end{aligned}$$

Similarly, with probability at least $1-\frac{\delta}{2m}$, for all $N \geq \frac{1.4\log\log(2N)+\log\frac{10.4m}{\delta}}{\beta\epsilon^2}$ simultaneously, it holds that $\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{k-1}})(i) \leq \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{k-1})(i) + 2.26\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{k-1})(i)N\epsilon$. A union bound concludes the proof. □

H.1.4 Estimators for $\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}$

We can directly invoke Lemma 17 in [89] to get guarantees for the estimator $\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i)$. In particular, we obtain the following result.

Lemma 13. *Let* $\left\|\mathbf{MV}^k_\theta - \widehat{\mathbf{M}}\mathbf{V}^k_\theta\right\|_\infty \leq \frac{\epsilon}{\gamma}$ *and* $\left|\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{k-1}})(i) - \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{k-1})(i)\right| \leq 2.26\epsilon\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{k-1}})(i)$*. Then, it holds that* $\left|\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i) - \mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\right| \leq 38\eta\epsilon\mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i) \leq 38\frac{\eta\epsilon}{\beta}$.

Proof. First, we notice that $\left\|\mathbf{MV}^k_\theta - \widehat{\mathbf{M}}\mathbf{V}^k_\theta\right\|_\infty \leq \frac{\epsilon}{\gamma}$ implies that $\widehat{\boldsymbol{\delta}}^{\,k}_{\mathbf{w},\boldsymbol{\theta}}(i) - \boldsymbol{\delta}^{\,k}_{\mathbf{w},\boldsymbol{\theta}}(i) \leq \epsilon$. Therefore, by Lemma 17 in [89] we get $\left|\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i) - \mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\right| \leq 38\eta\epsilon\mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)$. Moreover, it holds that

$$\mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i) = \frac{e^{-\eta\boldsymbol{\delta}^{\,k}_{\mathbf{w},\boldsymbol{\theta}}(i)}}{\sum_i^m \rho_{\boldsymbol{\phi}_i}(\pi_{k-1})e^{-\eta\boldsymbol{\delta}^{\,k}_{\mathbf{w},\boldsymbol{\theta}}(i)}} \leq \frac{e^{-\eta\boldsymbol{\delta}^{\,k}_{\mathbf{w},\boldsymbol{\theta}}(i)}}{\beta\sum_i^m e^{-\eta\boldsymbol{\delta}^{\,k}_{\mathbf{w},\boldsymbol{\theta}}(i)}} \leq \frac{1}{\beta}.$$

Therefore,

$$\left|\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i) - \mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\right| \leq \frac{38\eta\epsilon}{\beta}, \quad \text{and} \quad \widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i) \leq \mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\left(1+38\eta\epsilon\right) \leq \frac{1}{\beta}\left(1+38\eta\epsilon\right).$$

□

Corollary 8. *Let* $N_1 \geq \max\left(\mathcal{O}\left(\frac{\gamma^2 mD^2}{\beta\epsilon^2}\log(2/\delta)\right), \mathcal{O}\left(\frac{\gamma\sqrt{m}D}{\beta^3\epsilon}\log(2m/\delta)\right)\right)$ *and* $N_2 \geq \frac{1.4\log\log(2N_2)+\log\frac{20.8m}{\delta}}{\beta\epsilon^2}$*. Then, for* $\chi = \mathcal{O}\left(\frac{\log\frac{2m}{\delta}}{\beta N}\right)$*, and for* $N \geq \max(N_1, N_2)$*, with probability at least* $1-\delta$*, it holds that* $\left|\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i) - \mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\right| \leq 38\frac{\eta\epsilon}{\beta}$*, for all* $i \in [m]$.

Proof. By Corollary 5, we have that with $N \geq N_1$ it holds that $\left\|\mathbf{MV}^k_\theta - \widehat{\mathbf{M}}\mathbf{V}^k_\theta\right\|_\infty \leq \frac{\epsilon}{\gamma}$, with probability $1-\delta/2$. Furthermore, Lemma 12 gives that for $N \geq \frac{1.4\log\log(2N)+\log\frac{20.8m}{\delta}}{\beta\epsilon^2}$, it holds with probability $1-\delta/2$ that $\left|\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{k-1}})(i) - \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{k-1})(i)\right| \leq 2.26\epsilon\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{k-1}})(i)$, for all $i \in [m]$ simultaneously.

Therefore, a union bound gives that for $N \geq \max(N_1, N_2)$, with probability $1-\delta$, we have that $\left\|\mathbf{MV}^k_\theta - \widehat{\mathbf{M}}\mathbf{V}^k_\theta\right\|_\infty \leq \frac{\epsilon}{\gamma}$, and $\left|\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{k-1}})(i) - \boldsymbol{\rho}_{\boldsymbol{\Phi}}(\pi_{k-1})(i)\right| \leq 2.26\epsilon\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{k-1}})(i)$, for all $i \in [m]$. An application of Lemma 13 concludes the proof. □

H.1.5 Estimators for $\mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\boldsymbol{\Gamma}_k(i,j)$

We obtain an estimator for $\mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\boldsymbol{\Gamma}_k(i,j)$ simply as $\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\widehat{\boldsymbol{\Gamma}}_k(i,j)$. The next lemma gives guarantees for such an estimator.

Lemma 14. *Assume that for any* $(i,j) \in [m]^2$*, it holds that* $\left|\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i) - \mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\right| \leq \frac{38\eta\epsilon}{\beta}$ *and* $\left|\widehat{\mathbf{\Gamma}}_k(i,j) - \mathbf{\Gamma}_k(i,j)\right| \leq \epsilon$*. Then,* $\left|\mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\mathbf{\Gamma}_k(i,j) - \widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\widehat{\mathbf{\Gamma}}_k(i,j)\right| \leq \frac{\epsilon}{\beta}(1+(1+\epsilon)38\eta)$*, for all* $(i,j) \in [m]^2$.

Proof. We have that

$$
\begin{aligned}
\left|\mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\mathbf{\Gamma}_k(i,j) - \widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\widehat{\mathbf{\Gamma}}_k(i,j)\right| \leq & \mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\left|\widehat{\mathbf{\Gamma}}_k(i,j) - \mathbf{\Gamma}_k(i,j)\right| + \widehat{\mathbf{\Gamma}}_k(i,j)\left|\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i) - \mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\right| \\
\leq & \frac{1}{\beta}\left|\widehat{\mathbf{\Gamma}}_k(i,j) - \mathbf{\Gamma}_k(i,j)\right| + (1+\epsilon)\left|\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i) - \mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\right| \\
\leq & \frac{\epsilon}{\beta} + (1+\epsilon)\frac{38\eta\epsilon}{\beta} = \frac{\epsilon}{\beta}\big(1+(1+\epsilon)38\eta\big),
\end{aligned}
$$

where we used the bound $\hat{\mathbf{\Gamma}}_k(i,j) \leq \mathbf{\Gamma}_k(i,j) + \epsilon \leq 1+\epsilon$. $\square$

Lemma 15. *For* $\chi = \mathcal{O}\left(\frac{\log\frac{m}{\delta}}{\beta N}\right)$*, choose* $N \geq \max\left(N_1, N_2, \mathcal{O}\left(\frac{m}{\beta\epsilon^2}\log(m/\delta)\right), \mathcal{O}\left(\frac{\sqrt{m}}{\beta^3\epsilon}\log(m^2/\delta)\right)\right)$ *with* N_1 *and* N_2 *as defined in Corollary 8, then with probability* $1-\delta$*, for all* $(i,j) \in [m]^2$ *simultaneously:*

$$
\left|\mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\mathbf{\Gamma}_k(i,j) - \widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\widehat{\mathbf{\Gamma}}_k(i,j)\right| \leq \frac{\epsilon}{\beta}(1+(1+\epsilon)38\eta).
$$

Proof. By Corollary 8, when $\chi = \mathcal{O}\left(\frac{\log\frac{m}{\delta}}{\beta N}\right)$, and $N \geq \max(N_1, N_2)$, it holds with probability at least $1-\delta$ that

$$
\left|\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i) - \mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\right| \leq \frac{38\eta\epsilon}{\beta}, \text{ for all } i \in [m].
$$

Moreover, when $N \geq \max\left(\mathcal{O}\left(\frac{m}{\beta\epsilon^2}\log(m/\delta)\right), \mathcal{O}\left(\frac{\sqrt{m}}{\beta^3\epsilon}\log(m^2/\delta)\right)\right)$, by Corollary 7, with probability at least $1-\delta$, it holds that $\left\|\mathbf{\Gamma}_k(\cdot,j) - \widehat{\mathbf{\Gamma}}_k(\cdot,j)\right\|_2 \leq \epsilon$, for all $j \in [m]$ simultaneously.

Finally, a union bound and Lemma 14 give that with probability at least $1-\delta$, it holds that $\left|\mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\mathbf{\Gamma}_k(i,j) - \widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i)\widehat{\mathbf{\Gamma}}_k(i,j)\right| \leq \frac{\epsilon}{\beta}(1+(1+\epsilon)38\eta)$. $\square$

H.2 Properties of Stochastic gradients

Lemma 16. *Let* $N \geq \max\left(N_1, N_2, \mathcal{O}\left(\frac{m}{\beta\epsilon^2}\log(m/\delta)\right), \mathcal{O}\left(\frac{\sqrt{m}}{\beta^3\epsilon}\log(m^2/\delta)\right)\right)$ *with* N_1 *and* N_2 *with* N_1 *and* N_2 *as defined in Corollary 8. Then, with probability* $1-\delta$*, the following bounds on the stochastic gradient variance hold simultaneously:*

$$
\begin{aligned}
\left\|\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) - \mathop{\mathbb{E}}_{i_{k-1}^{(N+1)}}\left[\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})|\mathcal{F}_N\right]\right\|_\infty &\leq 2\frac{(1+38\epsilon\eta)}{\beta}(2+\epsilon) + 2(1-\gamma), \\
\left\|\widehat{\nabla}_{\mathbf{w}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) - \mathop{\mathbb{E}}_{i_{k-1}^{(N+1)}}\left[\widehat{\nabla}_{\mathbf{w}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})|\mathcal{F}_N\right]\right\|_\infty &\leq 2\left(1 + \frac{1+38\eta\epsilon}{\beta}\right).
\end{aligned}
$$

Furthermore, with probability at least $1-\delta$*, the following bounds on the stochastic gradient bias hold simultaneously:*

$$
\begin{aligned}
\left\|\mathbb{E}\left[\widehat{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})|\mathcal{F}_N\right] - \nabla_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})\right\|_1 &\leq m\frac{\epsilon}{\beta}\left(\gamma + 38\eta\left(1+\gamma(1+\epsilon)\right)\right), \\
\left\|\nabla_{\mathbf{w},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) - \mathbb{E}\left[\widehat{\nabla}_{\mathbf{w},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})|\mathcal{F}_N\right]\right\|_1 &\leq \frac{38\eta\epsilon}{\beta}.
\end{aligned}
$$

Proof. **Variance for gradient wrt $\boldsymbol{\theta}$.** Recall that by definition of the stochastic gradient we have that

$$\widehat{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) - (1-\gamma)\boldsymbol{\phi}_j(s_{k-1}^{(0)},a_{k-1}^{(0)}) = \widehat{\mathbf{B}}_{\mathbf{w},\boldsymbol{\theta}}^k(i_{k-1}^{(N+1)})\left[\gamma\widehat{\boldsymbol{\Gamma}}_k(i_{k-1}^{(N+1)},j) - \mathbb{1}\{i_{k-1}^{(N+1)}=j\}\right].$$

It then follows that

$$\left|\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) - (1-\gamma)\boldsymbol{\phi}_j(s_{k-1}^{(0)},a_{k-1}^{(0)})\right| \leq \gamma\left|\widehat{\mathbf{B}}_{\mathbf{w},\boldsymbol{\theta}}^k(i_{k-1}^{(N+1)})\widehat{\boldsymbol{\Gamma}}_k(i_{k-1}^{(N+1)},j)\right| + \left\|\widehat{\mathbf{B}}_{\mathbf{w},\boldsymbol{\theta}}^k(j)\right\|_\infty$$

Invoking Lemma 15, we have that if $N \geq \max\left(N_1, N_2, \mathcal{O}\left(\frac{m}{\beta\epsilon^2}\log(m/\delta)\right), \mathcal{O}\left(\frac{\sqrt{m}}{\beta^3\epsilon}\log(m^2/\delta)\right)\right)$ with N_1 and N_2 as defined in Corollary 8, then with probability $1-\delta$,

$$\left|\widehat{\mathbf{B}}_{\mathbf{w},\boldsymbol{\theta}}^k(i_{k-1}^{(N+1)})\widehat{\boldsymbol{\Gamma}}_k(i_{k-1}^{(N+1)},j)\right| \leq \frac{1}{\beta} + \frac{\epsilon}{\beta}(1+38(1+\epsilon)\eta) = \frac{1}{\beta}(1+\epsilon(1+38(1+\epsilon)\eta)).$$

Similarly, by Corollary 8, for $N \geq \max(N_1, N_2)$, we have that with probability $1-\delta$,

$$\widehat{\mathbf{B}}_{\mathbf{w},\boldsymbol{\theta}}^k(i_{k-1}^{(N+1)}) \leq \mathbf{B}_{\mathbf{w},\boldsymbol{\theta}}^k(i_{k-1}^{(N+1)})\left(1+38\eta\epsilon\right) \leq \frac{1}{\beta}\left(1+38\eta\epsilon\right).$$

Hence, a union bound gives that with probability $1-\delta$,

$$\left|\widehat{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) - (1-\gamma)\boldsymbol{\phi}_j(s_{k-1}^{(0)},a_{k-1}^{0})\right| \leq \gamma\frac{(1+38\epsilon\eta)}{\beta}(1+\epsilon) + \frac{(1+38\epsilon\eta)}{\beta}.$$

This implies that

$$\left|\widehat{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})\right| \leq \gamma\frac{(1+38\epsilon\eta)}{\beta}(1+\epsilon) + \frac{(1+38\epsilon\eta)}{\beta} + (1-\gamma).$$

Therefore, by introducing a filtration $\mathcal{F}_N = \sigma\left(\{(s_{k-1}^{(n)},a_{k-1}^{(n)},s_{k-1}'^{(n)})\}_{n=1}^N\right)$, and noticing that $\widehat{\mathbf{B}}_{\mathbf{w},\boldsymbol{\theta}}^k$ and $\widehat{\boldsymbol{\Gamma}}_k$ are $\mathcal{F}_N$-measurable, we get

$$\begin{aligned}\left|\mathop{\mathbb{E}}_{i_{k-1}^{(N+1)}}\left[\widehat{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})|\mathcal{F}_N\right]\right| &\leq \mathop{\mathbb{E}}_{i_{k-1}^{(N+1)}}\left[\left|\widehat{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})\right||\mathcal{F}_N\right]\\ &\leq \gamma\frac{(1+38\epsilon\eta)}{\beta}(1+\epsilon) + \frac{(1+38\epsilon\eta)}{\beta} + (1-\gamma)\end{aligned}$$

At this point, we can simply notice that

$$\begin{aligned}\left|\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) - \mathop{\mathbb{E}}_{i_{k-1}^{(N+1)}}\left[\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})|\mathcal{F}_N\right]\right| &\leq 2\left[\gamma\frac{(1+38\epsilon\eta)}{\beta}(1+\epsilon) + \frac{(1+38\epsilon\eta)}{\beta} + 2(1-\gamma)\right]\\ &\leq 2\frac{(1+38\epsilon\eta)}{\beta}(2+\epsilon) + 2(1-\gamma).\end{aligned}$$

Therefore, with probability $1-\delta$, it holds that

$$\left\|\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) - \mathop{\mathbb{E}}_{i_{k-1}^{(N+1)}}\left[\widehat{\nabla}_{\boldsymbol{\theta}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})|\mathcal{F}_N\right]\right\|_\infty \leq 2\frac{(1+38\epsilon\eta)}{\beta}(2+\epsilon) + 2(1-\gamma).$$

Variance for gradient wrt $\mathbf{w}$. Similarly with Corollary 8, we obtain that if $N \geq \max(N_1, N_2)$, then with probability at least $1-\delta$,

$$\left\|\widehat{\nabla}_{\mathbf{w}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})\right\|_\infty \leq 1 + \frac{1+38\eta\epsilon}{\beta}.$$

This implies that

$$\left\|\widehat{\nabla}_{\mathbf{w}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta}) - \mathop{\mathbb{E}}_{i_{k-1}^{(N+1)}}\left[\widehat{\nabla}_{\mathbf{w}}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})|\mathcal{F}_N\right]\right\|_\infty \leq 2\left(1+\frac{1+38\eta\epsilon}{\beta}\right).$$

Bias for gradient wrt $\boldsymbol{\theta}$. By using the unbiased estimator $\widetilde{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})$ in Equation (25), we get

$$\begin{aligned}\left|\widetilde{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})-\widehat{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})\right| &\leq \left|\gamma\left(\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i_{k-1}^{(N+1)})\widehat{\boldsymbol{\Gamma}}_k(i_{k-1}^{(N+1)},j)-\mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i_{k-1}^{(N+1)})\boldsymbol{\Gamma}_k(i_{k-1}^{(N+1)},j)\right)\right|\\ &\quad+\left|\mathbb{1}\{i_{k-1}^{(N+1)}=j\}\left(\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i_{k-1}^{(N+1)})-\mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i_{k-1}^{(N+1)})\right)\right|\\ &\leq\gamma\left|\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i_{k-1}^{(N+1)})\widehat{\boldsymbol{\Gamma}}_k(i_{k-1}^{(N+1)},j)-\mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i^{(N+1)})\boldsymbol{\Gamma}_k(i_{k-1}^{(N+1)},j)\right|.\end{aligned}$$

By choosing χ and N as in Lemma 15 and Corollary 8, and by a union bound, we have that with probability $1-\delta$,

$$\begin{aligned}\left|\widetilde{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})-\widehat{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})\right| &\leq\gamma\frac{\epsilon}{\beta}(1+(1+\epsilon)38\eta)+\frac{38\epsilon\eta}{\beta}\\ &=\frac{\epsilon}{\beta}\left(\gamma+38\eta\left(1+\gamma(1+\epsilon)\right)\right).\end{aligned}$$

Using that $\widetilde{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})$ is an unbiased estimator of $\nabla_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})$, we get

$$\begin{aligned}\left|\mathbb{E}\left[\widehat{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})|\mathcal{F}_N\right]-\nabla_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})\right| &=\left|\mathbb{E}\left[\widehat{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})-\widetilde{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})\right]\right|\\ &\leq\mathbb{E}\left[\left|\widehat{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})-\widetilde{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})\right|\right]\\ &\leq\frac{\epsilon}{\beta}\left(\gamma+38\eta\left(1+\gamma(1+\epsilon)\right)\right).\end{aligned}$$

Hence, we have that $\left\|\mathbb{E}\left[\widehat{\nabla}_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})|\mathcal{F}_N\right]-\nabla_{\boldsymbol{\theta},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})\right\|_1\leq m\frac{\epsilon}{\beta}\left(\gamma+38\eta\left(1+\gamma(1+\epsilon)\right)\right)$.

Bias bound for the gradient wrt $\mathbf{w}$. Similarly, we can notice that with probability at least $1-\delta$, it holds that

$$\begin{aligned}\left|\widetilde{\nabla}_{\mathbf{w},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})-\widehat{\nabla}_{\mathbf{w},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})\right| &=\left|\mathbb{1}\{i_{k-1}^{(N+1)}=j\}\left(\widehat{\mathbf{B}}^k_{\mathbf{w},\boldsymbol{\theta}}(i_{k-1}^{(N+1)})-\mathbf{B}^k_{\mathbf{w},\boldsymbol{\theta}}(i_{k-1}^{(N+1)})\right)\right|\\ &\leq\frac{38\eta\epsilon}{\beta}.\end{aligned}$$

Since we have only one non-zero element, and by the unbiasedness of $\widetilde{\nabla}_{\mathbf{w},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})$, we get

$$\left\|\nabla_{\mathbf{w},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})-\mathbb{E}\left[\widehat{\nabla}_{\mathbf{w},j}\mathcal{G}_k(\mathbf{w},\boldsymbol{\theta})|\mathcal{F}_N\right]\right\|_1\leq\frac{38\eta\epsilon}{\beta}.$$

□

I Proof of Theorem 2

We first prove a generalization of the Azuma-Hoeffding inequality (Theorem 3.14 in [75]) that holds when the martingale difference sequence is bounded with high probability but not almost surely.

Lemma 17 (Modified Azuma-Hoeffding)**.** *Let $\{Y_i\}_i^n$ be a martingale difference sequence adapted to $\mathcal{F}_i$, such that for each i, $|Y_i|\leq c_i$ with probability at least $1-\delta_2$. Then, it holds that*

$$\mathbb{P}\left[\sum_{i=1}^n Y_i\geq\epsilon\right]\leq\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)+n\delta_2. \tag{31}$$

Proof. Define the events $E_i=\{Y_i\leq c_i\}$ and the intersection $E=\cap_{i=1}^n\{E_i\}$, and notice that $\mathbb{P}[E^c]=\mathbb{P}[\cup_{i=1}^n\{E_i^c\}]\leq\sum_{i=1}^n\mathbb{P}[E_i^c]=n\delta_2$. We then have the following decomposition:

$$\begin{aligned}\mathbb{P}\left[\sum_{i=1}^n Y_i\geq\epsilon\right] &=\mathbb{P}\left[\{\sum_{i=1}^n Y_i\geq\epsilon\}\cap E\right]+\mathbb{P}\left[\{\sum_{i=1}^n Y_i\geq\epsilon\}\cap E^c\right]\\ &\leq\mathbb{P}\left[\{\sum_{i=1}^n Y_i\geq\epsilon\}\cap E\right]+\mathbb{P}[E^c]\\ &\leq\mathbb{P}\left[\sum_{i=1}^n Y_i\geq\epsilon\big|E\right]\underbrace{\mathbb{P}[E]}_{\leq 1}+n\delta_2\leq\exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2}\right)+n\delta_2,\end{aligned}$$

where in the last step we noticed that under the event E, the martingale difference sequence is bounded almost surely, therefore we can apply the standard Azuma-Hoeffding inequality. $\square$

Corollary 9. *Let $\{Y_i\}_i^n$ be a martingale difference sequence adapted to $\mathcal{F}_i$, such that for each i, $|Y_i| \leq c_i$ with probability at least $1-\delta_2$. Then, with probability $1-\delta_1$ (with $\delta_1 > n\delta_2$), it holds that*

$$\mathbb{P}\left[\sum_{i=1}^{n} Y_i \geq \sqrt{\frac{\left(\sum_{i=1}^{n} c_i^2\right)\log\left(1/\left(\delta_1 - n\delta_2\right)\right)}{2}}\right] \leq \delta_1 + n\delta_2.$$

Proof of Theorem 2. We fix a policy evaluation step $k \in [K]$, i.e., we study the k-th iteration of the outer loop of Algorithm 1. Similarly to the proof of Lemma 19 in [89], the biased SGD subroutine can be seen as an inexact gradient ascent scheme with updates

$$\mathbf{w}_{t+1}^k = \Pi_{\mathcal{W}}\Big(\mathbf{w}_t^k + \beta_t(\nabla_{\mathbf{w}} f(\mathbf{w}_t^k, \boldsymbol{\theta}_t^k) + b_{\mathbf{w},t}^k + \epsilon_{\mathbf{w},t}^k\Big), \tag{32}$$

$$\boldsymbol{\theta}_{t+1}^k = \Pi_{\Theta}\Big(\boldsymbol{\theta}_t^k + \beta_t(\nabla_{\boldsymbol{\theta}} f(\mathbf{w}_t^k, \boldsymbol{\theta}_t^k) + b_{\boldsymbol{\theta},t}^k + \epsilon_{\boldsymbol{\theta},t}^k\Big), \tag{33}$$

with

$$\epsilon_{\boldsymbol{\theta},t}^k \triangleq \widehat{\nabla}_{\boldsymbol{\theta}} \mathcal{G}_k(\mathbf{w}_t^k, \boldsymbol{\theta}_t^k) - \mathbb{E}\left[\widehat{\nabla}_{\boldsymbol{\theta}} \mathcal{G}_k(\mathbf{w}_t^k, \boldsymbol{\theta}_t^k) \mid \mathcal{F}_{t-1}\right], \tag{34}$$

$$\epsilon_{\mathbf{w},t}^k \triangleq \widehat{\nabla}_{\mathbf{w}} \mathcal{G}_k(\mathbf{w}_t^k, \boldsymbol{\theta}_t^k) - \mathbb{E}\left[\widehat{\nabla}_{\mathbf{w}} \mathcal{G}_k(\mathbf{w}_t^k, \boldsymbol{\theta}_t^k) \mid \mathcal{F}_{t-1}\right], \tag{35}$$

$$b_{\boldsymbol{\theta},t}^k \triangleq \mathbb{E}\left[\widehat{\nabla}_{\boldsymbol{\theta}} \mathcal{G}_k(\mathbf{w}_t^k, \boldsymbol{\theta}_t^k) \mid \mathcal{F}_{t-1}\right] - \nabla_{\boldsymbol{\theta}} \mathcal{G}_k(\mathbf{w}_t^k, \boldsymbol{\theta}_t^k), \tag{36}$$

$$b_{\mathbf{w},t}^k \triangleq \mathbb{E}\left[\widehat{\nabla}_{\mathbf{w}} \mathcal{G}_k(\mathbf{w}_t^k, \boldsymbol{\theta}_t^k) \mid \mathcal{F}_{t-1}\right] - \nabla_{\mathbf{w}} \mathcal{G}_k(\mathbf{w}_t^k, \boldsymbol{\theta}_t^k). \tag{37}$$

By Lemma 16, and a union bound, we get that for $n(t) \geq \max\left\{\mathcal{O}\left(\frac{\gamma^2 m D^2}{\beta \xi_t^2}\log(\frac{Tm}{\delta}),\right), \mathcal{O}\left(\frac{m}{\beta \xi_t^2}\log(\frac{Tm}{\delta}),\right)\right\}$, with probability at least $1-\delta/2$, for all $t = 1, \ldots, T$ simultaneously, it holds that

$$\|\epsilon_{\boldsymbol{\theta},t}^k\|_1 \leq 2m\frac{(1+38\xi_t\eta)}{\beta}(2+\xi_t) + 2(1-\gamma) \leq \frac{6m}{\beta}(1+38\eta) + 2, \tag{38}$$

$$\|\epsilon_{\mathbf{w},t}^k\|_1 \leq 2m\left(1 + \frac{1+38\eta\xi_t}{\beta}\right) \leq 2m(1+\frac{1+38\eta}{\beta}), \tag{39}$$

$$\|b_{\boldsymbol{\theta},t}^k\|_1 \leq m\frac{\xi_t}{\beta}\left(\gamma + 38\eta\left(1+\gamma(1+\xi_t)\right)\right) \leq \frac{m}{\beta}(1+114\beta), \tag{40}$$

$$\|b_{\mathbf{w},t}^k\|_1 \leq \frac{38\eta\xi_t}{\beta} \leq \frac{38\eta}{\beta}, \tag{41}$$

where we used that $\{\xi_t\}_{t=1}^T \cup \{\gamma\} \subset (0,1)$.

Moreover, by Hölder's inequality, we get

$$|\left\langle \epsilon_{\boldsymbol{\theta},t}^k, \boldsymbol{\theta}_t^k - \boldsymbol{\theta}_k^\star \right\rangle| \leq \|\epsilon_{\boldsymbol{\theta},t}^k\|_1 \|\boldsymbol{\theta}_t^k - \boldsymbol{\theta}_k^\star\|_\infty \leq \frac{12Dm}{\beta}(1+38\eta) + 2 \triangleq M_1, \tag{42}$$

$$|\left\langle \epsilon_{\mathbf{w},t}^k, \mathbf{w}_t^k - \mathbf{w}_k^\star \right\rangle| \leq \|\epsilon_{\mathbf{w},t}^k\|_1 \|\mathbf{w}_t^k - \mathbf{w}_k^\star\|_\infty \leq 2m(1+\frac{1+38\eta}{\beta}) \triangleq M_2, \tag{43}$$

where we used that by the triangle inequality and Proposition 3, it holds that $\|\boldsymbol{\theta}_t^k - \boldsymbol{\theta}_k^\star\|_\infty \leq 2\frac{1+|\log\beta|}{1-\gamma} \triangleq 2D$. We recall that $D \triangleq \frac{1+\log\left(\frac{1}{\beta}\right)}{1-\gamma} \geq 1$.

Since $\left\{X_{\boldsymbol{\theta},t}^k \triangleq \left\langle \epsilon_{\boldsymbol{\theta},t}^k, \boldsymbol{\theta}_t^k - \boldsymbol{\theta}_k^\star \right\rangle\right\}_{t=1}^\infty$ and $\left\{X_{\mathbf{w},t}^k \triangleq \left\langle \epsilon_{\mathbf{w},t}^k, \mathbf{w}_t^k - \mathbf{w}_k^\star \right\rangle\right\}_{t=1}^\infty$ are martingale differences, by using Corollary 9 and a simple union bound, we get that with probability at least $1-\delta/2$,

$$-\sum_{t=1}^{T}\left\langle \epsilon_{\boldsymbol{\theta},t}^k, \boldsymbol{\theta}_t^k - \boldsymbol{\theta}_k^\star \right\rangle \leq 2M_1\sqrt{T\log(\frac{16T}{\delta})}, \tag{44}$$

$$-\sum_{t=1}^{T}\left\langle \epsilon_{\mathbf{w},t}^k, \mathbf{w}_t^k - \mathbf{w}_k^\star \right\rangle \leq 2M_2\sqrt{T\log(\frac{16T}{\delta})}. \tag{45}$$

Furthermore, note that $\mathcal{G}_k$ is $\eta+\alpha$-smooth with respect to the $\|\cdot\|_\infty$-norm, and so by Lemma 12 in [89], we can bound the $\|\cdot\|_1$-norm of its gradients. In particular, we have

$$\|\nabla_{\boldsymbol{\theta}}\mathcal{G}(\mathbf{w}_t^k,\boldsymbol{\theta}_t^k)\|_1+\|\nabla_{\mathbf{w}}\mathcal{G}(\mathbf{w}_t^k,\boldsymbol{\theta}_t^k)\|_1\le 2(\eta+\alpha)(D+1). \tag{46}$$

This in turn implies that

$$\|\nabla_{\boldsymbol{\theta}}\mathcal{G}(\mathbf{w}_t^k,\boldsymbol{\theta}_t^k)\|_2^2+\|\nabla_{\mathbf{w}}\mathcal{G}(\mathbf{w}_t^k,\boldsymbol{\theta}_t^k)\|_2^2\le 4(\eta+\alpha)^2(D+1)^2. \tag{47}$$

By smoothness and concavity of the objective $\mathcal{G}_k$, we can apply Lemma 9 in [89]. In particular, by Equations (38)–(47), a union bound, and by summing over t in the bound of Lemma 9 in [89], we have the following guarantee for our inexact gradient scheme:

If $n(t)\ge\max\left\{\mathcal{O}\left(\frac{\gamma^2mD^2}{\beta\xi_t^2}\log(\frac{Tm}{\delta}),\right),\mathcal{O}\left(\frac{m}{\beta\xi_t^2}\log(\frac{Tm}{\delta}),\right)\right\}$, and $\beta_t\le\frac{2}{\alpha+\eta}$, then with probability at least $1-\delta$, it holds that

$$\sum_{t=1}^T\Big(\mathcal{G}_k(\mathbf{w}_k^\star,\boldsymbol{\theta}_k^\star)-\mathcal{G}_k(\mathbf{w}_t^k,\boldsymbol{\theta}_t^k)\Big) \tag{48}$$

$$\le\sum_{t=1}^T\frac{\|\mathbf{w}_t^k-\mathbf{w}_k^\star\|_2^2+\|\mathbf{w}_{t+1}^k-\mathbf{w}_k^\star\|_2^2}{2\beta_t}+\sum_{t=1}^T\frac{\|\boldsymbol{\theta}_t^k-\boldsymbol{\theta}_k^\star\|_2^2+\|\boldsymbol{\theta}_{t+1}^k-\boldsymbol{\theta}_k^\star\|_2^2}{2\beta_t} \tag{49}$$

$$+2\sum_{t=1}^T\beta_t\left(\|\nabla_{\boldsymbol{\theta}}\mathcal{G}(\mathbf{w}_t^k,\boldsymbol{\theta}_t^k)\|_2^2+\|\nabla_{\mathbf{w}}\mathcal{G}(\mathbf{w}_t^k,\boldsymbol{\theta}_t^k)\|_2^2\right) \tag{50}$$

$$+5\sum_{t=1}^T\beta_t\left(\|b_{\mathbf{w},t}^k\|_2^2+\|b_{\boldsymbol{\theta},t}^k\|_2^2+\|\epsilon_{\mathbf{w},t}^k\|_2^2+\|\epsilon_{\boldsymbol{\theta},t}^k\|_2^2\right) \tag{51}$$

$$+\sum_{t=1}^T\left(\|b_{\mathbf{w},t}^k\|_1+\|b_{\boldsymbol{\theta},t}^k\|_1\right)\max\{\|\mathbf{w}_t^k-\mathbf{w}_k^\star\|_\infty,\|\boldsymbol{\theta}_t^k-\boldsymbol{\theta}_k^\star\|_\infty\} \tag{52}$$

$$-\sum_{t=1}^T\left\langle\epsilon_{\boldsymbol{\theta},t}^k,\boldsymbol{\theta}_t^k-\boldsymbol{\theta}_k^\star\right\rangle-\sum_{t=1}^T\left\langle\epsilon_{\mathbf{w},t}^k,\mathbf{w}_t^k-\mathbf{w}_k^\star\right\rangle \tag{53}$$

$$\le\sum_{t=1}^T\frac{\|\mathbf{w}_t^k-\mathbf{w}_k^\star\|_2^2+\|\mathbf{w}_{t+1}^k-\mathbf{w}_k^\star\|_2^2}{2\beta_t}+\sum_{t=1}^T\frac{\|\boldsymbol{\theta}_t^k-\boldsymbol{\theta}_k^\star\|_2^2+\|\boldsymbol{\theta}_{t+1}^k-\boldsymbol{\theta}_k^\star\|_2^2}{2\beta_t} \tag{54}$$

$$+\sum_{t=1}^T\Big(\beta_tL_1+2DL_2\xi_t\Big)+2(M_1+M_2)\sqrt{T\log(\frac{16T}{\delta})}, \tag{55}$$

where

$$L_1=\mathcal{O}\Big((\eta+\alpha)^2D^2+\frac{\max\{\eta,1\}^2m^2}{\beta^2}\Big), \tag{56}$$

$$L_2=\mathcal{O}\Big(\frac{\eta+m}{\beta}\Big), \tag{57}$$

$$M_1=\mathcal{O}\Big(\frac{\max\{\eta,1\}m}{\beta}\Big), \tag{58}$$

$$M_2=\mathcal{O}\Big(\frac{\max\{\eta,1\}Dm}{\beta}\Big), \tag{59}$$

$$\tag{60}$$

We choose $\beta_t=\frac{L}{\sqrt{t}}$, for some constant L. Then a telescoping sum gives

$$\sum_{t=1}^T\Big(\frac{\|\mathbf{w}_t^k-\mathbf{w}_k^\star\|_2^2+\|\mathbf{w}_{t+1}^k-\mathbf{w}_k^\star\|_2^2}{2\beta_t}+\frac{\|\boldsymbol{\theta}_t^k-\boldsymbol{\theta}_k^\star\|_2^2+\|\boldsymbol{\theta}_{t+1}^k-\boldsymbol{\theta}_k^\star\|_2^2}{2\beta_t}\Big)\le\frac{1}{2L}(D^2+1)\sqrt{T}. \tag{61}$$

Moreover, $\sum_{t=1}^{T} \beta_t L_1 \le 2L_1 L\sqrt{T}$. By combining this inequality with Equations (55) and (61), we get that

$$\sum_{t=1}^{T}\Big(\mathcal{G}_k(\mathbf{w}_k^\star,\boldsymbol{\theta}_k^\star)-\mathcal{G}_k(\mathbf{w}_t^k,\boldsymbol{\theta}_t^k)\Big) \le \frac{1}{2L}(D^2+1)\sqrt{T}+2L_1L\sqrt{T}+2DL_2\sum_{t=1}^{T}\xi_t \tag{62}$$

$$+2(M_1+M_2)\sqrt{T\log(\frac{16T}{\delta})} \tag{63}$$

The optimal choice for L is $L=\frac{\sqrt{1+D^2}}{2\sqrt{L_1}}$. In addition, by setting $\xi_t=\sqrt{\frac{L_1}{t}}$, we conclude that

$$\sum_{t=1}^{T}\Big(\mathcal{G}_k(\mathbf{w}_k^\star,\boldsymbol{\theta}_k^\star)-\mathcal{G}_k(\mathbf{w}_t^k,\boldsymbol{\theta}_t^k)\Big) \le 4\max\Big\{\sqrt{1+D^2},\,2DL_2\Big\}\sqrt{L_1}\sqrt{T} \tag{64}$$

$$+2(M_1+M_2)\sqrt{T\log(\frac{16T}{\delta})}. \tag{65}$$

Therefore, by combining Equations (56)–(59) and Equation (64), and by Jensen's inequality, we get that if $n(t)\ge\max\Big(\mathcal{O}\Big(\frac{\gamma^2mDt}{(\eta+\alpha)^2\beta}\log\frac{Tm}{\delta}\Big),\mathcal{O}\Big(\frac{mt}{\beta}\log\frac{Tm}{\delta}\Big)\Big)$, and $\beta_t=\mathcal{O}(\frac{1}{\sqrt{t}})$, then, with probability at least $1-\delta$, it holds that $\mathcal{G}_k(\mathbf{w}_k^\star,\boldsymbol{\theta}_k^\star)-\mathcal{G}_k(\mathbf{w}_k,\boldsymbol{\theta}_k)\le\mathcal{O}\Big(\frac{\max\{\eta,1\}mD}{\beta\sqrt{T}}\Big)$.

□

I.1 Proof of Corollary 1

Proof of Corollary 1. We plug the upper bound for ϵ_k given by Theorem 2 in the error propagation analysis of Theorem 1. In particular, from Theorem 1, with probability at least $1-\delta_1$, it holds that

$$d_{\mathcal{C}}(\widehat{\pi},\pi_{\mathrm{E}})\le\frac{1}{K}\Big(\frac{D(\boldsymbol{\lambda}^*||\boldsymbol{\Phi}^\intercal\mathbf{d}_0)}{\eta}+\frac{H(\mathbf{d}^*||\mathbf{d}_0)}{\alpha}+C(\eta,\alpha)\sum_k\sqrt{\epsilon_k}+\sum_k\epsilon_k\Big)+\varepsilon.$$

where we replaced we made explicit the fact the constant (wrt to K and T) C depends on α and η (See Theorem 1 for the exact expression). By plugging in the bound for ϵ_k given by Theorem 2, and a union bound, we get that and if we use $n(t)\ge\max\Big(\mathcal{O}\Big(\frac{\gamma^2mDt}{\beta}\log\frac{Tm}{\delta_2}\Big),\mathcal{O}\Big(\frac{mt}{\beta}\log\frac{Tm}{\delta_2}\Big)\Big)$ samples per iteration, then with probability at least $1-\delta_1-\delta_2$, it holds that

$$d_{\mathcal{C}}(\widehat{\pi},\pi_{\mathrm{E}})\le\frac{1}{K}\left(\frac{D(\boldsymbol{\lambda}^*||\boldsymbol{\Phi}^\intercal\mathbf{d}_0)}{\eta}+\frac{H(\mathbf{d}^*||\mathbf{d}_0)}{\alpha}+C(\eta,\alpha)\sum_k\mathcal{O}\left(\sqrt{\frac{\eta mD}{\beta\sqrt{T}}}\right)+\sum_k\mathcal{O}\left(\frac{\eta mD}{\beta\sqrt{T}}\right)\right)+\varepsilon.$$

Setting $\eta=\alpha=1$, letting $C_1\triangleq C(1,1)$ and keeping only the dominant terms, we obtain

$$d_{\mathcal{C}}(\widehat{\pi},\pi_{\mathrm{E}})\le\frac{D(\boldsymbol{\lambda}^*||\boldsymbol{\Phi}^\intercal\mathbf{d}_0)+H(\mathbf{d}^*||\mathbf{d}_0)}{K}+\mathcal{O}\left(C_1\frac{mD}{\beta\sqrt[4]{T}}\right)+\varepsilon$$

Then, choosing $K=\frac{D(\boldsymbol{\lambda}^*||\boldsymbol{\Phi}^\intercal\mathbf{d}_0)+H(\mathbf{d}^*||\mathbf{d}_0)}{\epsilon}$ and $T=\Omega\left(\frac{m^4D^4}{\beta^4C_1^4\epsilon^4}\right)$, we can ensure that $d_{\mathcal{C}}(\widehat{\pi},\pi_{\mathrm{E}})\le\epsilon+\varepsilon$. The overall sample complexity is $Kn(T)=\Omega(KT)=\Omega\left(\epsilon^{-5}\right)$. Notice that the corollary improves upon the sample complexity bound of $\Omega\left(\epsilon^{-8}\right)$ derived in [89]. □

J Offline imitation learning version

Inspecting Equation (5), one can notice that estimating the empirical logistic Bellman evaluation objective $\mathcal{G}_k$ or its gradients requires sampling from $\mathbf{d}_{k-1}$. Hence, the algorithm needs interactions with the environment at every iteration k. It is possible to alleviate this requirement, changing the center point for the relative entropy. This is akin to smoothing [82] choosing a convenient center point. In particular, we replace Equation (2) with the following update:

$$(\boldsymbol{\lambda}_1,\mathbf{d}_1)=\operatorname*{arg\,min}_{\boldsymbol{\lambda}\in\Delta_{[m]},\mathbf{d}\in\Delta_{\mathcal{S}\times\mathcal{A}}}\left\langle\mathbf{y}^\star,\mathbf{A}\begin{bmatrix}\boldsymbol{\lambda}\\\mathbf{d}\end{bmatrix}+\widehat{\mathbf{b}}\right\rangle+\frac{1}{\eta}D(\boldsymbol{\lambda}||\boldsymbol{\Phi}^\intercal\boldsymbol{\mu}_{\pi_{\mathrm{E}}})+\frac{1}{\alpha}H(\mathbf{d}||\mathbf{d}_0). \tag{66}$$

Algorithm 3 Offline Proximal Point Imitation Learning ($\texttt{OP}^2\texttt{IL}$)

Input: Feature matrix $\mathbf{\Phi}$, number of iterations K, step sizes η, α, and β
Input: Expert demonstrations $\mathcal{D}_{\mathrm{E}}^{n_{\mathrm{E}},H}$
∘ Initialize π_0 as uniform distribution over $\mathcal{A}$, and set $\mathbf{w}_0 = \frac{1}{m}\mathbf{1}$
∘ Compute the empirical FEV $\widehat{\boldsymbol{\rho}_{\mathbf{\Phi}}}(\pi_{\mathrm{E}})$ using expert demonstrations
∘ Sample $\{(s^{(n)}, a^{(n)}, s'^{(n)})\}_{n=1}^N$ with $s^{(n)}, a^{(n)}$ sampled i.i.d. from $\boldsymbol{\mu}_{\pi_E}$ and $s'^{(n)} \sim P(\cdot|s^{(n)}, a^{(n)})$ and compute the empirical offline logistic Bellman error by

$$\widehat{\mathcal{G}}(\mathbf{w}, \boldsymbol{\theta}) = -\left\langle \boldsymbol{\rho}_{\mathbf{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w} \right\rangle - \frac{1}{\eta}\log\left(\frac{1}{N}\sum_{n=1}^{N} e^{-\eta\widehat{\boldsymbol{\delta}}_{\mathbf{w},\boldsymbol{\theta}}(s^{(n)},a^{(n)},s'^{(n)})}\right) + (1-\gamma)\left\langle \boldsymbol{\nu}_0, V_{\boldsymbol{\theta}} \right\rangle$$

```
// policy evaluation & cost update
```

∘ Find an approximate maximizer of the negative empirical logistic Bellman error

$$(\mathbf{w}_1, \boldsymbol{\theta}_1) \approx \operatorname{argmax}_{\mathbf{w},\boldsymbol{\theta}} \widehat{\mathcal{G}}(\mathbf{w}, \boldsymbol{\theta})$$

```
// policy improvement
```

Policy update:

$$\pi_{\mathbf{d}_1}(a|s) \propto \pi_{\mathbf{d}_0}(a|s)\, e^{-\alpha Q_{\boldsymbol{\theta}_1}(s,a)}$$

Output: Policy $\pi_{\mathbf{d}_1}$

Note that we have removed the iteration index k, since the offline version does not require to iteratively collect new samples from the environment. Changing the reference distribution from $\mathbf{\Phi}^\intercal \mathbf{d}_k$ to $\mathbf{\Phi}^\intercal \boldsymbol{\mu}_{\pi_{\mathrm{E}}}$ gives Algorithm 3. In this case, the logistic Bellman evaluation objective takes the form

$$\mathcal{G}(\mathbf{w}, \boldsymbol{\theta}) \triangleq -\frac{1}{\eta}\log\sum_{i=1}^{m}(\mathbf{\Phi}^\intercal \boldsymbol{\mu}_{\pi_E})(i) e^{-\eta\delta_{\mathbf{w},\boldsymbol{\theta}}(i)} + (1-\gamma)\left\langle \boldsymbol{\nu}_0, \mathbf{V}_{\boldsymbol{\theta}} \right\rangle - \left\langle \boldsymbol{\rho}_{\mathbf{\Phi}}(\widehat{\pi_{\mathrm{E}}}), \mathbf{w} \right\rangle, \tag{67}$$

The difference with the online variant is that in the first term we have the expert occupancy measure instead of the occupancy measure induced by the current policy. We describe the corresponding empirical estimate in Algorithm 3. Furthermore, we suppress the index k, since the offline algorithm does not require multiple iterations.

J.1 Theoretical guarantees for the offline case

With minor modifications of the error propagation analysis given in Theorem 1, one can prove the following result.

Theorem 4. *Under the same assumptions as in Theorem 1, and by choosing* $\alpha = \left(\frac{2H(\mathbf{d}^\star||\mathbf{d}_0)}{3w_{\max}}\sqrt{\frac{1-\gamma}{2\epsilon}}\right)^{2/3}$*, we obtain*

$$d_{\mathcal{C}}(\pi_{\mathrm{E}}, \pi_{\mathbf{d}_1}) - d_{\mathcal{C}}(\pi_{\mathrm{E}}, \pi_{\mathbf{d}^\star}) \leq \frac{D(\boldsymbol{\lambda}^\star||\mathbf{\Phi}^\intercal \boldsymbol{\mu}_{\pi_{\mathrm{E}}})}{\eta} + \left(\frac{243H(\mathbf{d}^\star||\mathbf{d}_0)w_{\max}^2}{2(1-\gamma)}\right)^{1/3}\epsilon_1^{1/3} + \epsilon_1 + \varepsilon. \tag{68}$$

where ϵ_1 is the error in the maximization of the logistic Bellman error, i.e. $\epsilon = \max_{\mathbf{w}\in\mathcal{W},\boldsymbol{\theta}} \mathcal{G}(\mathbf{w}, \boldsymbol{\theta}) - \mathcal{G}(\mathbf{w}_1, \boldsymbol{\theta}_1)$ and ε is the error in estimating the expert feature expectation vector as in Lemma 6.

Proof. Following exactly the same steps in the proof of Theorem 1 for the special case of $K = 1$, we get

$$d_{\mathcal{C}}(\pi_E, \pi_{\mathbf{d}_1}) - d_{\mathcal{C}}(\pi_E, \pi_{\mathbf{d}^\star}) \leq \frac{D(\boldsymbol{\lambda}^\star||\mathbf{\Phi}^\intercal \boldsymbol{\mu}_{\pi_E})}{\eta} + \frac{H(\mathbf{d}^\star||\mathbf{d}_0)}{\alpha} + 3w_{\max}\left\|\mathbf{d}_1 - \mathbf{d}_1^\star\right\|_1 + \epsilon_1 + \varepsilon. \tag{69}$$

By using the bound $\left\|\mathbf{d}_1 - \mathbf{d}_1^\star\right\|_1 \leq \sqrt{\frac{2\alpha\epsilon_1}{1-\gamma}}$, we have

$$d_{\mathcal{C}}(\pi_E, \pi_{\mathbf{d}_1}) - d_{\mathcal{C}}(\pi_E, \pi_{\mathbf{d}^\star}) \leq \frac{D(\boldsymbol{\lambda}^\star||\mathbf{\Phi}^\intercal \boldsymbol{\mu}_{\pi_E})}{\eta} + \frac{H(\mathbf{d}^\star||\mathbf{d}_0)}{\alpha} + 3w_{\max}\sqrt{\frac{2\alpha\epsilon_1}{1-\gamma}} + \epsilon_1 + \varepsilon. \tag{70}$$

Therefore, by choosing α as stated in the theorem we conclude the proof. $\square$

Notice that if the expert is nearly optimal, the step size η can be taken small, ensuring low bias in the gradients. This allows to use the original empirical logistic Bellman error analysis, proposed in [14], where one can control the bias by choosing η appropriately small. To this end, we need to relate the logistic bellman error in the feature space to the one in the state-action space. As we will show, this introduces an additional bias of order $\mathcal{O}(\eta)$. The statement is made precise in Theorem 5. Thanks to this result and Theorem 2 in [14], we have that $\epsilon \leq (8+e)\eta B^2 + 56\sqrt{\frac{m\log(1+4BN)\delta}{N}}$ where N is the number of expert transitions in the dataset. We have the following result.

Corollary 10. *Let* $C_1 = \left(\frac{243H(\mathbf{d}^\star||\mathbf{d}_0)w_{\max}^2}{2(1-\gamma)}\right)^{1/3}$, $\eta = \mathcal{O}\left(\frac{D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal\boldsymbol{\mu}_{\pi_{\mathrm{E}}})^{3/4}}{(C_1 B)^{1/4}}\right)$ *and* $N = \tilde{\mathcal{O}}(m\epsilon^{-6}\log(1/\delta))$. *Then, with probability* $1-\delta$, *it holds that*

$$d_{\mathcal{C}}(\pi_{\mathrm{E}}, \pi_{\mathbf{d}_1}) - d_{\mathcal{C}}(\pi_{\mathrm{E}}, \pi_{\mathbf{d}^\star}) \leq \mathcal{O}\left(C_1^{1/4}B^{1/4}D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal\boldsymbol{\mu}_{\pi_{\mathrm{E}}})^{1/4}\right) + \mathcal{O}(\epsilon). \tag{71}$$

Remark 2. *We notice that the optimal choice of η is smaller as the expert is closely optimal, i.e. $D(\boldsymbol{\lambda}^\star||\boldsymbol{\Phi}^\intercal\boldsymbol{\mu}_{\pi_{\mathrm{E}}})$ is small. In this condition, we can use the empirical objective estimator proposed in[14] ensuring small bias. This means that estimating the objective from sample is feasible in the offline setting. It is still an open question if this is viable for the online setting improving the error propagation analysis.*

Next, we present an important result showing that it is possible to replace the minimization of $\mathcal{G}$, with its counterpart in the state-action space defined as

$$\mathcal{G}^{\mathcal{S},\mathcal{A}}(\boldsymbol{\theta},\mathbf{w}) = -\frac{1}{\eta}\log\sum_{s,a}\boldsymbol{\mu}_{\pi_E}(s,a)e^{-\eta\boldsymbol{\delta}^{\mathcal{S},\mathcal{A}}_{\mathbf{w},\boldsymbol{\theta}}(s,a)} + (1-\gamma)\left\langle\boldsymbol{\nu}_0,\mathbf{V}_{\boldsymbol{\theta}}\right\rangle - \left\langle\boldsymbol{\rho}_{\boldsymbol{\Phi}}(\widehat{\pi_{\mathrm{E}}}),\mathbf{w}\right\rangle,$$

where we introduced $\boldsymbol{\delta}^{\mathcal{S},\mathcal{A}}_{\mathbf{w},\boldsymbol{\theta}} = \boldsymbol{\Phi}\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}$.

Theorem 5. *Let* $B \triangleq 1 + 2\frac{1+|\log\beta|}{1-\gamma}$. *Suppose η is chosen such that $\eta B \leq 1$. Then, it holds that*

$$\left|\mathcal{G}(\boldsymbol{\theta},\mathbf{w}) - \mathcal{G}^{\mathcal{S},\mathcal{A}}(\boldsymbol{\theta},\mathbf{w})\right| \leq e\eta B^2.$$

Proof. From Proposition 3, we have that $\|\boldsymbol{\theta}\|_\infty \leq \frac{1+|\log\beta|}{1-\gamma}$ and $\|\mathbf{V}_{\boldsymbol{\theta}}\|_\infty \leq \frac{1+|\log\beta|}{1-\gamma}$, for all $\theta\in\mathbb{R}^m$. It follows that for any $(\mathbf{w},\boldsymbol{\theta}) \in \mathcal{W}\times\mathbb{R}^m$, it holds that $\|\boldsymbol{\delta}_{\boldsymbol{\theta},\mathbf{w}}\|_\infty = \|\mathbf{w}+\gamma\mathbf{M}\mathbf{V}_{\boldsymbol{\theta}}-\boldsymbol{\theta}\|_\infty \leq 1+2\frac{1+|\log\beta|}{1-\gamma} = B$. Hence, it holds that $\eta\|\boldsymbol{\delta}_{\boldsymbol{\theta},\mathbf{w}}\|_\infty \leq \eta B \leq 1$. First, we recall the assumption that the rows of $\boldsymbol{\Phi}$ are probability distributions, i.e., $\boldsymbol{\phi}(s,a)\in\Delta_{[m]}$, for all (s,a). We then have

$$\boldsymbol{\delta}^{\mathcal{S},\mathcal{A}}_{\mathbf{w},\boldsymbol{\theta}}(s,a) = (\boldsymbol{\Phi}\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}})(s,a) = \sum_{i=1}^m\boldsymbol{\phi}_i(s,a)\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}(i) = \mathop{\mathbb{E}}_{i\sim\boldsymbol{\phi}(s,a)}\left[\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}(i)\right]. \tag{72}$$

Moreover, we have

$$\mathcal{G}(\mathbf{w},\boldsymbol{\theta}) - \mathcal{G}^{\mathcal{S},\mathcal{A}}(\mathbf{w},\boldsymbol{\theta}) = -\underbrace{\frac{1}{\eta}\log\left(\sum_{i=1}^m(\boldsymbol{\Phi}^\intercal\boldsymbol{\mu}_{\pi_E})(i)e^{-\eta\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}(i)}\right)}_{\triangleq W} + \underbrace{\frac{1}{\eta}\log\left(\sum_{s,a}\boldsymbol{\mu}_{\pi_E}(s,a)e^{-\eta\boldsymbol{\delta}^{\mathcal{S},\mathcal{A}}_{\mathbf{w},\boldsymbol{\theta}}(s,a)}\right)}_{\triangleq W^{\mathcal{S},\mathcal{A}}}$$

We can then lower bound W as

$$\begin{aligned}
W &= \frac{1}{\eta}\log\left(\sum_{i=1}^m\sum_{s,a}\boldsymbol{\phi}_i(s,a)\boldsymbol{\mu}_{\pi_E}(s,a)e^{-\eta\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}(i)}\right) \\
&= \frac{1}{\eta}\log\left(\mathop{\mathbb{E}}_{(s,a)\sim\boldsymbol{\mu}_{\pi_E}}\left[\mathop{\mathbb{E}}_{i\sim\boldsymbol{\phi}(s,a)}\left[e^{-\eta\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}(i)}\right]\right]\right) \\
&\geq \frac{1}{\eta}\log\left(\mathop{\mathbb{E}}_{(s,a)\sim\boldsymbol{\mu}_{\pi_E}}\left[e^{-\eta\mathop{\mathbb{E}}_{i\sim\boldsymbol{\phi}(s,a)}[\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}(i)]}\right]\right), \\
&= W^{\mathcal{S},\mathcal{A}},
\end{aligned}$$

where the inequality follows by Jensen's inequality for expectations.

We will now upper bound the term W. Thanks to the choice of η such that $\eta B \leq 1$, we have that $\eta \leq \frac{1}{B} \leq \frac{1}{|\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}(i)|}$, for all i. Therefore, we can apply the inequality $e^x \leq 1 + x + x^2$ for $x = -\eta \boldsymbol{\delta}^{\mathcal{SA}}_{\mathbf{w},\boldsymbol{\theta}}(i) \leq 1$ and obtain

$$\begin{aligned}
\mathop{\mathbb{E}}_{i\sim\boldsymbol{\phi}(s,a)}\left[e^{-\eta\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}(i)}\right] &\leq \mathop{\mathbb{E}}_{i\sim\boldsymbol{\phi}(s,a)}\left[1-\eta\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}(i)+(\eta\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}(i))^2\right] \\
&\leq 1-\mathop{\mathbb{E}}_{i\sim\boldsymbol{\phi}(s,a)}\left[\eta\boldsymbol{\delta}_{\mathbf{w},\boldsymbol{\theta}}(i)\right]+(\eta B)^2 \\
&= 1-\eta\boldsymbol{\delta}^{\mathcal{SA}}_{\mathbf{w},\boldsymbol{\theta}}(s,a)+(\eta B)^2 \\
&\leq e^{-\eta\boldsymbol{\delta}^{\mathcal{SA}}_{\mathbf{w},\boldsymbol{\theta}}(s,a)}+(\eta B)^2,
\end{aligned}$$

where in the third line we used Equation (72), and in the last line we used the inequality $1 - x \leq e^{-x}$ for $x = \boldsymbol{\delta}^{\mathcal{SA}}_{\mathbf{w},\boldsymbol{\theta}}(s,a)$. By taking expectations with respect to $\boldsymbol{\mu}_{\pi_E}$ and logarithms on both sides, we get

$$W^{\mathcal{SA}} \leq W \leq \frac{1}{\eta}\log \mathop{\mathbb{E}}_{(s,a)\sim\boldsymbol{\mu}_{\pi_E}}\left[e^{-\eta\boldsymbol{\delta}^{\mathcal{SA}}_{\mathbf{w},\boldsymbol{\theta}}(s,a)}+(\eta B)^2\right].$$

Subtracting W yields

$$\begin{aligned}
0 \leq W - W^{\mathcal{SA}} &\leq \frac{1}{\eta}\log \mathop{\mathbb{E}}_{(s,a)\sim\boldsymbol{\mu}_{\pi_E}}\left[e^{-\eta\boldsymbol{\delta}^{\mathcal{SA}}_{\mathbf{w},\boldsymbol{\theta}}(s,a)}+(\eta B)^2\right] - W^{\mathcal{SA}} \\
&= \frac{1}{\eta}\log\left(1+\frac{(\eta B)^2}{\mathop{\mathbb{E}}_{(s,a)\sim\boldsymbol{\mu}_{\pi_E}}\left[e^{-\eta\boldsymbol{\delta}^{\mathcal{SA}}_{\mathbf{w},\boldsymbol{\theta}}(s,a)}\right]}\right) \\
&\leq \frac{\eta B^2}{\mathop{\mathbb{E}}_{(s,a)\sim\boldsymbol{\mu}_{\pi_E}}\left[e^{-\eta\boldsymbol{\delta}^{\mathcal{SA}}_{\mathbf{w},\boldsymbol{\theta}}(s,a)}\right]} \\
&\leq \frac{\eta B^2}{\mathop{\mathbb{E}}_{(s,a)\sim\boldsymbol{\mu}_{\pi_E}}\left[e^{-\eta B}\right]} \\
&\leq e\eta B^2,
\end{aligned}$$

where in the third line we used the inequality $\log(1+x) \leq x$ for $x = \frac{(\eta B)^2}{\mathop{\mathbb{E}}_{(s,a)\sim\boldsymbol{\mu}_{\pi_E}}\left[e^{-\eta\boldsymbol{\delta}^{\mathcal{SA}}_{\mathbf{w},\boldsymbol{\theta}}(s,a)}\right]}$, while in the last line we used that $\eta B \leq 1$. This concludes the proof. $\square$

After having established with Theorem 5 that $\mathcal{G}^{\mathcal{S},\mathcal{A}}$ can be used as biased estimate of $\mathcal{G}$, we can proceed as in [14]. In particular, we maximize the empirical objective $\widehat{\mathcal{G}}$ (see Algorithm 3) that is a biased estimate of $\mathcal{G}^{\mathcal{S},\mathcal{A}}$ ([14, Theorem 2]). Then, we compute unbiased gradients of $\widehat{\mathcal{G}}$, recurring to the Donsker-Varadhan formula [18, Corollary 4.15] that implies the following result.

Theorem 6. *Given a batch of expert data* $\{\widetilde{S}_n, \widetilde{A}_n, \widetilde{S}'_n\}_{n=1}^N \sim \boldsymbol{\mu}_{\pi_{\mathrm{E}}} \times \mathbf{P}$*, the following is true:*

$$\max_\theta \max_w \widehat{\mathcal{G}}(\theta, w) = \max_\theta \max_w \min_z \mathcal{S}(\theta, w, z) \tag{73}$$

with:

$$\begin{aligned}
\mathcal{S}(\theta,w,z) = &-\frac{1}{N}\sum_{n=1}^N \boldsymbol{\mu}_{\pi_{\mathrm{E}}}(\tilde{S}_n,\tilde{A}_n)\sum_{i=1}^m \mathbf{w}_i\boldsymbol{\phi}_i(\tilde{S}_n,\tilde{A}_n) &(74)\\
&+\frac{1}{N}\sum_{n=1}^N z(n)\left(\widehat{\delta}_{\mathbf{w},\boldsymbol{\theta}}(\widetilde{S}_n,\widetilde{A}_n,\widetilde{S}'_n)+\frac{1}{\eta}\log(Nz(n))\right) &(75)\\
&+(1-\gamma)\langle\boldsymbol{\nu}_0,\mathbf{V}_{\boldsymbol{\theta}}\rangle &(76)
\end{aligned}$$

and the minimum attained at $z^\star \propto \frac{1}{N}e^{-\eta\widehat{\delta}_{\mathbf{w},\boldsymbol{\theta}}(\widetilde{S}_n,\widetilde{A}_n,\widetilde{S}'_n)}$

Hence, in the deep learning implementation we update the cost and the value networks backpropagating through $\mathcal{S}(\boldsymbol{\theta}, \mathbf{w}, z^\star)$.

J.2 Practical implementation

We test a practical relaxation of Algorithm 3 that uses two separate neural networks for cost and value function approximation. We use a two layers neural network with 128 units per layer with `ReLu` activation for the `CartPole-v1` environment. Whereas, for `Acrobot-v1` and `LunarLander-v2` we used a 3 layers architecture with 64 units per layer.

K Mirror Descent versus Proximal Point

To highlight an important message of our work, in this section, we briefly discuss a mirror descent scheme with alternating updates, and we compare it to our proximal point algorithm in Figure 6. Note that in contrast to the classical RL setting, where proximal point and mirror descent coincide because of the linear objective, in imitation learning this is not the case.

The updates for the mirror descent scheme involve alternation between updating the occupancy measure $\mathbf{d}_k$ and the feature expectation vector $\boldsymbol{\lambda}_k$ in one stage and the cost weights in a second stage. That is,

$$(\boldsymbol{\lambda}_k, \mathbf{d}_k) = \underset{(\boldsymbol{\lambda},\mathbf{d})\in\mathcal{M}_{\boldsymbol{\Phi}}}{\arg\min} \langle \boldsymbol{\mu}, \mathbf{c}_{\mathbf{w}_k}\rangle + \tfrac{1}{\eta} D(\boldsymbol{\lambda}||\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1}) + \tfrac{1}{\alpha} H(\mathbf{d}||\mathbf{d}_{k-1}), \tag{77}$$

$$\mathbf{w}_{k+1} = \underset{\mathbf{w}\in\Delta_{[m]}}{\arg\min} \left\langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}} - \mathbf{d}_k, \mathbf{c}_{\mathbf{w}}\right\rangle + \tfrac{1}{\beta} D(\mathbf{w}||\mathbf{w}_k). \tag{78}$$

One can notice that the update in Equation (77) corresponds to one update of Logistic Q-Learning [14]. Therefore, it can be implemented by maximizing the negative logistic Bellman error that is now a function only of the variable $\boldsymbol{\theta}$ and not of both $(\boldsymbol{\theta}, \mathbf{w})$ as in PPM. The next proposition is the counterpart of Proposition 2 for the mirror descent scheme.

Proposition 5. *For a parameter $\boldsymbol{\theta} \in \mathbb{R}^m$, we define the state-action logistic value function $\mathbf{Q}_{\boldsymbol{\theta}} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ by $\mathbf{Q}_{\boldsymbol{\theta}} \triangleq \boldsymbol{\Phi}\boldsymbol{\theta}$, and the k-step state logistic value function $\mathbf{V}^k_{\boldsymbol{\theta}} \in \mathbb{R}^{|\mathcal{S}|}$ by*

$$V^k_{\boldsymbol{\theta}}(s) \triangleq -\frac{1}{\alpha} \log\left(\sum_a \pi_{\mathbf{d}_{k-1}}(a|s) e^{-\alpha Q_{\boldsymbol{\theta}}(s,a)}\right).$$

Moreover, for a fixed cost $\mathbf{c} = \mathbf{c}_{\mathbf{w}}$, we define the k-step Bellman error function $\boldsymbol{\delta}^k_{\boldsymbol{\theta},\mathbf{w}}$ by $\boldsymbol{\delta}^k_{\boldsymbol{\theta},\mathbf{w}} \triangleq \mathbf{w} + \gamma \mathbf{M}\mathbf{V}^k_{\boldsymbol{\theta}} - \boldsymbol{\theta}$. Then, the unique solution of the aforementioned problem is given by

$$\lambda_k(i) \propto (\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1})(i)\, e^{-\eta \delta^k_{\boldsymbol{\theta}_k,\mathbf{w}_k}(i)}, \tag{79}$$

$$\pi_{\mathbf{d}_k}(a|s) \propto \pi_{\mathbf{d}_{k-1}}(a|s)\, e^{-\alpha Q_{\boldsymbol{\theta}_k}(s,a)}, \tag{80}$$

$$w_{k+1,i} \propto w_{k,i}\, e^{-\beta \langle \boldsymbol{\phi}_i\,,\, \boldsymbol{\mu}_{\pi_{\mathrm{E}}} - \mathbf{d}_k\rangle}, \tag{81}$$

where $\boldsymbol{\theta}_k$ is the maximizer of the negative k-step logistic Bellman error function

$$\mathcal{G}_k(\boldsymbol{\theta}) \triangleq -\frac{1}{\eta} \log \sum_{i=1}^m (\boldsymbol{\Phi}^\intercal \mathbf{d}_{k-1})(i) e^{-\eta \delta^k_{\boldsymbol{\theta},\mathbf{w}_k}(i)} + (1-\gamma)\left\langle \boldsymbol{\nu}_0, \mathbf{V}^k_{\boldsymbol{\theta}}\right\rangle.$$

Proposition 5 leads to an actor critic scheme that has three separate and alternating updates: (i) policy update stage, (ii) policy evaluation update, and (iii) cost weights update. Similar actor critic-schemes for different MDP models, and different policy evaluation objectives (e.g., minimizing the squared Bellman error) have been also proposed in [122, 70, 105]. Contrary to these schemes, in our proximal imitation learning algorithm, the policy evaluation step involves optimization of a single objective over both cost and Q-functions. In this way, we avoid instability or poor convergence in optimization due to nested policy evaluation and cost update steps. In section L.5, we verify numerically that PPM outperforms Mirror Descent in simple tabular environments (see Figure 6).

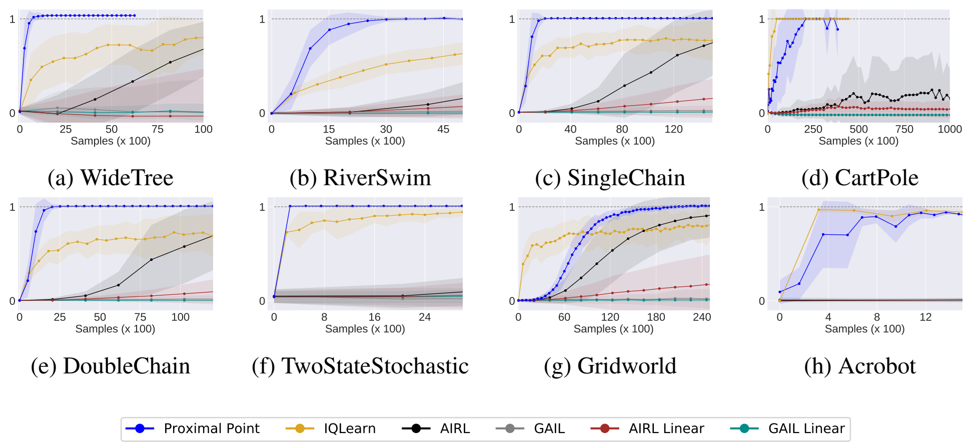


Figure 5: **Extended Online IL Experiments**. We show the total returns vs the number of env steps. We report the results of some environments omitted in the main text.

L Experimental Details

L.1 Refereences for environments description

In the tabular case we used the environments (`DoubleChain` [39], `SingleChain` [39], `RiverSwim` [108], `WideTree` [8], `Two States Deterministic` [9], `Two States Stochastic` [14] and `WindyGrid` [110]). While for the offline setting, we used the environments `CartPole` [12], `Acrobot` [42] and `LunarLander` [19]. The curves are averaged over 50 seeds. For the environments `Cartpole` and `Acrobot`, we used a three layer neural network to approximate the value function. In these cases we averaged 5 seeds.

L.2 Hyperparameters

We report the hyperparameters for the tabular online experiments in Table 1 and for the offline experiments in Table 2

Environment	`n-trajs`	`lr w`	`lr` $\boldsymbol{\theta}$	η	α	`optimizer`
TwoStateStochastic-v0	25	0.5	0.5	10	1	`FoRB`
TwoStateStochastic-v0	25	0.5	0.5	10	1	`Adam`
WideTree-v0	25	0.5	0.5	10	1	`FoRB`
RiverSwim-v0	50	0.2	0.2	10	1	`FoRB`
WindyGrid-v0	50	0.5	0.01	10	1	`FoRB`
SingleChainProblem-v0	50	0.3	0.005	10	1	`Adam`
DoubleChainProblem-v0	50	0.5	0.005	10	1	`Adam`

Table 1: Hyperparameters for proximal point imitation learning in tabular experiments. `FoRB` stands for Forward Reflected Backward [72].

L.3 On the data sampling

In all the experiments, we perform a relaxation of our theoretical scheme. In particular, to increase the sample efficiency we sample state action pairs from the Markovian stream of experience. Analyzing this setting is an open problem.

Environment	lr w	lr $\boldsymbol{\theta}$	η	α	optimizer
CartPole-v1	$5e-3$	$5e-3$	10	1	Adam
Acrobot-v1	$5e-3$	$5e-3$	10	1	Adam
LunarLander-v2	$1e-4$	$1e-4$	10	0.01	Adam

Table 2: Hyperparameters for offline experiments

Environment	n-trajs	lr w	lr $\boldsymbol{\theta}$	η	α
TwoStateStochastic-v0	25	0.5	0.5	10	1
TwoStateProblem-v0	25	0.5	0.5	10	1
WideTree-v0	25	0.5	0.5	10	1
RiverSwim-v0	25	0.5	0.01	10	1
WindyGrid-v0	50	0.5	0.0006	10	1
SingleChainProblem-v0	50	0.03	0.05	10	1
DoubleChainProblem-v0	50	0.03	0.025	10	1

Table 3: Hyperparameters for primal dual mirror descent imitation learning in tabular experiments. As optimizer, we used OGD in all cases.

L.4 Offline experiments setting

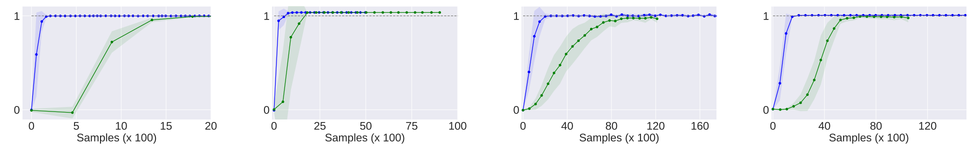
We consider a training environment and a test environment with different random seeds. We train both IQLearn and Proximal Point for $2e5$ environment steps and we evaluate the policy running 10 episodes on the evaluation environment every $1e3$ steps. We report the maximum evaluation result achieved at the end of training. We average the seeds from 0 to 10 for the results shown in Figure 2. We use two separate instances of the same architecture as function approximation for the Q-values and cost respectively. Finally, since the algorithm operates offline it has no access to the distribution $\boldsymbol{\nu}_0$. In order to approximate the term $\langle \boldsymbol{\nu}_0, \mathbf{V} \rangle$, we use the Bellman flow constraints and the fact that the expert occupancy measure is feasible, i.e. $(1-\gamma)\langle \boldsymbol{\nu}_0, \mathbf{V} \rangle = \left\langle \boldsymbol{\mu}_{\pi_E}, -\gamma \mathbf{P}\mathbf{V} + \mathbf{B}\mathbf{V} \right\rangle$ where the last term can be estimated from the expert samples.

L.5 Comparison with mirror descent

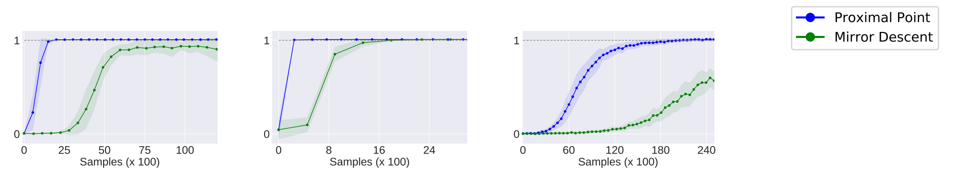
We designed also a mirror descent scheme with alternating updates for imitation learning, briefly described in Appendix K. The best hyperparameters are given in Table 3. Furthermore, we show a comparison with our proximal point scheme in Figure 6. It is interesting to notice that mirror descent and proximal point have been used interchangeably in the RL literature. Indeed, in that case the objective is linear therefore the two algorithms coincide. However, when considering the max-form objective in imitation learning the equivalence between mirror descent and proximal point does not longer hold true. We verify numerically that PPM outperforms mirror descent in simple tabular environments (see Figure 6).

L.6 Hyperparameters for Pong (Atari)

We use a convolutional neural network to learn the Q values instead of the linear function approximation class we considered in the theoretical analysis. We set the parameter α to $1e-3$ and η to $8e-2$, we used expert samples to approximate expectation with respect to the initial distribution. For optimizing the network we used Adam [61] with learning rate $1e-4$ and defaults value for β_1, β_2 Instead of hard constraints on the euclidean norm of the elements of $\mathcal{W}$ we consider a ℓ_2 penalty to the loss function. As expert trajectories we used the dataset released by [40]. This is the only hyperparameters configuration we tried using a single seed (using the seed 0) on our method because of the high computation requirements of this environment.



(a) Two State Deterministic (b) WideTree (c) RiverSwim (d) SingleChain



(e) DoubleChain (f) Two State Stochastic (g) Gridworld (h) Legend

Figure 6: **Proximal Point vs Mirror Descent.** Comparison of proximal point and mirror descent in tabular domains. Averages of 10 seeds.

L.7 Hyperparameters for MuJoCo (continuous control)

The policy network outputs a distribution over continuous action and is parametrized by independent gaussian distributions for every component of the continuous action vector. We use a three layer neural network to estimate their means and variances. We used as center point in the divergence D the expert feature expectation vector. With further modifications our method can extend also to continuous control tasks in MuJoCo [113]. The main challenge is that the policy improvement step can not be computed in closed form. We therefore approximate it with a SAC architecture as proposed in [40]. We set α to $1e-3$, η to $8e-2$, the SAC actor learning rate to $3e-5$ using Adam as optimizer using default values of β_1, β_2,for the critic we used again Adam with learning rate $3e-4$ and default values for β_1, β_2. The actor training of SAC is performed using a transition buffer containing expert and learner data in equal proportion. We used samples from the expert policy to estimate expectations wrt the initial distribution. We avoid using target networks. We tested our algorithm on both the environment `Ant` and `HalfCheetah` using either the data provided in [40] or fresh expert data that we generated training experts with PPO [101]. The results are averaged across 5 seeds. For `Hopper`, we used a larger SAC actor learning rate equal to $2e-4$ and $\alpha = 1e-2$. In addition, we notice that for this environment having a large β_1 in Adam was harmful. Hence, we used $\beta_1 = 0$.

For `Walker`, we set the actor learning to $1e-4$.

L.8 Acknowledging existing assets and license.

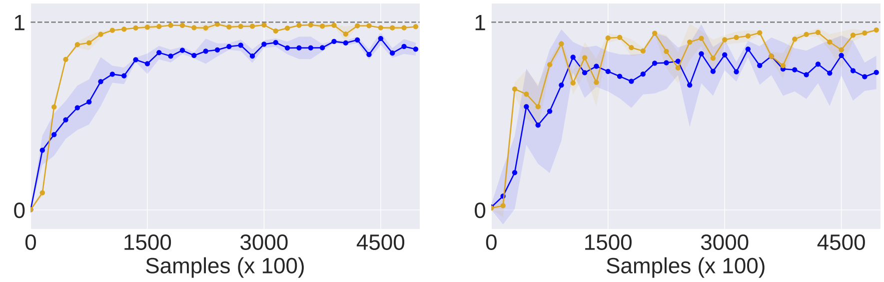
We built on the code and expert data provided in [40]. They are open sourced for academic scope according to their GitHub page `https://github.com/Div99/IQ-Learn/blob/main/LICENSE.md`.

L.9 On the importance of the dataset

We observed that the performance of our imitation learning algorithm and IQ-Learn can be affected by the choice of the expert data. In particular, in Appendix L.9, we show that IQ-Learn works better with the expert data provided in [40].

L.10 Hardware

We ran the experiments on our internal cluster.



(a) HalfCheetah-v2

(b) Ant-v2

Figure 7: Experiments in the MuJoCo environments with the expert data provided by [40]. The blue line is proximal point while the yellow line is IQLearn.

M Recovered Costs

A unique algorithmic feature of the proposed methodology is that we can explicitly recover a cost along with the Q-function without requiring adversarial training. In Figures 8 and 3, we visualize our recovered costs in several simple tabular environments (`River Swim`, `Single Chain`, `Double Chain`, and `Gridworld`, respectively). Most importantly, we verify that the recovered costs induce nearly optimal policies w.r.t. the unknown true cost function. Compared to IQ-Learn, the we do not require knowledge or further interaction with the environment. Therefore, the recovered cost functions show promising transfer capability to new dynamics.

We experimented with a transfer reward setting on a `Gridworld` (Figure 4). We consider two different Gridworld MDP environments, say M and $\widetilde{M}$, with opposite action effects. This means that action `Down` in $\widetilde{M}$ corresponds to action `Left` in M and vice versa. Similarly, the effects of `Up` and `Right` are swapped between $\widetilde{M}$ and M. We denote by $\mathbf{V}^{\pi}_{\widetilde{M},\mathbf{c}_{\text{true}}}$ (resp. $\mathbf{V}^{\star}_{\widetilde{M},\mathbf{c}_{\text{true}}}$) the value function of policy π (resp. optimal value function) in the MDP environment $\widetilde{M}$ with cost function $\mathbf{c}_{\text{true}}$. Moreover, we denote by $\pi^{\star}_{M,\mathbf{c}}$ the optimal policy in the MDP environment M under cost function $\mathbf{c}$. We notice that the recovered cost induces an optimal policy for the new dynamics while the imitating policy fails. Albeit, cost transfer is successful in this experiment we do not expect this fact to be true in general because we do not tackle the issue of cost shaping [87].

M.1 Preliminary theoretical arguments

We have some preliminary theoretical arguments justifying the near optimality of the recovered costs/rewards. We present briefly the reasoning.

For brevity, we consider the case $\mathcal{W} = B_1^m$. Then π_{E} is optimal for the IL problem. Moreover, for simplicity, we consider the case $\mathbf{\Phi} = \mathbf{I}$. Otherwise, in the following derivations, we replace $\mathbf{Q}$-values by parameterized $\mathbf{Q}_{\boldsymbol{\theta}}$.

Let $(\widehat{\mathbf{w}}_K, \widehat{\mathbf{Q}}_K)$ be the output (average iterate) of `P`^{2}`IL` after K outer loop iterations. We give a sketch of proof that $\widehat{\mathbf{w}}_K$ converges to an optimal solution to the inverse problem as $K \to \infty$, i.e., $\widehat{\mathbf{w}}_K$ converges to some $\mathbf{w}_{\text{A}} \in \mathcal{W}$ such that π_{E} is optimal for $\mathbf{c}_{\mathbf{w}_{\text{A}}}$. To this end, we first introduce the following definition.

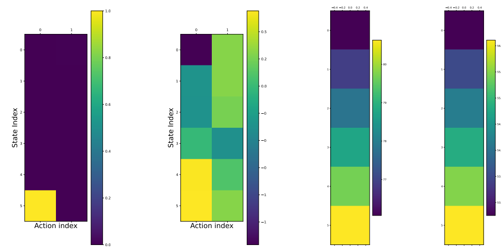
Definition 1. *We say that $\mathbf{w} \in \mathcal{W}$ is ε_1-optimal and ε_2-feasible for the (Dual) program if-f there exists $\mathbf{V} \in \mathbb{R}^{|\mathcal{S}|}$, such that*

$$\begin{aligned} \langle \boldsymbol{\mu}_{\pi_{\text{E}}}, \mathbf{c}_{\mathbf{w}} \rangle - (1-\gamma)\langle \boldsymbol{\nu}_0, \mathbf{V} \rangle &\leq \varepsilon_1, & (82)\\ \mathbf{c}_{\mathbf{w}} - (\mathbf{B} - \gamma \mathbf{P})\mathbf{V} &\geq -\varepsilon_2 \mathbf{1}. & (83) \end{aligned}$$

In this case, $\mathbf{V} \in \mathbb{R}^{|\mathcal{S}|}$ is called a certificate.

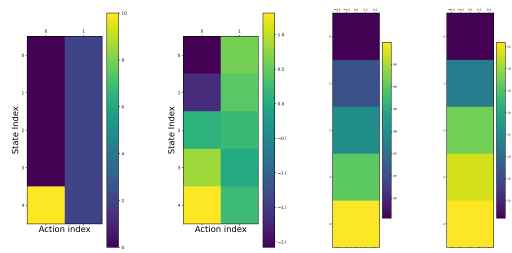
Note that the definition of ε_1-optimality for the (Dual) program follows from the fact that the dual optimal value is $\zeta^\star = 0$. Moreover, in the definition of ε_2-feasibility we have relaxed the nonnegativity constraint in the dual program (Dual). We make the following conjecture.

River Swim



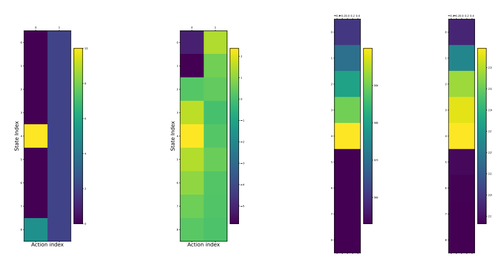
(a) $-\mathbf{c}_{\text{true}}$ (b) $-\mathbf{c}_K$ (c) - $V^\star_{\mathbf{c}_{\text{true}}}$ (d) -$V^\star_{\mathbf{c}_K}$

Single Chain



(e) $-\mathbf{c}_{\text{true}}$ (f) $-\mathbf{c}_K$ (g) - $V^\star_{\mathbf{c}_{\text{true}}}$ (h) -$V^\star_{\mathbf{c}_K}$

Double Chain



(i) $-\mathbf{c}_{\text{true}}$ (j) $-\mathbf{c}_K$ (k) - $V^\star_{\mathbf{c}_{\text{true}}}$ (l) -$V^\star_{\mathbf{c}_K}$

Figure 8: **Recovered Costs.** Comparison between the true cost $\mathbf{c}_{\text{true}}$ and the cost $\mathbf{c}_K$ recovered by $\texttt{P}^2\texttt{IL}$. We notice that the optimal value functions $V^\star_{\mathbf{c}_{\text{true}}}$ and $V^\star_{\mathbf{c}_K}$ present the same pattern. Hence, the optimal policy with respect to $\mathbf{c}_K$ is nearly optimal with respect to $\mathbf{c}_{\text{true}}$.

Conjecture: For a sufficiently large number of samples $N = \mathcal{O}\Big(\text{poly}\big(\frac{1}{\varepsilon}, \log(\frac{1}{\delta}), m\big)\Big)$, with probability at least $1-\delta$, the output cost weight $\widehat{\mathbf{w}}_K$ is ε-optimal and ε-feasible for the (Dual) program, with certificate the corresponding logistic value function $\mathbf{V}_{\widehat{\mathbf{Q}}_K}$.

This is easy to show for the exact PPM updates, since $(\mathbf{d}_{\widehat{\pi}_K}, \mathbf{d}_{\widehat{\pi}_K}, \widehat{\mathbf{w}}_K, \mathbf{V}_{\widehat{\mathbf{Q}}_K}, \widehat{\mathbf{Q}}_K)$ is a saddle-point of the (SPP). The proof needs much more effort for the inexact updates used in the sampling-based algorithm.

Lemma 18. *Assume that $\widetilde{\mathbf{w}}$ is ε_1-optimal and ε_2-feasible for the (Dual) program. Then, π_{E} is $(\varepsilon_1+\varepsilon_2)$-optimal for $\mathbf{c}_{\widetilde{\mathbf{w}}}$.*

Proof. There exists $\widetilde{\mathbf{V}} \in \mathbb{R}^{|\mathcal{S}|}$, such that

$$\begin{aligned} \langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\widetilde{\mathbf{w}}} \rangle - (1-\gamma)\langle \boldsymbol{\nu}_0, \widetilde{\mathbf{V}} \rangle &\leq \varepsilon_1, & (84)\\ \mathbf{c}_{\widetilde{\mathbf{w}}} - (\mathbf{B} - \gamma \mathbf{P})\widetilde{\mathbf{V}} &\geq -\varepsilon_2 \mathbf{1}. & (85) \end{aligned}$$

Let $\widetilde{\pi}$ be an optimal policy for $\mathbf{c}_{\widetilde{\mathbf{w}}}$. Then, we have that

$$\left\langle \boldsymbol{\mu}_{\widetilde{\pi}}, \mathbf{c}_{\widetilde{\mathbf{w}}} - (\mathbf{B} - \gamma \mathbf{P})\widetilde{\mathbf{V}} \right\rangle \geq -\varepsilon_2 \langle \boldsymbol{\mu}_{\widetilde{\pi}}, \mathbf{1} \rangle = -\varepsilon_2.$$

By using that $(\mathbf{B} - \gamma \mathbf{P})^{\mathsf{T}} \boldsymbol{\mu}_{\widetilde{\pi}} = (1-\gamma)\boldsymbol{\nu}_0$, we equivalently that

$$\langle \boldsymbol{\mu}_{\widetilde{\pi}}, \mathbf{c}_{\widetilde{\mathbf{w}}} \rangle - (1-\gamma)\langle \boldsymbol{\nu}_0, \widetilde{\mathbf{V}} \rangle \geq -\varepsilon_2.$$

Therefore,

$$\langle \boldsymbol{\mu}_{\mathrm{E}}, \mathbf{c}_{\widetilde{\mathbf{w}}} \rangle \leq (1-\gamma)\langle \boldsymbol{\nu}_0, \widetilde{\mathbf{V}} \rangle + \varepsilon_1 \leq \langle \boldsymbol{\mu}_{\widetilde{\pi}}, \mathbf{c}_{\widetilde{\mathbf{w}}} \rangle + \varepsilon_1 + \varepsilon_2.$$

Thus, π_{E} is $(\varepsilon_1 + \varepsilon_2)$-optimal for $\mathbf{c}_{\widetilde{\mathbf{w}}}$. □

Claim: As $K \to \infty$ one may approach as closely as desired an optimal solution to the inverse problem.

Proof for the ideal PPM updates. We recall that by Proposition 4, the set of such solutions is characterized as the set of $\mathbf{w}$-optimizers to (Dual).

Let $\widehat{\mathbf{V}}_K = \mathbf{V}_{\widehat{\mathbf{Q}}_K}$. By the conjecture, for all K, we have

$$\begin{aligned} \langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\widehat{\mathbf{w}}_K} \rangle - (1-\gamma)\langle \boldsymbol{\nu}_0, \widehat{\mathbf{V}}_K \rangle &\leq \varepsilon_K, & (86)\\ \mathbf{c}_{\widehat{\mathbf{w}}_K} - (\mathbf{B} - \gamma \mathbf{P})\widehat{\mathbf{V}}_K &\geq -\varepsilon_K \mathbf{1}, & (87) \end{aligned}$$

for some sequence $\{\varepsilon_K\}_{K=1}^{\infty}$ such that $\lim_{K\to\infty} \varepsilon_K = 0$. The sequence $\{\widehat{\mathbf{w}}_K\}_{K=1}^{\infty} \subset \mathcal{W}$ is bounded and so there exists a subsequence $\{\widehat{\mathbf{w}}_{K_l}\}_{l=1}^{\infty}$, such that $\lim_{l\to\infty} \widehat{\mathbf{w}}_{K_l} = \mathbf{w}_{\mathrm{A}}$, for some $\mathbf{w}_{\mathrm{A}} \in \mathcal{W}$. Similarly, by Proposition 3 the sequence $\{\widehat{\mathbf{V}}_{K_l}\}_{l=1}^{\infty}$ is bounded and so there exists a subsequence $\{\widehat{\mathbf{V}}_{K_{l_n}}\}_{n=1}^{\infty}$, such that $\lim_{n\to\infty} \widehat{\mathbf{V}}_{K_{l_n}} = \mathbf{V}_{\mathrm{A}}$, for some $\mathbf{V}_{\mathrm{A}}$. By Equations(86)–(87), we have that for all $n \in \mathbb{N}$,

$$\begin{aligned} \langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\widehat{\mathbf{w}}_{K_{l_n}}} \rangle - (1-\gamma)\langle \boldsymbol{\nu}_0, \widehat{\mathbf{V}}_{K_{l_n}} \rangle &\leq \varepsilon_{K_{l_n}}, & (88)\\ \mathbf{c}_{\widehat{\mathbf{w}}_{K_{l_n}}} - (\mathbf{B} - \gamma \mathbf{P})\widehat{\mathbf{V}}_{K_{l_n}} &\geq -\varepsilon_{K_{l_n}} \mathbf{1}. & (89) \end{aligned}$$

Taking $n \to \infty$, we end up that

$$\begin{aligned} \langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}_{\mathrm{A}}} \rangle - (1-\gamma)\langle \boldsymbol{\nu}_0, \mathbf{V}_{\mathrm{A}} \rangle &\leq 0, & (90)\\ \mathbf{c}_{\mathbf{w}_{\mathrm{A}}} - (\mathbf{B} - \gamma \mathbf{P})\mathbf{V}_{\mathrm{A}} &\geq 0. & (91) \end{aligned}$$

Equivalently,

$$\begin{aligned} \langle \boldsymbol{\mu}_{\pi_{\mathrm{E}}}, \mathbf{c}_{\mathbf{w}_{\mathrm{A}}} \rangle - (1-\gamma)\langle \boldsymbol{\nu}_0, \mathbf{V}_{\mathrm{A}} \rangle &= 0, & (92)\\ \mathbf{c}_{\mathbf{w}_{\mathrm{A}}} - (\mathbf{B} - \gamma \mathbf{P})\mathbf{V}_{\mathrm{A}} &\geq 0. & (93) \end{aligned}$$

Therefore, by Proposition 4, π_{E} is optimal for $\mathbf{c}_{\mathbf{w}_{\mathrm{A}}}$.

□