Learning without Smoothness and Strong Convexity

THÈSE N^{O} 8765 (2018)

PRÉSENTÉE LE 13 JUILLET 2018

À LA FACULTÉ DES SCIENCES ET TECHNIQUES DE L'INGÉNIEUR
LABORATOIRE DE SYSTÈMES D'INFORMATION ET D'INFÉRENCE
PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Yen-Huan LI

acceptée sur proposition du jury:

Prof. P. Thiran, président du jury
Prof. V. Cevher, directeur de thèse
Prof. J. Bolte, rapporteur
Prof. P. Ravikumar, rapporteur
Prof. M. Jaggi, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2018

Acknowledgements

First, I would like to thank my advisor, Volkan Cevher. He gave me the freedom to explore diverse topics in machine learning and data science, and helped me in almost every aspect throughout my PhD. Volkan shaped an interdisciplinary lab that solves interesting research problems in machine learning, signal processing, applied mathematics, theoretical computer science, etc.; I enjoyed this interdisciplinary style very much, and have benefited a lot from it. I am honored to have Jérôme Bolte, Martin Jaggi, Pradeep Ravikumar, and Patrick Thiran as members of my thesis defense committee. I am grateful for their time and valuable discussions.

I have learned a lot from my colleagues in EPFL. Quoc Tran-Dinh patiently answered every question of mine in convex optimization; he is the main person who helped me step in the field of convex optimization. Ilija Bogunovic introduced to me the fantastic Online Learning Summer School in Copenhagen in 2015, during which I realized the deepness of machine learning theory; this results in a rapid change of my research interests in the last three years of my PhD. Ya-Ping Hsieh and I had frequent *very long* research discussions; these discussions helped me clarify my ideas and introduced to me many interesting research problems. Baran Gözcü and Carlos Riofrío taught me practical details about magnetic resonance imaging and quantum state tomography, respectively. Jonathan Scarlett is a very nice person, and productive and sharp in research; he is basically my model of a perfect researcher. Junhong Lin pointed out to me important statistical learning literature. Paul Rolland helped me work out the French abstract of this thesis.

I would also like to thank Pradeep for hosting my visit to Carnegie Mellon University in this spring, during which I learned a lot about zeorth-order optimization. I would like to thank Felix Krahmer for inviting me to have a short visit to Technische Universität München, and taught me the idea of the small-ball method.

I am grateful to Volkan, Chen-Mou Cheng, Po-Sen Huang, Anastasios Kyrillidis, Chia-Han Lee, Lin-shan Lee, Kamalaruban Parameswaran, Jon, I-Hsiang Wang, and Ping-Cheng Yeh, for their advices regarding my academic career after graduation.

Finally, I would like to thank my family, and all friends I met in Taiwan, Switzerland, and conferences.

Lausanne, 30 June 2018 YHL

Abstract

Recent advances in statistical learning and convex optimization have inspired many successful practices. Standard theories assume smoothness—bounded gradient, Hessian, etc.—and strong convexity of the loss function. Unfortunately, such conditions may not hold in important real-world applications, and sometimes, to fulfill the conditions incurs unnecessary performance degradation. Below are three examples.

1. The standard theory for variable selection via ℓ_1-penalization only considers the linear regression model, as the corresponding quadratic loss function has a constant Hessian and allows for exact second-order Taylor series expansion. In practice, however, nonlinear regression models are often chosen to match data characteristics.

2. The standard theory for convex optimization considers almost exclusively smooth functions. Important applications such as portfolio selection and quantum state estimation, however, correspond to loss functions that violate the smoothness assumption; existing convergence guarantees for optimization algorithms hence do not apply.

3. The standard theory for compressive magnetic resonance imaging (MRI) guarantees the *restricted isometry property (RIP)*—a smoothness and strong convexity condition on the quadratic loss restricted on the set of sparse vectors—via random uniform sampling. The random uniform sampling strategy, however, yields unsatisfactory signal reconstruction performance empirically, in comparison to heuristic sampling approaches.

In this thesis, we provide rigorous solutions to the three examples above and other related problems. For the first two problems above, our key idea is to instead consider weaker *localized* versions of the smoothness condition. For the third, our solution is to propose a new theoretical framework for compressive MRI: We pose compressive MRI as a statistical learning problem, and solve it by empirical risk minimization. Interestingly, the RIP is not required in this framework.

Keywords: Smoothness, strong convexity, statistical learning, convex optimization, variable selection, lasso, quantum state estimation, mirror descent, compressive MRI

Résumé

Les récents progrès en apprentissage statistique et en optimisation convexe ont inspiré de nombreux pratiques avec succès. Les théories standards suppose une certaine régularité—gradient borné, Hessienne bornée, etc.—et la forte convexité de la fonction de coût. Malheureusement, ces conditions ne sont parfois pas satisfaites dans d'importantes applications réelles, et parfois, remplir ces conditions implique une perte de performance non nécessaire. Ci-dessous nous présentons trois exemples :

1. La théorie standard pour la selection de variables via une pénalisation l1 ne considère que le modèle de régression linéaire, car la fonction de coût quadratique associée possède un matrice Hessienne constante, ce qui permet une expansion de Taylor exacte au second ordre. En pratique, cependant, les modèles de régression non-linéaires sont souvent choisis pour correspondre aux charactéristiques des données.

2. La théorie standard d'optimisation convexe ne considère presque exclusivement que des fonctions régulières. Cependant, d'importantes applications telles que la sélection de portfolio ou l'estimation d'état quantique, correspondent à des fonctions de coût qui violent cette supposition de régularité ; les garanties de convergences existantes des algorithmes d'optimisation ne s'appliquent donc pas à ces cas.

3. La théorie standard de l'imagerie par résonance magnétique (IRM) compressive garantit la propriété d'isométrie restreinte (RIP)—une condition de régularité et de convexité forte sur le coût quadratique restreint sur l'ensemble de vecteurs parcimonieux— par échantillonnage uniforme aléatoire. Cependant, il a été observé que la stratégie d'échantillonnage aléatoire uniforme, cependant produit une performance de récupération d'image insatisfaisante, en comparaison avec des approches d'échantillonnage heuristiques.

Dans cette thèse, nous fournissons des solutions rigoureuses aux trois exemples ci-dessus et d'autres problèmes connexes. Pour les deux premiers problèmes ci-dessus, notre idée clé est de considérer plutôt des versions localisées plus faibles de la condition de lissage. Pour le troisième, notre solution est de proposer un nouveau cadre théorique pour l'IRM compressive : Nous posons l'IRM compressive comme un problème d'apprentissage statistique, et le résolvons par minimisation du risque empirique. Fait intéressant, le RIP n'est pas requis dans ce cadre.

Mots clés : Régularité, forte convexité, apprentissage statistique, optimisation convexe, sélection de variables, lasso, estimation d'état quantique, descente miroir, IRM compressive

Contents

List of Figures

List of Tables

1 Introduction

The interplay between statistical learning and convex optimization has been critical in developing efficient learning algorithms. For example, a standard approach to binary classification is empirical risk minimization (ERM) with the $0-1$ loss, which outputs 1 if the classification fails, and 0 otherwise. The corresponding statistical performance is quite well-studied—the expected loss is inversely proportional to the square root of the sample size and cannot be improved (see, e.g., [163, Chapter 6]). To compute the learned classification rule, however, requires solving a non-convex optimization problem, NP-hard in general [94]. Therefore, the use of convex surrogate functions to approximate the $0-1$ loss was considered, resulting in famous algorithms [190, 14]: logistic regression, adaptive boosting (AdaBoost), and the support vector machine.

Another example is compressive sensing. Suppose that we would like to estimate a vector $\beta^* \in \mathbb{R}^p$ given a matrix $X \in \mathbb{R}^{n\times p}$ and $y := X\beta^* \in \mathbb{R}^n$, where $n < p$. The linear equation is under-determined; a solution is to introduce the sparsity assumption: The vector β^* has s exactly zero entries, and s is significantly smaller than the dimension p. A natural approach is to find the sparsest vector $\tilde{\beta} \in \mathbb{R}^p$ that satisfies $y = X\tilde{\beta}$. This approach actually succeeds (i.e., $\tilde{\beta} = \beta$ holds) if $n = \Omega(s)$, under technical conditions; however, computing $\tilde{\beta}$ is NP-hard in general [133]. Convexifying the formulation for $\tilde{\beta}$, we arrive at the *basis pursuit* estimator, which outputs the vector $\hat{\beta}_{\mathrm{BP}} \in \mathbb{R}^p$ with the minimal ℓ_1-norm satisfying $y = X\hat{\beta}_{\mathrm{BP}}$. Given a technical condition called the *restricted isometry property (RIP)*, the basis pursuit estimator succeeds if $n = \tilde{\Omega}(s)$ ignoring logarithmic dependence on p [45, 89].

This thesis presents results in statistical learning, convex optimization, and their interplay, for machine learning problems without *smoothness* and *strong convexity*.

1.1 Importance of smoothness and strong convexity

In both fields of statistical learning and convex optimization, smoothness and strong convexity play important roles. Before continuing discussion, let us recall the definitions.

Definition 1.1 (Smoothness). *A real-valued function f is said to be k-smooth on a set $\mathscr{X}$ for a natural number k, if its $(k-1)$-th order derivative is Lipschitz on $\mathscr{X}$. The 0-th order derivative is defined as the function itself.*

Definition 1.2 (Strong convexity). *A real-valued function f is said to be strongly convex on a set $\mathscr{X}$, if there is some $\mu > 0$, such that*

$$(1-\alpha)f(x)+\alpha f(y) \geq f((1-\alpha)x+\alpha y)+\alpha(1-\alpha)\frac{\mu}{2}\|y-x\|_2^2, \quad \forall x, y \in \mathscr{X} \text{ and } \alpha \in [0,1]. \tag{1.1}$$

Notice that the terminology "smoothness" is defined differently in statistical learning and convex optimization. The definition in this thesis is closer to notion of a Hölder class in non-parametric statistics (see, e.g., [175]); we choose this definition simply for convenience. The *smoothness condition* in convex optimization—the gradient being Lipschitz continuous—corresponds to 2-smoothness here.

The importance of smoothness and strong convexity is well-known in convex optimization, as is illustrated by the following classical results.

- Convergence of the mirror descent can be established given 1-smoothness of the loss function [135, 19][1].
- Convergence of the gradient descent can be established given 2-smoothness of the loss function [136].
- Faster convergence rates can be guaranteed for the two cases above, if in addition strong convexity holds [136, 90, 100].

Notice that without smoothness, even for a convex function f, it is NP-hard to decide whether there exists some point y such that $f(y) < f(x)$ given a point x [137]. Existing minimax results show that without strong convexity, the convergence rates of mirror descent and gradient descent are provably slower [135, 136].

Arguably, the importance of smoothness and strong convexity is less obvious in statistical learning, as the formulations are typically more complicated and sometimes hidden in the technical derivations. Below are three examples.

- The RIP for compressive sensing is indeed a 2-smoothness and strong convexity condition for the loss function

$$f(\beta) := \frac{1}{2}\|y - A\beta\|_2^2, \quad \forall \beta \in \mathbb{R}^d, \tag{1.2}$$

[1] To be precise, the convergence results were proved assuming boundedness of the subgradients, but the non-smooth case is not the focus of this thesis.

restricted on the set of vectors whose ℓ_0-norms are smaller than a given integer, where $y \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times p}$ are given. We provide a formal proof in Chapter 7.

- In general, for possibly non-linear high-dimensional ($n \ll p$) regression problems, estimation consistency can be established given the *restricted strong convexity (RSC)* condition [25, 50, 177, 134]. The condition requires strong convexity of the loss function (1.2), restricted on a cone or cone-like set.

- The *irrepresentability condition* is a sufficient condition for the ℓ_1-penalized least squares estimator to achieve successful *variable selection*—identifying the positions of non-zero elements of the true weight vector—in the linear regression model [185, 191]. As we will see in Chapter 2, the sufficiency relies on the fact that the second-order derivative of the empirical risk given in (1.2) is a constant—a strong 3-smoothness condition.

The requirement for strong convexity is easier to understand: If the empirical risk function is a constant, the empirical risk minimizer can be arbitrarily far from the true weight vector. Strong convexity prevents such an undesirable situation, though this condition may be weakened. The necessity of smoothness is subtle, but oftentimes it enables us to derive more satisfactory results. For example, without smoothness, a convex optimization problem can still be solved by proximal point-type methods [65, 157]; however, in each iteration, these methods require solving an optimization sub-problem that is not essentially easier than the original one, rendering their applications limited in practice.

1.2 Challenges due to lack of smoothness and/or strong convexity

Although smoothness and strong convexity play important roles in statistical learning and convex optimization, logically speaking, they may not be necessary. Indeed, requiring or forcing them creates huge gaps between practice and theory. This thesis is mainly devoted to addressing the gaps in the following applications.

1.2.1 Variable selection consistency of ℓ_1-penalized estimators

As mentioned in the previous section, the ℓ_1-penalized least squares estimator achieves successful variable selection in the linear regression model, given the irrepresentability condition. In general, however, non-quadratic loss functions are often used to match the possibly non-linear statistical models [127]. Regarding the existing result for the linear regression model, a natural approach is to introduce an ℓ_1-penalty term in the corresponding maximum-likelihood (ML) estimators. While such an approach may provide satisfactory empirical performances, the existing variable selection consistency result for linear regression does not apply, as its proof is specific to the fact that the third-order derivative of the quadratic loss is exactly zero.

Challenge 1. *Can one develop a unified theory for the variable selection consistency of ℓ_1-penalized estimators in possibly non-linear statistical models?*

1.2.2 Non-asymptotic analysis of the (constrained) lasso

As mentioned in the previous section, under the high-dimensional setting, estimation consistency of the ℓ_1-penalized least squares estimator can be established given the RSC condition. A closely related formulation is the ℓ_1-constrained least squares estimator, called the least absolute shrinkage and selection operator (lasso) [170]. Unfortunately, one can easily verify that the RSC condition cannot hold for the lasso, except when the ℓ_1-norm constraint is exact, i.e., when we know the exact ℓ_1-norm of the true weight vector—typically impossible in practice. Therefore, the estimation error guarantee of the ℓ_1-penalized least squares estimator does not apply.

Challenge 2. *Can one establish a non-asymptotic estimation error guarantee for the lasso under the high-dimensional setting?*

1.2.3 Rigorous and fast exp-linear minimization

A loss function f is called exp-linear, if it takes the form $f(x) = -\log\langle a, x\rangle$ for some vector a, or $f(X) = -\log\mathrm{Tr}(AX)$ for some matrix A. Such a loss appears in growth-optimal portfolio selection, positron emission tomography, quantum state estimation (QSE), positive linear inverse problems, Poisson phase retrieval, etc., where it is asked to minimize sums of exp-linear losses on the probability simplex or spectahedron [30, 180, 96, 34, 143]. It is easily verified that an exp-linear loss is non-smooth on the probability simplex or spectahedron, as $\langle a, x\rangle$ (or $\mathrm{Tr}(AX)$ in the matrix case) can be arbitrarily close to zero. Therefore, standard convergence results for the mirror descent and gradient descent do not directly apply, while the gradient descent has been adopted in some QSE researches [28, 164].

Challenge 3. *Can one develop a fast algorithm for exp-linear minimization that has a convergence guarantee?*

1.2.4 Design of a compressive MRI system

Compressive MRI is one of the most important applications of compressive sensing, where one would like to recover an unknown image given a subset of its Fourier coefficients [124]. The standard theory suggests that the subset should be chosen randomly following the uniform distribution, as then the RIP holds with high probability [45]. In practice, however, uniform random sampling yields obviously worse performance in comparison to heuristic non-uniform sampling strategies, as observed in [124]. While the superiority of non-uniform sampling can be demonstrated in theory, if we introduce more delicate assumptions on the unknown image, it is unclear whether the assumptions always hold or not [1].

Challenge 4. *Can one develop an algorithm to find a "good" sampling strategy for compressive MRI, with a guarantee on the recovery performance and without any unverifiable assumption on the image?*

1.3 Contributions

In this thesis, we provide rigorous solutions to the four challenges mentioned in the preceding section. For the first three challenges, we develop theories that work with weaker formulations of smoothness and strong convexity; for the last challenge, we eliminate the need for the RIP.

- In Chapter 2, we show that a novel *local structured smoothness condition*, together with a general formulation of the irrepresentability condition, suffice to guarantee the variable selection consistency for possibly non-quadratic loss in general statistical models. We provide a unified framework to establish the variable selection consistency of lasso-type methods. We derive novel sharp sample complexity bounds for several applications.
- In Chapter 3, we derive sharp non-asymptotic estimation error bounds for the lasso, showing that the lasso is minimax optimal when the true weight vector is exactly sparse, and when the true weight vector is weakly sparse if its exact ℓ_1-norm is accessible.
- In Chapter 4, we prove that the Frank-Wolfe algorithm indeed converges for exp-linear minimization, with a slightly modified step size selection rule; moreover, the $O(1/k)$ convergence rate (see, e.g, [98]) for the standard Frank-Wolfe algorithm also holds.
- In Chapter 5, we prove that the mirror descent with Armijo line search is always guaranteed to converge, for a large class of functions that satisfy a novel *locally relatively smoothness* condition. With this convergence result, we demonstrate that the exponentiated gradient method (a.k.a. entropic mirror descent) with Armijo line search is the fastest guaranteed-to-converge algorithm for QST, empirically on real data-sets.
- In Chapter 6, we study the convergence of the exponentiated gradient method with Armijo line search, under the very weak assumption that the loss function is convex and differentiable. We prove that, as long as the set of iterates has a strictly positive limit point, the exponentiated gradient method with Armijo line search is always guaranteed to converge. A byproduct is an improved Peierls-Bogoliubov inequality based on self-concordant likeness.
- In Chapter 7, we develop a completely new framework for compressive MRI: We pose compressive MRI as a statistical learning problem, and find a good sampling strategy via ERM. We derive a rigorous bound on the generalization error, without any assumption (e.g., sparsity) on the image to be recovered. Training and image recovery can be done in almost linear time. The empirical performance is comparable to existing computationally much more expensive methods. Interestingly, the necessity of the RIP vanishes in our framework.

1.4 Notation

Let $v \in \mathbb{R}^p$ and $M \in \mathbb{R}^{p_1 \times p_2}$. We denote the transposes of v and M by $v^\top$ and $M^\top$, respectively. If M is invertible, its inverse is denoted by M^{-1}.

We will frequently deal with sub-vectors and sub-matrices. Let $\mathscr{E}$ be a subset of $\{1,\dots,p\}$. We denote by $|\mathscr{E}|$ the cardinality of $\mathscr{E}$. We define $\mathscr{E}^{\mathrm{c}} := \{1,\dots,p\} \setminus \mathscr{E}$. We denote by $v_{\mathscr{E}}$ the sub-vector of v, consisting of elements of v indexed by $\mathscr{E}$. Let $\mathscr{E}_1$ be a subset of $\{1,\dots,p_1\}$, and $\mathscr{E}_2$ be a subset of $\{1,\dots,p_2\}$. We denote by $M_{\mathscr{E}_1,\mathscr{E}_2}$ the sub-matrix of M, with rows indexed by $\mathscr{E}_1$ and columns indexed by $\mathscr{E}_2$. When we only want to pick a few columns while keep all rows, we write $M_{\mathscr{E}_2}$ for $M_{\{1,\dots,p\},\mathscr{E}_2}$ to simplify the notation. Also to simplify the notation, we write v_i and $M_{i,j}$ for the i-th element of v and (i,j)-th element of M, respectively.

We write $\|v\|_q$ for the ℓ_q-norm of v for $q \in [0,+\infty]$, i.e.,

$$\|v\|_q := \begin{cases} \left(\sum_{i=1}^p |v_i|^q\right)^{1/q} & q > 0, \\ |\{i : v_i \neq 0\}| & q = 0, \\ \max_i \{|v_i| \mid 1 \le i \le p\} & q = \infty. \end{cases}$$

The unit ℓ_q-norm ball is denoted by $\mathscr{B}_q$.

We write $\|M\|_q$ for the operator norm of M induced by the ℓ_q-norm; in particular, $\|M\|_2$ corresponds to the spectral norm of M, and

$$\|M\|_\infty = \max_i \left\{ \sum_{j=1}^{p_2} |M_{i,j}| \;\middle|\; 1 \le i \le p_1 \right\}.$$

We write $\|M\|_*$ for the nuclear norm of M, which corresponds to the sum of singular values of M. Let A, B be two matrices. The expression $A \succeq B$ means that the matrix $(A-B)$ is positive semi-definite, and $A \succ B$ means that the matrix $(A-B)$ is positive definite.

The notation $\operatorname{supp} v$ denotes the set of indices for which the corresponding element of v is non-zero; that is,

$$\operatorname{supp} v := \{i \mid v_i \neq 0\}.$$

The notation $\operatorname{sign} v$ denotes the vector $(\operatorname{sign} v_1, \dots, \operatorname{sign} v_p)^\top$, where $\operatorname{sign} v_i := v_i/|v_i|$ if $v_i \neq 0$, and $\operatorname{sign} v_i = 0$ otherwise.

We will consider the first-, second-, and third-order derivatives of a function f. The k-th order derivative of f at at a point x is denoted by $D^k f[x]$, which is a k-linear symmetric form. It suffices to keep in mind the following facts.

- For any $v \in \mathbb{R}^p$, we have $Df[x](v) = \langle \nabla f(x), v \rangle$, where ∇f denotes the gradient of f.
- For any $u, v \in \mathbb{R}^p$, we have $D^2 f[x](u,v) = \langle u, \nabla^2 f(x) v \rangle$, where $\nabla^2 f$ denotes the Hessian

of f.

- For any $u \in \mathbb{R}^p$, we have

$$D^3 f(x)[u] = \lim_{t \to 0} \frac{\nabla^2 f(x + tu) - \nabla^2 f(x)}{t} .$$

Moreover, for any $v, w \in \mathbb{R}^p$, we have

$$D^3 f(x)[u, v, w] = \langle v, \left(D^3 f(x)[u]\right) w \rangle .$$

The 1-linear form (vector) $D^3 f(x)[u, v]$ is then defined as the unique vector satisfying

$$D^3 f(x)[u, v, w] = \langle D^3 f(x)[u, v], w \rangle , \quad \forall w \in \mathbb{R}^p .$$

We write $\mathsf{P}\mathscr{E}$ for the probability that the event $\mathscr{E}$ happens. For example, $\mathsf{P}\{x \leq 1\}$ denotes the probability that the value of the random variable x is smaller or equal to 1. We write $\mathsf{E}\, x$ for the expectation of the random variable x.

This thesis contains results in statistical learning and convex optimization, respectively. To respect the convention in the two fields, in the first two chapters about statistical learning, we use β^* to denote the unknown true parameter to be learned, while in Chapter 4–6, we use $f^\star$ to denote the minimum value of the function f to be minimized.

2 Variable selection consistency of ℓ_1-penalized M-estimators

The first two chapters consider statistical learning in the *high-dimensional* setting, where the dimension of the unknown parameter can be much larger than and scale with the sample size. The high-dimensional setting allows one to use more flexible statistical models (with higher parameter dimension) with more data. In this setting, there are three main topics.

1. Prediction: Does a statistical estimator guarantee small expected loss?
2. Estimation: Does a statistical estimator recover the unknown, true parameter?
3. Variable selection: Does a statistical estimator identify relevant coordinates of the unknown, true parameter?

The three topics have been arguably well-studied for the ℓ_1-penalized least squares estimator in the linear regression model—the prediciton, estimation, and variable selection consistency has been established, and the corresponding sample compelxity bounds are known to be sharp (see, e.g., [32, 80, 87, 105]). The prediction issue was addressed with great generality in, e.g., [15, 83, 129]. Existing results for estimation and variable selection, however, do not apply if the setup is slightly modified.

In this chapter, we develop a general framework to establish the variable selection consistency of ℓ_1-penalized M-estimators, in possibly non-linear statistical models.

This chapter is based on the joint work with Jonathan Scarlett, Pradeep Ravikumar, and Volkan Cevher [122].

2.1 Introduction

Consider the linear regression model

$$y_i = \langle x_i, \beta^* \rangle + w_i, \quad i = 1, 2, \ldots, n,$$

for given $\{x_i\} \subset \mathbb{R}^p$ and some unknown *weight vector* $\beta^* \in \mathbb{R}^p$, where w_i are independent and identically distributed (i.i.d.) σ^2-subgaussian random variables (r.v.'s) for some $\sigma > 0$. Suppose that β^* is sparse, i.e.,

$$s := |\operatorname{supp} \beta^*| \ll p\,.$$

The task of *variable selection* asks one to identify $\operatorname{supp} \beta^*$, given the data $\{(x_1, y_1), \ldots, (x_n, y_n)\}$.

The ℓ_1-penalized least squares estimator is given by

$$\hat{\beta}_n \in \operatorname*{arg\,min}_{\beta} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \langle x_i, \beta \rangle\right)^2 + \tau_n \|\beta\|_1 \;\middle|\; \beta \in \mathbb{R}^p \right\}, \tag{2.1}$$

for some penalization coefficient $\tau_n > 0$. It is known that under standard assumptions, we have

$$\lim_{n\to\infty} \mathsf{P}\left\{ \operatorname{supp} \hat{\beta}_n \neq \operatorname{supp} \beta^* \right\} = 0\,,$$

as long as $n \gg s \log p$ [185]. Therefore, we say that the ℓ_1-penalized least squares estimator is *consistent in variable selection* under the high-dimensional setting.

The key assumption that enables the variable selection consistency result is the *irrepresentability condition*.

Definition 2.1 (Irrepresentability condition). *Define $X \in \mathbb{R}^{n\times p}$ as the matrix whose i-th row is given by $x_i^\top$, $\mathcal{S} := \operatorname{supp} \beta^*$, and $\mathcal{S}^{\mathrm{c}} := \{1, 2, \ldots, p\} \setminus \mathcal{S}$. We say that the irrepresentability condition holds for some $\alpha \in (0, 1)$, if*

$$\left\| X_{\mathcal{S}^{\mathrm{c}}}^\top X_{\mathcal{S}} \left(X_{\mathcal{S}}^\top X_{\mathcal{S}} \right)^{-1} \right\|_\infty < 1 - \alpha\,. \tag{2.2}$$

where $X_{\mathcal{S}}$ and $X_{\mathcal{S}^{\mathrm{c}}}$ denotes the sub-matrix consisting of columns indexed by $\mathcal{S}$ and $\mathcal{S}^{\mathrm{c}}$, respectively. The matrix norm $\|\cdot\|_\infty$ is defined as the largest ℓ_1-norm of the rows.

The irrepresentability condition is not only sufficient but almost necessary—if the left-hand side of (2.2) is strictly larger than one, variable selection consistency cannot hold [185].

Now, consider the general ℓ_1-penalized M-estimator

$$\hat{\theta}_n \in \operatorname*{arg\,min}_{\beta} \left\{ L_n(\beta) + \tau_n \|\beta\|_1 \;\middle|\; \beta \in \mathbb{R}^p \right\}, \tag{2.3}$$

of an unknown parameter β^* in a possibly non-linear statistical model, for some convex loss function L_n. Notice that Definition 2.1 may not be meaningful under this general setting, as the statistical model is not necessarily of the regression type. The aim of this chapter is to answer the question:

Under what conditions can $\hat{\beta}_n$ be consistent in variable selection?

2.1.1 Applications of variable selection

Suppose that the data is given as a set of pairs $\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R}^p \times \mathbb{R}$, and the probability distribution of y_i given x_i is solely determined by $\langle x_i, \beta^* \rangle$, for some unknown $\beta^* \in \mathbb{R}^p$—this is actually the case in most regression models (see, e.g., [127, 180]). Furthermore, assume that

$$s := |\operatorname{supp} \beta^*| \ll p .$$

Then it suffices to only keep $\{((x_1)_{\mathscr{S}}, y_1), \ldots, ((x_n)_{\mathscr{S}}, y_n)\} \subset \mathbb{R}^s \times \mathbb{R}$ and discard the elements of x_i's indexed by $\mathscr{S}^{\mathrm{c}}$. If s is much smaller than p, doing so significantly reduces the cost in storage memory, and accelerates any further data processing tasks. Notice that doing so does not incur any loss of information, as $(x_i)_{\mathscr{S}}$ suffices to determine the probability distribution of y_i for every i.

Another application is Gaussian graphical model selection. Let $x \in \mathbb{R}^p$ be a r.v. following the multivariate Gaussian distribution of zero mean and covariance $\Sigma^* \in \mathbb{R}^{p \times p}$. Suppose that the data is a set of n i.i.d. random vectors $x_1, \ldots, x_n$, following the same distribution as x. The task of graphical model selection asks one to identify the positions of the non-zero elements of the *concentration matrix* $\Theta^* := (\Sigma^*)^{-1}$. Given these positions, a graph consisting of p-nodes, where the nodes i and j are connected if and only if $\Theta_{i,j} \neq 0$, reveals the conditional independence relation among the elements in x—if two nodes are not connected, the corresponding two elements are conditionally independent, given all other elements [113]. Notice that since the statistical model is not of the regression type, the definition of the irrepresentability condition (Definition 2.1) does not apply.

2.1.2 Related work

For the specific case of sparse linear regression, the ℓ_1-penalized least squares estimator has received considerable attention. With respect to variable selection consistency, results have been obtained for both the noiseless case (e.g., [37, 66, 67]) and the noisy case [128, 185, 191]. While variable selection consistency results have been obtained for ℓ_1-penalized M-estimators on some *specific* non-linear models such as logistic regression and Gaussian graphical model selection [8, 33, 112, 128, 152, 153], *general* techniques with broad applicability are largely lacking.

To the best of our knowledge, the first work to study the variable selection consistency of a broad class of models was that of [72] for generalized linear models; however, the technical assumptions therein appear to be difficult to check for specific models, thus making their application difficult. Another related work is [114]; in Section 2.6, we compare it with our work, and discuss a key advantage of our approach.

2.1.3 Contributions

In this chapter, we introduce a novel condition called the *local structured smoothness condition* (LSSC) (Definition 2.2), which controls the smoothness of the objective function in a particular structured set. We illustrate how the LSSC enables us to address a broad set of variable selection results in a unified fashion, including logistic regression, gamma regression, and graphical model selection. We explicitly check the LSSC for ℓ_1-penalized maximum likelihood (ML) estimation in these statistical models. We then establish the variable selection consistencies of these ℓ_1-penalized ML estimators, and derive sample complexity bounds. To the best of our knowledge, the sample complexity bounds are currently the sharpest, and our results for gamma regression and Gaussian graphical model selection without a degree bound are the first in literature.

2.2 Local structured smoothness condition

The following definition provides the key property of convex functions that will be exploited in the subsequent variable selection consistency analysis.

Definition 2.2 (Local Structured Smoothness Condition (LSSC)). *Let $f \in C^3(\operatorname{dom} f)$ for some open $\operatorname{dom} f \subseteq \mathbb{R}^p$. Fix $x^* \in \operatorname{dom} f$, and let $\mathcal{N}_{x^*}$ be an open set in $\operatorname{dom} f$ containing x^*. The function f satisfies the $(x^*, \mathcal{N}_{x^*})$-LSSC with parameter $K \geq 0$, if*

$$\|D^3 f(x^* + \delta)[u,u]\|_\infty \leq K\|u\|_2^2,$$

for all $\delta \in \mathbb{R}^p$ such that $x^* + \delta \in \mathcal{N}_{x^*}$, and for all $u \in \mathbb{R}^p$ such that $u_{\mathcal{S}^c} = 0$, where $\mathcal{S} := \operatorname{supp} x^*$.

Note that $D^3 f(x^* + \delta)[u,u]$ is a 1-linear form, so $\|\cdot\|_\infty$ in Definition 2.2 is the vector ℓ_∞-norm. The following equivalent characterization follows immediately.

Proposition 2.3. *The function f satisfies the $(x^*, \mathcal{N}_{x^*})$-LSSC with parameter $K \geq 0$ if and only if*

$$|D^3 f(x^* + \delta)[u,u,e_j]| \leq K\|u\|_2^2, \tag{2.4}$$

for all $\delta \in \mathbb{R}^p$ such that $x^* + \delta \in \mathcal{N}_{x^*}$, for all $u \in \mathbb{R}^p$ such that $u_{\mathcal{S}^c} = 0$, where $\mathcal{S} := \operatorname{supp} x^*$, and for all $j \in \{1, \ldots, p\}$, where e_j is the standard basis vector with 1 in the j-th position and 0s elsewhere.

As we will see in the next section, this equivalent characterization is useful when verifying the LSSC for a given M-estimator.

Since differentiation is a linear operator, the LSSC is preserved under linear combinations with positive coefficients, as is stated formally in the following lemma.

Lemma 2.4. *Let f_1 satisfy the $(x, \mathcal{N}_1)$-LSSC with parameter K_1, and f_2 satisfy the $(x, \mathcal{N}_2)$-LSSC with parameter K_2. Let α and β be two positive real numbers. The function $f := \alpha f_1 + \beta f_2$ satisfies the $(x, \mathcal{N}_x)$-LSSC with parameter K, where $\mathcal{N}_x := \mathcal{N}_1 \cap \mathcal{N}_2$, and $K := \alpha K_1 + \beta K_2$.*

We conclude this section by briefly discussing the connection of the LSSC with other conditions. The following result, Proposition 9.1.1 of [140], will be useful here and throughout the chapter.

Proposition 2.5. *Let A be a 3-linear symmetric form on $(\mathbb{R}^p)^3$, and B be a positive-semidefinite 2-linear symmetric form on $(\mathbb{R}^p)^2$. If*

$$|A[u,u,u]| \leq B[u,u]^{3/2}$$

for all $u \in \mathbb{R}^p$, then

$$|A[u,v,w]| \leq B[u,u]^{1/2} B[v,v]^{1/2} B[w,w]^{1/2}$$

for all $u, v, w \in \mathbb{R}^p$.

This proposition shows that the condition in (2.4) *without structural constraints on u and e_j* is equivalent to the statement that

$$|D^3 f(x^* + \delta)[u,v,w]| \leq K \|u\|_2 \|v\|_2 \|w\|_2 \tag{2.5}$$

for all $u, v, w \in \mathbb{R}^p$. In Section 2.A, we show that (2.5) holds for all $\delta \in \mathbb{R}^p$ such that $x^* + \delta \in \mathcal{N}_{x^*}$ if and only if

$$\|D^2 f(x^* + \delta) - D^2 f(x^*)\|_2 \leq K \|\delta\|_2, \tag{2.6}$$

for all $\delta \in \mathbb{R}^p$ such that $x^* + \delta \in \mathcal{N}_{x^*}$. The latter condition is simply the local Lipschitz continuity of the Hessian of f. This is why we consider our condition a *local structured smoothness* condition, with structural constraints on the inputs of the $D^3 f(x^* + \delta)$ operator.

The preceding observations reveal that (2.5), or the equivalent formulation (2.6), is more restrictive than the LSSC. That is, (2.5) implies the LSSC, while the reverse is not true in general.

2.3 Examples

In this section, we provide some examples of functions that satisfy the LSSC.

Example 2.6. *Suppose that $f(\beta) := \|y - X\beta\|_2^2$ for some fixed $y \in \mathbb{R}^p$ and $X \in \mathbb{R}^{n \times p}$. Since $D^3 f(\beta) \equiv 0$ everywhere, the function f satisfies the $(\beta^*, \mathcal{N}_{\beta^*})$-LSSC with parameter $K = 0$ for any $\beta^* \in \mathbb{R}^p$ and any open set $\mathcal{N}_{\beta^*} \subseteq \mathbb{R}^p$ that contains β^*. This function appears in the linear regression model.*

Example 2.7. *Let $f(\beta) := \langle x, \beta\rangle - \log\langle x, \beta\rangle$ for some fixed $x \in \mathbb{R}^p$. We show that, for any fixed $\beta^* \in \operatorname{dom} f$ such that $\beta^*_{\mathscr{S}^c} = 0$, there exists some non-negative K and some open set $\mathscr{N}_{\beta^*}$ such that f satisfies the $(\beta^*, \mathscr{N}_{\beta^*})$-LSSC with parameter K. This function appears as the negative log-likelihood in gamma regression with the canonical link function.*

By a direct differentiation, we obtain for all $u \in \mathbb{R}^p$ that

$$|D^3 f(\beta^* + \delta)[u,u,u]| = 2|1+\gamma|^{-3}\left\{D^2 f(\beta^*)[u,u]\right\}^{3/2}, \tag{2.7}$$

where

$$\gamma := \frac{\langle x, \delta\rangle}{\langle x, \beta^*\rangle},$$

Combining this with Proposition 2.5, we have for each standard basis vector e_j that

$$\begin{aligned}|D^3 f(\beta^* + \delta)[u,u,e_j]| &\le 2|1+\gamma|^{-3} D^2 f(\beta^*)[u,u]\left\{D^2 f(\beta^*)[e_j,e_j]\right\}^{1/2} \\ &\le 2\left(1-|\gamma|\right)^{-3} D^2 f(\beta^*)[u,u]\left\{D^2 f(\beta^*)[e_j,e_j]\right\}^{1/2},\end{aligned}$$

if $|\gamma| \le 1$. Now define $\mathscr{S} := \operatorname{supp}\beta^*$, and suppose that $u_{\mathscr{S}^c} = \delta_{\mathscr{S}^c} = 0$, and that

$$\|\delta\|_2 \le \frac{\langle x, \beta^*\rangle}{(1+\kappa)\|x_{\mathscr{S}}\|_2}$$

for some $\kappa > 0$. By the Cauchy-Schwartz inequality, it immediately follows that $|\gamma| \le (1+\kappa)^{-1} < 1$, and thus $\beta^* + \delta \in \operatorname{dom} f$. Moreover, using this bound on $|\gamma|$, we can further upper bound $|D^3 f|$ as

$$|D^3 f(\beta^* + \delta)[u,u,e_j]| \le 2\left(1+\kappa^{-1}\right)^3 \lambda_{\max} d_{\max}^{1/2} \|u\|_2^2,$$

where $\lambda_{\max}$ is the maximum restricted eigenvalue of $D^2 f(\beta^*)$ defined as

$$\lambda_{\max} := \sup_u \left\{ D^2 f(\beta^*)[u,u] \;\middle|\; \|u\|_2 \le 1, u_{\mathscr{S}^c} = 0 \right\},$$

and $d_{\max}$ denotes the maximum diagonal entry of $\nabla^2 f(\beta^*)$. Therefore, f satisfies the $(\beta^*, \mathscr{N}_{\beta^*})$-LSSC with parameter $K := 2(1+\kappa^{-1})^3 \lambda_{\max} d_{\max}^{1/2}$, where

$$\mathscr{N}_{\beta^*} := \left\{ \beta^* + \delta : \|\delta\|_2 \le \frac{\langle x, \beta^*\rangle}{(1+\kappa)\|x_{\mathscr{S}}\|_2}, \delta \in \mathbb{R}^p \right\}.$$

Example 2.8. *Consider the function $f(\Theta) = \operatorname{Tr} X\Theta - \log\det\Theta$ with a fixed $X \in \mathbb{R}^{p\times p}$, and with $\operatorname{dom} f := \{\Theta \in \mathbb{R}^{p\times p} : \Theta > 0\}$. We show that, for any fixed $\Theta^* \in \operatorname{dom} f$, there exists some non-negative K and some open set $\mathscr{N}_{\Theta^*}$ such that f satisfies the $(\Theta^*, \mathscr{N}_{\Theta^*})$-LSSC with parameter K. This function appears as the negative log-likelihood in the Gaussian graphical learning problem.*

Note that the previous definitions (in particular, Definition 2.2), should be interpreted here as

being taken with respect to the vectorizations of the relevant matrices.

It is already known that f is standard self-concordant [136]; that is,

$$|D^3 f(\Theta^* + \Delta)[U,U,U]| \le 2\left\{D^2 f(\Theta^* + \Delta)[U,U]\right\}^{3/2},$$

for all $U \in \mathbb{R}^{p\times p}$ and all $\Delta \in \mathbb{R}^{p\times p}$ such that $\Theta^* + \Delta \in \operatorname{dom} f$. This implies, by Proposition 2.5,

$$|D^3 f(\Theta^* + \Delta)[U,U,V]| \le 2\left\{D^2 f(\Theta^* + \Delta)[U,U]\right\}\left\{D^2 f(\Theta^* + \Delta)[V,V]\right\}^{1/2},$$

for all $U, V \in \mathbb{R}^{p\times p}$, and all $\Delta \in \mathbb{R}^{p\times p}$ such that $\Theta^* + \Delta \in \operatorname{dom} f$.

Moreover, by a direct differentiation,

$$\|D^2 f(\Theta^* + \Delta)\|_2 = \|(\Theta^* + \Delta)^{-1} \otimes (\Theta^* + \Delta)^{-1}\|_2 = \|\left(\Theta^* + \Delta\right)^{-1}\|_2^2.$$

Fix a positive constant κ, and suppose that we choose Δ such that $\|\Delta\|_F \le (1+\kappa)^{-1}\rho_{\min}$, where $\rho_{\min}$ denotes the smallest eigenvalue of Θ^*. Since $\|\Delta\|_2 \le \|\Delta\|_F$, it follows that $\|\Delta\|_2 \le (1+\kappa)^{-1}\rho_{\min}$, and, by Weyl's theorem [95],

$$\|\left(\Theta^* + \Delta\right)^{-1}\|_2 \ge \frac{\kappa}{1+\kappa}\rho_{\min}.$$

Combining the preceding observations, it follows that f satisfies the $(\Theta^*, \mathcal{N}_{\Theta^*})$-LSSC with parameter $K := 2\kappa^{-3}(1+\kappa)^3\rho_{\min}^{-3}$, where

$$\mathcal{N}_{\Theta^*} = \left\{\Theta^* + \Delta : \|\Delta\|_F < \frac{1}{1+\kappa}\rho_{\min}, \Delta = \Delta^\top, \Delta \in \mathbb{R}^{p\times p}\right\}.$$

Here we have not exploited the special structure of U in Definition 2.2 (namely, $u_{\mathscr{S}^c} = 0$), though conceivably the constant K could improve by doing so. Note that $\mathcal{N}_{\Theta^*} \subset \operatorname{dom} f$ and $\mathcal{N}_{\Theta^*}$ is convex.

2.4 Sufficient conditions

We are now in a position to state the main result in this chapter, whose proof can be found in Section 2.B.

Let $\beta^* \in \mathbb{R}^p$ be the true parameter, and define $\mathscr{S} := \operatorname{supp}\beta^*$. Define the "genie-aided" estimator with access to $\operatorname{supp}\beta^*$:

$$\check{\beta}_n \in \arg\min_{\beta}\left\{L_n(\beta) + \tau_n\|\beta\|_1 \;\middle|\; \beta \in \mathbb{R}^p, \beta_{\mathscr{S}^c} = 0\right\}. \tag{2.8}$$

Theorem 2.9. *Suppose that $\check{\beta}_n$ is uniquely defined. Then the ℓ_1-penalized estimator $\hat{\beta}_n$ defined in (2.3) uniquely exists, successfully recovers the sign pattern, i.e.,* $\operatorname{sign}\hat{\beta}_n = \operatorname{sign}\beta^*$*, and satisfies*

the error bound

$$\left\|\hat{\beta}_n - \beta^*\right\|_2 \leq r_n := \frac{\alpha+4}{\lambda_{\min}}\sqrt{s}\tau_n, \tag{2.9}$$

if the following conditions hold true.

1. (Local structured smoothness condition) *L_n is convex, three times continuously differentiable, and satisfies the $(\beta^*, \mathscr{N}_{\beta^*})$-LSSC with parameter $K \geq 0$, for some convex $\mathscr{N}_{\beta^*} \subseteq \operatorname{dom} L_n$.*

2. (Positive definite restricted Hessian) *The restricted Hessian at β^* satisfies*

   $$\left[\nabla^2 L_n(\beta^*)\right]_{\mathscr{S},\mathscr{S}} \geq \lambda_{\min} I,$$

   for some $\lambda_{\min} > 0$.

3. (Irrepresentablility condition) *For some $\alpha \in (0,1]$, it holds that*

   $$\left\|\left[\nabla^2 L_n(\beta^*)\right]_{\mathscr{S}^c,\mathscr{S}}\left[\nabla^2 L_n(\beta^*)\right]^{-1}_{\mathscr{S},\mathscr{S}}\right\|_\infty < 1-\alpha. \tag{2.10}$$

4. (Beta-min condition) *The smallest non-zero entry of β^* satisfies*

   $$\beta_{\min} := \min\left\{\left|(\beta^*)_k\right| \,\middle|\, k \in \mathscr{S}\right\} > r_n, \tag{2.11}$$

   where r_n is defined in (2.9).

5. *The penalization parameter τ_n satisfies*

   $$\tau_n < \frac{\lambda^2_{\min}}{4(\alpha+4)^2}\frac{\alpha}{Ks}. \tag{2.12}$$

6. *The gradient of L_n at β^* satisfies*

   $$\left\|\nabla L_n(\beta^*)\right\|_\infty \leq \frac{\alpha}{4}\tau_n. \tag{2.13}$$

7. *The relation $\mathscr{B}_{r_n} \subseteq \mathscr{N}_{\beta^*}$ holds, where*

   $$\mathscr{B}_{r_n} := \left\{\beta \in \mathbb{R}^p : \left\|\beta_n - \beta^*\right\|_2 \leq r_n, \beta_{\mathscr{S}^c} = 0\right\}$$

   and r_n is defined in (2.9).

As mentioned previously, the first condition is the key assumption permitting us to perform a general analysis. The second, third, and forth assumptions are analogous to those appearing

in the literature for sparse linear regression. We refer to [32] for a systematic discussion of these conditions[1].

The remaining conditions determine the interplay between τ_n , n, p, and s. Whether the relation $\mathscr{B}_{r_n} \subseteq \mathscr{N}_{\beta^*}$ holds depends on the specific $\mathscr{N}_{\beta^*}$ that one can derive for the given loss function L_n. Whether the upper bound on $\left\|\nabla L_n(\beta^*)\right\|_\infty$ holds depends on the concentration of measure behavior of $\nabla L_n(\beta^*)$, which usually concentrates around 0. In the next section, we will give concrete examples for the high-dimensional setting, where p and s scale with n.

Of course, $\operatorname{sign}\hat{\beta}_n = \operatorname{sign}\beta^*$ implies that $\operatorname{supp}\hat{\beta}_n = \operatorname{supp}\beta^*$, i.e. successful variable selection.

2.5 Applications

In this section, we provide several applications of Theorem 2.9, presenting concrete bounds on the sample complexity in each case. We defer the full proofs of the results in this section to Section 2.C. However, in each case, we present here the most important step of the proof, namely, verifying the LSSC.

2.5.1 Linear regression

Recall the linear regression model and the ℓ_1-penalized least squares estimator defined in Section 2.1. The ℓ_1-penalized least squares estimator is simply an ℓ_1-penalized M-estimator, where the loss function is given by

$$L_n(\beta) = \frac{1}{2n}\sum_{i=1}^{n}\left(y_i - \langle x_i, \beta\rangle\right)^2, \quad \forall \beta \in \mathbb{R}^p.$$

We consider the *fixed design* case, where the vectors x_i are given and deterministic; furthermore, we assume that

$$\sum_j (x_i)_j^2 \le n, \quad \forall i = 1,\dots,n, \tag{2.14}$$

to normalize the vectors as in, e.g., [25] and [185]. We relax a little bit the assumption on the additive noise; we assume that $w_1,\dots,w_n$ are independent mean-zero subgaussian r.v.'s of unit subgaussian norm.

As shown in the first example of Section 2.3, the loss function L_n satisfies the LSSC with parameter $K = 0$ everywhere in $\mathbb{R}^p$. Therefore, the condition on τ_n in (2.12) is trivially satisfied, as is the final condition listed in the theorem.

Corollary 2.10. *For the linear regression problem described above, suppose that Assumptions 2–4 of Theorem 2.9 hold for some $\lambda_{\min}$ and α bounded away from zero.*[2] *If $s\log p \ll n$, and we*

[1] Equation (2.10) is sometimes called the *incoherence condition* [185].

[2] For all of the examples in this section, these assumptions are independent of the data, and we can thus talk

choose $\tau_n \gg (n^{-1}\log p)^{1/2}$, then the ℓ_1-penalized maximum likelihood estimator is consistent in variable selection.

This corollary recovers the sample complexity result given in [185].

2.5.2 Logistic regression

In the logistic regression model, the data is given by a set of independent r.v.'s

$$\{(x_1, y_1), \ldots, (x_n, y_n)\} \subset \mathbb{R}^p \times \{0, 1\}\,.$$

As in Section 2.5.1, we assume that the vectors x_i are given and they are properly normalized (cf. (2.14)). Each r.v. y_i follows the probability distribution

$$\mathsf{P}\{Y_i = 1\} = 1 - \mathsf{P}\{Y_i = 0\} = \frac{1}{1 + \mathrm{e}^{-\langle x_i, \beta^* \rangle}}\,.$$

The ℓ_1-penalized ML estimator corresponds to (2.3) with

$$L_n(\beta) := \frac{1}{n}\sum_{i=1}^{n} \log\left[1 + \mathrm{e}^{-(2y_i - 1)\langle x_i, \beta\rangle}\right].$$

Define

$$\ell_i(\beta) = \log\left[1 + \mathrm{e}^{-(2y_i - 1)\langle x_i, \beta\rangle}\right] \quad \forall\, i = 1, \ldots, n\,.$$

The cases $y_i = 0$ and $y_i = 1$ are handled similarly, so we focus on the latter. A direct differentiation yields the following (this is most easily verified for $u = v$):

$$\begin{aligned} |D^3\ell_i(\beta^* + \delta)[u, u, v]| &= \frac{\left|1 - \mathrm{e}^{-\langle x_i, \beta^* + \delta\rangle}\right|}{1 + \mathrm{e}^{-\langle x_i, \beta^* + \delta\rangle}} |\langle x_i, v\rangle|\, D^2\ell_i(\beta^* + \delta)[u, u] \\ &\le |\langle x_i, v\rangle|\, D^2\ell_i(\beta^* + \delta)[u, u]\,, \quad \forall \delta, u, v \in \mathbb{R}^p\,, \end{aligned}$$

and

$$\begin{aligned} D^2\ell_i(\beta)[u, u] &= \frac{\mathrm{e}^{-\langle x_i, \beta\rangle}\,\langle x_i, u\rangle^2}{\left(1 + \mathrm{e}^{-\langle x_i, \beta\rangle}\right)^2} \\ &\le \frac{1}{4}\langle x_i, u\rangle^2\,, \quad \forall \beta, u \in \mathbb{R}^p\,. \end{aligned}$$

The last inequality follows since the function $\frac{z}{(1+z)^2}$ has a maximum value of $\frac{1}{4}$ for $z \ge 0$. It

about them being satisfied *deterministically*.

follows that

$$|D^3\ell_i(\beta^*+\delta)[u,u,v]| \le \frac{1}{4}|\langle x_i,v\rangle||\langle x_i,u\rangle|^2$$
$$\le \frac{1}{4}\|(x_i)_{\mathscr{S}}\|_2^2\|x_i\|_\infty\|u\|_2^3,$$

for any $u\in\mathbb{R}^p$ such that $u_{\mathscr{S}^c}=0$, and for any v equal to some standard basis vector e_j. Hence, L_n satisfies the $(\beta^*,\mathscr{N}_{\beta^*})$-LSSC with parameter $K=(1/4)\nu_n^2\gamma_n$, where

$$\nu_n := \max_i \|(x_i)_{\mathscr{S}}\|_2, \quad \gamma_n := \max_i \|x_i\|_\infty.$$

The neighborhood $\mathscr{N}_{\beta^*}$ can be any fixed open convex neighborhood of β^* in $\mathbb{R}^p$.

Corollary 2.11. *For the logistic regression problem described above, suppose that Assumptions 2–4 of Theorem 2.9 hold for some $\lambda_{\min}$ and α bounded away from zero. If we choose $\tau_n \gg (n^{-1}\log p)^{1/2}$, and s and p such that $s^2(\log p)\nu_n^4\gamma_n^2 \ll n$, then the ℓ_1-penalized maximum-likelihood estimator is sparsistent.*

In [33], a sample complexity bound $s \ll \frac{\sqrt{n}}{(\log n)^2}$ is given, but the result is restricted to the case that p grows polynomially with n. The result in [8] yields the sample complexity bound $s^2(\log p)\overline{\nu_n}^2 \ll n$, where $\overline{\nu}_n := \max\{\|x_i\|_2\}$. It should be noted that $\overline{\nu}_n$ is generally significantly larger than ν_n and γ_n; for example, for i.i.d. Gaussian vectors, these scale on average as $O(\sqrt{p})$, $O(\sqrt{s})$ and $O(1)$, respectively. Our result recovers the same dependence of n on s and p as that in [8], but removes the dependence on $\overline{\nu}_n$. Of course, we do not restrict p to grow polynomially with n.

2.5.3 Gamma regression

In the gamma regression model, the data is given by a set of independent r.v.'s

$$\{(x_1,y_1),\ldots,(x_n,y_n)\}\subset\mathbb{R}^p\times[0,+\infty).$$

As in Section 2.5.1, we assume that the vectors x_i are given and they are properly normalized (cf. (2.14)). For every i, the r.v. y_i follows the gamma distribution with known shape parameter $k>0$ and unknown scale parameter

$$\theta_i = \frac{1}{k\langle x_i,\beta^*\rangle},$$

for some unknown $\beta^*\in\mathbb{R}^p$. The corresponding density function is of the form

$$\frac{1}{\Gamma(k)\theta_i^k}y^{k-1}\mathrm{e}^{-y_i/\theta_i},$$

where Γ denotes the gamma function. We assume that

$$\langle x_i, \beta^* \rangle \geq \mu_n, \quad \forall i = 1, \ldots, n, \tag{2.15}$$

for some $\mu_n > 0$, so θ_i is always well-defined.

The ℓ_1-penalized maximum-likelihood estimator is given by (2.3) with

$$L_n(\beta) := \frac{1}{n} \sum_{i=1}^{n} \left(-\log \langle x_i, \beta \rangle + y_i \langle x_i, \beta \rangle \right).$$

Note that θ_i only enters the log-likelihood via constant terms not containing β; these have been omitted, as they do not affect the estimation.

Defining

$$\ell_i(\beta) = -\log \langle x_i, \beta \rangle + y_i \langle x_i, \beta \rangle, \quad \forall i = 1, \ldots, n,$$

we obtain the following for all $u \in \mathbb{R}^p$ such that $u_{\mathscr{S}^c} = 0$, using the Cauchy-Schwartz inequality and (2.15):

$$D^2 \ell_i(\beta^*)[u,u] = \frac{\langle x_i, u \rangle^2}{\langle x_i, \beta^* \rangle^2} \leq \frac{\|(x_i)_{\mathscr{S}}\|_2^2}{\langle x_i, \beta^* \rangle^2} \|u\|_2^2 \leq \frac{1}{\mu_n^2} \|u\|_2^2 \|(x_i)_{\mathscr{S}}\|_2^2.$$

Thus, the largest restricted eigenvalue of $D^2 \ell_i(\beta^*)$ is upper bounded by $\mu_n^{-2} \nu_n^2$, where

$$\nu_n := \max_i \left\{ \|(x_i)_{\mathscr{S}}\|_2 \,|\, i = 1, \ldots, n \right\},$$

Similarly, we obtain

$$D^2 \ell_i(\beta^*)[e_j, e_j] \leq \frac{1}{\mu_n^2} \|x_i\|_\infty^2,$$

for any standard basis vector e_j. Thus, the largest diagonal entry of $D^2 \ell_i(\beta^*)$ is upper bounded by $\mu_n^{-2} \gamma_n^2$, where $\gamma_n = \max_i \|x_i\|_\infty$.

Fix $\kappa > 0$. By Example 2.7 and Lemma 2.4, L_n satisfies the $(\beta^*, \mathscr{N}_{\beta^*})$-LSSC with parameter $K = 2(1+\kappa^{-1})^3 \mu_n^{-3} \nu_n^2 \gamma_n$, and

$$\mathscr{N}_{\beta^*} = \left\{ \beta^* + \delta : \|\delta\|_2 < \frac{\mu_n}{(1+\kappa)\nu_n}, \delta \in \mathbb{R}^p \right\}.$$

Corollary 2.12. *Consider the gamma regression problem as described above, and suppose that Assumptions 2–4 of Theorem 2.9 hold for some $\lambda_{\min}$, and α bounded away from zero. If $\tau_n \gg \sqrt{n}^{-1} \log p$ and $s^2 (\log p)^2 \mu_n^{-6} \nu_n^4 \gamma_n^2 \ll n$, then the ℓ_1-penalized maximum likelihood estimator is consistent in variable selection.*

To the best of our knowledge, this is the first variable selection consistency result for gamma regression.

2.5.4 Graphical model selection

We consider a setup slightly more general than the one described in Section 2.1.1. Let $\Theta^* \in \mathbb{R}^{p\times p}$ be a positive-definite matrix. We assume there are at most s non-zero entries in Θ^*, and let $\mathscr{S}$ denote its support set. Let $X_1,\dots,X_n$ be independent p-dimensional random vectors generated according to a common distribution with mean zero and covariance matrix $\Sigma^* := (\Theta^*)^{-1}$. We are interested in recovering the support of Θ^* given $X_1,\dots,X_n$.

We assume that each $\left(\Sigma_{i,i}\right)^{-1/2} X_{i,i}$ is subgaussian with parameter $c > 0$, and that $\Sigma_{i,i}$ is bounded above by a constant κ_{Σ^*}, for all $i \in \{1,\dots,p\}$. Let $\rho_{\min}$ denote the smallest eigenvalue of Θ^*.

We consider the ℓ_1-penalized M-estimator of the form (2.3), given by

$$\hat{\Theta}_n := \arg\min_{\Theta} \left\{ L_n(\Theta) + \tau_n |\Theta|_1 \,|\Theta > 0, \Theta \in \mathbb{R}^{p\times p} \right\}.$$

Here $|\Theta|_1$ denotes the entry-wise ℓ_1-norm, i.e.,

$$|\Theta|_1 = \sum_{(i,j)\in\{1,\dots,p\}^2} \left|\Theta_{i,j}\right|,$$

and

$$L_n(\Theta) = \mathrm{Tr}\left(\hat{\Sigma}_n\Theta\right) - \log\det\Theta,$$

where

$$\hat{\Sigma}_n := \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$$

is the sample covariance matrix.

Fix $\kappa > 0$. By Example 2.8, we know that L_n satisfies the $(\Theta^*, \mathscr{N}_{\Theta^*})$-LSSC with parameter $2\kappa^{-3}(1+\kappa)^3\rho_{\min}^{-3}$, where

$$\mathscr{N}_{\Theta^*} := \left\{ \Theta^* + \Delta \;\middle|\; \|\Delta\|_F < \frac{1}{1+\kappa}\rho_{\min}, \Delta = \Delta^T, \Delta \in \mathbb{R}^{p\times p} \right\},$$

where $\rho_{\min}$ denotes the smallest eigenvalue of Θ^*.

The beta-min condition can be written as

$$\min\left\{ \Theta^*_{i,j} \;\middle|\; \Theta^*_{i,j} \neq 0, (i,j) \in \{1,\dots,p\}^2 \right\} > r_n.$$

We now have the following.

Corollary 2.13. *Consider the graphical model selection problem described above, and suppose the above assumptions and assumptions 2 to 4 of Theorem 2.9 hold for some c, κ_{Σ^*}, $\rho_{\min}$, $\lambda_{\min}$, and α bounded away from zero. If $\tau_n \gg (n^{-1}\log p)^{1/2}$ and $s^2 \log p \ll n$, the ℓ_1-penalized M-estimator $\hat{\Theta}_n$ is consistent in variable selection.*

Corollary 2.13 is for graphical learning on general sparse networks, as we only put a constraint on s. Several previous works have instead imposed structural constraints on the maximum degree of each node; e.g. see [153]. Since this model requires additional structural assumptions beyond sparsity alone, it is outside the scope of our theoretical framework.

2.6 Discussions

Our work bears some resemblance to the independent work of [114]. The smoothness condition therein is in fact the *non-structured* condition in (2.6). From the discussion in Section 2.2, we see that our condition is less restrictive. As a consequence, both analyses lead to scaling laws of the form $n \gg K^2 s^2 (\log p)^\gamma$ for some $\gamma > 0$ for generalized linear models, but the corresponding definitions of K differ significantly. Eliminating the dependence of K on p requires additional non-trivial extensions of the framework in [114], whereas in our framework the desired independence is immediate (e.g. see the logistic and gamma regression examples).

The framework presented here considers general sparse parameters. It is of great theoretical and practical importance to sharpen this framework for structured sparse parameters, e.g., group sparsity, and graphical model learning for networks with bounded degrees.

The sample complexity results we have derived are worst-case with respect to all s-sparse parameters. As our notion of the LSSC is local, it is interesting to explore the possibility of deriving sample complexity bounds that depend not only on the sparsity but also other characteristics of the true parameter.

2.A Auxiliary result for the non-structured case

In this sub-section, we prove the following claim made in Section 3. Note that, in contrast to the main definition of the LSSC, the vectors here are *not* necessarily structured.

Proposition 2.14. *Consider a function $f \in \mathscr{C}^3(\mathrm{dom\, f})$ with domain $\mathrm{dom}\, f \subseteq \mathbb{R}^p$. Fix $x^* \in \mathrm{dom}\, f$, and let $\mathscr{N}_{x^*}$ be an open set in $\mathrm{dom}\, f$ containing x^*. Let $K \geq 0$. The following statements are equivalent.*

1. *$D^2 f(x)$ is locally Lipschitz continuous with respect to x^*; that is,*

$$\|D^2 f(x^* + \delta) - D^2 f(x^*)\|_2 \leq K\|\delta\|_2, \tag{2.16}$$

for all $\delta \in \mathbb{R}^p$ *such that* $x^* + \delta \in \mathcal{N}_{x^*}$.

2. $D^3 f(x)$ *is locally bounded; that is,*

$$|D^3 f(x^* + \delta)[u, v, w]| \leq K\|u\|_2\|v\|_2\|w\|_2 \tag{2.17}$$

for all $\delta \in \mathbb{R}^p$ *such that* $x^* + \delta \in \mathcal{N}_{x^*}$, *and for all* $u, v, w \in \mathbb{R}^p$.

Proof. Suppose that (2.16) holds. By Proposition 2.5, it suffices to prove that

$$|D^3 f(x^* + \delta)[u, u, u]| \leq K\|u\|_2^3$$

for all $u \in \mathbb{R}^p$. By definition, we have

$$|D^3 f(x^* + \delta)[u, u, u]| = |\langle u, Hu\rangle| \leq \|H\|_2\|u\|_2^2,$$

where

$$H := \lim_{t\to 0} \frac{D^2 f(x^* + \delta + tu) - D^2 f(x^* + \delta)}{t}.$$

We therefore have (2.17) since $\|H\|_2 \leq K\|\delta\|_2$ by (2.16).

Conversely, suppose that (2.17) holds. We have the following Taylor expansion [188]:

$$D^2 f(x^* + \delta) = D^2 f(x^*) + \int_0^1 D^3 f(x_t)[\delta]\, dt,$$

where $x_t := x^* + t\delta$. We also have from (2.17) and the definition of the spectral norm that $\|D^3 f(x^* + \delta)[\delta]\|_2 \leq K\|u\|_2$, and hence

$$\begin{aligned}\|D^2 f(x^* + \delta) - D^2 f(x^*)\|_2 &= \|\int_0^1 D^3 f(x_t)[\delta]\, dt\|_2 \\ &\leq K\|\delta\|_2.\end{aligned}$$

This completes the proof. □

2.B Proof of Theorem 2.9

The proof is based on the optimality conditions on $\hat{\beta}$ for the original problem, and those on $\check{\beta}$ for the restricted problem. We first observe that $\check{\beta}_n$ exists, since the function $x \mapsto \|x\|_1$ is coercive. Recall that $\check{\beta}_n$ is assumed to be uniquely defined.

To achieve variable selection consistency, it suffices that $\hat{\beta}_n = \check{\beta}_n$ and $\operatorname{supp} \check{\beta}_n = \operatorname{supp} \beta^*$. We derive sufficient conditions for $\hat{\beta}_n = \check{\beta}_n$ in Lemma 2.15, and make this sufficient condition explicitly dependent on the problem parameters in Lemma 2.16. This lemma will require that $\left\|\check{\beta}_n - \beta^*\right\|_2 \leq R_n$ for some $R_n > 0$. We will derive an estimation error bound of the form

$\left\|\check{\beta}_n - \beta^*\right\|_2 \leq r_n$ in Lemma 2.18. We will then conclude that $\hat{\beta}_n = \check{\beta}_n$ if $r_n \leq R_n$ and the assumptions in Lemma 2.16 are satisfied, from which it will follow that $\operatorname{sign}\check{\beta} = \operatorname{sign}\beta^*$ provided that $\beta_{\min} \geq r_n$.

The following lemma is proved via an extension of the techniques of [185].

Lemma 2.15. *We have $\hat{\beta}_n = \check{\beta}_n$ if*

$$\left\|\left[\nabla L_n(\check{\beta}_n)\right]_{\mathscr{S}^c}\right\|_\infty < \tau_n. \tag{2.18}$$

Proof. Recall that L_n is convex by assumption. Also recall that $\check{\beta}_n$ is assumed to be uniquely defined, and hence it is the only vector the satisfies the corresponding optimality condition:

$$\left[\nabla L_n(\check{\beta}_n)\right]_{\mathscr{S}} + \tau_n \check{z}_{\mathscr{S}} = 0 \tag{2.19}$$

for some $\check{z}_{\mathscr{S}}$ such that $\|\check{z}_{\mathscr{S}}\|_\infty \leq 1$. Moreover, the fact that (2.18) is satisfied means that there exists $\check{z}_{\mathscr{S}^c}$ such that $\|\check{z}_{\mathscr{S}^c}\|_\infty < 1$ and

$$\nabla L_n(\check{\beta}_n) + \tau_n \check{z} = 0,$$

where $\check{z} := (\check{z}_{\mathscr{S}}, \check{z}_{\mathscr{S}^c})$. Therefore, $\check{\beta}_n$ is a minimizer of the original optimization problem in $\mathbb{R}^p$.

We now address the uniqueness of $\hat{\beta}$. By a similar argument to Lemma 1 in [152] (see also Lemma 1(b) in [185]), any minimizer $\tilde{\beta}$ of the original optimization problem satisfies $\tilde{\beta}_{\mathscr{S}^c} = 0$. Thus, since $\check{\beta}$ is the only optimal vector for the restricted optimization problem, we conclude that $\hat{\beta}_n = \check{\beta}_n$ uniquely. □

We now combine Lemma 2.15 with the assumptions of Theorem 2.9 to obtain the following.

Lemma 2.16. *Under assumptions 1, 2, 3 and 6 of Theorem 2.9, we have $\hat{\beta}_n = \check{\beta}_n$ if $\check{\beta} \in \mathscr{N}_{\beta^*} \cap \mathscr{B}_{R_n}$, where*

$$\mathscr{B}_{R_n} := \left\{\beta \in \mathbb{R}^p \;\middle|\; \left\|\beta - \beta^*\right\|_2 \leq R_n, \beta_{\mathscr{S}^c} = 0\right\}$$

with

$$R_n = \frac{1}{2}\sqrt{\frac{\alpha\tau_n}{K}}. \tag{2.20}$$

Proof. Applying a Taylor expansion at β^*, and noting that both β^* and $\check{\beta}_n$ are supported on $\mathscr{S}$, we obtain

$$\left[\nabla L(\check{\beta}_n)\right]_{\mathscr{S}^c} = \left[\nabla L_n(\beta^*)\right]_{\mathscr{S}^c} + \left[\nabla^2 L_n(\beta^*)\right]_{\mathscr{S}^c,\mathscr{S}}\left(\check{\beta}_n - \beta^*\right)_{\mathscr{S}} + (\epsilon_n)_{\mathscr{S}^c}, \tag{2.21}$$

where the remainder term is given by

$$\epsilon_n = \int_0^1 (1-t) D^3 L_n(\beta_t)[\check{\beta} - \beta^*, \check{\beta} - \beta^*]\,\mathrm{d}t,$$

where $\beta_t := \beta^* + t(\check{\beta} - \beta^*)$ (see, e.g., [188, Section 4.5]), and thus satisfies

$$\|\epsilon_n\|_\infty \leq \sup_t \left\{ \|D^3 L_n(\beta_t)[\check{\beta} - \beta^*, \check{\beta} - \beta^*]\|_\infty \,\middle|\, t \in [0,1] \right\}. \tag{2.22}$$

Recall the optimality condition for $\check{\beta}$ in (2.19). Again using a Taylor expansion, we can write this condition as

$$\left[\nabla L_n(\beta^*)\right]_{\mathscr{S}} + \left[\nabla^2 L_n(\beta^*)\right]_{\mathscr{S},\mathscr{S}} \left(\check{\beta}_n - \beta^*\right)_{\mathscr{S}} + (\epsilon_n)_{\mathscr{S}} + \tau_n \check{z}_{\mathscr{S}} = 0. \tag{2.23}$$

Recall that $\left[\nabla^2 L_n(\beta^*)\right]_{\mathscr{S},\mathscr{S}}$ is invertible by the second assumption of Theorem 2.9. Solving for $\left(\check{\beta}_n - \beta^*\right)_{\mathscr{S}}$ in (2.23) and substituting the solution into (2.21), we obtain

$$\begin{aligned}
\left[\nabla L_n(\check{\beta}_n)\right]_{\mathscr{S}^c} = &-\tau_n \left[\nabla^2 L_n(\beta^*)\right]_{\mathscr{S}^c,\mathscr{S}} \left[\nabla^2 L_n(\beta^*)\right]^{-1}_{\mathscr{S},\mathscr{S}} \check{z}_{\mathscr{S}} \\
&+ \left[\nabla L(\beta^*)\right]_{\mathscr{S}^c} \\
&- \left[\nabla^2 L_n(\beta^*)\right]_{\mathscr{S}^c,\mathscr{S}} \left[\nabla^2 L_n(\beta^*)\right]^{-1}_{\mathscr{S},\mathscr{S}} \left[\nabla L_n(\beta^*)\right]_{\mathscr{S}} \\
&+ (\epsilon_n)_{\mathscr{S}^c} \\
&- \left[\nabla^2 L_n(\beta^*)\right]_{\mathscr{S}^c,\mathscr{S}} \left[\nabla^2 L_n(\beta^*)\right]^{-1}_{\mathscr{S},\mathscr{S}} (\epsilon_n)_{\mathscr{S}}.
\end{aligned}$$

Using the irrepresentability condition (assumption 3 of Theorem 2.9) and the triangle inequality, we have $\left\|\left[\nabla L_n(\check{\beta}_n)\right]_{\mathscr{S}^c}\right\|_\infty < \tau_n$ provided that

$$\max\left\{\left\|\nabla L_n(\beta^*)\right\|_\infty, \|\epsilon_n\|_\infty\right\} \leq \frac{\alpha}{4}\tau_n.$$

The first requirement $\left\|\nabla L_n(\beta^*)\right\|_\infty \leq (\alpha/4)\tau_n$ is simply assumption 6 of Theorem 2.9, so it remains to determine a sufficient condition for $\|\epsilon_n\|_\infty \leq (\alpha/4)\tau_n$. Since L_n satisfies the $(\beta^*, \mathscr{N}_{\beta^*})$-LSSC with parameter K, we have from (2.22) that

$$\|\epsilon_n\|_\infty \leq K \left\|\check{\beta} - \beta^*\right\|_2^2,$$

provided that $\check{\beta} \in \mathscr{N}_{\beta^*}$ (since $\mathscr{N}_{\beta^*}$ is convex by assumption, this implies $\beta_t \in \mathscr{N}_{\beta^*}$). Thus, to have $\|\epsilon_n\|_\infty \leq \frac{\alpha}{4}\tau_n$, it suffices that

$$\left\|\check{\beta} - \beta^*\right\|_2 \leq \frac{1}{2}\sqrt{\frac{\alpha\tau_n}{K}}$$

and $\check{\beta} \in \mathscr{N}_{\beta^*}$. □

To bound the distance $\left\|\check{\beta} - \beta^*\right\|_2$, we adopt an approach from [152, 159]. We begin with an auxiliary lemma.

Lemma 2.17. *Let $g : \mathbb{R}^p \to \mathbb{R}$ be a convex function, and let $z \in \mathbb{R}^p$ be such that $g(z) \leq 0$. Let $\mathscr{B} \subset \mathbb{R}^p$ be a closed set, and let $\partial\mathscr{B}$ be its boundary. If $g > 0$ on $\partial\mathscr{B}$ and $g(b) \leq 0$ for some*

$b \in \mathscr{B} \setminus \partial\mathscr{B}$, then $x \in \mathscr{B}$.

Proof. We use a proof by contradiction. Suppose that $z \notin \mathscr{B}$. We first note that there exists some $t^* \in (0,1)$ such that $b + t^*(z-b) \in \partial\mathscr{B}$; if such a t^* did not exist, then we would have $z_t := b + t(z-b) \to z$ as $t \to 1$, which is impossible since $z \notin \mathscr{B}$ and $\mathscr{B}$ is closed.

We now use the convexity of g to write

$$g(b + t^*(x-b)) \leq (1-t^*)g(b) + t^* g(x) \leq 0,$$

which is a contradiction since $g > 0$ on $\partial\mathscr{B}$. □

The following lemma presents the desired bound on $\left\|\check{\beta}_n - \beta^*\right\|_2$; note that this can be interpreted as the estimation error in the $n > p$ setting, considering $\beta^*_{\mathscr{S}}$ as the parameter to be estimated.

Lemma 2.18. *Define the set*

$$\mathscr{B}_{r_n} := \left\{ \beta \in \mathbb{R}^p \;\middle|\; \left\|\beta - \beta^*\right\|_2 \leq r_n, \beta_{\mathscr{S}^c} = 0 \right\},$$

where

$$r_n := \frac{\alpha + 4}{\lambda_{\min}} \sqrt{s}\tau_n. \tag{2.24}$$

Under assumptions 1, 2, 6 and 7 of Theorem 2.9, if

$$\tau_n < \frac{3\lambda_{\min}^2}{2(\alpha+4)Ks}, \tag{2.25}$$

then $\check{\beta}_n \in \mathscr{B}_{r_n}$.

Proof. Set $s = |\mathscr{S}|$, and for $\beta \in \mathbb{R}^s$ let $Z(\beta) = (\beta, 0) \in \mathbb{R}^p$ be the zero-padding mapping, where $(\beta, 0)$ denotes the vector that equals to β on $\mathscr{S}$ and 0 on $\mathscr{S}^c$. Then we have

$$\check{\beta}_{\mathscr{S}} = \arg\min_{\beta} \left\{ (L_n \circ Z)(\beta) + \tau_n \left\|\beta\right\|_1 \;\middle|\; \beta \in \mathbb{R}^s \right\}.$$

For $\delta \in \mathbb{R}^s$, define

$$g(\delta) = (L_n \circ Z)(\beta^*_{\mathscr{S}} + \delta) - (L_n \circ Z)(\beta^*_{\mathscr{S}}) + \tau_n \left(\left\|\beta^*_{\mathscr{S}} + \delta\right\|_1 - \left\|\beta^*_{\mathscr{S}}\right\|_1 \right).$$

We trivially have $g(0) = 0$, and thus $g(\delta^*) \leq g(0) = 0$, where $\delta^* := \check{\beta}_{\mathscr{S}} - \beta^*_{\mathscr{S}}$. Now our goal is prove that $g > 0$ on the boundary of $(\mathscr{B}_{r_n})_{\mathscr{S}} := \left\{\delta \in \mathbb{R}^s : \|\delta\|_2 \leq r_n\right\}$, thus permitting the application of Lemma 2.17.

We proceed by deriving a lower bound on $g(\delta)$. We define

$$\varphi(t) := (L_n \circ Z)(\beta^*_{\mathscr{S}} + t\delta),$$

and write the following Taylor expansion:

$$\begin{aligned}(L_n \circ Z)(\beta^*_{\mathscr{S}} + \delta) - (L_n \circ Z)(\beta^*_{\mathscr{S}}) &= \varphi(1) - \varphi(0) \\ &= \varphi'(0) + \frac{1}{2}\varphi''(0) + \frac{1}{6}\varphi'''(\tilde{t}),\end{aligned}$$

for some $\tilde{t} \in [0,1]$ (recall that L_n is three times differentiable by assumption). We bound the term $\varphi'(0)$ as follows:

$$\begin{aligned}\left|\varphi'(0)\right| &= \left|\langle \left[\nabla L_n(\beta^*)\right]_{\mathscr{S}}, \delta\rangle\right| \\ &\leq \sqrt{s}\left\|\left[\nabla L_n(\beta^*)\right]_{\mathscr{S}}\right\|_\infty \|\delta\|_2 \\ &\leq \frac{\alpha\tau_n}{4}\sqrt{s}\,\|\delta\|_2,\end{aligned}$$

where the first step follows from Hölder's inequality and the inequality $\|z\|_2 \leq \sqrt{s}\|z\|_1$, and the second step uses Assumption 6 of Theorem 2.9. To bound the term $\varphi''(0)$, we use the second assumption of Theorem 2.9 and write

$$\varphi''(0) = \delta^T \left[\nabla^2 L_n(\beta^*)\right]_{\mathscr{S},\mathscr{S}} \delta \geq \lambda_{\min}\|\delta\|_2^2.$$

We now turn to the term $\varphi'''(\tilde{t})$. Again using the fact that L_n satisfies the $(\beta^*, \mathscr{N}_{\beta^*})$-LSSC with parameter K, it immediately follows that $(L_n \circ Z)$ satisfies the $(\beta^*_{\mathscr{S}}, \left(\mathscr{N}_{\beta^*}\right)_{\mathscr{S}})$-LSSC with parameter K, where $\left(\mathscr{N}_\beta\right)_{\mathscr{S}} = \left\{\beta_{\mathscr{S}} | \beta \in \mathscr{N}_{\beta^*}\right\}$. Hence, and also making use of Hölder's inequality and the fact that $\|z\|_1 \leq \sqrt{s}\|z\|_2$ ($z \in \mathbb{R}^s$), we have

$$\begin{aligned}\left|\varphi'''(\tilde{t})\right| &= \left|D^3(L_n \circ Z)(\beta^*_{\mathscr{S}} + \tilde{t}\delta)[\delta,\delta,\delta]\right| \\ &\leq \|\delta\|_1 \|D^3(L_n \circ Z)(\beta^*_{\mathscr{S}} + \tilde{t}\delta)[\delta,\delta]\|_\infty \\ &\leq K\sqrt{s}\,\|\delta\|_2^3\end{aligned}$$

provided that $\beta^*_{\mathscr{S}} + \tilde{t}\delta \in \left(\mathscr{N}_\beta\right)_{\mathscr{S}}$. Since $\mathscr{B}_{r_n} \subseteq \mathscr{N}_{\beta^*}$ by assumption 7 of Theorem 2.9, the latter condition holds provided that $\delta \in (\mathscr{B}_{r_n})_{\mathscr{S}}$.

Using the triangle inequality, we have

$$\left|\left\|\beta^*_{\mathscr{S}} + \delta\right\|_1 - \left\|\beta^*_{\mathscr{S}}\right\|_1\right| \leq \|\delta\|_1 \leq \sqrt{s}\,\|\delta\|_2.$$

Hence, and combining the preceding bounds, we have $g(\delta) \geq f(\|\delta\|_2)$, where

$$f(x) = -\frac{\alpha\tau_n}{4}\sqrt{s}x + \frac{\lambda_{\min}}{2}x^2 - \frac{K\sqrt{s}}{6}x^3 - \sqrt{s}\tau_n x.$$

Observe that if the inequality

$$0 < x < \frac{3\lambda_{\min}}{2K\sqrt{s}}. \tag{2.26}$$

holds, then we can bound the coefficient to x^3 in terms of that of x^2 to obtain

$$f(x) > \frac{\lambda_{\min}}{4} x^2 - \left(1 + \frac{\alpha}{4}\right) \sqrt{s} \tau_n x. \tag{2.27}$$

By a direct calculation, this lower bound has roots at 0 and r_n (see (2.24)), and hence $f(r_n) > 0$ provided that $x = r_n$ satisfies (2.26). By a direct substitution, this condition can be ensured by requiring that

$$\tau_n < \frac{3\lambda_{\min}^2}{2(\alpha+4)Ks}. \tag{2.28}$$

□

Recalling that $g(\delta) \geq f(\|\delta\|_2)$, we have proved that g satisfies the conditions of Lemma 2.17 with $z = \delta^*$, $b = 0$, and $\mathscr{B} = (\mathscr{B}_{r_n})_{\mathscr{S}}$, and we thus have $\delta^* \in (\mathscr{B}_{r_n})_{\mathscr{S}}$, or equivalently $\check{\beta}_n \in \mathscr{B}_{r_n}$.

We now combine the preceding lemmas to obtain Theorem 2.9. We require $r_n \leq R_n$ so the assumption that $\left\|\check{\beta} - \beta^*\right\|_\infty \leq R_n$ in Lemma 2.16 is satisfied. From the definitions in (2.20) and (2.24), this is equivalent to requiring

$$\tau_n \leq \frac{\lambda_{\min}^2}{4(\alpha+4)^2} \frac{\alpha}{Ks},$$

which is true by assumption 5 of the theorem. This assumption also implies that (2.25) holds, since $\frac{\alpha}{4(\alpha+4)} \leq \frac{3}{2}$ for any $\alpha \geq 0$. Finally, by the conclusion of Lemma 2.18, we have successful sign pattern recovery if $\beta_{\min} \geq r_n$, thus recovering assumption 4 of the theorem.

2.C Proofs of the results in Section 2.5

2.C.1 Proof of Corollary 2.10

By a direct calculation, we have

$$\nabla L_n(\beta^*) = \frac{1}{n} \sum_{i=1}^n (y_i - \mathsf{E}\, y_i) x_i.$$

By the union bound and the Hoeffding-type inequality for subgaussian r.v.'s (cf. Theorem A.5),

$$\begin{aligned} \mathsf{P}\left\{ \left\|\nabla L_n(\beta^*)\right\|_\infty \geq \frac{\alpha\tau_n}{4} \right\} &\leq \sum_{i=1}^p \mathsf{P}\left\{ \left|\left[\nabla L_n(\beta^*)\right]_i\right| \geq \frac{\alpha\tau_n}{4} \right\} \\ &\leq 2\mathrm{e}p\mathrm{e}^{-nt^2}\Big|_{t=\frac{\alpha\tau_n}{4}}. \end{aligned}$$

Since $[D^2 L_n(\beta)]_{\mathscr{S},\mathscr{S}} = [D^2 L_n(\beta^*)]_{\mathscr{S},\mathscr{S}}$ is positive definite for all $\beta \in \mathbb{R}^p$ by the second assumption of Theorem 2.9, $\check{\beta}_n$ uniquely exists, and Theorem 2.9 is applicable. Choosing τ_n sufficiently large that the above bound decays to zero, we obtain the corollary.

2.C.2 Proof of Corollary 2.11

By a direct differentiation, we obtain for $j \in \{1,\dots,p\}$ that

$$\left[\nabla L_n(\beta^*)\right]_j = -\sum_{i=1}^{n} \varepsilon_i (x_i)_j,$$

where $\varepsilon_i = n^{-1}\left(y_i - \mathsf{E}\, y_i\right)$.

Define $\xi_i := x_i y_i$ for all i. Notice that every ξ_i is a bounded r.v. taking values in $[0, x_i]$. By Hoeffding's inequality (cf. Theorem A.6) and the union bound, we obtain

$$\begin{aligned}\mathsf{P}\left\{\left\|\nabla L_n(\beta^*)\right\|_\infty \geq \frac{\alpha\tau_n}{4}\right\} &\leq \sum_{j=1}^{p} \mathsf{P}\left\{\left|\left[\nabla L_n(\beta^*)\right]_j\right| \geq \frac{\alpha\tau_n}{4}\right\} \\ &\leq 2\exp\left(\log p - 2nt^2\right)\Big|_{t=\frac{\alpha\tau_n}{4}}.\end{aligned}$$

This decays to zero provided that $\tau_n \gg (n^{-1}\log p)^{1/2}$. Substituting this scaling into the fifth condition of Theorem 2.9, we obtain the condition $s^2\left(\log p\right)v_n^4\gamma_n^2 \ll n$. The required uniqueness of $\check{\beta}$ can be proved by showing that the composition $L_n \circ Z$ (with Z being the zero-padding of a vector in $\mathbb{R}^s$) is strictly convex, given the second condition of Theorem 2.9. One way to prove this is via self-concordant like inequalities [173]; we omit the proof here for brevity.

2.C.3 Proof of Corollary 2.12

By a direct differentiation, we obtain

$$\left[\nabla L_n(\beta^*)\right]_j = \sum_{i=1}^{n} \varepsilon_i\,(x_i)_j$$

for $j \in \{1,\dots,p\}$, where $\varepsilon_i := n^{-1}\left(y_i - \mathsf{E}\, y_i\right)$.

To study the concentration of measure behavior of $\nabla L_n(\beta^*)$, we use Bernstein's inequality (cf. Theorem A.7). We first check the moment conditions using the following fact.

Lemma 2.19. *Let η be a gamma r.v. with shape parameter $k > 0$ and scale parameter θ. Then*

$$\mathsf{E}\,\eta^q = \frac{\Gamma(q+k)}{\Gamma(k)}\theta^q, \quad \forall\, q \in \mathbb{N},$$

where Γ denotes the gamma function.

Fix $j \in \{1, \ldots, p\}$, and define $\xi_i := n^{-1}(x_i)_j y_i$. We have

$$\begin{aligned}\sum_{i=1}^{n} \mathsf{E}\,\xi_i^2 &= \sum_{i=1}^{n} \frac{(x_i)_j^2}{n^2} \mathsf{E}\, y_i^2 \\ &= \sum_{i=1}^{n} \frac{(x_i)_j^2}{n^2} \frac{\Gamma(k+2)}{\Gamma(k)} \theta_i^2.\end{aligned}$$

Recall that $\theta_i = k^{-1} \langle x_i, \beta^* \rangle^{-1}$. Using the first displayed equation in Section 2.5.3, we have

$$\theta_i \leq \left(k\mu_n\right)^{-1}, \tag{2.29}$$

and thus

$$\begin{aligned}\sum_{i=1}^{n} \mathsf{E}\,\xi_i^2 &\leq \frac{1}{(n\mu_n)^2} \frac{\Gamma(k+2)}{k^2\Gamma(k)} \sum_{i=1}^{n} (x_i)_j^2 \\ &\leq \frac{1}{n\mu_n^2} \frac{\Gamma(k+2)}{k^2\Gamma(k)},\end{aligned}$$

where we have applied the assumption $\sum_{i=1}^{n} (x_i)_j^2 \leq n$. Using the identity $\Gamma(k+2) = k(k+1)\Gamma(k)$, we obtain

$$\sum_{i=1}^{n} \mathsf{E}\,\xi_i^2 \leq \frac{k+1}{n\mu_n^2 k}.$$

As for the moments of higher orders, we have

$$\begin{aligned}\sum_{i=1}^{n} \mathsf{E}\,|\xi_i|^q &= \sum_{i=1}^{n} \frac{\left|(x_i)_j\right|^q}{n^q} \mathsf{E}\left|y_i\right|^q \\ &= \sum_{i=1}^{n} \frac{\left|(x_i)_j\right|^q}{n^q} \frac{\Gamma(k+q)}{\Gamma(k)} \theta_i^q.\end{aligned}$$

With the upper bound (2.29) on θ_i, we have

$$\begin{aligned}\sum_{i=1}^{n} \mathsf{E}\,|\xi_i|^q &\leq \frac{\Gamma(k+q)}{(kn\mu_n)^q\Gamma(k)} \sum_{i=1}^{n} \left|(x_i)_j\right|^q \\ &= \frac{\Gamma(k+q)}{(kn\mu_n)^q\Gamma(k)} \left\|((x_1)_j, \ldots, (x_n)_j)\right\|_q^q.\end{aligned}$$

Using the identity $\|z\|_q \leq \|z\|_2$ for $q \geq 2$, and the assumption $\sum_{i=1}^{n} (x_i)_j^2 \leq n$, we obtain

$$\sum_{i=1}^{n} \mathsf{E}\,|\xi_i|^q \leq \frac{\Gamma(k+q)}{(k\sqrt{n}\mu_n)^q\Gamma(k)}.$$

For $k \in (0,1]$, we have $\frac{\Gamma(k+q)}{\Gamma(q)} \leq q!$, and hence by a direct substitution it suffices to choose

$$v = \frac{k+1}{n\mu_n^2 k^2}, \quad c = \frac{1}{k\sqrt{n}\mu_n}. \tag{2.30}$$

For $k \in (1,\infty)$, we have by induction on q that $\frac{\Gamma(k+q)}{\Gamma(q)} \leq q!k^q$. Thus, for $k \in (1,\infty)$, it suffices that

$$v = \frac{2k}{n\mu_n^2}, \quad c = \frac{1}{\sqrt{n}\mu_n}. \tag{2.31}$$

Thus, applying Bernstein's inequality and the union bound, we obtain

$$\begin{aligned}\mathsf{P}\left\{\left\|\nabla L_n(\beta^*)\right\|_\infty \geq \frac{\alpha\tau_n}{4}\right\} &\leq \sum_{i=1}^{p} \mathsf{P}\left\{\left|\left[\nabla L_n(\beta^*)\right]_i\right| \geq \frac{\alpha\tau_n}{4}\right\} \\ &\leq 2\exp\left[\log p - \frac{t^2}{2(v+ct)}\right]\Bigg|_{t=\frac{\alpha\tau_n}{4}}.\end{aligned}$$

Since L_n is self-concordant and $\left[D^2 L_n(\beta^*)\right]_{\mathscr{S},\mathscr{S}}$ is positive definite by assumption, the composition $L_n \circ Z$ with the padding operator Z is strictly convex [140] and thus $\check{\beta}_n$ uniquely exists. Therefore, we can apply Theorem 2.9. The scaling laws on τ_n and (p,n,s) follow via the same argument to that in the proof of Corollary 2.11. Note that the final condition of Theorem 2.9 also imposes conditions on (p,n,s), but for this term even the weaker condition $s^2(\log p)v_n^2 \ll n$ suffices.

2.C.4 Proof of Corollary 2.13

By a direct differentiation, we obtain

$$\nabla L_n(\Theta^*) = \hat{\Sigma}_n - \left(\Theta^*\right)^{-1} = \hat{\Sigma}_n - \Sigma.$$

We apply the following lemma from [153] to study the concentration behavior of $\nabla L_n(\Theta^*)$.

Lemma 2.20. *Let Σ and $\hat{\Sigma}_n$ be defined as in Section 6.4. We have*

$$\mathsf{P}\left\{\left|\left(\hat{\Sigma}_n\right)_{i,j} - \Sigma_{i,j}\right| > t\right\} \leq 4\exp\left[-\frac{nt^2}{128(1+4c^2)^2\kappa_{\Sigma^*}^2}\right],$$

for all $t \in (0, 8\kappa_{\Sigma^*}(1+c)^2)$.

Using the union bound, we have

$$\mathsf{P}\left\{\left\|\nabla L_n(\Theta^*)\right\|_\infty \leq \frac{\alpha\tau_n}{4}\right\} \leq 4p^2 \exp\left[-\frac{nt^2}{128(1+4\sigma^2)^2\kappa^2_{\Sigma^*}}\right]\Bigg|_{t=\frac{\alpha\tau_n}{4}},$$

provided that $\tau_n \to 0$, and that n is large enough so that the upper bound on t in the lemma is satisfied.

Define

$$\check{\Theta}_n \in \arg\min_{\Theta}\left\{ L_n(\Theta) + \tau_n|\Theta|_1 \;\middle|\; \Theta > 0, \Theta_{\mathscr{S}^c} = 0, \Theta \in \mathrm{R}^{p\times p}\right\}. \tag{2.32}$$

Since L_n is self-concordant and $\left[D^2 L_n(\Theta^*)\right]_{\mathscr{S},\mathscr{S}}$ is positive definite by assumption, the composition $L_n \circ Z$ with the padding operator Z is strictly convex [140] and thus $\check{\Theta}_n$ uniquely exists. Therefore, we can apply Theorem 2.9. The scaling laws on τ_n and (p,n,s) follow via the same arguments as the preceding examples.

3 Estimation error of the lasso

In the previous chapter, we have studied variable selection in the high-dimensional setting. In this chapter, we will focus on statistical estimation in the same setting. In particular, our aim is to establish the estimation consistency of the least absolute shrinkage and selection operator (lasso). The standard approach to establishing estimation consistency in the high-dimensional setting relies on the *restricted strong convexity (RSC)* of the loss function, but the RSC does not hold for the lasso in general. Via *relaxing* the RSC, we obtain sharp estimation error bounds, and establish the minimax optimality of the lasso in some scenarios.

This chapter is based on the joint work with Nissim Zerbib, Ya-Ping Hsieh, and Volkan Cevher [189].

3.1 Introduction

Let us revisit the linear regression model:

$$y_i = \langle x_i, \beta^* \rangle + w_i\,, \quad i = 1, \dots, n\,,$$

for a given set of vectors $\{x_i\} \subset \mathbb{R}^p$ and an unknown weight vector β^*, where $w_1, \dots, w_n$ are i.i.d. mean-zero subgaussian random variables (r.v.'s). For convenience, we define the *design matrix* $X \in \mathbb{R}^p$, the i-th row of which is given by $x_i^\top$, and write

$$y = X\beta^* + w\,,$$

where $y := (y_1, \dots, y_n)$ and $w := (w_1, \dots, w_n)$. The lasso is defined as

$$\hat{\beta}_n \in \operatorname*{arg\,min}_{\beta} \left\{ f_n(\beta) \,\middle|\, \beta \in c\mathscr{B}_1 \subset \mathbb{R}^p \right\}, \tag{3.1}$$

for some $c > 0$ [170], where

$$f_n(\beta) := \frac{1}{2n} \sum_{i=1}^{n} \left(y_i - \langle x_i, \beta \rangle \right)^2 ,$$

and $\mathscr{B}_1$ denotes the unit ℓ_1-norm ball. One may view the ℓ_1-penalized least squares estimator studied in Chapter 2,

$$\hat{\beta}_{n,\mathrm{PLS}} \in \operatorname*{arg\,min}_{\beta} \left\{ f_n(\beta) + \tau_n \|\beta\|_1 \;\middle|\; \beta \in \mathbb{R}^p \right\}$$

for some $\tau_n > 0$, as the Lagrangian formulation of the lasso.

Interestingly, while the lasso is conceptually not very different from its Lagrangian form, existing non-asymptotic estimation error analyses for the latter cannot be directly applied to the former. The essential reason is that existing analyses require the restricted strong convexity (RSC) condition [25, 134], which does not hold for the lasso in general.

In this chapter, we introduce a novel *relaxed* RSC condition, based on which we derive non-asymptotic estimation error bounds for the lasso. The result shows that the lasso is minimax optimal if β^* is sparse, or if β^* is weakly sparse and $c = \|\beta^*\|_1$.

3.1.1 Related work

If $c = \|\beta^*\|_1$, and the noise w_i are independent and identically distributed (i.i.d.) standard normal random variables, the lasso is known to satisfy

$$\|\hat{\beta}_n - \beta^*\|_2 \le L\sigma \sqrt{\frac{s \log p}{n}}, \tag{3.2}$$

with high probability for some constant $L > 0$, where s is the number of non-zero entries in β^* [87]. The bound (3.2) shows that the lasso automatically adapts to β^*—the sparser β^* is, the smaller the estimation error bound.

This error bound (3.2), however, is not true in general when $c \neq \|\beta^*\|_1$. While (3.2) provides an $O((\sigma^2 n^{-1} \log p)^{\frac{1}{2}})$ error decaying rate, the minimax result in [151] shows that *no estimator* can achieve an error decaying rate better than $O((\sigma^2 n^{-1} \log p)^{\frac{1}{4}})$ for all $\beta^* \in c\mathscr{B}_1$. This gap is due to the possibility that β^* may lie strictly in $c\mathscr{B}_1$ or, in other words, $c > \|\beta^*\|_1$.

Therefore, a more general estimation error bound for the lasso is needed. Especially, a satisfactory estimation error bound for the lasso should be 1) *sharp* enough to recover (3.2) that varies with the sparsity of β^*, and 2) able to characterize the effect of the quantity $c - \|\beta^*\|_1$ on the estimation error.

Existing results do not provide such a satisfactory error bound. The proof in [87] for (3.2) fails when c is strictly larger than $\|\beta^*\|_1$. While the results in [148, 183] are valid as long as $c \ge \|\beta^*\|_1$,

the derived bounds are independent of β^*, and hence not sharp enough to recover (3.2). The small-ball approach yields an estimation error bound that depends on β^* [129, Theorem 4.6], but the dependence is implicit, and even whether it can recover (3.2) is unclear. The results in [120, 147] recover (3.2) when $c = \|\beta^*\|_1$; the dependence on $c - \|\beta^*\|_1$, however, is also vague.

The paper [151] assumed that β^* lies in an ℓ_q-norm ball $\mathscr{B}_q$, $q \in [0,1]$, and derived an estimation error bound for a *lasso-like* estimator, for which the ℓ_1-norm constraint in (3.1) is replaced by the corresponding ℓ_q-norm constraint. In contrast to [151], this paper will also consider the same assumption on β^*, but analyze the estimation performance of the lasso defined by (3.1) where an ℓ_1 norm constraint is used (cf. Corollary 3.15).

We are not aware of any existing work that discusses the estimation error of the lasso when $c < \|\beta^*\|_1$, though the analysis in [120] can be easily extended to this case, and yield an estimation error bound that is implicitly dependent on $\|\beta^*\|_1$. Note that in this case, the lasso cannot be consistent, i.e., the estimation error is always bounded away from zero no matter how large the sample size n is, because β^* is not a feasible solution of the optimization problem (3.1).

We note that while there are many well-studied estimators closely related to the lasso, such as the penalized lasso, Dantzig selector, square-root lasso, and basis pursuit-type estimators [22, 25, 50, 134, 177], the analysis techniques in the cited works cannot be directly applied to study the lasso when $c > \|\beta^*\|_1$. See Section 3.3.2 for a detailed discussion.

3.1.2 Contributions

The main result of this paper, Theorem 6.1, provides a *non-asymptotic* estimation error bound that is valid for any $c \geq \|\beta^*\|_1$, and for the case when β^* is not exactly sparse. It is sharp as it recovers (3.2) when $c = \|\beta^*\|_1$ (cf. Corollary 3.12). For the general case, it shows the following (cf. Corollary 3.15).

- For estimating any $\beta^* \in c\mathscr{B}_1$, the lasso is minimax optimal as long as $c \geq \|\beta^*\|_1$. The worst case (with respect to where β^* lies in $c\mathscr{B}_1$) error decaying rate is

$$\|\hat{\beta}_n - \beta^*\|_2 = O\left(\left(\frac{\sigma^2 \log p}{n}\right)^{\frac{1}{4}}\right).$$

- For estimating any *weakly sparse* $\beta^* \in c\mathscr{B}_1$ that is to mean it has bounded ℓ_q-norm for some $q \in (0,1]$, the lasso is minimax optimal if $c = \|\beta^*\|_1$. The worst case error decaying rate is

$$\|\hat{\beta}_n - \beta^*\|_2 = O\left(\left(\frac{\sigma^2 \log p}{n}\right)^{\frac{1}{2}-\frac{1}{4}q}\right).$$

Formal statements can be found in Section 3.4.

The results in this paper are non-asymptotic, i.e., the error bounds (and the corresponding probability bounds) are valid for all finite values of the sample size n, parameter dimension p, sparsity level s, and other parameters that will be specified in Section 3.4.

3.2 Preliminaries

Fix $\mathcal{K} \subseteq \mathbb{R}^p$ and $\lambda \in \mathbb{R}$. The notations $\mathcal{K} - v$ and $\lambda\mathcal{K}$ denote the sets $\{u - v : u \in \mathcal{K}\}$ and $\{\lambda u : u \in \mathcal{K}\}$, respectively. The notation $\overline{\mathcal{K}}$ denotes the conic hull of $\mathcal{K}$, i.e.,

$$\overline{\mathcal{K}} := \left\{ \rho v \;\middle|\; v \in \mathcal{K}, \rho \geq 0 \right\}.$$

The following notions about r.v.'s are necessary for our proof.

Definition 3.1. *A random vector $\eta \in \mathbb{R}^p$ is* isotropic*, if for any $v \in \mathbb{R}^p$,*

$$\mathsf{E} \langle \eta, v \rangle^2 = \|v\|_2^2.$$

Definition 3.2. *A random vector $\eta \in \mathbb{R}^p$ is* subgaussian*, if the r.v. $\langle \eta, v \rangle$ is subgaussian for all $v \in \mathbb{R}^p$. The* subgaussian norm *of a subgaussian random vector η is defined as*

$$\|\eta\|_{\psi_2} := \sup \left\{ \| \langle \eta, v \rangle \|_{\psi_2} \;\middle|\; v \in \mathbb{R}^p, \|v\|_2 = 1 \right\}.$$

Example 3.3. *A vector of either i.i.d. standard normal (Gaussian with zero mean and unit variance) or i.i.d. Rademacher (random sign) r.v.'s is subgaussian.*

The Gaussian width is useful when studying a collection of subgaussian r.v.'s indexed by a subset in the metric space $(\mathbb{R}^p, \|\cdot\|_2)$ [169, Theorem 2.4.1].

Definition 3.4 (Gaussian width). *The* Gaussian width *of a set $\mathcal{K} \subseteq \mathbb{R}^p$ is given by*

$$w(\mathcal{K}) := \mathsf{E} \sup \{ \langle g, v \rangle : v \in \mathcal{K} \},$$

where g is a vector of i.i.d. standard normal r.v.'s.

By Proposition 3.10 below, the Gaussian width of a set of the form $\mathcal{C} \cap \mathcal{B}_2$, where $\mathcal{C} \subseteq \mathbb{R}^p$ is a closed convex cone, characterizes the sample size required for the lasso to have a small estimation error. We always have $w(\mathcal{C} \cap \mathcal{B}_2) \leq \sqrt{p}$. By Proposition 3.10 and Theorem 3.11, this implies the possibility of doing estimation when $n < p$.

Proposition 3.5. *We have the following:*

1. *If $\mathcal{K}_1 \subseteq \mathcal{K}_2$, then $w(\mathcal{K}_1) \leq w(\mathcal{K}_2)$.*

2. *If* $\mathscr{K} = \mathbb{R}^p$, *then* $w(\mathscr{K} \cap \mathscr{B}_2) = \sqrt{p}$.

Proof. The first assertion is obvious by definition. The second assertion is because

$$w(\mathbb{R}^p \cap \mathscr{B}_2) = w(\mathscr{B}_2) = (1/\sqrt{p})\mathsf{E}\,\|g\|_2^2 = \sqrt{p},$$

where g is a vector of i.i.d. standard normal r.v.'s. □

3.3 Relaxed restricted strong convexity

The key notion for deriving the results in this paper is the *relaxed restricted strong convexity (RSC) condition* introduced in our unpublished work [120]. This section provides a brief discussion on the relaxed RSC condition, specialized for the lasso.

3.3.1 Definition of the relaxed RSC condition

Conventionally, linear regression is solved by the least-squares (LS) estimator, which works as long as the Hessian matrix $H_n := \nabla^2 f_n(\beta^*) \equiv n^{-1}X^T X$ is non-singular. Under the high-dimensional setting where $n < p$, however, the Hessian matrix H_n is always singular, and the LS approach fails, as illustrated by [43, Fig. 1].

The idea of the relaxed RSC condition is to require, *only in some directions*, that the Hessian matrix H_n behaves like a non-singular matrix.

Definition 3.6 (Feasible Set). *The* feasible set *is defined as*

$$\mathscr{F} := c\mathscr{B}_1 - \beta^* = \left\{ \beta - \beta^* \;\middle|\; \beta \in c\mathscr{B}_1 \right\}.$$

That is, the feasible set is the set of all possible error vectors.

Definition 3.7 (Relaxed RSC [120]). *The* (μ, t_n)-relaxed RSC condition *holds for some* $\mu > 0$ *and* $t_n \geq 0$, *if and only if for all* $v \in \mathscr{F} \setminus t_n\mathscr{B}_2$,

$$\langle \nabla f_n(\beta^* + v) - \nabla f_n(\beta^*), v \rangle \geq \mu \|v\|_2^2.$$

Remark 3.8. *The parameter* t_n *in general can scale with the sample size* n; *therefore the subscript* n *is added.*

Proposition 3.9. *The* (μ, t_n)*-relaxed RSC condition is equivalent to requiring*

$$\min \left\{ \frac{v^T H_n v}{\|v\|_2^2} \;\middle|\; v \in \mathscr{F} \setminus t_n\mathscr{B}_2 \right\} \geq \mu,$$

i.e., it requires the restricted smallest eigenvalue *of* H_n *with respect to* $\mathscr{F} \setminus t_n\mathscr{B}_2$ *is bounded below by* μ.

Proof. By direct calculation, we obtain

$$\langle \nabla f_n(\beta^* + v) - \nabla f_n(\beta^*), v \rangle = v^T H_n v. \qquad \square$$

The validity of assuming the relaxed RSC condition is verified by the following proposition, which shows that as long as the sample size n is sufficiently large (while it can be still less than p), the relaxed RSC condition can hold with high probability.

Proposition 3.10. *Suppose that the rows of the design matrix X are i.i.d., isotropic, and subgaussian with subgaussian norm $\alpha > 0$. There exist constants $c_1, c_2 > 0$ such that for any $\delta \in (0,1)$, if*

$$\sqrt{n} \geq c_1^2 \alpha^2 w(\overline{\mathscr{F} \setminus t\mathscr{B}_2} \cap \mathscr{B}_2), \tag{3.3}$$

for some $t \geq 0$, the $(1-\delta, t)$-relaxed RSC condition holds with probability at least $1 - \mathrm{e}^{-c_2 \delta^2 n / \alpha^4}$.

Proof. Assume that (3.3) is satisfied. By [130, Theorem 2.3], with probability at least $1 - \mathrm{e}^{-c_2 \delta^2 n}$, we have

$$\frac{\|Xv\|_2^2}{n} = \frac{\langle v, H_n v \rangle}{n} \geq (1-\delta)\|v\|_2^2,$$

for any $v \in \mathscr{F} \setminus t\mathscr{B}_2$. The proposition follows by Proposition 3.9. $\square$

3.3.2 Discussions

One interesting special case of Proposition 3.10 is when β^* has only $s < p$ non-zero entries and $c = \|\beta^*\|_1$. In this case, we can simply choose $t_n \equiv 0$; then $\overline{\mathscr{F} \setminus t_n \mathscr{B}_2}$ reduces to $\overline{\mathscr{F}}$, called the *tangent cone* in [50]. By [50, Proposition 3.10], the inequality (3.3) can be guaranteed, if

$$\sqrt{n} \geq c_1^2 \alpha^2 \sqrt{2s \log\left(\frac{p}{s}\right) + \frac{5}{4}s}.$$

Notice that the right-hand side can be much smaller than $\sqrt{p}$.

This observation is the main idea behind existing works on high-dimensional sparse parameter estimation in [177, 22, 25, 50, 134], to cite a few. Roughly speaking, the approach in the cited works can be summarized as follows.

1. Identify a convex cone $\mathscr{K}$ (possibly with a controlled small perturbation [134, 150]) in which the error vector $\tilde{\beta}_n - \beta^*$ lies, where $\tilde{\beta}_n$ denotes the estimator under consideration.

2. Derive a lower bound on the sample size n, such that the RSC (relaxed RSC with $t_n \equiv 0$, not necessary with respect to the ℓ_2-norm [177]) with respect to $\mathscr{K}$ holds with high probability.

3. Given that the RSC condition holds, the Hessian $H_n = n^{-1}X^TX$ behaves like a non-singular matrix with respect to the error vector, and classical approaches for analyzing the estimation error for the LS estimator applies.

While this existing approach is valid for analyzing the penalized least squares estimator, Dantzig selector, square-root lasso, and basis pursuit-type estimators as shown in [22, 177, 25, 50, 134], it is not applicable to the lasso. When $c > \|\beta^*\|_1$, the conic hull of the set of all possible error vectors of the lasso, $\hat{\beta}_n - \beta^*$, is the whole space $\mathbb{R}^p$, and hence requiring the relaxed RSC condition with t_n=0 is equivalent to requiring the non-singularity of the Hessian H_n, which cannot hold when $n \ll p$.

The next section shows that the relaxed RSC condition with a non-zero t_n suffices for deriving minimax optimal estimation error bounds for the lasso.

3.4 Main result and its implications

The main theorem requires the following assumptions to be satisfied.

Assumption 1. *The noise w is a vector of i.i.d. mean-zero subgaussian r.v.'s of unit subgaussian norm.*

Assumption 2. *The design matrix X is normalized, i.e., $\sum_j X_{i,j}^2 \le n$ for all $i \le p$.*

Assumption 3. *The (μ, t_n)-relaxed RSC condition holds for some $\mu, t_n > 0$.*

The first assumption on the noise is valid in the standard Gaussian linear regression model, where w is a vector of i.i.d. standard normal r.v.'s, and the persistence framework in [129], where w is a vector of i.i.d. mean-zero bounded r.v.'s. The second assumption is standard; recall that we have introduced this assumption in Chapter 2. We had discussed the validity of the third assumption in Section 3.3.

Theorem 3.11. *If Assumptions 1–3 are satisfied, then there exists a constant $c_3 > 0$ such that, for any $\tau > 0$ and $\mathcal{S} \subseteq \{1, \ldots, p\}$,*

$$\|\hat{\beta}_n - \beta^*\|_2 \le \max\left\{ t_n, \frac{c_3\sqrt{1+\tau}}{\mu} \cdot \sigma\sqrt{\frac{\log p}{n}}\gamma(t_n; \beta^*, \mathcal{S}) \right\}$$

with probability at least $1 - \mathrm{e}p^{-\tau}$, where

$$\gamma(t_n; \beta^*, \mathcal{S}) := 2\sqrt{|\mathcal{S}|} + \frac{2\|\beta^*_{\mathcal{S}^c}\|_1 + (c - \|\beta^*\|)}{t_n}. \tag{3.4}$$

Proof. See Section 3.A. □

Theorem 3.11 immediately recovers the well-known result (3.2) up to a constant scaling.

Corollary 3.12. *Suppose that β^* has s non-zero entries, and $c = \|\beta^*\|_1$ in* (3.1)*. Then if Assumptions 1–3 are satisfied, there exists a constant $c_3 > 0$ such that, for any $\tau > 0$, we have*

$$\|\hat{\beta}_n - \beta^*\|_2 \le \frac{2c_3\sqrt{1+\tau}}{\mu} \cdot \sigma\sqrt{\frac{s\log p}{n}},$$

with probability at least $1 - \mathrm{e}p^{-\tau}$.

Proof. Recall that in this case (cf. Section 3.3), the relaxed RSC can hold with $t_n \equiv 0$, as discussed in Section 3.3. Choosing $t_n \equiv 0$ and $\mathscr{S}$ as the support set of β^* in Theorem 3.11 completes the proof. □

In general, β^* may not be exactly sparse, and in practice, c can hardly be chosen as exactly $\|\beta^*\|_1$.

Definition 3.13 (Weak sparsity [134]). *A vector $v \in \mathbb{R}^p$ is q-*weakly sparse *for some $q \in [0,1]$, if and only if there exists some $C_q > 0$ such that $\|v\|_q := \sum_i |v_i|^q \le C_q$.*

Remark 3.14. *A 0-weakly sparse parameter is exactly sparse.*

Corollary 3.15. *Assume that β^* is q-weakly sparse for some $q \in [0,1]$, $\log p \ll n$, and Assumptions 1–3 are satisfied with*

$$t_n = \begin{cases} \Theta\left(\sqrt{C_q}\left(\frac{(1+\tau)\sigma^2\log p}{\mu^2 n}\right)^{\frac{1}{2}-\frac{1}{4}q}\right) & \text{if } c = \|\beta^*\|_1, \\ \Theta\left(\sqrt{\delta + C_q}\left(\frac{(1+\tau)\sigma^2\log p}{\mu^2 n}\right)^{\frac{1}{4}}\right) & \text{if } c > \|\beta^*\|_1. \end{cases} \tag{3.5}$$

where $\delta := c - \|\beta^*\|_1$ and $C_q := \|\beta^*\|_q$. Then we have, with probability at least $1 - \mathrm{e}p^{-\tau}$,

$$\|\hat{\beta}_n - \beta^*\|_2 = O(t_n)$$

for any $\tau \in (0,1)$.

Proof. See Section 3.B. □

Remark 3.16. *If t_n converges too fast to zero with respect to increasing n, the sample complexity bound* (3.3) *may not hold, and the validity of Assumption 3 in Corollary 3.15 would be in question. However, since*

$$w(\overline{\mathscr{F} \setminus t_n\mathscr{B}_2} \cap \mathscr{B}_2) = \frac{w(\overline{\mathscr{F} \setminus \mathscr{B}_2} \cap \mathscr{B}_2)}{t_n} = \Theta\left(\frac{1}{t_n}\right),$$

the sample complexity bound (3.3) *can hold as long as $t_n = O(n^{-1/2})$, which is satisfied in Corollary 3.15.*

The minimax error bound in [151, Theorem 3] shows that *no estimator* can achieve a better error decaying rate than

$$O\left(\sqrt{C_q}\left(\frac{\sigma^2 \log p}{n}\right)^{\frac{1}{2}-\frac{1}{4}q}\right)$$

with probability larger than 1/2 in the worst case, for estimating a q-weakly sparse parameter, $q \in (0,1]$. According to Corollary 3.15, this implies:

- The lasso with $c \geq \|\beta^*\|_1$ is minimax optimal (up to a constant scaling) for estimating a parameter with bounded ℓ_1-norm.
- The lasso with $c = \|\beta^*\|_1$ is minimax optimal (up to a constant scaling) for estimating a q-weakly sparse parameter, $q \in (0,1]$.

Note that the error decaying rates in the two assertions are for the worst case. It is possible to have a better error decaying rate in special cases, as shown by Corollary 3.12.

3.5 Discussions

We have focused on the case where the design matrix X has subgaussian rows and the noise w has subgaussian entries. This is simply for convenience of presentation, and the analysis framework can be easily extended to more general cases.

Proposition 3.10, which shows the validity of the relaxed RSC condition, can be easily extended for design matrices whose rows are not necessarily subgaussian, with a possibly worse sample complexity bound compared to (3.3). The interested reader is referred to [108, 145, 166] for the details.

Theorem 3.11 can be easily extended for possibly non-subgaussian noise. One only needs to replace the Hoeffding-type inequality in the proof of Proposition 3.20 by Bernstein's inequality [126] or other appropriate concentration inequalities for sums of independent r.v.'s. Note that the obtained estimation error bound may be worse, as shown in [122].

Finally, we remark that by Proposition 3.10 and the union bound, Theorem 3.11 also implies an estimation error bound for the *random design* case, where the design matrix X is a random matrix independent of the noise w. Such an error bound can be useful for compressive sensing, where the design matrix is not given, but can be chosen by the practitioner.

Corollary 3.17. *Suppose the rows of the design matrix X are i.i.d., isotropic, and subgaussian with subgaussian norm $\alpha > 0$, and X is independent of the noise w. Then there exist constants $c_1, c_2, c_3 > 0$ such that, if* (3.3) *and Assumptions 2 and 3 are satisfied, for any $\tau > 0$ and $\mathscr{S} \subseteq$*

$\{1,\dots,p\}$, *we have*

$$\|\hat{\beta}_n - \beta^*\|_2 \le \max\left\{ t_n, \frac{c_3\sqrt{1+\tau}\sigma}{1-\delta}\sqrt{\frac{\log p}{n}}\gamma(t_n;\beta^*,\mathscr{S}) \right\}$$

with probability at least $1 - \mathrm{e}p^{-\tau} - \exp(-c_2\delta^2 n/\alpha^4)$ *(with respect to the design matrix* X *and the noise* w*), where* $\gamma(t_n;\beta^*,\mathscr{S})$ *is defined as in* (3.4).

Corollary 3.15 can be extended for the random design case in the same manner.

3.A Proof of Theorem 3.11

Define $\Delta_n := \hat{\beta}_n - \beta^*$ for convenience.

By definition, Δ_n lies in either $t_n\mathscr{B}_2$ or $\mathscr{F} \setminus t_n\mathscr{B}_2$. In the former case, it holds trivially that $\|\Delta_n\|_2 \le t_n$. We now consider the latter case.

Proposition 3.18. *If the* (μ, t_n)*-relaxed RSC condition holds for some* $\mu, t > 0$*, and if* $\Delta_n \in \mathscr{F} \setminus t\mathscr{B}_2$*, then we have*

$$\|\Delta_n\|_2 \le \frac{1}{\mu}\frac{\|\Delta_n\|_1}{\|\Delta_n\|_2}\frac{\langle -\nabla f_n(\beta^*), \Delta_n\rangle}{\|\Delta_n\|_1}. \tag{3.6}$$

Proof. By the relaxed RSC condition, we have

$$\langle \nabla f_n(\hat{\beta}_n) - \nabla f_n(\beta^*), \Delta_n\rangle \ge \mu\|\Delta_n\|_2. \tag{3.7}$$

Since (3.1) defines a convex optimization problem, we have, by the optimality condition of $\hat{\beta}_n$ [136],

$$\langle -\nabla f_n(\hat{\beta}_n), \Delta_n\rangle \ge 0. \tag{3.8}$$

Summing up (3.7) and (3.8), we obtain

$$\langle -\nabla f_n(\beta^*), \Delta_n\rangle \ge \mu\|\Delta_n\|_2^2,$$

which implies

$$\|\Delta_n\|_2 \le \frac{1}{\mu}\frac{\|\Delta_n\|_1}{\|\Delta_n\|_2}\frac{\langle -\nabla f_n(\beta^*), \Delta_n\rangle}{\|\Delta_n\|_1}.$$

This completes the proof. □

The rest of this subsection is devoted to deriving an upper bound of the right-hand side of (3.6), which is independent of Δ_n.

We first derive a bound on $(\|\Delta_n\|_1 / \|\Delta_n\|_2)$.

Proposition 3.19. *The estimation error satisfies*

$$\|\Delta_n\|_1 \le 2(\|(\Delta_n)_{\mathscr{S}}\|_1 + \|\beta^*_{\mathscr{S}^c}\|_1) + (c - \|\beta^*\|),$$

for any $\mathscr{S} \subseteq \{1,\dots,p\}$, *where* $\mathscr{S}^c := \{1,\dots,p\} \setminus \mathscr{S}$.

Proof. By definition, we have $\hat{\beta}_n \in c\mathscr{B}_1$, and hence

$$\begin{aligned} c \ge \|\hat{\beta}_n\|_1 &= \|(\beta^* + \Delta_n)_{\mathscr{S}} + (\beta^* + \Delta_n)_{\mathscr{S}^c}\|_1 \\ &\ge \|\beta^*_{\mathscr{S}} + (\Delta_n)_{\mathscr{S}^c}\|_1 - \|\beta^*_{\mathscr{S}^c} + (\Delta_n)_{\mathscr{S}}\|_1 \\ &= \|\beta^*_{\mathscr{S}}\|_1 + \|(\Delta_n)_{\mathscr{S}^c}\|_1 - \|\beta^*_{\mathscr{S}^c}\|_1 - \|(\Delta_n)_{\mathscr{S}}\|_1 \\ &= \|\beta^*\|_1 - 2\|\beta^*_{\mathscr{S}^c}\|_1 + \|\Delta_n\|_1 - 2\|(\Delta_n)_{\mathscr{S}}\|_1, \end{aligned}$$

which proves the proposition. □

By Proposition 3.19, we obtain

$$\begin{aligned} \frac{\|\Delta_n\|_1}{\|\Delta_n\|_2} &\le 2\frac{\|(\Delta_n)_{\mathscr{S}}\|_1}{\|\Delta_n\|_2} + \frac{\|2\beta^*_{\mathscr{S}^c}\|_1 + (c - \|\beta^*\|)}{\|\Delta_n\|_2} \\ &\le 2\frac{\|(\Delta_n)_{\mathscr{S}}\|_1}{\|(\Delta_n)_{\mathscr{S}}\|_2} + \frac{2\|\beta^*_{\mathscr{S}^c}\|_1 + (c - \|\beta^*\|)}{t_n} \\ &\le 2\sqrt{|\mathscr{S}|} + \frac{2\|\beta^*_{\mathscr{S}^c}\|_1 + (c - \|\beta^*\|)}{t_n}, \end{aligned} \tag{3.9}$$

if $\Delta_n \in \mathscr{F} \setminus t_n\mathscr{B}_2$.

Now we bound the term $\langle -\nabla f_n(\beta^*), \Delta_n\rangle / \|\Delta_n\|_1$.

Proposition 3.20. *If the design matrix* X *is normalized, i.e.,* $\sum_j X^2_{i,j} \le n$ *for all* $i \le p$, *there exists a universal constant* $c_3 > 0$ *such that for any* $\tau > 0$, *we have*

$$\frac{\langle -\nabla f_n(\beta^*), \Delta_n\rangle}{\|\Delta_n\|_1} \le c_3\sigma\sqrt{\frac{(1+\tau)\log p}{n}},$$

with probability at least $1 - \mathrm{e}p^{-\tau}$.

Proof. We note that

$$\begin{aligned} \frac{\langle -\nabla f_n(\beta^*), \Delta_n\rangle}{\|\Delta_n\|_1} &\le \sup\left\{ \langle -\nabla f_n(\beta^*), \nu\rangle \;\middle|\; \|\nu\|_1 = 1 \right\} \\ &= \| -\nabla f_n(\beta^*)\|_\infty. \end{aligned}$$

By a direct calculation, we obtain

$$(\nabla f_n(\beta^*))_i = \frac{1}{n}\sum_{j=1}^{n} X_{i,j} w_j ,$$

for all $i \le p$; hence, by the Hoeffding-type inequality for subgaussian r.v.'s (cf. Theorem A.5), there exists a universal constant $L > 0$ such that for any $\varepsilon > 0$,

$$\mathsf{P}\left\{ |(\nabla f_n(\beta^*))_i| \ge \varepsilon \right\} \le \mathrm{e} \cdot \exp\left(-\frac{L\varepsilon^2 n}{\sigma^2} \right).$$

By the union bound, this implies

$$\begin{aligned} \mathsf{P}\left\{ \|\nabla f_n(\beta^*)\|_\infty \ge \varepsilon \right\} &\le \sum_{i=1}^{p} \mathsf{P}\left\{ |(\nabla f_n(\beta^*))_i| \ge \varepsilon \right\} \\ &\le \mathrm{e} \cdot \exp\left(-\frac{L\varepsilon^2 n}{\sigma^2} + \log p \right). \end{aligned}$$

Choosing

$$\varepsilon = \sigma \sqrt{\frac{(1+\tau)\log p}{Ln}}$$

completes the proof. □

Theorem 3.11 follows by combining (3.9) and Proposition 3.20.

3.B Proof of Corollary 3.15

Define

$$\mathscr{S}_n := \left\{ i \;\middle|\; |\beta_i^*| \ge \rho_n \right\},$$

for some $\rho_n > 0$. Then we have

$$|\mathscr{S}_n| \le C_q \rho_n^{-q},$$

as

$$C_q \ge \sum_{i \in \mathscr{S}_n} |\beta_i^*|^q \ge |\mathscr{S}_n| \rho_n^q.$$

Moreover, we have

$$\begin{aligned}\|\beta^*_{\mathscr{S}_n^c}\|_1 &= \sum_{i\in\mathscr{S}_n^c} |\beta_i^*|^q |\beta_i^*|^{1-q} \\ &\le \sum_{i\in\mathscr{S}_n^c} |\beta_i^*|^q \rho_n^{1-q} \\ &\le C_q \rho_n^{1-q}.\end{aligned}$$

Applying Theorem 3.11 with $\mathscr{S} = \mathscr{S}_n$, we obtain

$$\begin{aligned}\|\hat{\beta}_n - \beta^*\|_2 &\le \max\left\{ t, \frac{c_3\sqrt{1+\tau}\sigma}{\mu}\sqrt{\frac{\log p}{n}}\gamma_n \right\} \\ &\le t_n + \frac{c_3\sqrt{1+\tau}\sigma}{\mu}\sqrt{\frac{\log p}{n}}\gamma_n, \end{aligned} \tag{3.10}$$

with probability at least $1 - \mathrm{e}p^{-\tau}$, where

$$\gamma_n := 2\sqrt{C_q \rho_n^{-q}} + \frac{2C_q\rho_n^{1-q} + (c - \|\beta^*\|_1)}{t_n}.$$

The corollary follows by optimizing over t_n and ρ_n by the inequality for arithmetic and geometric means on (3.10). Specifically, the best possible error decaying rate can be achieved when

$$\rho_n = \Theta\left(\left(\frac{(1+\tau)\sigma^2 \log p}{\mu^2 n}\right)^{\frac{1}{2}}\right),$$

and t_n is chosen as in (3.5).

4 A Frank-Wolfe algorithm for Poisson phase retrieval

In this and the following two chapters, we address convex optimization problems where the objective functions are not smooth with respect to Definition 1.1. In particular, we will mainly focus on the exp-linear function: $f(x) = -\log\langle a, x\rangle$ for some vector a or $f(X) = -\log\mathrm{Tr}(AX)$ for some matrix A. Notice that an exp-linear function is not smooth, if $\langle a, x\rangle$ (or $\mathrm{Tr}(AX)$) is allowed to be arbitrarily close to zero; therefore, many existing analyses of convex optimization algorithms do not apply. Section 4.3 provides a detailed discussion, regarding why existing algorithms are not guaranteed to converge for exp-linear functions.

As discussed in Chapter 1, exp-linear functions appear in many applications. In this chapter, we consider the specific application of phase retrieval with Poisson noise. Adopting the idea of PhaseLift, the maximum-likelihood estimator is computed via minimizing a sum of exp-linear losses on a nuclear norm ball. In practice, the dimension of the parameter is typically so high that computing the projection onto the nuclear norm ball is computationally expensive. The Frank-Wolfe algorithm then becomes a competitive choice, as it avoids the projection step, unlike most existing optimization algorithms.

Unfortunately, existing convergence guarantees for the Frank-Wolfe algorithm require the objective function to be smooth. We prove in this chapter that, with a slightly modified step size selection rule, the Frank-Wolfe algorithm provably converges for a prototype convex optimization problem, of which Poisson phase retrieval is a special case.

This chapter is based on the joint work with Gergely Odor *et al.* [143].

4.1 Introduction

Phase retrieval is the problem of estimating a complex-valued signal from intensity measurements, which arises in many applications such as X-ray crystallography, diffraction imaging, astronomical imaging, and many others [165].

We focus on the Poisson noise model in this paper. Formally speaking, we are interested

in estimating a signal $x^\natural \in \mathbb{C}^p$, given $a_1, \ldots, a_n \in \mathbb{C}^p$ and measurement outcomes $y_1, \ldots, y_n$, modeled as independent random variables following the Poisson distribution:

$$\mathbb{P}\{y_i = y\} = \frac{\mathrm{e}^{-\lambda_i}\lambda_i^y}{y!}, \quad y \in \{0\} \cup \mathbb{N},$$

where $\lambda_i := |\langle a_i, x^\natural \rangle|^2$ for all i. In practice, each y_i represents the number of photons detected by the sensor [73].

The maximum-likelihood (ML) estimation approach yields a non-convex optimization problem that is difficult to solve. A recent approach to circumvent this computational issue is PhaseLift [39, 44]. The PhaseLift approach casts the phase retrieval problem as a low rank matrix recovery problem, for which we can apply any convex optimization-based estimator, such as the basis pursuit like estimator [154], nuclear-norm penalized estimator [42], and lasso-like estimator [63].

Following the PhaseLift approach, we show in Section 4.2 that we can recover $x^\natural$ by solving

$$\hat{X} \in \operatorname*{arg\,min}_X \left\{ f(X) \;\middle|\; X \in \mathscr{X} \right\}, \tag{4.1}$$

where

$$f(X) := \sum_{i=1}^n \left\{ -y_i \log\left[\mathrm{Tr}\,(A_i X)\right] + \mathrm{Tr}\,(A_i X) \right\}, \tag{4.2}$$

$$\mathscr{X} := \left\{ X \in \mathbb{C}^{p\times p} \;\middle|\; X \geq 0,\ \|X\|_* \leq c \right\}, \tag{4.3}$$

for some $c > 0$, $A_i := a_i a_i^{\mathrm{H}}$. The notation $\|\cdot\|_*$ denotes the nuclear norm—the sum of singular values. A rule of thumb for choosing the parameter c is presented in Section 4.2.1. We then find an eigenvector associated with the largest eigenvalue of $\hat{X}$ as our estimate of $x^\natural$.

It is easy to check that (4.1) is a convex optimization problem. Existing convex optimization tools, however, are not directly applicable due to two issues.

1. Most existing algorithms, such as [172], are computationally expensive for nuclear norm constraints, as they require computing the eigenvalue decomposition of a matrix in $\mathbb{C}^{p\times p}$ at each iteration.

2. While Frank-Wolfe-type algorithms are relatively scalable for nuclear norm constraints [98], existing theoretical convergence guarantees for these Frank-Wolfe-type algorithms are not valid for our loss function in (4.1).

We will address the issues in detail in Section 4.3.

In this chapter, we show that the standard Frank-Wolfe algorithm converges for the optimization problem (4.1), with a properly chosen parameter to be explicitly specified in Theorem 4.2.

Our theorem guarantees that the Frank-Wolfe algorithm converges at the rate $O(1/t)$ globally, where t is the iteration counter. Numerical experiments show that the empirical convergence rate can be even faster. The algorithm shares the same merit of the standard Frank-Wolfe algorithm, in the sense that it is scalable when dealing with a nuclear norm constraint.

To the best of our knowledge, this is the first theoretical guarantee for the Frank-Wolfe algorithm applied to a non-Hölder (and hence non-Lipschitz) continuous gradient objective function.

4.2 Poisson phase retrieval by convex optimization

For the Poisson noise model, the ML estimator of $x^\natural$ is given by

$$\hat{x}_{\mathrm{ML}} \in \arg\min_x \left\{ L(x) \;\middle|\; x \in \mathbb{C}^p \right\}, \tag{4.4}$$

where L is the negative log-likelihood function (up to a constant shift):

$$L(x) := \sum_{i=1}^{n} \left[-y_i \log\left(|\langle a_i, x\rangle|^2 \right) + |\langle a_i, x\rangle|^2 \right].$$

The function L, unfortunately, is non-convex.

Motivated by the PhaseLift approach [39, 44], we can reformulate the non-convex optimization problem (4.4) as follows. Define $A_i := a_i a_i^{\mathrm{H}}$ for all i, and $X^\natural := x^\natural (x^\natural)^{\mathrm{H}}$. We have

$$|\langle a_i, x^\natural\rangle|^2 = \mathrm{Tr}\left(A_i X^\natural \right), \quad \forall\, i,$$

and hence we can rewrite the original optimization problem as

$$\hat{x}_{\mathrm{ML}} \in \arg\min_x \left\{ f(X) | X = x x^{\mathrm{H}}, x \in \mathbb{C}^p \right\},$$

where f is given in (4.2). This is equivalent to the optimization problem

$$\hat{X}_{\mathrm{ML}} \in \arg\min_X \left\{ f(X) | X \geq 0, \mathrm{rank}(X) = 1, X \in \mathbb{C}^{p\times p} \right\}.$$

Note that given $\hat{X}_{\mathrm{ML}}$, $\hat{x}_{\mathrm{ML}}$ can be recovered via the relation $\hat{X}_{\mathrm{ML}} = \hat{x}_{\mathrm{ML}} \hat{x}_{\mathrm{ML}}^{\mathrm{H}}$.

As the variable X is always of rank 1, we consider the convex relaxation given in (4.1). We then find an eigenvector associated with the largest eigenvalue of $\hat{X}$ as our estimate of $x^\natural$.

It is easily verified that (4.1) is a convex optimization problem.

4.2.1 A rule of thumb for setting the constraint

In the convex optimization formulation (4.1), we leave one parameter c unspecified. The ideal setting should be $c = \|X^\natural\|_* = \|x^\natural\|_2^2$. While this setting may not be practically feasible, we need $c > \|x^\natural\|_2^2$ to ensure that $X^\natural$ is in the constraint set $\mathscr{X}$.

The following theorem shows that choosing $c = (1/n)\sum_{i=1}^n y_i$ suffices, if the sampling scheme satisfies an isometry property with high probability.

Proposition 4.1. *Let $A \in \mathbb{C}^{n\times p}$, whose i-th row is given by a_i^H. Assume that there exists some $\varepsilon > 0$ such that*

$$(1-\varepsilon)\|x^\natural\|_2^2 \le \left\| \frac{1}{\sqrt{n}} Ax^\natural \right\|_2^2 \le (1+\varepsilon)\|x^\natural\|_2^2 \tag{4.5}$$

with probability at least $1 - p_\varepsilon$. Then we have, for any $t > 0$,

$$\bar{y} := \frac{1}{n}\sum_{i=1}^n y_i > (1+\varepsilon)\|x^\natural\|_2^2 + t$$

with probability at least $1 - p_\varepsilon - p_t$, where

$$p_t := \exp\left[-\frac{nt}{4} \log\left(1 + \frac{t}{2(1+\varepsilon)\|x^\natural\|_2^2} \right) \right].$$

We defer the proof to Section 4.A. If $x^\natural$ is sparse, the isometry condition (4.5) can be implied by the restricted isometry property (RIP) of A [38, 76, 160]. Even without sparsity, if n is sufficiently large, a matrix A of independent and identically distributed (i.i.d.) subgaussian random variables can also satisfy (4.5) with high probability (see, e.g., [76]).

While the isometry property of the Fourier measurement with a coded diffraction pattern is unclear currently, we show via numerical experiments in Section 4.5 that this rule of thumb works well on both synthetic and real-world data.

4.3 Review of convex optimization tools

We address why several existing convex optimization algorithms are not applicable to (4.1) in this section.

Note that (4.1) is a constrained convex minimization problem with a differentiable loss function, and there are many well-known algorithms for solving such a problem. State-of-the-art choices for large-scale applications include the proximal gradient-type methods [7, 20, 56, 137, 139, 172], alternating direction method of multipliers (ADMM) [69], and Frank-Wolfe-type algorithms (a.k.a. conditional gradient methods) [77, 79, 86, 98, 138, 187]. There are also well-developed MATLAB packages available on the Internet [21, 172]. Those seemingly

Algorithm 1 (Frank-Wolfe algorithm)

Choose an arbitrary $x_0 \in \mathscr{X}$
for $t = 0, \ldots, T$ **do**
  Compute $v_t \in \arg\min_s \{ \langle s, \nabla f(x_t) \rangle \mid s \in \mathscr{X} \}$
  Update $x_{t+1} = (1 - \tau_t) x_t + \tau_t v_t$
end for

ready-to-use convex optimization tools, however, are not desirable for solving our problem (4.1) for two issues.

The first issue is scalability. When applied to the problem (4.1), both proximal gradient-type methods and the ADMM require computing the *prox-mapping* given by

$$\mathrm{prox}(X) := \arg\min_S \{ \omega(S - X) : S \in \mathscr{X} \}$$

for a given strongly convex "distance generating function" (DGF) ω. A standard choice of DGF for matrix variables is $\omega(X) := (1/2)\|X\|_{\mathrm{F}}^2$, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. For a positive semi-definite matrix $X \in \mathbb{C}^{p \times p}$, whose eigenvalue decomposition is $X = U \mathrm{diag}(v) U^{\mathrm{H}}$, we have $\mathrm{prox}(X) = U \mathrm{diag}(\tilde{v}) U^{\mathrm{H}}$, where $\tilde{v}$ is the Euclidean projection of v onto the probability simplex in $\mathbb{R}^p$ scaled by c. While the prox-mapping is simple to describe, the eigenvalue decomposition renders the algorithm slow when the parameter dimension p is large, as its computational complexity is in general $O(p^3)$. Similar issues exist when we choose other DGFs. See the next chapter for an example, where ω is chosen as the quantum relative entropy.

Scalability is a major reason why Frank-Wolfe-type algorithms have been attracting attention in recent years. We summarize the Frank-Wolfe algorithm (when applied to (4.1)) in Algorithm 1, where $\{\tau_t\}$ is a sequence of real numbers in $(0, 1]$ to be specified. There is a slight abuse of notations; when applied to our specific problem (4.1), the variables $x_0, \ldots, x_t$ and $\nabla f(x_t)$ should be understood as their matrix counterparts $X_0, \ldots, X_t$ and $\nabla f(X_t)$, respectively.

The computational bottleneck is in computing v_t (or its matrix counterpart V_t). For the specific constraint set $\mathscr{X}$ given in (4.3) and any positive semi-definite matrix X_t, it can be easily verified that V_t is a scaled rank-one approximation of $\nabla f(X_t)$, and hence can be efficiently computed by the Lanczos method [98]. More precisely, let $u_t \in \mathbb{C}^p$ be an eigenvector of $\nabla f(X_t)$ associated with the largest eigenvalue. We have $V_t = c(u_t u_t^{\mathrm{H}})$.

Unfortunately, the second issue arises: none of the existing theoretical convergence guarantees for Frank-Wolfe-type algorithms, to the best of our knowledge, is valid for the specific loss function (4.2). The result in [98] requires a bounded curvature condition; [77, 79, 86] require the gradient of the objective function to be Lipschitz continuous; [138] requires a weaker condition that the gradient is Hölder continuous; the Frank-Wolfe like algorithm in [187] requires the gradient of the conjugate of the objective function to be Hölder continuous. However, the objective function given in (4.2) does not satisfy the Hölder gradient conditions.

The second issue also exists for proximal gradient-type methods and the ADMM, as [7, 20, 56, 69, 137, 139] also require the Lipschitz continuity of the gradient. The only exception is SCOPT—a proximal gradient method for composite self-concordant minimization—proposed in [172]. Notice that the logarithmic function is a typical example of self-concordant functions.

Recently, there have been some other computationally efficient approaches to phase retrieval under the noiseless or additive-noise setting [11, 41, 53, 81, 141]. The theoretical guarantees therein do not directly extend for the Poisson noise case. After we finished this work, a gradient descent-type algorithm aiming at directly computing the ML estimator was proposed in [54].

4.4 Convergence guarantee

In this section, we provide a convergence guarantee for the Frank-Wolfe algorithm in Algorithm 1, for the prototype constrained convex optimization optimization problem:

$$g^\star := \min_X \left\{ g(X) \;\middle|\; X \in \mathscr{C} \right\} \tag{4.6}$$

where $\mathscr{C}$ is a nuclear norm ball in $\mathbb{R}^{p\times p}$, and

$$g(X) := \mathrm{Tr}(\Psi X) - \sum_{i=1}^{n} \eta_i \log \mathrm{Tr}(\Phi_i X) \tag{4.7}$$

for some $\Psi \in \mathbb{R}^{p\times p}$, non-negative integers $\eta_1,\ldots,\eta_n$, and positive semi-definite matrices $\Phi_1,\ldots,\Phi_n \in \mathbb{R}^p$.

We start with some definitions. Define $d_{\mathscr{C}}$ as the diameter of $\mathscr{C}$, i.e.,

$$d_{\mathscr{C}} := \max_{X,Y} \{ \|X - Y\|_2 \mid X, Y \in \mathscr{C} \},$$

where $\|\cdot\|_2$ denotes the spectral norm—the operator norm induced by the ℓ_2-norm. Let $d_\Phi := \max_i \|\Phi_i\|$ and $d_\Psi := \|\Psi\|$. Furthermore, we define

$$\bar{\mu} := \max_{i,x} \{ \mathrm{Tr}(\Phi_i X) \mid 1 \le i \le n, X \in \mathscr{C} \},$$
$$\underline{\mu} := \min_i \{ \mathrm{Tr}(\Phi_i X_0) \mid 1 \le i \le n \}.$$

Notice that we need to choose the initial iterate X_0 such that $\underline{\mu} > 0$, due to the presence of logarithmic functions in g.

Our main theoretical result is the following theorem:

Theorem 4.2. *Consider the optimization problem (4.6). The iterates $(X_t)_{t\ge 0}$ given by Algorithm 1 with*

$$\tau_t := \frac{2}{t+3}$$

satisfies

$$g(X_t) - g^\star < \frac{8\gamma^2 d_\Phi^2 d_{\mathscr{C}}^2}{t+2} + \frac{2d_{\mathscr{C}}\|\nabla g(X_0)\|}{\underline{\mu}(t+1)(t+2)}$$

The quantity $\gamma := \max\{\gamma_1, \gamma_2, \gamma_3\}$ *is a constant independent of* t*, where*

$$\gamma_1 := \frac{2d_\Psi d_{\mathscr{C}}}{\underline{\mu}}, \quad \gamma_2 := 2\frac{nd_\eta}{\underline{\mu}}\left(\frac{4n\bar{\mu}d_\eta}{\underline{\mu}} + 1\right)^2,$$
$$\gamma_3 := \frac{64n^2\bar{\mu}^2 d_\eta^2}{\underline{\mu}^3}\left(\frac{4n\bar{\mu}d_\eta}{\underline{\mu}} + 1\right).$$

Consequently, we have $g(X_t) - g^\star = O(1/t)$.

Remark 4.3. *Our choice of* τ_t *is slightly different from the standard one in [98, 138], where* $\tau_t := 2/(t+2)$*. This is due of technical concerns in the proof.*

Theorem 4.2 establishes the validity of using the Frank-Wolfe algorithm to solve (4.1). We note that this theorem is a *worst case* guarantee for all loss functions of the form (4.7). As we will see in the next section, empirically, the constants and the convergence rate can be much better.

We defer the proof to Section 4.B. The key idea in the proof is to show the boundedness of $\|\nabla g(X_{t+1}) - \nabla g(X_t)\|$ for all t, where $\|\cdot\|$ denotes the spectral norm. This bound, by the framework in [138], is sufficient to establish the convergence guarantee. This is simple if the gradient is Hölder continuous, since then

$$\|\nabla g(X_{t+1}) - \nabla g(X_t)\| \le L_\nu \|X_{t+1} - X_t\|_*^\nu \le L_\nu d_{\mathscr{C}}^\nu$$

for some $\nu \in (0,1]$ and $L_\nu > 0$. For the optimization problem (4.1) we consider, this issue can be reduced to the boundedness of

$$C_t := \sum_{i=1}^n \frac{\eta_i}{\mathrm{Tr}(\Phi_i X_t)}$$

for all t. We complete the proof by showing that C_t is bounded above by a constant for all t, if we choose $\tau_t = 2/(t+3)$.

4.5 Numerical results

In this section, we present numerical evidence to assess the convergence behaviour and the scalability of the proposed Frank-Wolfe algorithm.

Our numerical experiment is based on coded diffraction pattern measurements with the octonary modulation, which were considered in [41, 187] for the noiseless model. A similar

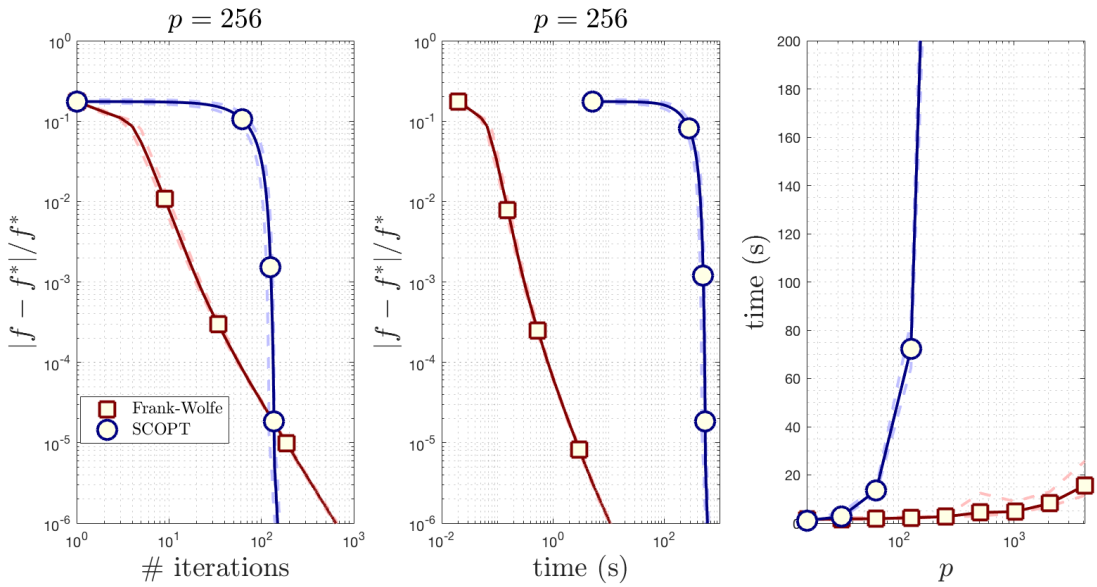


Figure 4.1 – Convergence behaviours of the algorithms for different dimensions in the first experiment. Solid lines show the average performance over 10 random trials, and the two dashed lines show the best and the worst performances, respectively.

setup was also considered also in [40] for the Poisson noise model.

In [40], the MATLAB package TFOCS [21] was used to solve a convex optimization problem similar to (4.1). The algorithm, however, is not guaranteed to converge for the problem under our consideration (cf. Section 4.3). We compare the Frank-Wolfe algorithm with SCOPT—a proximal gradient algorithm for composite self-concordant minimization—proposed in [172]. Recall that the objective function is self-concordant, and hence the algorithms in [172] are applicable.

In the first experiment, we consider the random Gaussian signal model: We generate $x^\natural$ as a random vector in $\mathbb{R}^p$, the real and imaginary parts of the elements of which are independent and identically standard normal random variables. We run both algorithms starting from the same Gaussian initial iterate, sampled from the same distribution as $x^\natural$. We keep track of the objective value and the elapsed time over the iterations, and compute the approximate relative objective residual ($|f - f^*|/|f^*|$) as the performance measure, where the actual optimum value $f^\star$ is approximated by f^*, the minimum objective value obtained by running 200 iterations of the SCOPT and/or 10000 iterations of the Frank-Wolfe algorithm.

In the second experiment, we test the scalability of the Frank-Wolfe approach, by recovering a real image as in [41, 187]. We choose the EPFL campus image of 1323×1984 pixels as the signal to be measured, which corresponds to a signal dimension $p = 2624832$. We apply the Frank-Wolfe algorithm to recover three color channels separately, and stop the algorithm when the recovery error ($\|x - x^\natural\|_\mathrm{F}/\|x^\natural\|_\mathrm{F}$) reached 10^{-2}.

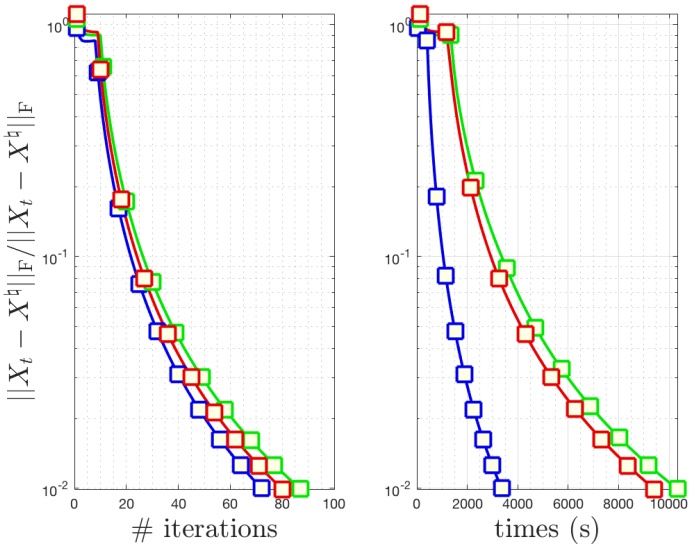


Figure 4.2 – Each color (blue, green, red) represents one color channel.

In both experiments, we set the constraint parameter c to the mean of the measurements, following the rule of thumb in Section 4.2.1, and we set the number of different modulating waveforms L to 20.

We implement Algorithm 1 in MATLAB and use the built-in `eigs` function, which is based on the Lanczos algorithm, with 10^{-3} relative error tolerance, to perform the minimization step of the Frank-Wolfe algorithm. In the weighting step, we adapt the efficient thin singular value decomposition updating method of [29] under low rank modifications, as explained in [187], in order to tame the memory growth.

We time our experiments on a computer cluster, and restricting the computational resource to 8 CPU of 2.40 GHz and 32 GB of memory space per simulation.

Figure 4.1 illustrates the convergence behaviour of the algorithms for different data sizes.

The first three plots on the left correspond to the first experiment. Solid lines show the average performance over 10 random trials, and the two dashed lines show the best and the worst instances, respectively. In the first two plots, we observe that the empirical rate of convergence is about $O(t^{-1.89})$, which is better than the theoretically guaranteed rate $O(t^{-1})$. In the third plot, we show the time required to reach a predefined accuracy level of 10^{-5} in terms of the relative objective residual, for different data sizes.

The last two plots of Figure 4.1 correspond to the second experiment, which also provides an empirical evidence for the estimation quality using the constraint parameter c. Each color (blue, green, red) represents one color channel.

Figure 4.3 – An EPFL image of size 1323 × 1984, reconstructed by 75 iterations of the Frank-Wolfe algorithm: PSNR = 44.92 dB.

Finally, Figure 4.3 shows the estimate x_t, after 75 iterations of the Frank-Wolfe method. The PSNR of the reconstructed image is 44.92dB.

Notice that, considering the lifted dimensions p^2 in the second experiment, even the generation of a simple iterate X_t would require approximately 7 TB of memory space, for a single color channel, when using the prox-mapping-based solver in SCOPT. By avoiding the computation of the prox-mapping, and adapting the efficient low rank updates, the Frank-Wolfe algorithm keeps a low memory footprint, and hence is more scalable compared to the self-concordant optimization method in SCOPT.

4.6 Discussions

The dependence of the convergence rate on the number of summands n is $O(n^6)$. This is unsatisfactory for applications where the sample size is large—in machine learning applications, typically, the number of summands n corresponds to the sample size. A future work is to study the tightness of the convergence rate guarantee.

While we focus on the Poisson phase retrieval problem in this chapter, the main contribution is in verifying the validity of applying the Frank-Wolfe algorithm to optimization problems of the form (4.1). Therefore, the application of Theorem 4.2 is not restricted to Poisson phase

retrieval. One interesting application is quantum state tomography [96]. We will see in the next chapter that, unfortunately, the per-iteration computational efficiency of the Frank-Wolfe algorithm does not compensate the relatively slow $O(1/t)$ convergence rate, empirically on real data-sets.

4.A Proof of Proposition 3.1

Notice that, conditioning on $a_1, \dots, a_n$, the random variable $n\bar{y}$ is Poisson with mean $\sum_{i=1}^n \lambda_i$. By the tail bound for Poisson random variables [27, 109], conditioning on $a_1, \dots, a_n$, we have for any $t > 0$,

$$\mathbb{P}\{\bar{y} - \mathbb{E}\,\bar{y} > t\} \le \exp\left[-\frac{nt}{4}\log\left(1 + \frac{t}{2\lambda}\right)\right],$$

where $\lambda := (1/n)\sum_{i=1}^n \lambda_i$.

Recall that $\lambda_i := |\langle a_i, x^\natural\rangle|^2$. By the assumption on A, we have $(1-\delta)\|x\|_2^2 \le \lambda \le (1+\delta)\|x\|_2^2$ with probability at least $1 - p_\delta$. Moreover, on this event, we have

$$\begin{aligned}\mathbb{P}\{\bar{y} - (1+\delta)\|x^\natural\|_2^2 > t\} &\le \mathbb{P}\{\bar{y} - \lambda > t\} \\ &\le \exp\left[-\frac{nt}{4}\log\left(1 + \frac{t}{2(1+\delta)\|x^\natural\|_2^2}\right)\right].\end{aligned}$$

This proves the proposition.

4.B Proof of Theorem 5.1

Let $\{\alpha_t\}$, $\alpha_0 \neq 0$, be a sequence of non-negative real numbers. We consider step sizes of the form

$$\tau_t = \alpha_{t+1}/S_{t+1}, \tag{4.8}$$

where $S_t := \sum_{k=0}^t \alpha_t$. Unless otherwise stated, $\{X_t\}$ refers to the sequence of iterates generated by Algorithm 1, with the step size chosen as in (4.8). Notice that then the convergence rate of the algorithm depend on the choice of $\{\alpha_t\}$.

By the convexity of $\mathscr{C}$, it is obvious that $X_t \in \mathscr{C}$ for all t. Due to the presence of the logarithmic function, we also need to verify that $X_t \in \operatorname{dom} g$ for all t.

Proposition 4.4. *The following hold.*

1. $\operatorname{Tr}(\Phi_i V_t) \ge 0$ *for all* i *and* t.
2. *If* $\operatorname{Tr}(\Phi_i X_0) > 0$, *then* $\operatorname{Tr}(\Phi_i X_t) > 0$ *for all* i *and* t.

Proof. See Section 4.B.1. □

Now we show the boundedness of C_t for all t, as stated in Section 5. Recall that

$$C_t := \sum_{i=1}^{n} (\eta_i / \mathrm{Tr}(\Phi_i X_t)).$$

Lemma 4.5. *For any T such that $1 - 4n(\bar{\mu}/\underline{\mu}) d_\eta \tau_T > 0$, we have $C_t \le C$, where C is a constant independent of t defined as*

$$C := \max\left\{ \frac{2 d_\Psi d_{\mathscr{C}}}{\underline{\mu}}, C_0 \prod_{i=0}^{T} \frac{1}{1-\tau_i}, \frac{64 n^2 \bar{\mu}^2 d_\eta^2}{\underline{\mu}^3 \left(1 - \frac{4n\bar{\mu} d_\eta}{\underline{\mu}} \tau_T\right)} \right\}.$$

Proof. See Section 4.B.2. □

The following lemma mimics [138, Theorem 1]. Define

$$B_t := \alpha_0 \max_X \left\{ \langle \nabla g(X_0), X_0 - X \rangle \;\middle|\; X \in \mathscr{X} \right\} + \left(\sum_{k=1}^{t} \frac{\alpha_k^2}{S_{k-1}} \right) \gamma,$$

where $\gamma := C^2 d_\Phi^2 d_{\mathscr{C}}^2$.

Lemma 4.6. *For any $t \ge 0$ and $X \in \mathscr{X}$, we have*

$$S_t g(X_t) \le \sum_{k=0}^{t} \left\{ \alpha_k \left[g(X_k) + \langle \nabla g(X_k), X - X_k \rangle \right] \right\} + B_t$$

Proof. See Section 4.B.3. □

Set $X = X^\star$, a minimizer, in Lemma 4.6, and notice that

$$g(X_k) + \langle \nabla g(X_k), X^\star - X_k \rangle \le g^\star$$

for all k. We immediately obtain a convergence guarantee for any $(a_t)_{t \ge 0}$.

Corollary 4.7. *We have $g(X_t) - g^\star \le (B_t / S_t)$.*

Now we consider the special case where $\alpha_t = t+1$. As then $S_t = (t+1)(t+2)/2$, this choice corresponds to $\tau_t = 2/(t+3)$ as in Theorem 6.1.

Proposition 4.8. *Choose $\alpha_t = t+1$. We have*

$$\frac{B_t}{S_t} < \frac{8 \left(\max\{\gamma_1, \gamma_2, \gamma_3\} \right)^2 d_\Phi^2 d_{\mathscr{C}}^2}{t+2} + \frac{2 d_{\mathscr{C}} \|\nabla g(X_0)\|}{(t+1)(t+2)},$$

where

$$\gamma_1 := \frac{2d_\Psi d_{\mathscr{C}}}{\underline{\mu}}, \quad \gamma_2 := 2\frac{nd_\eta}{\underline{\mu}}\left(\frac{4n\bar{\mu}d_\eta}{\underline{\mu}}+1\right)^2,$$
$$\gamma_3 := \frac{64n^2\bar{\mu}^2d_\eta^2}{\underline{\mu}^3}\left(\frac{4n\bar{\mu}d_\eta}{\underline{\mu}}+1\right).$$

Proof. See Section 4.B.4. □

4.B.1 Proof of Proposition 4.4

Recall that V_t is always a positive semi-definite matrix of rank 1, as discussed in Section 4. Since Φ_i is also positive semi-definite, this implies $\mathrm{Tr}(\Phi_i V_t) \geq 0$ for all i and t.

We prove the second claim by induction. The second claim holds true for $t = 0$ by assumption. Suppose $\mathrm{Tr}(\Phi_i X_t) > 0$ for some $t \geq 0$ for all i. Because of the assumption that $\alpha_0 \neq 0$, we always have $\tau_t < 1$ for all t. Then

$$\begin{aligned}\mathrm{Tr}(\Phi_i X_{t+1}) &= (1-\tau)\mathrm{Tr}(\Phi_i X_t) + \tau_t \mathrm{Tr}(\Phi_i V_t)\\ &\geq (1-\tau)\mathrm{Tr}(\Phi_i X_t) > 0,\end{aligned}$$

where the first inequality is by the first claim.

4.B.2 Proof of Lemma 4.5

Consider the sequence $(C_t)_{t\geq 0}$. Roughly speaking, the idea behind the proof is to show that there exists some $T > 0$, such that $C_{t+1} \leq C_t$ for all $t \geq T$; then we can bound C_t from above by C_T for all $t \geq T$, a constant independent of t. Notice that, however, the actual argument in this proof is slightly more delicate (cf. the proof of Proposition 4.11).

A simple bound on C_{t+1} is

$$\begin{aligned}C_{t+1} &= \sum_{i=1}^n \frac{\eta_i}{\mathrm{Tr}(\Phi_i X_{t+1})}\\ &\leq \frac{1}{(1-\tau_t)}\sum_{i=1}^n \frac{\eta_i}{\mathrm{Tr}(\Phi_i X_t)}\\ &= \frac{1}{1-\tau_t}C_t,\end{aligned} \tag{4.9}$$

obtained by the fact that

$$\mathrm{Tr}(\Phi_i X_{t+1}) \geq (1-\tau_t)\mathrm{Tr}(\Phi_i X_t).$$

This yields the following simple result.

Proposition 4.9. *We have* $C_t \le C_0 \prod_{i=0}^{t} (1-\tau_t)^{-1}$.

However, as $1-\tau_t < 1$, the upper bound (4.9) is not sharp enough for our purpose.

Notice that for any k, we have

$$\begin{aligned} C_{t+1} &= \sum_{i\neq k} \frac{\eta_i}{\mathrm{Tr}(\Phi_i X_{t+1})} + \frac{\eta_k}{\mathrm{Tr}(\Phi_k X_{t+1})} \\ &\le \sum_{i\neq k} \frac{\eta_i}{(1-\tau_t)\mathrm{Tr}(\Phi_i X_t)} + \frac{\eta_k}{\mathrm{Tr}(\Phi_k X_{t+1})} \\ &= \frac{C_t}{1-\tau_t} - \frac{\eta_k}{(1-\tau_t)\mathrm{Tr}(\Phi_k X_t)} + \frac{\eta_k}{\mathrm{Tr}(\Phi_k X_{t+1})} \\ &= \frac{C_t}{1-\tau_t} - \frac{\eta_k \tau_t \mathrm{Tr}(\Phi_k V_t)}{[(1-\tau_t)\mathrm{Tr}(\Phi_k X_t)]\,\mathrm{Tr}(\Phi_k X_{t+1})} \\ &\le \frac{C_t}{1-\tau_t} - \xi_k \end{aligned} \tag{4.10}$$

where

$$\xi_k := \frac{\tau_t \mathrm{Tr}(\Phi_k V_t)}{[(1-\tau_t)\mathrm{Tr}(\Phi_k X_t)]\,\mathrm{Tr}(\Phi_k X_{t+1})};$$

the last inequality is due to the fact that either $\eta_k = 0$ or $\eta_k \ge 1$ in the Poisson phase retrieval problem. This bound is sharper than (4.9), as ξ_k is always non-negative.

Proposition 4.10. *If* $C_t > 2\underline{\mu}^{-1} d_\Psi d_{\mathscr{C}}$, *then there exists some* $k \le n$ *such that*

$$\begin{aligned} \frac{1}{\mathrm{Tr}(\Phi_k X_t)} &\ge \frac{\underline{\mu} C_t}{4n\bar{\mu} d_\eta}, \\ \mathrm{Tr}(\Phi_k V_t) &\ge \frac{\underline{\mu}}{4}. \end{aligned}$$

Proof. We prove by contradiction. By the definition of V_t, we have

$$\langle V_t, \nabla g(X_t)\rangle \le \langle X_0, \nabla g X_t\rangle\,.$$

Hence,

$$\begin{aligned} \sum_{i=1}^{n} \frac{\eta_i \langle V_t, \Phi_i\rangle}{\langle X_t, \Phi_i\rangle} &\ge \sum_{i=1}^{n} \left(\frac{\eta_i \langle X_0, \Phi_i\rangle}{\langle X_t, \Phi_i\rangle} \right) + \langle \Psi, X_0 - V_t\rangle \\ &\ge \sum_{i=1}^{n} \frac{\eta_i \langle X_0, \Phi_i\rangle}{\langle X_t, \Phi_i\rangle} - d_\Psi d_{\mathscr{C}} \\ &\ge \underline{\mu} C_t - d_\Psi d_{\mathscr{C}} \ge \frac{\underline{\mu} C_t}{2}. \end{aligned}$$

Let Ω be the set of i's such that $\langle X_t, \Phi_i\rangle^{-1} \ge \underline{\mu} C_t/(4n\bar{\mu} d_\eta)$. Suppose the claim of the proposition

is false, i.e. for all $i \in \Omega$, $\langle V_t, \Phi_i \rangle < \underline{\mu}/4$. Then we have

$$\begin{aligned}\sum_{i=1}^{n} \frac{\eta_i \langle V_t, \Phi_i \rangle}{\langle X_t, \Phi_i \rangle} &= \sum_{i \in \Omega} \frac{\eta_i \langle V_t, \Phi_i \rangle}{\langle X_t, \Phi_i \rangle} + \sum_{i \notin \Omega} \frac{\eta_i \langle V_t, \Phi_i \rangle}{\langle X_t, \Phi_i \rangle} \\ &< \frac{\underline{\mu}}{4} C_t + n d_\eta \bar{\mu} \frac{\underline{\mu} C_t}{4 n \bar{\mu} d_\eta} \\ &= \frac{\underline{\mu} C_t}{2},\end{aligned}$$

□

a contradiction. This completes the proof. □

Assume $C_t > 2\underline{\mu}^{-1} d_\Psi d_{\mathscr{C}}$. By Proposition 4.10 and (4.10), we have

$$C_{t+1} \le C_t \left\{ \frac{1}{1-\tau_t} - \frac{\frac{\tau_t \underline{\mu}}{4}}{(1-\tau_t)\frac{4n\bar{\mu}d_\eta}{\underline{\mu}} \left[(1-\tau_t)\frac{4n\bar{\mu}d_\eta}{\underline{\mu} C_t} + \tau_t \frac{\underline{\mu}}{4} \right]} \right\}.$$

By direct calculation, we obtain $C_{t+1} \le C_t$, if

$$1 - \frac{4n\bar{\mu}d_\eta}{\underline{\mu}} \tau_t > 0, \tag{4.11}$$

$$C_t \ge \kappa_t := \frac{64(1-\tau_t) n^2 \bar{\mu}^2 d_\eta^2}{\underline{\mu}^3 \left(1 - \frac{4n\bar{\mu}d_\eta}{\underline{\mu}} \tau_t \right)}.$$

Proposition 4.11. *Assume that $C_t > 2\underline{\mu}^{-1} d_\Psi d_{\mathscr{C}}$. Choose T such that* (4.11) *holds for $t = T$. Then we have*

$$C_t \le \max \left\{ C_0 \prod_{i=0}^{T} \frac{1}{1-\tau_i}, \frac{64 n^2 \bar{\mu}^2 d_\eta^2}{\underline{\mu}^3 \left(1 - \frac{4n\bar{\mu}d_\eta}{\underline{\mu}} \tau_T \right)} \right\}.$$

Proof. Since $(\tau_t)_{t \ge 0}$ is a decreasing sequence, the inequality (4.11) holds for all $t \ge T$.

If $t \le T$, we can apply Proposition 4.9, and obtain

$$C_t \le C_0 \prod_{i=0}^{t} \frac{1}{1-\tau_i} \le C_0 \prod_{i=0}^{T} \frac{1}{1-\tau_i}.$$

Consider the case when $t > T$. Suppose $C_T \ge \kappa_T$. We have $C_{t+1} \le C_t \le C_T$, which can be bounded using Proposition 4.9, until some t^* such that $C_{t^*} < \kappa_{t^*}$. But then $C_{t+1} \le (1-\tau_t)^{-1} \kappa_t$ for all $t \ge t^*$. If $C_T < \kappa_T$, similarly, we also obtain $C_{t+1} \le (1-\tau_t)^{-1} \kappa_t$ for all $t \ge T$. The

proposition follows, as

$$\frac{1}{1-\tau_t}\kappa_t = \frac{64n^2\bar{\mu}^2 d_\eta^2}{\underline{\mu}^3\left(1-\frac{4n\bar{\mu}d_\eta}{\underline{\mu}}\tau_t\right)} \le \frac{64n^2\bar{\mu}^2 d_\eta^2}{\underline{\mu}^3\left(1-\frac{4n\bar{\mu}d_\eta}{\underline{\mu}}\tau_T\right)}. \qquad \square$$

If $C_t \le 2\underline{\mu}^{-1} d_\Psi d_{\mathscr{C}}$, then this is already a constant upper bound on C_t. This completes the proof.

4.B.3 Proof of Lemma 4.6

We prove by induction. The claim is obviously correct for $t = 0$. Suppose the claim holds for some $t \ge 0$. Then we have

$$\begin{aligned}
&\sum_{k=0}^{t+1} \alpha_k \left[g(X_k) + \langle \nabla g(X_k), X - X_k\rangle\right] + B_t \\
&\quad \ge S_t g(X_t) + \alpha_{k+1}\left[g(X_{t+1}) + \langle \nabla g(X_{t+1}), X - X_{t+1}\rangle\right] \\
&\quad = S_{t+1} g(X_{t+1}) + S_t\left[g(X_t) - g(X_{t+1})\right] + \langle \nabla g(X_{t+1}), \alpha_{t+1}(X - X_{t+1})\rangle \\
&\quad \ge S_{t+1} g(X_{t+1}) + \langle \nabla g(X_{t+1}), \alpha_{t+1}(X - X_{t+1}) + S_t(X_t - X_{t+1})\rangle \\
&\quad = S_{t+1} g(X_{t+1}) + \alpha_{t+1}\langle \nabla g(X_{t+1}), X - V_t\rangle \\
&\quad \ge S_{t+1} g(X_{t+1}) + \alpha_{t+1}\langle \nabla g(X_{t+1}) - \nabla g(X_t), X - V_t\rangle,
\end{aligned}$$

where the second inequality is due to convexity of g, and the third inequality is due to the fact that

$$\langle \nabla g(X_t), X - V_t\rangle \ge 0$$

for any $X \in \mathscr{C}$, as V_t minimizes $\langle \nabla g(X_t), \cdot\rangle$ on $\mathscr{C}$.

To complete the proof, we need to show that

$$\alpha_{t+1}\langle \nabla g(X_{t+1}) - \nabla g(X_t), X - V_t\rangle \ge B_t - B_{t+1} = -\frac{\alpha_{t+1}^2}{S_t}\gamma,$$

or

$$\langle \nabla g(X_{t+1}) - \nabla g(X_t), X - V_t\rangle \ge -\frac{\alpha_{t+1}}{S_t}\gamma. \tag{4.12}$$

By Hölder's inequality, we have

$$\begin{aligned}
|\langle \nabla g(X_{t+1}) - \nabla g(X_t), X - V_t\rangle| &\le \|\nabla g(X_{t+1}) - \nabla g(X_t)\| \|X - V_t\|_* \\
&\le \|\nabla g(X_{t+1}) - \nabla g(X_t)\| d_{\mathscr{C}},
\end{aligned}$$

where $\|\cdot\|$ denotes the spectral norm.

Now we bound the quantity $\|\nabla g(X_{t+1}) - \nabla g(X_t)\|$. By direct calculation, we obtain

$$\begin{aligned}
\|\nabla g(X_{t+1}) - \nabla g(X_t)\| &= \left\| \sum_{i=1}^{n} \frac{\eta_i \langle X_t - X_{t+1}, \Phi_i \rangle}{\langle X_t, \Phi_i \rangle \langle X_{t+1}, \Phi_i \rangle} \Phi_i \right\| \\
&\le d_\Phi \sum_{i=1}^{n} \frac{\eta_i |\langle X_t - X_{t+1}, \Phi_i \rangle|}{\langle X_t, \Phi_i \rangle \langle X_{t+1}, \Phi_i \rangle} \\
&= \tau_t d_\Phi \sum_{i=1}^{n} \frac{\eta_i |\langle X_t - V_t, \Phi_i \rangle|}{\langle X_t, \Phi_i \rangle \langle X_{t+1}, \Phi_i \rangle} \\
&\le \tau_t d_\Phi \sum_{i=1}^{n} \frac{\eta_i \|X_t - V_t\|_* \|\Phi_i\|}{\langle X_t, \Phi_i \rangle \langle X_{t+1}, \Phi_i \rangle} \\
&\le \tau_t d_\Phi^2 d_{\mathscr{C}} \sum_{i=1}^{n} \frac{\eta_i}{\langle X_t, \Phi_i \rangle \langle X_{t+1}, \Phi_i \rangle}.
\end{aligned}$$

Since either $\eta_i = 0$ or $\eta_i \ge 1$, we have

$$\begin{aligned}
\|\nabla g(X_{t+1}) - \nabla g(X_t)\| &\le \tau_t d_\Phi^2 d_{\mathscr{C}} \sum_{i=1}^{n} \frac{\eta_i^2}{\langle X_t, \Phi_i \rangle \langle X_{t+1}, \Phi_i \rangle} \\
&\le \frac{\tau_t d_\Phi^2 d_{\mathscr{C}}}{1 - \tau_t} \sum_{i=1}^{n} \left(\frac{\eta_i}{\langle X_t, \Phi_i \rangle} \right)^2 \\
&\le \frac{\tau_t}{1 - \tau_t} d_\Phi^2 d_{\mathscr{C}} \left(\sum_{i=1}^{n} \frac{\eta_i}{\langle X_t, \Phi_i \rangle} \right)^2 \\
&\le \frac{\alpha_{t+1}}{S_t} d_\Phi^2 d_{\mathscr{C}} \left(\sum_{i=1}^{n} \frac{\eta_i}{\langle X_t, \Phi_i \rangle} \right)^2.
\end{aligned}$$

By Lemma 4.5,

$$\|\nabla g(X_{t+1}) - \nabla g(X_t)\| \le \frac{\alpha_{t+1}}{S_t} d_\Phi^2 d_{\mathscr{C}} C^2.$$

Hence it suffices to choose $\gamma \ge C^2 d_\Phi^2 d_{\mathscr{C}}^2$.

4.B.4 Proof of Proposition 4.8

By Hölder's inequality, the first term in the definition of B_t can be bounded above by $\|\nabla g(X_0)\| d_{\mathscr{C}}$. The second term can be bounded as

$$\left(\sum_{k=1}^{t} \frac{\alpha_k^2}{S_{k-1}} \right) \gamma = \gamma \sum_{k=1}^{t} \left(2 + \frac{2}{k} \right) \le 4t\gamma.$$

Then we obtain

$$\begin{aligned}\frac{B_t}{S_t} &\leq \frac{8t\gamma}{(t+1)(t+2)} + \frac{2d_{\mathscr{C}}\|\nabla g(X_0)\|}{(t+1)(t+2)} \\ &< \frac{8\gamma}{t+2} + \frac{2d_{\mathscr{C}}\|\nabla g(X_0)\|}{(t+1)(t+2)} \\ &\leq \frac{8C^2 d_\Phi^2 d_{\mathscr{C}}^2}{t+2} + \frac{2d_{\mathscr{C}}\|\nabla g(X_0)\|}{(t+1)(t+2)}.\end{aligned}$$

The definition of C in Lemma 4.5 also involves τ_t. We notice that choosing $T = 8n(\bar{\mu}/\underline{\mu})d_\eta - 1$ suffices to ensure $1 - 4n(\bar{\mu}/\underline{\mu})d_\eta\tau_T \geq 0$. Then we obtain

$$\begin{aligned}\prod_{k=0}^{T}\frac{1}{1-\tau_k} &= \frac{(T+2)(T+3)}{2} \\ &< \frac{(T+3)^2}{2} = 2\left(\frac{4n\bar{\mu}d_\eta}{\underline{\mu}} + 1\right)^2.\end{aligned}$$

The quantity C_0 can be easily bounded as $C_0 \leq n\underline{\mu}^{-1}d_\eta$. Finally, we have

$$\frac{64n^2\bar{\mu}^2 d_\eta^2}{\underline{\mu}^3\left(1 - \frac{4n\bar{\mu}d_\eta}{\underline{\mu}}\tau_T\right)} = \frac{64n^2\bar{\mu}^2 d_\eta^2}{\underline{\mu}^3}\left(\frac{4n\bar{\mu}d_\eta}{\underline{\mu}} + 1\right).$$

5 Convergence of mirror descent under a weak smoothness condition

For this and the next chapters, the main motivation was quantum state tomography (QST). Numerically, QST corresponds to minimizing a sum of exp-linear functions, on the set of *quantum density matrices*—a matrix analogue of the probability simplex. In the previous chapter, we have proved that the Frank-Wolfe algorithm can be adopted for QST. Unfortunately, as the numerical results in Section 5.6 show, the $O(1/k)$ convergence rate of the Frank-Wolfe algorithm is unsatisfactory on real experimental data-sets. It is natural to expect that an optimization algorithm that does not only use the linear minimization oracle would converge faster. However, existing convergence results of first-order algorithms typically require smoothness of the objective function, but the loss function for QST is a sum of exp-linear functions that do not satisfy the smoothness condition.

In this chapter, we explore the possibility of proving convergence for a class of optimization algorithms, under weak assumptions on the objective function. Specifically, we prove that the mirror descent with Armijo line search always converges, if the objective function is *locally relatively smooth*.

This chapter is based on the joint work with Carlos Riofrío and Volkan Cevher [121].

5.1 Introduction

Consider a constrained convex optimization problem:

$$f^\star = \min_x \{ f(x) \mid x \in \mathscr{X} \}, \tag{P}$$

where f is a convex differentiable function, and $\mathscr{X}$ is a convex closed set in $\mathbb{R}^d$. We assume that $f^\star > -\infty$.

The mirror descent algorithm is standard for solving such a constrained convex optimization

problem [19, 135]. Given an initial iterate $x_0 \in \mathscr{X}$, the mirror descent algorithm iterates as

$$x_{k+1} = \arg\min_x \left\{ \langle \nabla f(x_k), x - x_k \rangle + \alpha_k D_h(x, x_k) \,\middle|\, x \in \mathscr{X} \right\}, \quad \forall k \in \mathbb{N}, \tag{5.1}$$

for some convex differentiable function h and a properly chosen sequence of step sizes $\{\alpha_k\}$, where D_h denotes the Bregman divergence induced by h:

$$D_h(z_2, z_1) := h(z_2) - [h(z_1) + \langle \nabla h(z_1), z_2 - z_2 \rangle], \quad \forall (z_2, z_2) \in \operatorname{dom} h \times \operatorname{dom} \nabla h.$$

In comparison to the standard projected gradient descent, the mirror descent algorithm can have an almost dimension-independent convergence rate guarantee, or lower per-iteration computational complexity. A famous example is the exponentiated gradient method, which enjoys both benefits [100]. The exponentiated gradient method corresponds to the mirror descent algorithm with h being the negative Shannon entropy.

Convergence of the mirror descent algorithm has been established under the following two conditions on the objective function.

1. Bounded gradient: There exists some $L > 0$, such that

   $$\|\nabla f(x)\| \le L, \quad \forall x \in \mathscr{X},$$

   for some norm $\|\cdot\|$ [19, 135].

2. Relative smoothness: There exist some $L > 0$ and a convex differentiable function h, such that

   $$f(y) \le f(x) + \langle \nabla f(x), y - x \rangle + L D_h(y, x), \quad \forall x, y \in \mathscr{X},$$

   where D_h denotes the Bregman divergence induced by h [7, 16, 123][1].

These conditions may not hold, or introduce undesirable computational burdens for some applications. Quantum state tomography is one such instance.

Example 5.1. Quantum state tomography (QST) *is the task of estimating the state of qubits (quantum bits) given measurement outcomes [146]; this task is essential to calibrating quantum computation devices. Numerically, it corresponds to minimizing the function*

$$f_{QST}(x) := -\sum_{i=1}^{n} \log \operatorname{Tr}(M_i x),$$

[1] Notice that the relative smoothness condition only involves the first-order derivative, and hence does not conform perfectly to the general definition of "smoothness" adopted in this thesis (cf. Definition 1.1). We decide to use the term "relative smoothness" here, to keep consistency with existing literature [123].

for given positive semi-definite matrices M_i, on the set of quantum density matrices

$$\mathscr{D} := \left\{ x \in \mathbb{C}^{d\times d} \;\middle|\; x \geq 0, \mathrm{Tr}(x) = 1 \right\}. \tag{5.2}$$

The dimension d equals 2^q, where q is the number of qubits.

Notice that the diagonal of a density matrix in $\mathbb{R}^{d\times d}$ must belong to the probability simplex in $\mathbb{R}^d$; therefore, a density matrix can be viewed as a matrix analogue of a probability distribution. Regarding this observation, it is natural to consider the matrix version of the exponentiated gradient method, for which the Shannon entropy is replaced by its matrix analogue called the von Neumann entropy [31, 174]. Unfortunately, we prove the following in Appendix 5.A.

Proposition 5.2. *The gradient of the function f_{QST} is not bounded. The function f_{QST} is not smooth relative to the von Neumann entropy.*

Another popular choice of the function h is Burg's entropy; the resulting mirror descent algorithm iterates as

$$x_{k+1} = \left(x_k^{-1} + \alpha_k \nabla f(x_k)\right)^{-1}, \quad \forall k \in \mathbb{N},$$

where α_k is chosen such that $\mathrm{Tr}(x_{k+1}) = 1$ [110]. The numerical search for α_k yields high per-iteration computational complexity of the mirror descent.

We note that in terms of the objective functions and constraint sets, positron emission tomography, optimal portfolio selection, and non-negative linear inverse problems are essentially vector analogues of QST [34, 58, 180]. The same issues we have discussed above remain in these applications, though the computational burden due to the Burg entropy may be relatively minor in the vector cases.

To address "non-standard" applications like QST, we relax the condition on the objective function. Specifically, we propose a novel localized version of the relative smoothness condition. The local relative smoothness condition does not involve any parameter, in comparison to the bounded gradient and (global) relative smoothness conditions. Therefore, we do not seek for a closed-form expression for the step sizes; instead, we consider selecting the step sizes adaptively by Armijo line search.

5.1.1 Related work

The mirror descent algorithm was introduced in [135]. The formulation (5.1) was proposed in [19], which is equivalent to the original one under standard assumptions. The interior gradient method studied in [7] is also of the form (5.1); the difference lies in the conditions on the loss function and the algorithm setup. Standard convergence analyses of the mirror descent, as discussed above, assumes either bounded gradient or relative smoothness [7, 16, 19, 123, 135].

The exponentiated gradient method was proposed in [103]; it is also known as the entropic mirror descent.

For quantum state tomography, there are few guaranteed-to-converge optimization algorithms. The $R\rho R$ algorithm was proposed as an analogue of the expectation maximization algorithm [96], but does not always converge [184]. The diluted $R\rho R$ algorithm is a variant of the $R\rho R$ algorithm, which guarantees convergence by exact line search [184]. The preceding chapter shows that the Frank-Wolfe algorithm converges with a slightly different step size selection rule. The SCOPT algorithm proposed in [172], a proximal gradient method for composite self-concordant minimization, also converges, as the logarithmic function is a standard instance of a self-concordant function. The numerical results in Section 5.6, unfortunately, showed that the convergence speeds of the diluted $R\rho R$, Frank-Wolfe, and SCOPT algorithms are not satisfactory on real data-sets.

For the vector analogues of QST mentioned above, the standard approach is based on expectation-maximization-type methods developed in [58, 62]. See also [35] for a modern introduction. The numerical results in Section 5.6 showed that this standard approach is slow on real data-sets for portfolio selection.

Armijo line search was proposed in [6], for minimizing 2-smooth functions. The formulation of Armijo line search studied in this chapter is the generalized version proposed in [23].

5.1.2 Contributions

Our main result is Theorem 5.11, which establishes convergence of mirror descent with Armijo line search under the relative smoothness condition. Numerical results showed that, because of Theorem 5.11, the exponentiated gradient method with Armijo line search was the fastest guaranteed-to-converge algorithm for QST, empirically on real data-sets. To the best of our knowledge, even for globally relatively smooth objective functions, convergence of mirror descent with Armijo line search has not been proven; Theorem 5.11 provides the first convergence guarantee for this setup.

5.2 Mirror descent with Armijo line search

Let h be a convex differentiable function strictly convex on $\mathscr{X}$. The corresponding Bregman divergence is given by

$$D_h(z_2, z_1) := h(z_2) - [h(z_1) + \langle \nabla h(z_1), z_2 - z_2 \rangle], \quad \forall (z_2, z_2) \in \operatorname{dom} h \times \operatorname{dom} \nabla h.$$

Because of the strict convexity of h, it holds that $D_h(z_2, z_1) \geq 0$, and $D_h(z_2, z_1) = 0$ if and only if $z_2 = z_1$.

Define $\tilde{\mathscr{X}} := \mathscr{X} \cap \operatorname{dom} \nabla f \cap \operatorname{dom} \nabla h$. The corresponding mirror descent algorithm starts with

Algorithm 2 Mirror Descent with Armijo Line Search

Require: $\bar{\alpha} > 0$, $r \in (0,1)$, $\tau \in (0,1)$, $x_0 \in \mathcal{X}_h$

for $k = 1, 2, \ldots$ **do**
  $\alpha_k \leftarrow \bar{\alpha}$
  while $\tau \left\langle \nabla f(x_{k-1}), x_{k-1}(\alpha_k) - x_{k-1} \right\rangle + f(x_{k-1}) < f(x_{k-1}(\alpha_k))$ **do**
    $\alpha_k \leftarrow r\alpha_k$
  end while
  $x_k \leftarrow x_{k-1}(\alpha_k)$
end for

some $x_0 \in \tilde{\mathcal{X}}$, and iterates as

$$x_k = x_{k-1}(\alpha_k) := \operatorname*{arg\,min}_x \{\alpha_k \langle \nabla f(x_{k-1}), x - x_{k-1}\rangle + D_h(x, x_{k-1}) \mid x \in \mathcal{X}\}, \quad \forall k \in \mathbb{N},$$

where α_k denotes the step size. To ensure that the mirror descent algorithm is well-defined, we will assume the following throughout this paper.

Assumption 4. *For every $x \in \tilde{\mathcal{X}}$ and $\alpha \geq 0$, $x(\alpha)$ is uniquely defined and lies in $\tilde{\mathcal{X}}$.*

There are several sufficient conditions that guarantee Assumption 4, but in practice, it is typically easier to directly check Assumption 4. The interested reader is referred to, e.g., [16, 17] for the details.

We consider choosing the step sizes by the Armijo rule. Let $\bar{\alpha} > 0$ and $r, \tau \in (0,1)$. The Armijo rule outputs $\alpha_k = r^j \bar{\alpha}$ for every k, where j is the least non-negative integer such that

$$f(x_{k-1}(r^j \bar{\alpha})) \leq f(x_{k-1}) + \tau \langle \nabla f(x_{k-1}), x_{k-1}(r^j \bar{\alpha}) - x_{k-1}\rangle .$$

The Armijo rule can be easily implemented by a while-loop, as shown in Algorithm 2.

5.3 Local relative smoothness

In this section, we introduce the local relative smoothness condition, and provide a detailed discussion. In particular, we provide some practical approaches to checking the local relative smoothness condition, alone with concrete examples illustrating when the practical approaches can and cannot be applied.

Roughly speaking, the local relative smoothness condition asks that for every point, there exists a neighborhood on which f is relatively smooth.

Definition 5.3. *We say that f is locally smooth relative to h on $\mathcal{X}$, if for every $x \in \mathcal{X} \cap \operatorname{dom} f$, there exist some $L_x > 0$ and $\varepsilon_x > 0$, such that*

$$f(z_2) \leq f(z_1) + \langle \nabla f(z_1), z_2 - z_1\rangle + L_x D_h(z_2, z_1), \quad \forall z_1, z_2 \in \mathcal{B}_{\varepsilon_x}(x) \cap \tilde{\mathcal{X}}, \tag{5.3}$$

where $\mathscr{B}_{\varepsilon_x}(x)$ denotes the ball centered at x of radius ε_x with respect to a norm.

If we set $h : x \mapsto (1/2)\|x\|_2^2$, then (5.3) becomes

$$f(z_2) \le f(z_1) + \langle \nabla f(z_1), z_2 - z_1 \rangle + \frac{L_x}{2}\|z_2 - z_1\|_2^2, \quad \forall z_1, z_2 \in \mathscr{B}_{\varepsilon_x}(x) \cap \tilde{\mathscr{X}},$$

This is indeed the the locally Lipschitz gradient condition in literature.

Lemma 5.4. *The following two statements are equivalent.*

1. *The function f is locally smooth relative to $h : x \mapsto (1/2)\|x\|_2^2$ on $\mathscr{X}$.*
2. *Its gradient ∇f is locally Lipschitz on $\operatorname{int}\tilde{\mathscr{X}}$; that is, for every $x \in \mathscr{X} \cap \operatorname{dom} f$, there exists some $L_x > 0$ and $\varepsilon_x > 0$, such that*

$$\|\nabla f(z_2) - \nabla f(z_1)\|_2 \le L_x \|z_2 - z_1\|_2, \quad \forall z_1, z_2 \in \mathscr{B}_{\varepsilon_x}(x) \cap \tilde{\mathscr{X}}.$$

The proof of Lemma 5.4 is standard; we give it in Appendix 5.B.

It is already known that the local Lipschitz gradient condition lies strictly between the following two conditions.

1. The function f is differentiable.
2. The gradient of f is (globally) Lipschitz; that is, the function f is 2-smooth.

See [93, 97] for the details.

The following result provides a practical approach to checking the local Lipschitz gradient condition.

Proposition 5.5. *Suppose that* $\operatorname{dom} f \cap \mathscr{X}$ *is relatively open in $\mathscr{X}$, and f is twice continuously differentiable on* $\operatorname{dom} f \cap \mathscr{X}$. *Then f is locally smooth relative to $h(\cdot) := (1/2)\|\cdot\|_2^2$ on $\mathscr{X}$.*

Proof. Recall the definition of relative openness: For every x in $\operatorname{dom} f \cap \mathscr{X}$, there exists some ε_x such that $\mathscr{B}_{\varepsilon_x}(x) \cap \mathscr{X} \subseteq \operatorname{dom} f \cap \mathscr{X}$. Notice that the largest eigenvalue of $\nabla^2 f$ is a continuous function on $\mathscr{B}_{\varepsilon_x}(x) \cap \mathscr{X}$; by the extreme value theorem, there exists some L_x such that $\nabla^2 f(z) \le L_x I$ for every $z \in \mathscr{B}_{\varepsilon_x}(x) \cap \mathscr{X}$. For every $z_1, z_2 \in \mathscr{B}_{\varepsilon_x} \cap \tilde{\mathscr{X}}$, we use Taylor's formula with the integral remainder and write

$$\begin{aligned} f(z_2) &= f(z_1) + \langle \nabla f(z_1), z_2 - z_1 \rangle + \int_0^1 \int_0^t \langle \nabla^2 f(z_1 + \tau(z_2 - z_1))(z_2 - z_1), z_2 - z_1 \rangle \, \mathrm{d}\tau \, \mathrm{d}t \\ &\le f(z_1) + \langle \nabla f(z_1), z_2 - z_1 \rangle + \int_0^1 \int_0^t L_x \|z_2 - z_1\|_2^2 \, \mathrm{d}\tau \, \mathrm{d}t \\ &= f(z_1) + \langle \nabla f(z_1), z_2 - z_1 \rangle + \frac{L_x}{2}\|z_2 - z_1\|_2^2, \end{aligned}$$

which proves the proposition. □

Corollary 5.6. *If f is twice continuously differentiable on $\mathscr{X}$, then it is locally smooth relative to $h(\cdot) := (1/2)\|\cdot\|_2^2$ on $\mathscr{X}$.*

Indeed, under the setting of Corollary 5.6, the function f has a bounded Hessian by the extreme value theorem, and hence is smooth relative to $h(\cdot) := (1/2)\|\cdot\|_2^2$, i.e., the function satisfies the standard smoothness assumption in literature [136]; then most existing convergence results for first-order optimization algorithms apply. To derive an upper bound of the Lipschitz parameter, however, may be non-trivial. Moreover, there are cases where Corollary 5.6 does not apply, while Proposition 5.5 is applicable. Below is an example.

Example 5.7. *Set $f(x) := -\log(x_1) - \log(x_2)$ for every $x := (x_1, x_2) \in \mathbb{R}^2$. Set $\mathscr{X}$ to be the positive orthant. Then f is not twice continuously differentiable on $\mathscr{X}$; for example, $\nabla^2 f(1,0)$ does not exist. However, Proposition 5.5 is applicable—*$\operatorname{dom} f \cap \mathscr{X}$ *is relatively open in $\mathscr{X}$ as* $\operatorname{dom} f$ *is open, and it is easily checked that f is twice continuously differentiable on* $\operatorname{dom} f \cap \mathscr{X}$.

Note that the local Lipschitz gradient condition is not always applicable.

Example 5.8. *Set $f(x) := x_1 \log(x_1) + x_2 \log(x_2)$ for every $x := (x_1, x_2) \in \mathbb{R}^2$. Set $\mathscr{X}$ to be the probability simplex in $\mathbb{R}^2$. Then f is not locally smooth relative to $h(\cdot) := (1/2)\|\cdot\|_2^2$. For example, the point $x = (0,1)$ lies in* $\operatorname{dom} f \cap \mathscr{X}$*, while ∇f is unbounded around $(0,1)$. However, it is obvious that f is locally smooth relative to the negative Shannon entropy—indeed, f itself is the negative Shannon entropy function.*

A standard setting for the mirror descent algorithm requires the following [7, 19, 100].

Assumption 5. *The function h is strongly convex with respect to a norm $\|\cdot\|$ on $\mathscr{X}$; that is, there exists some $\mu > 0$, such that*

$$D_h(z_2, z_1) \geq \frac{\mu}{2}\|z_2 - z_1\|^2, \quad \forall (z_2, z_1) \in (\operatorname{dom} h \cap \mathscr{X}) \times (\operatorname{dom} \nabla h \cap \mathscr{X}).$$

If f is locally smooth relative to $h(x) := (1/2)\|x\|_2^2$, it is also locally smooth relative to any function $\tilde{h}$ strongly convex on $\mathscr{X}$ with respect to a norm $\|\cdot\|$—if for some $L > 0$ and $z_1, z_2 \in \operatorname{dom} \nabla \tilde{h} \times \operatorname{dom} \tilde{h}$, it holds that

$$f(z_2) \leq f(z_1) + \langle \nabla f(z_1), z_2 - z_1 \rangle + \frac{L}{2}\|z_2 - z_1\|_2^2,$$

then we have

$$f(z_2) \leq f(z_1) + \langle \nabla f(z_1), z_2 - z_1 \rangle + \frac{CL}{\mu} D_{\tilde{h}}(z_2, z_1),$$

for some $C > 0$ such that $\|\cdot\|_2 \leq C\|\cdot\|$, which exists because all norms on a finite-dimensional space are equivalent. Therefore, with Assumption 5, it suffices to check for local smoothness relative to $h(x) := (1/2)\|x\|_2^2$.

Example 5.9. *Suppose that the constraint set $\mathscr{X}$ is the probability simplex. By Pinsker's inequality, the negative Shannon entropy is strongly convex on $\mathscr{X}$ with respect to the ℓ_1-norm [61]. By the discussion above and Corollary 5.6, any convex objective function that is twice continuously differentiable on $\mathscr{X}$ is locally smooth relative to the negative Shannon entropy.*

It is possible that Assumption 5 does not hold, while we have local relative smoothness.

Example 5.10. *Consider the function f as defined in Example 5.7. Set $h := f$, the Burg entropy. Then obviously, f is smooth—and hence locally smooth—relative to h. However, if we set $\mathscr{X}$ to be the positive orthant, h is not strongly convex on $\mathscr{X}$.*

5.4 Main result

The main result of this chapter, the following theorem, says that the mirror descent algorithm with Armijo line search is well-defined, and guaranteed to converge, given assumptions discussed above.

Theorem 5.11. *Suppose that Assumption 4 holds. Suppose that* $\operatorname{dom} f \cap \mathscr{X} \subseteq \operatorname{dom} h \cap \mathscr{X}$*, and f is locally smooth relative to h. Then the following hold.*

1. *The Armijo line search procedure terminates in finite steps.*
2. *The sequence $\{f(x_k)\}$ is non-increasing.*
3. *The sequence $\{f(x_k)\}$ converges to $f^\star$, if $\{x_k\}$ is bounded.*

Boundedness of the sequence $\{x_k\}$ holds, for example, when the constraint set $\mathscr{X}$ or level set $\{x \in \mathscr{X} \mid f(x) \leq f(x_0)\}$ is bounded. A sufficient condition for the latter case is *coercivity*—a function is called coercive, if for every sequence $\{x_k\}$ such that $\|x_k\| \to +\infty$, we have $f(x_k) \to +\infty$ (see, e.g., [18]).

5.5 Proof of Theorem 5.11

The proof of Theorem 5.11 stems from standard arguments (see, e.g., [7]), showing that the mirror descent algorithm converges, as long as the step sizes α_k are bounded away from zero. However, without any global parameter of the objective function, we are not able to provide a lower bound for all step sizes as in [7]. We solve this difficulty by proving the *existence* of a strictly positive lower bound, for *all but a finite number* of the step sizes.

The following result shows that for every $x \in \tilde{\mathscr{X}}$, $x(\alpha)$ can be arbitrarily close to x by setting α very small. This result is so fundamental in our analysis that we will use it without explicitly mentioning it.

Lemma 5.12. *The function $x(\alpha)$ is continuous in α for every $x \in \tilde{\mathscr{X}}$.*

Proof. Apply Theorem 7.41 in [158]. □

For ease of presentation, we put the proofs of some technical lemmas in Section 5.C.

5.5.1 Proof of Statement 1

Statement 1 follows from the following lemma.

Lemma 5.13. *For every $x \in \tilde{\mathscr{X}}$, there exists some $\alpha_x > 0$, such that*

$$f(x(\alpha)) \leq f(x) + \tau \langle \nabla f(x), x(\alpha) - x \rangle, \quad \forall \alpha \in (0, \alpha_x]. \tag{5.4}$$

Proof. We write (5.4) equivalently as

$$f(x(\alpha)) - \left[f(x) + \langle \nabla f(x), x(\alpha) - x \rangle \right] \leq -(1-\tau) \langle \nabla f(x), x(\alpha) - x \rangle, \quad \forall \alpha \in (0, \alpha_x].$$

By the local relative smoothness condition, it suffices to check

$$L_x D_h(x(\alpha), x) \leq -(1-\tau) \langle \nabla f(x), x(\alpha) - x \rangle, \quad \forall \alpha \in (0, \alpha_x].$$

By Lemma 5.20, it suffices to check

$$\alpha L_x D_h(x(\alpha), x) \leq (1-\tau) D_h(x(\alpha), x), \quad \forall \alpha \in (0, \alpha_x].$$

If $D_h(x(\alpha), x) > 0$, it suffices to set $\alpha_x = L_x^{-1}(1-\tau)$. Otherwise, we have $x = x(\alpha)$; then Lemma 5.19 implies that x is a minimizer, and Lemma 5.13 follows with any $\alpha_x > 0$. □

5.5.2 Proof of Statements 2 and 3

We start with the following theorem.

Theorem 5.14. *Let $\{x_k\}$ be a sequence in $\tilde{\mathscr{X}}$. Suppose that the assumptions in Theorem 5.11 hold. Then the sequence $\{f(x_k)\}$ monotonically converges to $f^\star$, if the following hold.*

1. *There exists some $\tau \in (0,1)$, such that*

$$f(x_k) \leq f(x_{k-1}) + \tau \langle \nabla f(x_{k-1}), x_k - x_{k-1} \rangle, \quad \forall k \in \mathbb{N}.$$

2. *The sum of step sizes diverges, i.e., $\sum_{k=1}^{\infty} \alpha_k = +\infty$.*

Theorem 5.14 is essentially a restatement of Theorem 4.1 in [7]. We give a proof in Appendix 5.D for completeness.

The first condition in Theorem 5.11 is automatically satisfied by the definition of Armijo line search. The second condition is verified by the following lemma.

Lemma 5.15. *Suppose that the assumptions in Theorem 5.11 hold. If none of the iterates is a solution to* (P)*, it holds that* $\sum_{k=1}^{\infty} \alpha_k = +\infty$.

Proof. We prove by contradiction. Suppose that $\liminf\{\alpha_k\} = 0$. Then there exists a sub-sequence $\{\alpha_k \mid k \in \mathcal{K} \subseteq \mathbb{N}\}$ converging to zero. By the boundedness of $\{x_k\}$, there exists a sub-sequence $\{x_k \mid k \in \mathcal{K}' - 1\}$ converging to a limit point x_∞, for some $\mathcal{K}' \subseteq \mathcal{K}$. Notice that $\{\alpha_k \mid k \in \mathcal{K}'\}$ converges to zero. For large enough $k \in \mathcal{K}' - 1$, we have, by the definition of Armijo line search, that

$$f(x_{k-1}(r^{-1}\alpha_k)) > f(x_{k-1}) + \tau \langle \nabla f(x_{k-1}), x_{k-1}(r^{-1}\alpha_k) - x_{k-1} \rangle .$$

This implies

$$\begin{aligned} &f(x_{k-1}(r^{-1}\alpha_k)) - \left[f(x_{k-1}) + \langle \nabla f(x_{k-1}), x_{k-1}(r^{-1}\alpha_k) - x_{k-1} \rangle \right] \\ &\quad > -(1-\tau) \langle \nabla f(x_{k-1}), x_{k-1}(r^{-1}\alpha_k) - x_{k-1} \rangle . \end{aligned}$$

By the local relative smoothness condition and Lemma 5.20, we write

$$r^{-1}\alpha_k L_{x_\infty} D_h(x_{k-1}(r^{-1}\alpha_k), x_{k-1}) > (1-\tau) D_h(x_{k-1}(r^{-1}\alpha_k), x_{k-1}).$$

If $x_{k-1}(r^{-1}\alpha_k) \neq x_{k-1}$, we get

$$\alpha_k > \frac{r(1-\tau)}{L_{x_\infty}},$$

for large enough $k \in \mathcal{K}' - 1$, a contradiction. Therefore, $\liminf\{\alpha_k\}$ is strictly positive, and the lemma follows. □

Proof of Statements 2 and 3 of Theorem 5.11. If none of the iterates is a solution to (P), Theorem 5.14 and Lemma 5.15 imply that the sequence $\{f(x_k)\}$ converges to $f^\star$. Otherwise, if x_k is a solution, Lemma 5.19 implies that $x_{k'} = x_k$ for every $k' > k$. Monotonicity of the sequence $\{f(x_k)\}$ follows from Corollary 5.21 in Section 5.C. □

5.6 Numerical results

We illustrate applications of Theorem 5.11 in this section.

5.6.1 Portfolio selection

Consider long-term investment in a market of d stocks under the discrete-time setting. At the beginning of the t-th day, $t \in \mathbb{N}$, the investor distributes his total wealth to the stocks following a vector x_t in the probability simplex $\mathscr{P} \subset \mathbb{R}^d$. Denote the price relatives—(possibly negative) returns the investor would receive at the end of the day with one-dollar investment—of the stocks by a vector $a_t \in [0, +\infty)^d$. Then, if the investor has one dollar at the beginning of the first day, the wealth at the end of the t-th day is $\Pi_{i=1}^{t} \langle a_i, x_i \rangle$. For every $t \in \mathbb{N}$, the *best constant rebalanced portfolio* $x_t^\star$ up to the t-th day is defined as a solution of the optimization problem [59]

$$x^\star \in \operatorname*{arg\,min}_x \left\{ -\sum_{i=1}^{t} \log \langle a_i, x \rangle \;\middle|\; x \in \mathscr{P} \right\}. \tag{BCRP}$$

The wealth incurred by the best constant rebalanced portfolio is a benchmark for on-line portfolio selection algorithms [59, 60, 91].

Denote the objective function in (BCRP) by f_{BCRP}. As f_{BCRP} is simply a vector analogue of f_{QST}, most existing convergence guarantees in convex optimization does not hold. The optimization problem (BCRP) was addressed by an expectation-maximization (EM)-type method developed by Cover [58]. Given an initial iterate $x_0 \in \mathscr{P} \cap \mathrm{dom}(f_{\mathrm{BCRP}})$, Cover's algorithm iterates as

$$x_k = -x_{k-1} \cdot \nabla f_{\mathrm{BCRP}}(x_{k-1}), \quad \forall k \in \mathbb{N},$$

where the symbol "$\cdot$" denotes element-wise multiplication. The algorithm possesses a guarantee of convergence but not the convergence rate [58, 62].

Now we show that the optimization problem (BCRP) can be also solved by the exponentiated gradient method with Armijo line search.

Proposition 5.16. *The function f_{BCRP} is locally smooth relative to the (negative) Shannon entropy on the constraint set $\mathscr{P}$.*

Proof. Note that $\mathrm{dom}(f_{\mathrm{BCRP}})$ is open, and hence $\mathrm{dom}(f_{\mathrm{BCRP}}) \cap \mathscr{X}$ is relatively open in $\mathscr{X}$. It is easily checked that f_{BCRP} is twice continuously differentiable on $\mathrm{dom}(f_{\mathrm{QST}})$, and hence on $\mathrm{dom}(f_{\mathrm{BCRP}}) \cap \mathscr{X}$. By Proposition 5.5, the function f_{BCRP} is locally smooth relative to $h(\cdot) := (1/2)\|\cdot\|_2^2$. By Pinsker's inequality [61], the Shannon entropy is strongly convex on $\mathscr{P}$ with respect to the ℓ_1-norm. As all norms on a finite-dimensional space are equivalent, the proposition follows. □

Therefore, the exponentiated gradient method—mirror descent with the Shannon entropy—is guaranteed to converge for solving (BCRP). The iteration rule has a closed-form:

$$x(\alpha) = c^{-1} x \cdot \exp(-\alpha \nabla f_{\mathrm{BCRP}}(x)), \quad \forall x \in \mathscr{P}, \alpha \geq 0,$$

where we set $\exp(v) := (\mathrm{e}^{v_1}, \ldots, \mathrm{e}^{v_d})$ for any $v = (v_1, \ldots, v_p) \in \mathbb{R}^p$.

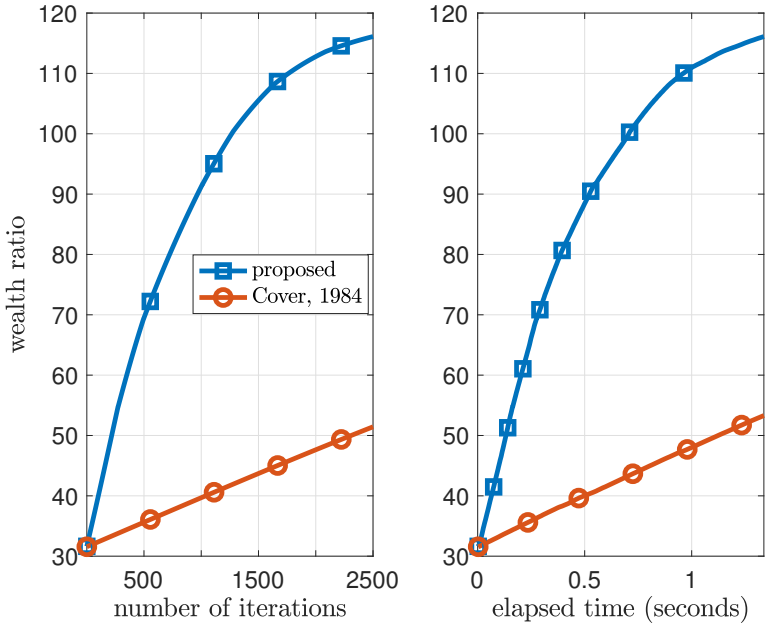


Figure 5.1 – Wealth yielded by different algorithms for the NYSE data.

We compare the convergence speeds of Cover's algorithm and the exponentiated gradient method with Armijo line search, for the New York Stock Exchange (NYSE) data during January 1st, 1985–June 30th, 2010 [116]. The corresponding dimensions are $n = 6431$ and $d = 23$. We set $\overline{\alpha} = 10$, $r = 0.5$, and $\tau = 0.8$ for the Armijo line search procedure. The numerical experiments were done in MATLAB R2018a, on a MacBook Pro with an Intel Core i7 2.8GHz processor and 16GB DDR3 memory.

The numerical result is presented in Figure 5.1, where we plot the total wealth yielded by the algorithm iterates, with an initial wealth of one dollar. The proposed approach—exponentiated gradient method with Armijo line search—was obviously faster than Cover's algorithm. For example, fixing the budget of the computation time to be one second, the proposed approach yields more than twice of the wealth yielded by Cover's algorithm.

5.6.2 Quantum state tomography

Quantum state tomography (QST) is the task of estimating the state of qubits (quantum bits), given measurement outcomes. Numerically, QST corresponds to solving a convex optimization problem specified in Example 5.1. Recall that in the introduction, we have shown that the corresponding objective function, f_{QST}, does not satisfy the bounded gradient condition and is not smooth relative to the von Neumann entropy, while mirror descent with the Burg entropy has high per-iteration computational complexity.

The following proposition is a matrix analogue to Proposition 5.16. A proof is provided in

Section 5.E.

Proposition 5.17. *The function f_{QST} is locally smooth relative to the von Neumann entropy on the constraint set $\mathscr{D}$.*

Therefore, the *(matrix) exponentiated gradient method*—mirror descent with the von Neumann entropy—with Armijo line search is guaranteed to converge, by Theorem 5.11. The corresponding iteration rule has a closed-form expression [31, 174]:

$$x(\alpha) = c^{-1} \exp(\log(x) - \alpha \nabla f(x)),$$

for every $x \in \tilde{\mathscr{X}}$ and $\alpha \geq 0$, where c is a positive real normalizing the trace of $x(\alpha)$. The functions exp and log denote matrix exponential and logarithm, respectively.

We test the empirical performance of the exponentiated gradinet method with Armijo line search, on real experimental data generated following the setting in [85]. To the best of our knowledge, the diluted $R\rho R$ algorithm [184], SCOPT [172], and the Frank-Wolfe algorithm studied in the previous chapter [143] are the only existing algorithms that are guaranteed to converge. We will also consider the $R\rho R$ algorithm [96]; it does not always converge [184], but is typically much faster than the diluted $R\rho R$ algorithm in practice.

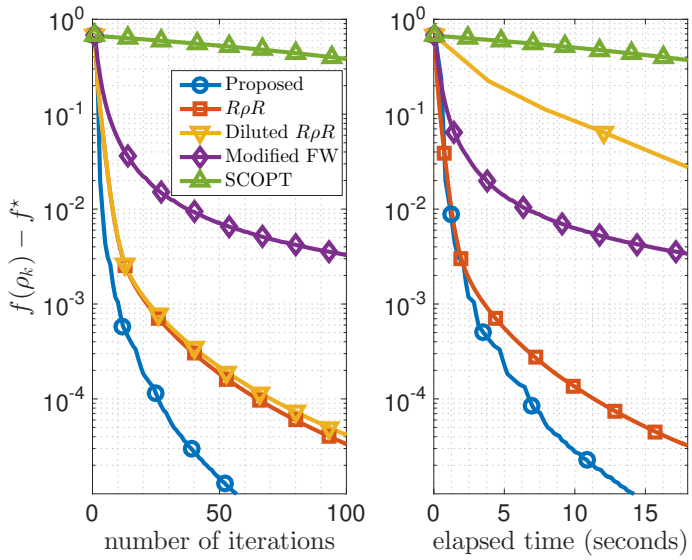


Figure 5.2 – The 6-qubit case.

We compare the convergence speeds for the 6-qubit ($d = 2^6$) and 8-qubit ($d = 2^8$) cases, in Fig. 5.2 and 5.3, respectively. The corresponding "sample sizes" (number of summands in f_{QST}) are $n = 60640$ and $n = 460938$, respectively. The numerical experiments were done in MATLAB R2015b, on a MacBook Pro with an Intel Core i7 2.8GHz processor and 16GB DDR3 memory. We set $\alpha = 10$, and $\gamma = \tau = 0.5$ in Algorithm 2 for both cases. In both figures, $f^\star$ denotes the minimum value of f_{QST} found by the five algorithms in 120 iterations.

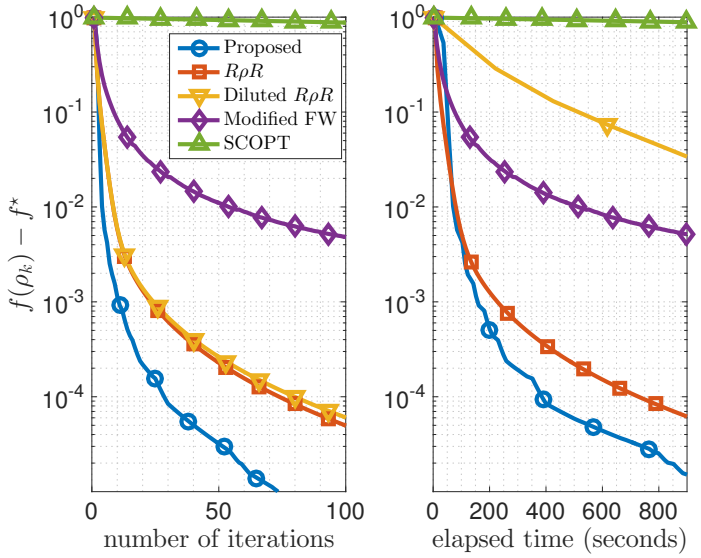


Figure 5.3 – The 8-qubit case.

One can observe that the exponentiated gradient method with Armijo line search is the fastest, in terms of the actual elapsed time. The slowness of the other algorithms is explainable.

1. The diluted $R\rho R$ algorithm, using the notation of this paper, iterates as

$$x_{k+1} = c_k^{-1} \left[I + \beta_k f'(x_k) \right]^{\mathrm{H}} \rho_k \left[I + \beta_k f'(x_k) \right],$$

   where c_k normalizes the trace of x_{k+1}. To guarantee convergence, the step sizes β_k are computed by exact line search. The exact line search procedure renders the algorithm slow.

2. SCOPT is a projected gradient method for minimizing self-concordant functions [136, 140]. Notice that projection onto $\mathscr{D}$ typically results in a low-rank output; hence, it is possible that $\mathrm{Tr}(M_i x_k) = 0$ for some low-rank M_i and iterate x_k, but then x_k is not a feasible solution because $\log(0)$ is not defined[2]. This issue was pointed out by us, and called the stalling problem in [104]. Luckily, self-concordance of f_{QST} ensures that if an iterate x_k lies in $\mathrm{dom}\, f_{\mathrm{QST}}$, and the next iterate x_{k+1} lies in a small enough *Dikin ellipsoid* centered at x_k, then x_{k+1} also lies in $\mathrm{dom}\, f_{\mathrm{QST}}$. It is easily checked that f_{QST} is a self-concordant function of parameter $2\sqrt{n}$. Following the theory in [136, 140], the radius of the Dikin ellipsoid shrinks at the rate $O(n^{-1/2})$, so SCOPT becomes slow when n is large.

3. The Frank-Wolfe algorithm suffers for a sub-linear convergence rate when the solution is near an extreme point of the constraint set (see, e.g., [111] for an illustration in the

[2] In a standard implementation of quantum state tomography, the matrices M_i are single-rank [85].

vector case). Notice that the set of extreme points of $\mathscr{D}$ is the set of single-rank positive semi-definite matrices of unit trace. In the experimental data we have, the density matrix to be estimated is indeed close to a single-rank matrix (which is called a *pure state* in quantum mechanics). Therefore, the ML estimate—the minimizer of f_{QST} on $\mathscr{D}$—is expected to be also close to a single-rank matrix.

5.7 Concluding remarks

We have proved convergence of the mirror descent under a novel local relative smoothness condition, which is satisfied in many applications. Indeed, according to our discussion in Section 5.3, the local relative smoothness condition lies strictly between two existing conditions: the objective function being differentiable, and the objective function has a locally Lipschitz gradient. Our numerical results showed that the exponentiated gradient method with Armijo line search is a rigorous and efficient solution to portfolio selection and quantum state tomography.

We point out two future research directions. First, our result only considers convergence in function value. It would be good to also establish convergence in iterates, or show that such convergence cannot hold in general without modifying the algorithm. Note that for portfolio selection and quantum state tomography, uniqueness of the minimizer is easily checked via self-concordance of the objective function [140, 136]; hence, convergence in function value implies convergence in iterates. Second, it is of both practical and theoretical importance to study convergence for possibly non-convex objective functions, under weak smoothness conditions. This topic was recently studied in [68, 142], using the notion of a *growth function*. The results therein require in addition convergence in iterates and/or other conditions on the objective function, verifying which is an issue in general.

5.A Proof of Proposition 5.2

Consider the two-dimensional case, where $x = (x_{i,j})_{1\le i,j\le 2} \in \mathbb{C}^{2\times 2}$. Define $e_1 := (1,0)$ and $e_2 := (0,1)$. Suppose that there are only two summands, with $M_1 = e_1 \otimes e_1$ and $M_2 = e_2 \otimes e_2$. Then we have $f(x) = -\log(x_{1,1}) - \log(x_{2,2})$. It suffices to disprove all properties on the set of diagonal density matrices. Hence, we will focus on the function $g(x,y) := -\log x - \log y$, defined for any (x,y) in the probability simplex $\mathscr{P} \subset \mathbb{R}^2$.

As either x or y can be arbitrarily close to zero, it is easily checked that the gradient of g is unbounded. Now we check the relative smoothness condition. As we only consider diagonal matrices, it suffices to check with respect to the (negative) Shannon entropy:

$$h(x,y) := -x\log x - y\log y + x + y, \quad \forall (x,y) \in \mathscr{P},$$

for which the convention $0\log 0 := 0$ is adopted.

Lemma 5.18 ([123]). *The function g is L-smooth relative to the Shannon entropy for some $L > 0$, if and only if $-Lh - g$ is convex.*

Therefore, we check the positive semi-definiteness of the Hessian of $-Lh - g$. A necessary condition for the Hessian to be positive semi-definite is that

$$-L\frac{\partial^2 h}{\partial x^2}(x,y) - \frac{\partial^2 g}{\partial x^2}(x,y) = \frac{L}{x} - \frac{1}{x^2} \geq 0,$$

for all $x \in (0,1)$, but the inequality cannot hold for $x < (1/L)$, for any fixed $L > 0$.

5.B Proof of Lemma 5.4

(Statement 2 $\Rightarrow$ Statement 1) Let $x \in \mathscr{X} \cap \operatorname{dom} f$, and $z_1, z_2 \in \mathscr{B}_{\varepsilon_x}(x) \cap \tilde{\mathscr{X}}$. Define, for every $\tau \in [0,1]$, $z_\tau := z_1 + \tau(z_2 - z_1)$. We write

$$\begin{aligned} f(z_2) - \big[f(z_1) + \langle \nabla f(z_1), z_2 - z_1\rangle\big] &= \int_0^1 \langle \nabla f(z_\tau) - \nabla f(z_1), z_2 - z_1\rangle \,\mathrm{d}\tau \\ &\leq \int_0^1 \|\nabla f(z_\tau) - \nabla f(z_1)\|_2 \|z_2 - z_1\|_2 \,\mathrm{d}\tau \\ &\leq \int_0^1 L_x \tau \|z_2 - z_1\|_2^2 \,\mathrm{d}\tau \\ &= \frac{L_x}{2}\|z_2 - z_1\|_2^2, \end{aligned}$$

where we have applied the Cauchy-Schwarz inequality for the first inequality, and the local smoothness condition for the second inequality. Note that $\mathscr{B}_{\varepsilon_x} \cap \tilde{\mathscr{X}}$ is the intersection of convex sets, and hence is convex; therefore, $z_\tau \in \mathscr{B}_{\varepsilon_x} \cap \tilde{\mathscr{X}}$ for every $\tau \in [0,1]$.

(Statement 1 $\Rightarrow$ Statement 2) Let $x \in \mathscr{X} \cap \operatorname{dom} f$, and $z_1, z_2 \in \mathscr{B}_{\varepsilon_x}(x) \cap \tilde{\mathscr{X}}$. Define $\varphi(z) := f(z) - \langle \nabla f(z_1), z\rangle$. Then $\nabla\varphi$ is locally Lipschitz on $\tilde{\mathscr{X}}$; moreover, since $\nabla\varphi(z_1) = 0$, the point z_1 is a global minimizer of φ. Therefore, we obtain

$$\varphi(z_1) \leq \varphi(z_2 - \frac{1}{L_x}\nabla\varphi(z_2)) \leq \varphi(z_2) - \frac{1}{2L_x}\|\nabla\varphi(z_2)\|^2;$$

that is,

$$f(z_2) \geq f(z_1) + \langle \nabla f(z_1), z_2 - z_1\rangle + \frac{1}{2L_x}\|\nabla f(z_2) - \nabla f(z_1)\|_2^2.$$

Similarly, we get

$$f(z_1) \geq f(z_2) + \langle \nabla f(z_2), z_1 - z_2\rangle + \frac{1}{2L_x}\|\nabla f(z_1) - \nabla f(z_2)\|_2^2.$$

Summing up the two inequalities; we obtain

$$\langle \nabla f(z_2) - \nabla f(z_1), z_2 - z_1 \rangle \geq \frac{1}{L_x} \|\nabla f(z_2) - \nabla f(z_1)\|_2^2.$$

This implies, by the Cauchy-Schwarz inequality,

$$\|\nabla f(z_2) - \nabla f(z_1)\|_2 \leq L_x \|z_2 - z_1\|_2.$$

5.C Auxiliary technical lemmas for proving Theorem 5.11

Lemma 5.19. *If $x(\alpha) = x$ for some $x \in \tilde{\mathscr{X}}$, then x is a solution to* (P)*. If a point $x \in \tilde{\mathscr{X}}$ is a solution to* (P)*, then $x(\alpha) = x$ for all $\alpha \in [0, +\infty)$.*

Proof. That x is a solution to (P) is equivalent to the optimality condition

$$\langle \nabla f(x), z - x \rangle \geq 0, \quad \forall z \in \mathscr{X}.$$

We can equivalently write

$$\langle \alpha \nabla f(x) + \nabla h(x) - \nabla h(x), z - x \rangle \geq 0, \quad \forall z \in \mathscr{X},$$

which is the optimality condition of

$$x(\alpha) = \arg\min_z \{ \alpha \langle \nabla f(x), z - x \rangle + D_h(z, x) \mid z \in \mathscr{X} \}.$$ □

Lemma 5.20. *For every $x \in \tilde{\mathscr{X}}$ and $\alpha > 0$, it holds that*

$$\langle \nabla f(x(\alpha)), x(\alpha) - x \rangle \leq -\alpha^{-1} D(x(\alpha), x) \leq 0.$$

Proof. By definition, we have

$$\alpha \langle \nabla f(x(\alpha)), x(\alpha) - x \rangle + D(x(\alpha), x) \leq \alpha \langle \nabla f(x, x - x \rangle + D(x, x) = 0.$$ □

Corollary 5.21. *The sequence $\{x_k\}$ is non-increasing.*

Proof. The Armijo rule and Lemma 5.20 guarantee that

$$f(x_k) \leq f(x_{k-1}) + \tau \langle \nabla f(x_{k-1}), x_k - x_{k-1} \rangle \leq f(x_{k-1}).$$ □

5.D Proof of Theorem 5.14

For every $u \in \mathcal{X} \cap \operatorname{dom} f$, we write

$$\begin{aligned} f(x_{k-1}) - f(u) &\leq -\langle \nabla f(x_{k-1}), u - x_{k-1} \rangle \\ &= -\langle \nabla f(x_{k-1}), u - x_k \rangle - \langle \nabla f(x_{k-1}), x_k - x_{k-1} \rangle . \end{aligned}$$

The optimality condition for x_k implies

$$\langle \alpha_k \nabla f(x_{k-1}) + \nabla h(x_k) - \nabla h(x_{k-1}), u - x_k \rangle \geq 0.$$

Applying the *three-point identity*, we obtain

$$\begin{aligned} \langle \nabla f(x_{k-1}), u - x_k \rangle &\geq -\alpha_k^{-1} \langle \nabla h(x_k) - \nabla h(x_{k-1}), u - x_k \rangle \\ &= -\alpha_k^{-1} \left[D_h(u, x_{k-1}) - D_h(u, x_k) - D_h(x_k, x_{k-1}) \right] \\ &\geq -\alpha_k^{-1} \left[D_h(u, x_{k-1}) - D_h(u, x_k) \right] . \end{aligned}$$

Then we can write

$$\alpha_k \left[f(x_{k-1}) - f(u) \right] \leq \left[D_h(u, x_{k-1}) - D_h(u, x_k) \right] - \alpha_k \langle \nabla f(x_{k-1}), x_k - x_{k-1} \rangle .$$

Summing up the inequality for all $1 \leq k \leq n$, we get

$$-S_n f(u) + \sum_{k=1}^{n} \alpha_k f(x_{k-1}) \leq D(u, x_0) - \sum_{k=1}^{n} \alpha_k \langle \nabla f(x_{k-1}), x_k - x_{k-1} \rangle ,$$

where $S_n := \sum_{k=1}^{n} \alpha_k$. Corollary 5.21 says that the sequence $(f(x_k))_{k \in \mathbb{N}}$ is non-increasing; then we have

$$\sum_{k=1}^{n} \alpha_k f(x_{k-1}) \geq \sum_{k=1}^{n} \alpha_k f(x_n) = S_n f(x_n).$$

Therefore, we obtain

$$f(x_n) - f(u) \leq S_n^{-1} \left[D(u, x_0) - \sum_{k=1}^{n} \alpha_k \langle \nabla f(x_{k-1}), x_k - x_{k-1} \rangle \right] .$$

Note that by the Armijo rule, we have

$$\begin{aligned} f(x_0) - f^\star &\geq \lim_{k \to \infty} f(x_0) - f(x_k) \\ &= \sum_{j=1}^{\infty} \left[f(x_{j-1}) - f(x_j) \right] \\ &\geq -\tau \sum_{j=1}^{\infty} \langle \nabla f(x_{j-1}), x_j - x_{j-1} \rangle . \end{aligned}$$

Therefore, $\langle \nabla f(x_{k-1}), x_k - x_{k-1} \rangle$, which are non-negative by Lemma 5.15, must converge to zero. Theorem 5.14 then follows from the following lemma.

Lemma 5.22 ([149]). *Let $\{a_k\}$ be a sequence of real numbers, and $\{b_k\}$ be a sequence of positive real numbers. Define $c_n := \sigma_n^{-1} \sum_{k=1}^n b_k a_k$ for every $n \in \mathbb{N}$, where $\sigma_n := \sum_{k=1}^n b_k$. If $a_k \to 0$ and $\sigma_n \to +\infty$, then $c_n \to 0$.*

5.E Proof of Proposition 5.17

Note that $\mathrm{dom}(f_{\mathrm{QST}})$ is open, and hence $\mathrm{dom}(f_{\mathrm{QST}}) \cap \mathscr{X}$ is relatively open in $\mathscr{X}$. It is easily checked that f_{QST} is twice continuously differentiable on $\mathrm{dom}(f_{\mathrm{QST}})$, and hence on $\mathrm{dom}(f_{\mathrm{QST}}) \cap \mathscr{X}$. By Proposition 5.5, the function f_{QST} is locally smooth relative to $h(\cdot) := (1/2)\|\cdot\|_{\mathrm{F}}^2$, where $\|\cdot\|_{\mathrm{F}}$ denotes the Frobenius norm. By the quantum version of Pinsker's inequality [92], the von Neumann entropy is strongly convex on $\mathscr{D}$ with respect to the trace norm. As all norms on a finite-dimensional space are equivalent, the proposition follows.

6 A general convergence result for the exponentiated gradient method

In the previous chapter, we have proved that the mirror descent with Armijo line search converges, as long as the loss function is locally relatively smooth. Is it possible to further get rid of the local relative smoothness condition? In this chapter, we explore the possibility of deriving similar results for mirror descent-type algorithms. Specifically, we prove that the exponentiated gradient method—arguably the most well-known instance of mirror descent-type algorithms—always converges for the problem of minimizing a continuously differentiable convex function on the spectahedron, if the sequence of iterates possesses a strictly positive limit point. As a byproduct, we obtain an improved Peierls-Bogoliubov inequality, via the self-concordant likeness of a log-partition function.

This chapter is based on the joint work with Volkan Cevher [119].

6.1 Introduction

Consider the optimization problem

$$f^\star = \min\{f(\rho) \mid \rho \in \mathscr{D}\}, \tag{P}$$

where f is a convex function differentiable on $\operatorname{int}\operatorname{dom} f$, and $\mathscr{D}$ denotes the set of quantum density matrices, i.e.,

$$\mathscr{D} := \{\rho \in \mathbb{C}^{d\times d} \mid \rho \geq 0, \operatorname{Tr}\rho = 1\},$$

for some positive integer d. We assume that $f^\star > -\infty$.

This problem formulation (P) allows us to address two other constraints simultaneously:

- The probability simplex $\mathscr{P} := \{x \in \mathbb{R}^d_+ \mid \|x\|_1 = 1\}$.
- The spectahedron $\mathscr{S} := \{X \in \mathbb{R}^{d\times d} \mid X \geq 0, \operatorname{Tr} X = 1\}$.

Algorithm 3 Exponentiated Gradient Method with Armijo Line Search

Require: $\bar{\alpha} > 0$, $r \in (0,1)$, $\tau \in (0,1)$, $\rho_0 \in \mathscr{D}$ non-singular
1: **for** $k = 1, 2, \ldots$ **do**
2: $\quad \alpha_k \leftarrow \bar{\alpha}$
3: $\quad$ **while** $f(\rho_{k-1}(\alpha_k)) > f(\rho_{k-1}) + \tau \left\langle f'(\rho_{k-1}), \rho_{k-1}(\alpha_k) - \rho_{k-1} \right\rangle$ **do**
4: $\quad\quad \alpha_k \leftarrow r\alpha_k$
5: $\quad$ **end while**
6: $\quad \rho_k \leftarrow \rho_{k-1}(\alpha_k)$
7: **end for**

Optimization problems with a probability simplex, spectahedron, or quantum density matrix constraint appear in various applications, such as sparse regression [170], low-rank matrix estimation [106], and quantum state tomography [146], to mention a few; the corresponding objective functions are typically convex and differentiable.

Starting with some non-singular $\rho_0 \in \mathscr{D}$, the exponentiated gradient (EG) method iterates as

$$\rho_k = C_k^{-1} \exp\left[\log(\rho_{k-1}) - \alpha_k \nabla f(\rho_{k-1})\right], \quad k \in \mathbb{N}, \tag{6.1}$$

where C_k is a positive real number normalizing the trace of ρ_k, and $\alpha_k > 0$ denotes the step size. Equivalently, one may write

$$\rho_k \in \arg\min\{\alpha_k \langle \nabla f(\rho_{k-1}, \sigma - \rho_{k-1})\rangle + D(\sigma, \rho_{k-1}) \mid \sigma \in \mathscr{D}\}, \tag{6.2}$$

where D denotes the quantum relative entropy. Therefore, the EG method can be viewed as mirror descent with the von Neumann entropy [19, 135, 174], or a special case of the interior gradient method [7].

There are various approaches to selecting the step size. We focus on Armijo line search. Let $\bar{\alpha} > 0$ and $r, \tau \in (0,1)$. The Armijo line search procedure outputs $\alpha_k = r^j \bar{\alpha}$, where j is the least non-negative integer that satisfies

$$f(\rho_k) \leq f(\rho_{k-1}) + \tau \langle \nabla f(\rho_{k-1}), \rho_k - \rho_{k-1}\rangle;$$

the dependence on j lies implicitly in ρ_k. We give the pseudo codes in Algorithm 3, where we define

$$\rho_{k-1}(\alpha_k) := \tilde{C}_k^{-1} \exp\left[\log(\rho_{k-1}) - \alpha_k \nabla f(\rho_{k-1})\right], \quad \forall k \in \mathbb{N};$$

$\tilde{C}_k$ normalizes the trace of $\rho_{k-1}(\alpha_k)$.

Implementing Armijo line search does not require any parameter of the objective function, e.g., the Lipschitz constant of the objective function or its gradient. This observation shows the possibility of proving a convergence guarantee for the EG method with respect to a general class of objective functions. Indeed, we will only assume that the objective function is convex

and differentiable throughout this chapter.

6.1.1 Related work

Recall that, as discussed in the previous chapter, quantum state tomography, positron emission tomography, optimal portfolio selection, and non-negative linear inverse problems all require one to solve an optimization problem of the form (P), but existing results cannot guarantee convergence of the exponentiated gradient method. The essential reason is that existing results require either 1) the gradient ∇f is bounded, or 2) the objective function is smooth relative to the negative entropy, but none of the conditions holds in these applications.

There are some convergence guarantees that require mild differentiability conditions, but they are all for gradient descent-type methods. Bertsekas proved that the projected gradient descent with Armijo line search always converges for a differentiable objective function, when the constraint is a box or the positive orthant [23]. Gafni and Bertsekas generalized the previous result for any compact convex constraint [78]. Salzo proved the convergence of proximal variable metric methods with various line search schemes, assuming that ∇f is uniformly continuous on any compact set [161].

Our proof relies on the self-concordant likeness of a log-partition function. The notion of self-concordant likeness is closely related to that of self-concordance, the foundation of interior point methods [140]. Self-concordant likeness was proposed by Bach for statistical analyses [8]; it was introduced to the field of convex optimization to derive fast convergence rates for non-strongly convex functions in [10, 173]. A generalization of self-concordant likeness for non-Euclidean norms can be found in [55, 117]. A systematic study of self-concordant likeness and other variants of self-concordance can be found in [168].

6.1.2 Contributions

In comparison to existing results, we highlight the following contributions.

- To the best of our knowledge, we give the first convergence guarantee of a mirror descent-type method[1] that only requires differentiability.
- Our convergence analysis exploits the self-concordant likeness of the log partition function. As a by-product, we improve on the Peierls-Bogoliubov inequality, which is of independent interest (cf. Remark 6.9).

6.2 Main result

Our main result is the following theorem.

[1] Here we exclude the very standard projected gradient method.

Theorem 6.1. *Suppose that f is differentiable at every non-singular $\rho \in \mathscr{D}$. Then we have:*

1. *The Armijo line search procedure terminates in finite steps.*
2. *The sequence $(f(\rho_k))_{k\in\mathbb{N}}$ is non-increasing.*
3. *For any converging sub-sequence $(\rho_k)_{k\in\mathscr{K}}$, $\mathscr{K} \subseteq \mathbb{N}$, it holds that*

$$\liminf\{D(\rho_k(\beta), \rho_k) \mid k \in \mathscr{K}\} = 0,$$

for every $\beta > 0$, where $\rho_k(\beta)$ denotes the next iterate of ρ_k with step size β.

Remark 6.2. *Statement 3 is always meaningful—due to the compactness of $\mathscr{D}$, there exists at least one converging sub-sequence of $(\rho_k)_{k\in\mathbb{N}}$.*

Taking limit, we obtain the following convergence guarantee.

Corollary 6.3. *If the sequence $(\rho_k)_{k\in\mathbb{N}}$ possesses a non-singular limit point, the sequence $(f(\rho_k))_{k\in\mathbb{N}}$ monotonically converges to $f^\star$.*

Proof. Let $(\rho_k)_{k\in\mathscr{K}}$ be a sub-sequence converging to a non-singular matrix $\rho_\infty \in \mathscr{D}$. By Statement 3 of Theorem 6.1, there exists a sub-sequence $(\rho_k)_{k\in\mathscr{K}'}$, $\mathscr{K}' \subseteq \mathscr{K}$, such that $D(\rho_k(\beta), \rho_k) \to 0$ as $k \to \infty$ in $\mathscr{K}'$. As ρ_∞ is non-singular, we can take the limit and obtain $D(\rho_\infty(\beta), \rho_\infty) = 0$, showing that $\rho_\infty(\beta) = \rho_\infty$. Lemma 6.18 in the appendix then implies that ρ_∞ is a minimizer of f on $\mathscr{D}$. Since the sequence $(f(\rho_k))_{k\in\mathbb{N}}$ is non-increasing and bounded from below by $f^\star$, $\lim_{k\to\infty} f(\rho_k)$ exists. We write

$$f^\star \leq \lim_{k\to\infty} f(\rho_k) = \liminf\{f(\rho_k) \mid k \in \mathbb{N}\} \leq f(\rho_\infty) = f^\star. \qquad \square$$

It is currently unclear to us whether convergence to the optimum holds, when there does not exist a non-singular limit point; see Section 6.4.3 for a discussion. One way to get around is to consider solving

$$f_\lambda^\star = \min\{f(\rho) - \lambda \log\det\rho \mid \rho \in \mathscr{D}\}, \tag{P-λ}$$

where λ is a positive real number. As $-\log\det(\cdot)$ is a barrier function for the set of positive semi-definite matrices [140], every limit point must be non-singular; otherwise, monotonicity of the sequence $(f(\rho_k))_{k\in\mathbb{N}}$ (Statement 2 in Theorem 6.1) cannot hold.

Proposition 6.4. *It holds that* $\lim_{\lambda\downarrow 0} f_\lambda^\star = f^\star$.

Proof. Notice that $-\log\det(\cdot) > 0$ on $\mathscr{D}$. We write

$$\lim_{\lambda\downarrow 0} f_\lambda^\star = \inf_{\lambda>0} f_\lambda^\star = \inf_{\lambda>0}\inf_{\rho\in\mathscr{D}} f_\lambda(\rho) = \inf_{\rho\in\mathscr{D}}\inf_{\lambda>0} f_\lambda(\rho) = \inf_{\rho\in\mathscr{D}} f(\rho) = f^\star,$$

where $f_\lambda(\rho) := f(\rho) - \lambda\log\det\rho$. $\square$

Existence of a non-singular limit point can be easily verified in some applications. For example, hedged quantum state tomography corresponds to solving (P) with the objective function

$$f_{\mathrm{HQST}}(\rho) := f_{\mathrm{QST}}(\rho) - \lambda \log\det\rho,$$

for some $\lambda > 0$ [26]. As discussed above, all limit points of the iterates must be non-singular. Similarly in the probability simplex constraint case, if the optimization problem involves the Burg entropy as in [64], all limit points must be element-wisely strictly positive[2].

Notation

Let $A \in \mathbb{C}^{d\times d}$. We denote its largest and smallest eigenvalues by $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$, respectively. We denote its Schatten p-norm by $\|A\|_p$. We will only use the Hilbert-Schmidt inner product in this chapter; that is, $\langle A, B\rangle := \mathrm{Tr}(A^{\mathrm{H}}B)$ for any $A, B \in \mathbb{C}^{d\times d}$, where A^{H} denotes the Hermitian of A.

6.3 Proof of Theorem 6.1

The key to our analysis is the following proposition.

Proposition 6.5. *Let $\rho \in \mathscr{D}$ be non-singular. Suppose that*

$$\Delta := \lambda_{\max}(\nabla f(\rho)) - \lambda_{\min}(\nabla f(\rho)) > 0.$$

Then the mapping

$$\alpha \mapsto \frac{D(\rho(\alpha), \rho)}{\mathrm{e}^{\Delta\alpha}(\Delta\alpha - 1) + 1} \tag{6.3}$$

is non-increasing on $(0, +\infty)$.

Proposition 6.5 was inspired by a lemma due to Gafni and Bertsekas [78], which says that the mapping

$$\alpha \mapsto \frac{\|\Pi_{\mathscr{D}}(\rho - \alpha\nabla f(\rho)) - \rho\|_{\mathrm{F}}}{\alpha} \tag{6.4}$$

is non-increasing on $[0, +\infty)$, where $\Pi_{\mathscr{D}}$ denotes projection onto $\mathscr{D}$ with respect to the Frobenius norm $\|\cdot\|_{\mathrm{F}}$. The lemma of Gafni and Bertsekas was proved by an Euclidean geometric argument; see [24] for an illustration. In comparison, we will prove Proposition 6.5 by exploiting the self-concordant likeness of the log-partition function.

We prove Proposition 6.5 in Section 6.3.1. Then we prove the three statements in Theorem

[2] For any element-wisely strictly positive vector $v := (v_i)_{1\le i\le d}$, the Burg entropy is defined as $b(v) := -\sum_{i=1}^{d} \log v_i$.

6.1 separately in the following three sub-sections. To simplify the presentation, we put some necessary technical lemmas in Appendix 6.A.

6.3.1 Self-concordant likeness of the log-partition function and proof of Proposition 6.5

For any non-singular $\rho \in \mathscr{D}$ and $\alpha > 0$, define

$$\varphi(\alpha;\rho) := \log \mathrm{Tr} \exp\left[\log(\rho) - \alpha \nabla f(\rho)\right],$$

which, in statistical physics, is the log-partition function of the Gibbs state for the Hamiltonian $H_\alpha := -\log(\rho) + \alpha \nabla f(\rho)$ at temperature 1. We will simply write $\varphi(\alpha)$ instead of $\varphi(\alpha;\rho)$, when the corresponding ρ is clear from the context or irrelevant.

The log-partition function is indeed closely related to the EG method, as shown by the following lemma.

Lemma 6.6. *For any non-singular $\rho \in \mathscr{D}$ and $\alpha > 0$, it holds that*

$$D(\rho(\alpha), \rho) = \varphi(0) - \left[\varphi(\alpha) + \varphi'(\alpha)(0-\alpha)\right].$$

Proof. Direct calculation. □

We say that a three times continuously differentiable convex function g is *μ-self-concordant like*, if $|g'''(x)| \le \mu g''(x)$ for all x [8, 10, 173].

Lemma 6.7. *For any non-singular $\rho \in \mathscr{D}$, the function $\varphi(\alpha)$ is Δ-self-concordant like, where $\Delta := \lambda_{\max}(\nabla f(\rho)) - \lambda_{\min}(\nabla f(\rho))$.*

Proof. Lemma 6.19 shows that

$$\varphi''(\alpha) = \mathsf{E}\left(\eta_\alpha - \mathsf{E}\eta_\alpha\right)^2, \quad \varphi'''(\alpha) = \mathsf{E}\left(\eta_\alpha - \mathsf{E}\eta_\alpha\right)^3,$$

where η_α is a random variable taking values in $[-\lambda_{\max}(\nabla f(\rho)), -\lambda_{\min}(\nabla f(\rho))]$. The lemma follows. □

The following sandwich inequality follows from self-concordant likeness [173].

Lemma 6.8. *Suppose that $\Delta > 0$. For any non-singular $\rho \in \mathscr{D}$, it holds that either ρ is a solution of* (P)*, or*

$$\frac{\left(\mathrm{e}^{-\Delta\alpha} + \Delta\alpha - 1\right)}{\Delta^2}\varphi''(\alpha) \le \varphi(0) - \left[\varphi(\alpha) + \varphi'(\alpha)(0-\alpha)\right] \le \frac{\left(\mathrm{e}^{\Delta\alpha} - \Delta\alpha - 1\right)}{\Delta^2}\varphi''(\alpha).$$

Remark 6.9. *The lower bound improves upon the Peierls-Bogoliubov inequality [144], which says that*

$$0 \le \varphi(0) - \left[\varphi(\alpha) + \varphi'(\alpha)(0-\alpha)\right].$$

Notice that lower bound provided by Lemma 6.8 is always non-negative.

For completeness, we provide a proof in Section 6.B.

Now we are ready to prove Proposition 6.5.

Proof of Proposition 6.5. We look for a differentiable function $\chi : (0,+\infty) \to (0,+\infty)$, such that the mapping

$$g(\alpha) := \frac{D(\rho(\alpha), rho)}{\chi(\alpha)}$$

is non-increasing on $(0,+\infty)$. Note that g is non-increasing if and only if $g' \le 0$ on $(0,+\infty)$. Applying Lemma 6.6, a direct calculation gives

$$g'(\alpha) = \frac{\alpha\varphi''(\alpha)\chi(\alpha) - \left\{\varphi(0) - \left[\varphi(\alpha) + \varphi'(\alpha)(0-\alpha)\right]\right\}\chi'(\alpha)}{\left[\chi'(\alpha)\right]^2}.$$

Therefore, $g'(\alpha) \le 0$ if and only if the numerator is negative, i.e.,

$$(\log\chi)'(\alpha) \ge \frac{\alpha\varphi''(\alpha)}{\varphi(0) - \left[\varphi(\alpha) + \varphi'(\alpha)(0-\alpha)\right]},$$

where we have used the fact that $\chi'/\chi = (\log\chi)'$. By Lemma 6.8, we can set

$$(\log\chi)'(\alpha) = \frac{\Delta^2\alpha}{\mathrm{e}^{-\Delta\alpha} + \Delta\alpha - 1}.$$

Solving the equation gives $\chi(\alpha) := \mathrm{e}^{\Delta\alpha}(\Delta\alpha - 1) + 1$. □

For convenience, we will apply Proposition 6.5 via the following corollary.

Corollary 6.10. *Let $\rho \in \mathscr{D}$ be non-singular and $\bar{\alpha} > 0$. Suppose that $\Delta := \lambda_{\max}(\nabla f(\rho)) - \lambda_{\min}(\nabla f(\rho))$ is strictly positive. It holds that*

$$\frac{D(\rho(\alpha),\rho)}{\alpha^2} \ge \kappa D(\rho(\bar{\alpha}),\rho), \quad \forall \alpha \in (0,\bar{\alpha}],$$

where $\kappa := \left\{2\left[\mathrm{e}^{\Delta\bar{\alpha}}(\Delta\bar{\alpha} - 1) + 1\right]\right\}^{-1}\Delta^2$.

Proof. Define $g(\alpha) := \mathrm{e}^{\Delta\alpha}(\Delta\alpha - 1) + 1 - (\Delta^2/2)\alpha^2$. Then $g(0) = 0$, and

$$g'(\alpha) = \alpha\left[\mathrm{e}^{\Delta\alpha}\Delta^2 - \Delta^2\right] \ge \alpha(\Delta^2 - \Delta^2) = 0, \quad \forall \alpha \in (0,+\infty).$$

Therefore, $g(\alpha) \geq 0$ on $(0, +\infty)$, i.e.,

$$e^{\Delta\alpha}(\Delta\alpha - 1) + 1 \geq \frac{\Delta^2}{2}\alpha^2, \quad \forall \alpha \in (0, +\infty).$$

By Proposition 6.5, we write

$$\frac{D(\rho(\alpha), \rho)}{\frac{\Delta^2}{2}\alpha^2} \geq \frac{D(\rho(\alpha), \rho)}{e^{\Delta\alpha}(\Delta\alpha - 1) + 1} \geq \frac{D(\rho(\bar{\alpha}), \rho)}{e^{\Delta\bar{\alpha}}(\Delta\bar{\alpha} - 1) + 1}, \quad \forall \alpha \in (0, \bar{\alpha}]. \qquad \square$$

6.3.2 Proof of Statement 1

The first statement is a direct consequence of the following proposition.

Proposition 6.11. *For every non-singular $\rho \in \mathscr{D}$, there exists some $\alpha_\rho > 0$ such that*

$$f(\rho(\alpha)) \leq f(\rho) + \tau \langle \nabla f(\rho), \rho(\alpha) - \rho \rangle, \quad \forall \alpha \in [0, \alpha_\rho]. \tag{6.5}$$

Recall that τ is the parameter in Armijo line search.

Proof. If ρ is a minimizer, by Lemma 6.18, we have $\rho(\alpha) = \rho$ for all $\alpha \in [0, +\infty)$, and the proposition follows. Suppose that ρ is not a minimizer in the rest of this proof. By Lemma 6.18, we have $D(\rho(\alpha), \rho) > 0$ for all $\alpha \in (0, +\infty)$. By the mean-value theorem, we write

$$f(\rho(\alpha)) - f(\rho) = \langle \nabla f(\sigma), \rho(\alpha) - \rho \rangle,$$

for some σ in the line segment joining $\rho(\alpha)$ and ρ. Then (6.5) can be equivalently written as

$$\langle \nabla f(\sigma) - \nabla f(\rho), \rho(\alpha) - \rho \rangle \leq -(1 - \tau) \langle \nabla f(\rho), \rho(\alpha) - \rho \rangle, \quad \forall \alpha \in [0, \alpha_\rho]. \tag{6.6}$$

By Lemma 6.17, (6.6) holds if

$$\langle \nabla f(\sigma) - \nabla f(\rho), \rho(\alpha) - \rho \rangle \leq \frac{(1 - \tau) D(\rho(\alpha), \rho)}{\alpha}, \quad \forall \alpha \in [0, \alpha_\rho]. \tag{6.7}$$

Consider two cases.

- If $\lambda_{\max}(\nabla f(\rho)) = \lambda_{\min}(\nabla f(\rho))$, then $\nabla f(\rho)$ is a multiple of the identity. We have

  $$\langle \nabla f(\rho), \sigma - \rho \rangle = 0, \quad \forall \sigma \in \mathscr{D};$$

  showing that ρ is a minimizer. By Lemma 6.18, the proposition follows for every $\alpha_\rho > 0$.

- Otherwise, set $\alpha_\rho \leq \bar{\alpha}$. By Corollary 6.10, there exists some $\kappa > 0$, such that

  $$\frac{D(\rho(\alpha), \rho)}{\alpha} \geq \sqrt{D(\rho(\alpha), \rho)} \sqrt{\kappa D(\rho(\bar{\alpha}), \rho)}, \quad \forall \alpha \in [0, \alpha_\rho].$$

Applying Hölder's inequality and Pinsker's inequality, we write

$$\begin{aligned}\langle \nabla f(\sigma) - \nabla f(\rho), \rho(\alpha) - \rho \rangle &\leq \|\nabla f(\sigma) - \nabla f(\rho)\|_\infty \|\rho(\alpha) - \rho\|_1 \\ &\leq \|\nabla f(\sigma) - \nabla f(\rho)\|_\infty \sqrt{2D(\rho(\alpha), \rho)}.\end{aligned}$$

Then (6.7) holds if

$$\|\nabla f(\sigma) - \nabla f(\rho)\|_\infty \sqrt{2} \leq (1-\tau)\sqrt{\kappa D(\rho(\bar{\alpha}), \rho)}, \quad \forall \alpha \in [0, \alpha_\rho]$$

Recall that a convex differentiable function is continuously differentaible [156]. Notice that $\rho(\alpha)$ is continuous in α. As the right-hand side is a strictly positive constant by Lemma 6.18, the proposition follows for a small enough α_ρ. □

6.3.3 Proof of Statement 2

By the definition of Armijo line search and Lemma 6.17, we have

$$f(\rho_k) \leq f(\rho_{k-1}) + \tau \langle \nabla f(\rho_{k-1}), \rho_k - \rho_{k-1} \rangle \leq f(\rho_{k-1}) - \frac{\tau D(\rho_k, \rho_{k-1})}{\alpha_k}.$$

As the quantum relative entropy D is always non-negative, it follows that the sequence $(f(\rho_k))_{k\in\mathbb{N}}$ is non-increasing.

6.3.4 Proof of Statement 3

If ρ_k is a minimizer for some $k \in \mathbb{N}$, by Lemma 6.18, it holds that $\rho_{k'} = \rho_k$ for all $k' > k$, and the statement trivially follows. In the rest of this sub-section, we assume that ρ_k is not a minimizer for all k; then by Lemma 6.18, it holds that $\rho_k \neq \rho_{k-1}$ for all $k \in \mathbb{N}$.

Let $(\rho_k)_{k\in\mathcal{K}}$ be a sub-sequence converging to a limit point $\rho_\infty \in \mathscr{D}$, which exists due to the compactness of $\mathscr{D}$. Then ρ_∞ must lie in $\operatorname{int}\operatorname{dom} f$; otherwise, monotonicity of the sequence $(f(\rho_k))_{k\in\mathbb{N}}$ (Statement 2 of Theorem 6.1) cannot hold. As f is continuously differentiable, it holds that

$$\frac{\Delta_\infty}{2} \leq \lambda_{\max}(\nabla f(\rho_k)) - \lambda_{\min}(\nabla f(\rho_k)) \leq 2\Delta_\infty, \tag{6.8}$$

for large enough $k \in \mathcal{K}$, where $\Delta_\infty := \lambda_{\max}(\nabla f(\rho_\infty)) - \lambda_{\min}(\nabla f(\rho_\infty)$.

Lemma 6.12. *If* $\Delta_\infty = 0$, *then* $\liminf\{D(\rho_k(\beta), \rho_k) \mid k \in \mathcal{K}\} = 0$ *for every* $\beta \in [0, +\infty)$.

Proof. Define $\Delta_k := \lambda_{\max}(\nabla f(\rho_k)) - \lambda_{\min}(\nabla f(\rho_k))$; then $\Delta_k \to \Delta_\infty = 0$. Define $\varphi_k : \alpha \mapsto \varphi(\alpha; \rho_k)$.

By Lemma 6.8 and Corollary 6.20, we have

$$\begin{aligned}\varphi_k(0) - \left[\varphi_k(\beta) - \varphi_k'(\beta)(0-\beta)\right] &\leq \frac{\left(\mathrm{e}^{\Delta_k\beta} - \Delta_k\beta - 1\right)}{\Delta_k^2}\varphi_k''(\beta) \\ &\leq \frac{\left(\mathrm{e}^{\Delta_k\beta} - \Delta_k\beta - 1\right)}{4}.\end{aligned}$$

By Lemma 6.6, we obtain

$$\begin{aligned}0 &\leq \liminf\{D(\rho_k(\beta),\rho_k) \mid k \in \mathcal{K}\} \\ &= \liminf\{\varphi_k(0) - \left[\varphi_k(\beta) - \varphi_k'(\beta)(0-\beta)\right] \mid k \in \mathcal{K}\} \\ &\leq \frac{\mathrm{e}^0 - 0 - 1}{4} = 0.\end{aligned}$$

□

Suppose that $\Delta_\infty > 0$. We have the following analogy of Corollary 6.10 for large enough $k \in \mathcal{K}$:

Corollary 6.13. *Suppose that $\Delta_\infty > 0$ and ρ_k is not a minimizer for every $k \in \mathcal{K}$. There exists some $\kappa > 0$, such that*

$$\frac{D(\rho_k(\alpha),\rho_k)}{\alpha^2} \geq \kappa D(\rho_k(\bar{\alpha}),\rho_k), \quad \forall \alpha \in (0,\bar{\alpha}],$$

for large enough $k \in \mathcal{K}$.

Proof. Recall that (6.8) provides both upper and lower bounds of $\lambda_{\max}(\nabla f(\rho_k)) - \lambda_{\min}(\nabla f(\rho_k))$, for large enough $k \in \mathcal{K}$. With regard to Corollary 6.10, it suffices to set

$$\kappa = \frac{\Delta_\infty^2}{4\left[\mathrm{e}^{2\Delta_\infty\bar{\alpha}}(2\Delta_\infty\bar{\alpha} - 1) + 1\right]}.$$

□

Based on Corollary 6.13, we prove the following proposition.

Proposition 6.14. *Suppose that $\Delta_\infty > 0$ and ρ_k is not a minimizer for every $k \in \mathcal{K}$. It holds that* $\liminf\{D(\rho_k(\bar{\alpha}),\rho_k) \mid k \in \mathcal{K}\} = 0$.

The proof of Proposition 6.14 can be found in Section 6.C, which essentially follows the strategy of Gafni and Bertsekas [78] with necessary modifications.

To summarize, we have proved that for any converging sub-sequence $(\rho_k)_{k\in\mathcal{K}}$, there exists some $\gamma > 0$ such that

$$\liminf\{D(\rho_k(\gamma),\rho_k) \mid k \in \mathcal{K}\} = 0.$$

For the case where ρ_k is a minimizer for some $k \in \mathcal{K}$ or $\Delta_\infty = 0$, γ can be any strictly positive real number. Otherwise, we set $\gamma = \bar{\alpha}$ by Proposition 6.14.

By Lemma 6.6 and Lemma 6.8, it holds that

$$\begin{aligned}0 &\le \liminf\left\{ \frac{\left(\mathrm{e}^{-(1/2)\Delta_\infty\gamma} + (1/2)\Delta_\infty\gamma - 1\right)}{\gamma^2}\varphi_k''(\gamma) \,\middle|\, k \in \mathscr{K} \right\} \\ &\le \liminf\{ D(\rho_k(\gamma), \rho_k) \mid k \in \mathscr{K} \} = 0,\end{aligned}$$

showing that $\liminf\{\varphi_k''(\gamma) \mid k \in \mathscr{K}\} = 0$. Applying Lemma 6.6 and Lemma 6.8 again, we obtain

$$\begin{aligned}0 &\le \liminf\{ D(\rho_k(\beta), \rho_k) \mid k \in \mathscr{K} | k \in \mathscr{K} \} \\ &\le \liminf\left\{ \frac{\left(\mathrm{e}^{2\Delta_\infty\beta} - 2\Delta_\infty\beta - 1\right)}{\beta^2}\varphi_k''(\beta) \,\middle|\, k \in \mathscr{K} \right\} = 0,\end{aligned}$$

for any $\beta \in (0, +\infty)$. This proves Statement 3 of Theorem 6.1.

6.4 Concluding remarks

Assuming only differentiability of the objective function, we have proved that the EG method with Armijo line search monotonically converges to the optimum, if the sequence of iterates possesses a non-singular limit point. Our proof exploits the self-concordant likeness of the log-partition function, which is of independent interest; in particular, Lemma 6.8 improves upon the Peierls-Bogoliubov inequality.

6.4.1 Importance of self-concordant likeness

With regard to (6.4), one may suspect whether it suffices, for the convergence analysis, to prove the following: There exists some $\epsilon > 0$, such that the mapping $\alpha \mapsto \alpha^{-\epsilon} D(\rho(\alpha), \rho)$ is non-increasing on $(0, \bar{\alpha}]$ for every non-singular $\rho \in \mathscr{D}$. Indeed, following the proof strategy for Proposition 6.5, we obtain the following result *without self-concordant likeness*.

Proposition 6.15. *Let $\rho \in \mathscr{D}$ be non-singular. Define*

$$M := \sup\{\varphi''(\alpha; \rho) \mid \alpha \in (0, \bar{\alpha})\}, \quad m := \inf\{\varphi''(\alpha; \rho) \mid \alpha \in (0, \bar{\alpha})\}.$$

Suppose that $m > 0$. Then the mapping $\alpha \mapsto \alpha^{-\epsilon} D(\rho(\alpha), \rho)$ is non-increasing on $(0, \bar{\alpha})$, where $\epsilon := 2M/m$.

Remark 6.16. *For the case where $m = 0$, Lemma 6.19 implies that ∇f must be a multiple of the identity. Then it is easily checked that ρ is a minimizer as it verifies the optimality condition.*

Then in the proof of Proposition 6.11, for example, the condition we need to verify becomes:

$$\|\nabla f(\sigma) - \nabla f(\rho)\|_\infty \sqrt{2} \le (1 - \tau)\alpha^{\epsilon/2 - 1} \sqrt{\frac{D(\rho(\bar{\alpha}), \rho)}{\bar{\alpha}^2}}, \quad \forall \alpha \in [0, \alpha_\rho].$$

Notice that $\epsilon \geq 2$ by definition. Both sides can converge to zero as $\alpha \to 0$, so in general, there does not exist a small enough α_ρ that verifies the condition. Moreover, because $\alpha^\epsilon \leq \alpha^2$ for $\alpha \in [0,1]$, it is impossible to obtain an analogue of Corollary 6.10.

The point in our analysis is to show that there exists some $\chi(\alpha)$, bounded from below by α^2 for every α close to zero, such that the mapping $\alpha \mapsto D(\rho(\alpha),\rho)/\chi(\alpha)$ is non-increasing. This is where self-concordant likeness of the log-partition function comes into play.

6.4.2 Extensions for the probability simplex and spectahedron constraints

The EG method can be extended for the spectahedron and probability simplex constraints; in fact, the EG method is arguably better known for these two cases [7, 19, 103, 174]. For the former case, the iteration rule writes exactly the same as (6.1), and is equivalent to (6.2) with $\mathscr{D}$ replaced by the spectahedron $\mathscr{S}$. For the latter case, the iteration rule becomes element-wise (see, e.g., [19]) and is equivalent to (6.2), with $\mathscr{D}$ replaced by the probability simplex $\mathscr{P}$, and the quantum relative entropy replaced by the (classical) relative entropy. The Armijo line search rule applies without modification.

It is easily checked that our proof holds without modification for the spectahedron constraint. As a vector in $\mathbb{R}^d$ is equivalent to a diagonal matrix in $\mathbb{R}^{d\times d}$, it can be easily checked that the statements in Theorem 6.1 applies to the probability simplex constraint. Corollary 6.3 also holds true for these two constraints with slight modification—for the probability simplex constraint, non-singularity should be replaced by element-wise strict positivity.

6.4.3 Convergence with possibly singular limit points

Corollary 6.3 requires existence of at least one non-singular limit point. Regarding the result in the preceding chapter, if we introduce a slightly stronger condition that the objective function is locally smooth relative to the entropy function, convergence of the exponentiated gradient method also holds. In general without the local relative smoothness condition, we conjecture that convergence to the optimum cannot be guaranteed. However, we have not found a counter-example.

6.A Technical lemmas necessary for Section 6.3

Recall the definition:

$$\rho(\alpha) := C_\rho^{-1} \exp\left[\log(\rho) - \alpha \nabla f(\rho)\right],$$

for every non-singular $\rho \in \mathscr{D}$ and $\alpha \geq 0$, where C_ρ is the positive real number normalizing the trace of $\rho(\alpha)$.

Lemma 6.17. *For every non-singular $\rho \in \mathscr{D}$ and $\alpha > 0$, it holds that*

$$\langle \nabla f(\rho), \rho(\alpha) - \rho \rangle \leq -\frac{D(\rho(\alpha), \rho)}{\alpha}.$$

Proof. The equivalent formulation of the EG method, (6.2), implies that

$$\alpha \langle \nabla f(\rho), \rho(\alpha) - \rho \rangle + D(\rho(\alpha), \rho) \leq \alpha \langle \nabla f(\rho), \rho - \rho \rangle + D(\rho, \rho) = 0. \qquad \square$$

Lemma 6.18. *Let $\rho \in \mathscr{D}$ be non-singular. If ρ is a minimizer of f on $\mathscr{D}$, then $\rho(\alpha) = \rho$ for all $\alpha \geq 0$. If $\rho(\alpha) = \rho$ for some $\alpha > 0$, then ρ is a minimizer of f on $\mathscr{D}$.*

Proof. The optimality condition says that $\rho \in \operatorname{int}\mathscr{D}$ is a minimizer, if and only if

$$\langle \nabla f(\rho), \sigma - \rho \rangle \geq 0, \quad \forall \sigma \in \mathscr{D}.$$

For any $\alpha > 0$, we can equivalently write

$$\langle \alpha \nabla f(\rho) + \left[\nabla h(\rho) - \nabla h(\rho)\right], \sigma - \rho \rangle \geq 0, \quad \forall \sigma \in \mathscr{D}, \tag{6.9}$$

where h denotes the negative von Neumann entropy function, i.e.,

$$h(\rho) := \operatorname{Tr}(\rho \log \rho) - \operatorname{Tr} \rho.$$

Notice that the quantum relative entropy D is the Bregman divergence induced by the negative von Neumann entropy. It is easily checked, again by the optimality condition, that (6.9) is equivalent to

$$\rho = \arg\min \{ \alpha \langle \nabla f(\rho), \sigma - \rho \rangle + D(\sigma, \rho) \mid \sigma \in \mathscr{D} \} = \rho(\alpha). \qquad \square$$

For every non-singular $\rho \in \mathscr{D}$ and $\alpha \geq 0$, define

$$G := -\nabla f(\rho), \quad H_\alpha := \log \rho + \alpha G.$$

Let $G = \sum_j \lambda_j P_j$ be the spectral decomposition of G. Define η_α as a random variable satisfying

$$\mathsf{P}\left\{\eta_\alpha = \lambda_j\right\} = \frac{\operatorname{Tr}\left(P_j \exp(H_\alpha)\right)}{\operatorname{Tr} \exp(H_\alpha)}. \tag{6.10}$$

It is easily checked that $\mathsf{P}\left(\eta_\alpha = \lambda_j\right) > 0$ for all j, and the probabilities sum to one.

Lemma 6.19. *For any $\alpha \in \mathbb{R}$, it holds that*

$$\varphi'(\alpha) = \mathsf{E}\,\eta_\alpha, \quad \varphi''(\alpha) = \mathsf{E}\left(\eta_\alpha - \mathsf{E}\,\eta_\alpha\right)^2, \quad \varphi'''(\alpha) = \mathsf{E}\left(\eta_\alpha - \mathsf{E}\,\eta_\alpha\right)^3.$$

Proof. Notice that

$$\mathsf{E}\,\eta_\alpha^n = \frac{\operatorname{Tr}(G^n \exp(H_\alpha))}{\operatorname{Tr}\exp(H_\alpha)},$$

for any $n \in \mathbb{N}$. Define $\sigma_\alpha := \exp(H_\alpha)/\operatorname{Tr}\exp(H_\alpha)$. A direct calculation gives

$$\begin{aligned}
&\varphi'(\alpha) = \operatorname{Tr}(G\sigma_\alpha), \quad \varphi''(\alpha) = \operatorname{Tr}(G^2\sigma_\alpha) - (\operatorname{Tr}(G\sigma_\alpha))^2, \\
&\varphi'''(\alpha) = \operatorname{Tr}(G^3\sigma_\alpha) - 3\operatorname{Tr}(G^2\sigma_\alpha)\operatorname{Tr}(G\sigma_\alpha) + 2(\operatorname{Tr}(G\sigma_\alpha))^3.
\end{aligned}$$

The lemma follows. □

Since η_α is a bounded random variable, it follows that φ'' is bounded from above.

Corollary 6.20. *It holds that* $\varphi''(\alpha) \le (1/4)\Delta^2$, *where* $\Delta := \lambda_{\max}(\nabla f(\rho)) - \lambda_{\min}(\nabla f(\rho))$.

Proof. Recall that the variance of a random variable taking values in $[a,b]$ is bounded from above by $(b-a)^2/4$. □

6.B Proof of Lemma 6.8

Recall the random variable η_α defined in (6.10). Suppose that $\varphi''(\alpha) = 0$ for some $\alpha \in [0,+\infty)$. Then we have $\eta_\alpha = 0$ almost surely, but this implies that $\Delta = 0$, a contradiction. Therefore, we have $\varphi''(\alpha) > 0$ for all $\alpha \in [0,+\infty)$.

We prove a general result. Let $\psi : \mathbb{R} \to \mathbb{R}$ be a μ-self-concordant like function. Suppose that $\psi''(t) > 0$ for all t. Consider the function $\chi(t) := \log\big(\psi''(t)\big)$. We write, by the self-concordant likeness of ψ, that

$$|\chi'(t)| = \frac{|\psi'''(t)|}{\psi''(t)} \le \mu, \quad \forall\, t \in \mathbb{R}.$$

Then, for any $t_1, t_2 \in \mathbb{R}$, we have

$$|\chi(t_1) - \chi(t_2)| = \big|\log\big(\psi''(t_1)\big) - \log\big(\psi''(t_2)\big)\big| \le \mu|t_2 - t_1|;$$

that is,

$$\mathrm{e}^{-\mu|t_2-t_1|}\psi''(t_2) \le \psi''(t_1) \le \mathrm{e}^{\mu|t_2-t_1|}\psi''(t_2).$$

Applying the Newton-Leibniz formula, we obtain

$$\begin{aligned}\psi'(t_2)-\psi'(t_1) &= \int_0^1 \psi''(t_1+\tau(t_2-t_1))(t_2-t_1)\,\mathrm{d}\tau \\ &\le \int_0^1 \mathrm{e}^{\mu\tau|t_2-t_1|}\psi''(t_1)(t_2-t_1)\,\mathrm{d}\tau \\ &= \left(\frac{\mathrm{e}^{\mu|t_2-t_1|}-1}{\mu|t_2-t_1|}\right)\psi''(t_1)(t_2-t_1);\end{aligned}$$

similarly, we obtain

$$\psi'(t_2)-\psi'(t_1) \ge -\left(\frac{\mathrm{e}^{-\mu|t_2-t_1|}-1}{\mu|t_2-t_1|}\right)\psi''(t_1)(t_2-t_1).$$

Applying the Newton-Leibniz formula again, we obtain

$$\begin{aligned}\psi(t_2)-\psi(t_1) &= \int_0^1 \psi'(t_1+\tau(t_2-t_1))(t_2-t_1)\,\mathrm{d}\tau \\ &= \psi'(t_1)(t_2-t_1)+\int_0^1 \big(\psi'(t_1+\tau(t_2-t_1))-\psi'(t_1)\big)(t_2-t_1)\,\mathrm{d}\tau \\ &\le \psi'(t_1)(t_2-t_1)+\int_0^1 \left(\frac{\mathrm{e}^{\mu\tau|t_2-t_1|}-1}{\mu\tau|t_2-t_1|}\right)\psi''(t_1)\tau(t_2-t_1)^2\,\mathrm{d}\tau \\ &= \psi'(t_1)(t_2-t_1)+\frac{\big(\mathrm{e}^{\mu|t_2-t_1|}-\mu|t_2-t_1|-1\big)}{\mu^2}\psi''(t_1);\end{aligned}$$

similarly, we obtain

$$\psi(t_2)-\psi(t_1) \ge \psi'(t_1)(t_2-t_1)+\frac{\big(\mathrm{e}^{-\mu|t_2-t_1|}+\mu|t_2-t_1|-1\big)}{\mu^2}\psi''(t_1).$$

Lemma 6.8 follows from setting $\psi=\varphi$, $\mu=\Delta$, $t_2=0$, and $t_1=\alpha$.

6.C Proof of Proposition 6.14

Suppose that $\underline{\alpha} := \liminf\{\alpha_k \mid k\in\mathscr{K}\} > 0$. By the Armijo line search rule and Corollary 6.10, we write

$$\begin{aligned}f(\rho_k)-f(\rho_{k+1}) &\ge -\tau\langle\nabla f(\rho_k), f(\rho_{k+1})-f(\rho_k)\rangle \\ &\ge \tau\alpha_k^{-1}D(\rho_{k+1},\rho_k) \\ &= \tau\alpha_k\alpha_k^{-2}D(\rho_k(\alpha_k),\rho_k) \\ &\ge \tau\underline{\alpha}\kappa D(\rho_k(\bar{\alpha}),\rho_k) \\ &\ge 0,\end{aligned}$$

for large enough $k\in\mathscr{K}$. Taking limit, we obtain that $D(\rho_k(\bar{\alpha}),\rho_k)\to 0$ as $k\to\infty$ in $\mathscr{K}$.

Suppose that $\liminf\{\alpha_k \mid k \in \mathcal{K}\} = 0$. Let $(\alpha_k)_{k\in\mathcal{K}'}$, $\mathcal{K}' \subseteq \mathcal{K}$, be a sub-sequence converging to zero. According to the Armijo rule, we have

$$f(\rho_k(r^{-1}\alpha_k)) - f(\rho_k) > \tau \langle \nabla f(\rho_k), \rho(r^{-1}\alpha_k) - \rho(\alpha_k)\rangle, \tag{6.11}$$

for large enough $k \in \mathcal{K}$. The mean value theorem says that

$$f(\rho_k(r^{-1}\alpha_k)) - f(\rho_k) = \langle \nabla f(\sigma), \rho_k(r^{-1}\alpha_k) - \rho_k\rangle,$$

for some σ in the line segment jointing $\rho_k(r^{-1}\alpha_k)$ and ρ_k. Then (6.11) can be equivalently written as

$$\langle \nabla f(\sigma) - \nabla f(\rho_k), \rho_k(r^{-1}\alpha_k) - \rho_k\rangle > -(1-\tau)\langle \nabla f(\rho_k), \rho_k(r^{-1}\alpha_k) - \rho_k(\alpha_k)\rangle. \tag{6.12}$$

By Pinsker's inequality and Hölder's inequality, we obtain

$$\begin{aligned}\|\nabla f(\sigma) - \nabla f(\rho_k)\|_\infty \sqrt{2D(\rho_k(r^{-1}\alpha_k), \rho_k)} &\geq \|\nabla f(\sigma) - \nabla f(\rho_k)\|_\infty \|\rho_k(r^{-1}\alpha_k), \rho_k\|_1 \\ &\geq \langle \nabla f(\sigma) - \nabla f(\rho_k), \rho_k(r^{-1}\alpha_k) - \rho_k\rangle. \end{aligned} \tag{6.13}$$

for large enough $k \in \mathcal{K}$. Notice that $r^{-1}\alpha_k \leq \bar{\alpha}$ for large enough $k \in \mathcal{K}$. By Lemma 6.17 and Corollary 6.13, we obtain

$$\begin{aligned} -\langle \nabla f(\rho_k), \rho_k(r^{-1}\alpha_k) - \rho_k(\alpha_k)\rangle &\geq \frac{D(\rho_k(r^{-1}\alpha_k), \rho_k)}{r^{-1}\alpha_k} \\ &\geq \sqrt{\kappa D(\rho_k(\bar{\alpha}), \rho_k)}\sqrt{D(\rho_k(r^{-1}\alpha_k), \rho_k)}, \end{aligned} \tag{6.14}$$

for large enough $k \in \mathcal{K}$. Since $D(\rho_k(r^{-1}\alpha_k), \rho_k)$ is strictly positive for all $k \in \mathcal{K}'$ by assumption, (6.12), (6.13), and (6.14) imply

$$\|\nabla f(\sigma) - \nabla f(\rho_k)\|_\infty > (1-\tau)\sqrt{\frac{\kappa D(\rho_k(\bar{\alpha}), \rho_k)}{2}} \geq 0.$$

Taking limits, we obtain that $D(\rho_k(\bar{\alpha}), \rho_k) \to 0$ a $k \to \infty$ in $\mathcal{K}'$.

7 An agnostic PAC approach to compressive MRI

In the previous chapters, the focuses are either statistical or computational. In this chapter, we present a theoretical framework for compressive magnetic resonance imaging (MRI) that addresses both aspects. The framework is indeed a standard application of the agnostic probably approximately correct (PAC) learning theory—empirical risk minimization in particular. The result, however, is quite interesting: Subjective conditions, e.g., sparsity, is not required. The *restricted isometry property (RIP)*–a condition on the strong convexity and smoothness of the quadratic loss that ensures accurate signal reconstruction—is no longer required; competitive empirical results were achieved via computationally cheap (almost linear-time) algorithms.

This chapter is mainly based on the joint work with Volkan Cevher [118], supplemented by the results in [12, 82].

7.1 Introduction

Compressive MRI is essentially a linear inverse problem. The goal is to recover an unknown signal $x^\natural \in \mathbb{C}^p$, given a a sub-sampling pattern $\Omega \subset \{1, \dots, p\}$ with $|\Omega| = n$ for some $n < p$, and the outcome of compressive sampling:

$$y := P_\Omega \mathscr{F} x^\natural, \tag{7.1}$$

where $\mathscr{F} : \mathbb{C}^p \to \mathbb{C}^p$ is the Fourier transform matrix, and $P_\Omega : \mathbb{C}^p \to \mathbb{C}^n$ is a linear operator that only keeps entries of $\mathscr{F} x^\natural$ indexed by Ω. In practice, $x^\natural$ is usually a two- or three-dimensional object; then $\mathscr{F}$ should be replaced by the corresponding multi-dimensional Fourier transform.

The standard theory of compressive sampling (CS) assumes that $x^\natural$ possesses certain *structure*, and studies conditions on Ω such that $x^\natural$ can be recovered given y and Ω. For example, if $x^\natural$ is sparse, then as long as the matrix $A_\Omega := P_\Omega \mathscr{F}$ satisfies the RIP and n is sufficiently large, the basis pursuit estimator,

$$\hat{x}_{\mathrm{BP}} \in \arg\min_x \left\{ \|x\|_1 \;\middle|\; y = A_\Omega x, x \in \mathbb{C}^p \right\},$$

perfectly recovers $x^\natural$ [38, 43].

The RIP indeed requires the quadratic loss $f(x) := (1/2)\|y - A_\Omega x\|_2^2$ to be both strongly convex and 2-smooth, restricted on the set of sparse vectors. Recall the definition.

Definition 7.1 (Restricted isometry property (RIP) [38]). *We say that the (s,δ)-RIP holds for some $s \in \mathbb{N}$ and $\delta \geq 0$, if*

$$(1-\delta)\|x\|_2^2 \leq \|A_\Omega x\|_2^2 \leq (1+\delta)\|x\|_2^2, \quad \forall\, x \in \mathbb{C}^p \text{ such that } \|x\|_0 \leq 2s. \tag{7.2}$$

Proposition 7.2. *The RIP ensures that the function f is strongly convex with parameter μ, on the set*

$$\mathscr{X}_s := \{x \in \mathbb{C}^p \mid \|x\|_0 \leq s\}.$$

Proof. A direct calculation shows that strong convexity with parameter μ holds (cf. Definition 1.2), if and only if

$$\|A_\Omega(z-x)\|_2^2 \geq \mu\|z-x\|_2^2, \quad \forall\, x, z \in \mathscr{X}_s.$$

Notice that $\|z-x\|_0 \leq 2s$ for all $x, z \in \mathscr{X}_s$. Therefore, the desired inequality is guaranteed by the lower bound in (7.2), with $\mu = 1-\delta$. □

Similarly, it is easily checked that the upper bound in (7.2) implies that f is 2-smoothness with parameter $L = 1+\delta$ on $\mathscr{X}_s$, with respect to the following equivalent definition[1].

Definition 7.3. *We say that a function g is 2-smooth on a set $\mathscr{X}$, if there exits some $L > 0$ such that*

$$(1-\alpha)g(x) + \alpha g(z) \leq g((1-\alpha)x + \alpha z) + \alpha(1-\alpha)\frac{L}{2}\|z-x\|_2^2, \quad \forall\, x, z \in \mathscr{X} \text{ and } \alpha \in (0,1).$$

The RIP is guaranteed to hold for compressive MRI—if n is sufficiently large, and the set Ω is randomly chosen from all sub-sets of $\{1,\ldots,p\}$ following the uniform distribution, the RIP holds with high probability [45, 89, 160]. Similar theories have been developed for signal structures more complicated than mere sparsity, where the objective function in the basis pursuit estimator may be replaced by some non-differentiable stucture-promoting functions [1, 9, 13, 50, 71].

The standard theory of compressive MRI described above is theoretically sound, and has inspired many breakthroughs, such as matrix completion, phase retrieval, and numerical algorithms for non-differentiable optimization, to mention a few (see, e.g., [39, 139, 154]). However, it has three undesirable properties especially to MRI practitioners. First, to force

[1] Notice that f is a real-valued non-constant function of complex variables, so it violates the Cauchy-Riemann equation and is not differentiable. Therefore, here we use the definition of 2-smoothness that can be directly extended for the complex variable case [136, Theorem 2.1.5].

the RIP, the theory deviates from the practice—the uniformly random sub-sampling does not give satisfactory reconstruction results. An empirical evidence can be found in [124, Figure 6]. In fact, empirical observations have shown that low-frequency samples are more relevant; therefore, usually a random *variable density sampling* strategy that focuses more on low-frequency samples is adopted [51].

Second, the signal structure must be known in advance, and the sub-sampling pattern is chosen with respect the known structure. This results in difficulties in practice. Finding a sparse representation of the unknown signal is highly non-trivial [171], whereas we cannot guarantee in advance whether the unknown signal under a given representation is sparse enough for the RIP to hold, without any subjective assumption. Moreover, focusing on a specific *a priori* guess of the signal structure may result in a sub-optimal design of the sub-sampling pattern, due to overlooking other structures that the unknown signal *also* possesses.

Third, the signal reconstruction procedure requires computing a basis pursuit-like estimator, formulated as a convex optimization problem with a non-differentiable objective function. While there are a variety of convex optimization tools guaranteed to solve the optimization problem up to numerical accuracy (see, e..g, [76]), a significant amount of computation time is needed, especially when compared to the classical least squares (LS) approach.

In this chapter, we propose a new theoretical framework for compressive MRI, aiming to address the three issues raised above simultaneously.

7.1.1 Related work

Existing approaches to compressive MRI essentially follows the standard theory. The typical pattern is summarized as follows.

1. Find a transformation matrix $\Psi : \mathbb{C}^p \to \mathbb{C}^p$, such that $x^\natural = \Psi^{-1} z^\natural$ and $z^\natural$ possesses certain *structure*. For example, the sparsity of $z^\natural$ was considered in [43], the sparsity of $z^\natural$ and smoothness of $x^\natural$ were exploited in [124], the tree sparsity of $z^\natural$ was considered in [52], and the multi-level sparsity of $z^\natural$ wasconsidered in [1].

2. Choose a *random* sub-sampling pattern Ω and sample $\mathscr{F} x^\natural$ accordingly. The probability distribution may be constructed to ensure a RIP or RIP-like condition regarding the signal structure, or with respect to the practical wisdom that low-frequency samples are more informative (see, e.g., [1, 43, 51, 124, 181], to mention a few).

3. Finally, apply a *non-linear* signal reconstruction algorithm to reconstruct $x^\natural$. The standard basis pursuit estimator was considered in [1, 51]. A basis pursuit-like estimator minimizing a sum of the ℓ_1-norm and the total variation semi-norm was proposed in [124]. An LS estimator with the ℓ_1-norm and total variation semi-norm penalizations was considered in [186]. A similar penalized LS estimator with an additional penalization term for tree sparsity was introduced by [52].

The second and third issues we raised in the introduction—requirement of *a priori* knowledge of the signal structure and computationally expensive signal reconstruction—are obviously inherited. As for the sub-sampling pattern, the theoretically sound approaches only consider specific signal structures [1, 43, 51]; the heuristic approaches lack rigorous performance guarantees, and may involve many parameters to be properly tuned [124, 181].

7.1.2 Contributions

The main novelty is that we formulate the design of an optimal compressive MRI system as a *statistical learning problem*. We solve the learning problem via empirical risk minimization (ERM). We adopt the agnostic PAC perspective [88, 176, 179] and derive rigorous performance guarantees, which hold without any *a priori* knowledge of the signal structure. We show that, within our framework, even the classical LS reconstruction approach yields competitive empirical results on real MRI images, in comparison to seminal work [124]. Notice that the LS reconstruction approach is computationally cheap—the LS estimator is simply $\mathscr{F}^{\mathrm{H}} P_{\Omega}^{\top} y$ for any given sub-sampling pattern Ω. Indeed, we show that with the LS estimator, the ERM problem can be exactly solved via an almost linear-time algorithm; hence the training procedure is also computationally cheap.

7.2 Agnotic PAC framework

We briefly introduce relevant notions in PAC learning theory in this section. We start with a formal, abstract definition of the statistical learning model; then we provide an illustrating example. The interested reader is referred to, e.g., [132, 163] for more details.

7.2.1 Formal definition

Let $\mathscr{Z}$ be an abstract set. A statistical learning problem consists of three ingredients.

1. The *training data* is a set of independent and identically distributed (i.i.d.) random variables (r.v.'s) $\{Z_1, \ldots, Z_m\} \subset \mathscr{Z}$, following an *unknown* probability distribution Q.
2. The *hypothesis class* is a set of functions $\mathscr{H}$. An element of the hypothesis class is called a *hypothesis*.
3. The *loss functions* is a real-valued function $L : \mathscr{H} \times \mathscr{Z} \to \mathbb{R}$.

We assume that the loss function takes values in $[0,1]$ for simplicity.

The expected loss is called the *risk*, which is a function $R : \mathscr{H} \to \mathbb{R}$ given by

$$R(h) := \mathsf{E}\, L(h, Z), \quad \forall h \in \mathscr{H},$$

where Z is a r.v. following the probability distribution Q. The *optimal hypothesis* $h^\star$ minimizes the risk; that is,

$$h^\star \in \arg\min_h \{ R(h) \mid h \in \mathscr{H} \} .$$

As the probability distribution Q is unknown, we cannot evaluate the risk function exactly. The ERM approach approximates the risk by the *empirical risk*. The empirical risk $\hat{R}_m$ is simply the average of the loss on the data; that is,

$$\hat{R}_m(h) := \frac{1}{m} \sum_{i=1}^{m} L(h, Z_i) , \quad \forall h \in \mathscr{H} .$$

It is natural to expect that $\hat{R}_m$ is close to R when m is large enough, by the law of large numbers. We expect that the empirical risk minimizer $\hat{h}_m$ given by

$$\hat{h}_m \in \arg\min_h \left\{ \hat{R}_m(h) \,\middle|\, h \in \mathscr{H} \right\}$$

is a good approximation of the optimal hypothesis.

Definition 7.4. *We say that the hypothesis class $\mathscr{H}$ is agnotic PAC learnable via ERM, if for every $\varepsilon, \delta \in (0,1)$, there exists some positive integer $m_{\mathscr{H}}(\varepsilon, \delta)$, such that*

$$R(\hat{h}_m) \leq R(h^\star) + \varepsilon ,$$

with probability at least $(1-\delta)$.

Sometimes, we are interested in ensuring that the hypothesis $\tilde{h}_m$ found by a given training algorithm yields a small enough risk. Define the *generalization error*

$$G(h) := |R(h) - \hat{R}_m(h)| , \quad \forall h \in \mathscr{H} .$$

If the generalization error is guaranteed to be small, the empirical risk $\hat{R}_m(\tilde{h}_m)$ can serve as an estimate of the risk $R(\tilde{h}_m)$.

7.2.2 Example

Consider the problem of binary classification. In this problem, the training data is given by pairs as

$$Z_i = (X_i, Y_i) \in \mathbb{R}^p \times \{0,1\} =: \mathscr{Z} , \quad \forall i = 1, \ldots, m ,$$

For example, the vector X_i may be the numerical representation of an image; the corresponding *label* $Y_i = 1$ if the image contains a human face, and $Y_i = 0$, otherwise.

For the problem of binary classification, the hypothesis class consists of *classification rules*—functions that map a given image to the corresponding label. Suppose that the aim is to minimize the probability of false classification. It suffices to choose the loss function as the $0-1$ loss, given by

$$L(h,(X,Y)) = \begin{cases} 0 & \text{if } h(X) = Y\,, \\ 1 & \text{otherwise.} \end{cases}$$

It is easily checked that the risk corresponds to the probability of false classification.

7.3 Proposed framework for compressive sampling

The training data is absent in the standard theory of compressive MRI; at first glance, therefore, there is nothing from which we can *learn*. Indeed, the result of learning is implicitly encoded by the sparsity assumption.

Let us examine the origin of the sparsity assumption. Without any subjective assumption, the validity of the sparsity assumption can only be established by its empirical success: We have observed that many real-world signals are essentially sparse, if a proper representation is chosen [125]; therefore, the sparsity assumption seems to be reasonable. The discovery of the sparsity structure is hence the result of a learning procedure, perhaps not very principled, given real-world signals. The real-world signals that help us discover the sparsity structure are then the training data.

Our framework takes into presence of training data into account, in order to develop a principled approach to designing the sub-sampling pattern. Instead of modeling the unknown signal $x^\natural$ to be deterministic unknown, we adopt the statistical learning philosophy. We model $x^\natural$ as a random vector following some unknown probability distribution Q, and assume that we have access to m training signals $x_1,\dots,x_m \in \mathbb{C}^p$, which are i.i.d. r.v.'s following the same probability distribution Q. Notice that this is different from Bayesian compressive sampling [99], as the *prior* Q is unknown in our model.

Our theory indeed works with any unitary measurement matrix. Therefore, we consider the following measurement model in the rest of this chapter:

$$y = P_\Omega \Phi x^\natural\,,$$

for some unitary $\Phi \in \mathbb{C}^{p\times p}$.

For any given sub-sampling pattern Ω, the LS estimator has an explicit form:

$$\begin{aligned} \hat{x}_{\mathrm{LS}} &= \arg\min_x \left\{ \|y - P_\Omega \Phi x\|_2^2 \;\middle|\; x \in \mathbb{C}^p \right\} \\ &= \Phi^{\mathrm{H}} P_\Omega^\top y\,. \end{aligned}$$

Once the estimator is fixed, the only issue is to choose Ω that optimizes the resulting estimation performance.

The set of all possible sub-sampling patterns is then the hypothesis class. Define

$$\hat{x}_\Omega := \Phi^{\mathrm{H}} P_\Omega^\top P_\Omega \Phi x, \quad \forall x \in \mathbb{C}^p,$$

the LS estimate of x given the measurement outcomes using the sub-sampling pattern Ω. The loss function we consider is the normalized squared error

$$L(\Omega, x) := \frac{\|\hat{x}_\Omega - x\|_2^2}{\|x\|_2^2}, \quad \forall \Omega \subseteq \{1, \dots, p\} \text{ and } x \in \mathbb{C}^p.$$

Proposition 7.5. *The risk—expected squared error—is given by*

$$R(\Omega) := \mathsf{E}\, L(\Omega, x) = 1 - \mathsf{E}\, f_\Omega(x), \quad \forall \Omega \subset \{1, \dots, p\}, \tag{7.3}$$

where the expectation is with respect to $x \sim Q$, and

$$f_\Omega(x) := \frac{\|P_\Omega \Phi x\|_2^2}{\|x\|_2^2}.$$

Proof. In fact, the equality holds deterministically, as

$$\begin{aligned}
\|\hat{x}_\Omega - x^\natural\|_2^2 &= \|\hat{x}_\Omega\|_2^2 - 2\langle \hat{x}_\Omega, x^\natural \rangle + \|x^\natural\|_2^2 \\
&= \|\Phi^{\mathrm{H}} P_\Omega^\top P_\Omega \Phi x^\natural\|_2^2 - 2\langle \Phi^{\mathrm{H}} P_\Omega^\top P_\Omega \Phi x^\natural, x^\natural \rangle + \|x^\natural\|_2^2 \\
&= \|P_\Omega \Phi x^\natural\|_2^2 - 2\|P_\Omega \Phi x^\natural\|_2^2 + \|x^\natural\|_2^2.
\end{aligned}$$

In the third equality, we used the fact that $AA^\dagger A = A$ for any unitary matrix A and its Moore-Penrose generalized inverse $A^\dagger$, by setting $A := P_\Omega \Phi$. □

Fix a budget $n \in \mathbb{N}$ on the total number of measurements. The proposition above implies that the optimal sub-sampling pattern $\Omega^\star$, or the optimal hypothesis, is given by the following optimization problem:

$$\Omega^\star \in \operatorname*{arg\,max}_\Omega \left\{ \mathsf{E}\, f_\Omega(x) \,\middle|\, \Omega \subset \{1, \dots, p\}, |\Omega| = n \right\}. \tag{7.4}$$

We cannot evaluate the function $\mathsf{E}\, f_\Omega(x)$ exactly as Q is assumed unknown, so we adopt the ERM approach. The empirical risk minimizer $\hat{\Omega}_m$ is given by

$$\hat{\Omega}_m \in \operatorname*{arg\,max}_\Omega \left\{ \hat{\mathsf{E}}_m\, f_\Omega(x) \,\middle|\, \Omega \subset \{1, \dots, p\}, |\Omega| = n \right\}, \tag{7.5}$$

where $\hat{\mathsf{E}}_m$ denotes the expectation with respect to the empirical measure, i.e.,

$$\hat{\mathsf{E}}_m f_\Omega(x) := \frac{1}{m}\sum_{i=1}^{m} f_\Omega(x_i) = \frac{1}{m}\sum_{i=1}^{m} \frac{\|P_\Omega \Phi x_i\|_2^2}{\|x_i\|_2^2}.$$

Notice that $\hat{\Omega}_m$ is a random variable, as it depends on the random vectors $x_1, \dots, x_m$.

The overall compressive sampling system is summarized as follows.

1. Find a sub-sampling pattern $\hat{\Omega}_m$ by (7.5).
2. Sub-sample $x^\natural$ using $\hat{\Omega}_m$ and obtain the measurement outcome

$$y := P_{\hat{\Omega}_m} \Phi x^\natural.$$

3. Recover $x^\natural$ by

$$\hat{x} := \Phi^{\mathrm{H}} P_{\hat{\Omega}_m}^{\top} y.$$

7.3.1 On computing the empirical risk minimizer

Define $\varphi_i^\top$ as the i-th row of the matrix Φ. Interestingly, the optimization problem (7.5) can be exactly solved by the following algorithm.

1. Compute the values

$$v_i := \frac{1}{m}\sum_{j=1}^{m} \frac{|\langle \varphi_i, x_j\rangle|^2}{\|x_j\|_2^2}, \quad \forall i = 1, \dots, p.$$

2. Set $\tilde{\Omega}_m$ as the set of indices corresponding to n largest v_i's.

Proposition 7.6. *It holds that* $\tilde{\Omega}_m = \hat{\Omega}_m$.

Proof. Notice that, for any $\Omega = \{k_1, \dots, k_n\} \subset \{1, \dots, p\}$,

$$\begin{aligned}
\hat{\mathsf{E}}_m f_\Omega(x) &= \frac{1}{m}\sum_{i=1}^{m} \frac{\|P_\Omega \Phi x_i\|_2^2}{\|x_i\|_2^2} \\
&= \frac{1}{m}\sum_{i=1}^{m}\sum_{j=1}^{n} \frac{|\langle \varphi_{k_j}, x_i\rangle|^2}{\|x_i\|_2^2} \\
&= \sum_{j=1}^{n} v_{k_j}.
\end{aligned}$$

□

The first step of the algorithm requires computing the matrix-vector products Φx_i; hence, its computational complexity is $O(mp^2)$. The second step of the algorithm requires sorting p real

numbers, the computational complexity of which is $O(p \log p)$ [57]. The overall computational complexity is then $O(mp^2)$. If the matrix Φ is a suitably structured matrix, such as the Fourier and Hadamard matrix, the overall computational complexity becomes $O(mp \log p)$, almost linear-time [74].

7.4 Performance Analysis

Formulating compressive sampling as a statistical learning problem, we can derive a performance guarantee of the agnostic PAC type (cf. Definition 7.4).

Proposition 7.7. *For any $\varepsilon, \delta \in (0,1)$, if the number of training signals m satisfies*

$$m \geq \frac{2 \log \left[\binom{p}{n} / \delta \right]}{\varepsilon^2},$$

it holds that

$$R(\hat{\Omega}_m) \leq R(\Omega^\star) + \varepsilon,$$

with probability at least $(1-\delta)$. Recall that R is the expected normalized reconstruction error (cf. (7.3)).

The proof is standard. Define the empirical risk

$$\hat{R}_m(\Omega) := \frac{1}{m} \sum_{i=1}^{m} L(\Omega, x_i), \quad \forall \Omega \subseteq \{1, \ldots, p\}.$$

We first prove the uniform convergence of the empirical risk $\hat{R}_m$ to the risk R on the hypothesis class

$$\mathscr{C}_n := \{\Omega \subseteq \{1, \ldots, p\} \mid |\Omega| = n\}.$$

Lemma 7.8. *For any $\delta > 0$, we have*

$$\mathsf{P}\left\{ R(\Omega) - \hat{R}_m(\Omega) \leq t_\delta \ \forall \Omega \in \mathscr{C}_n \right\} \geq 1 - \delta,$$
$$\mathsf{P}\left\{ \hat{R}_m(\Omega) - R(\Omega) \leq t_\delta \ \forall \Omega \in \mathscr{C}_n \right\} \geq 1 - \delta,$$

where

$$t_\delta := \sqrt{\frac{\log \left[\binom{p}{n} / \delta \right]}{2m}}. \tag{7.6}$$

Proof. Notice that the normalized reconstruction error satisfies

$$L(\Omega, x) \in [0,1], \quad \forall \Omega \in \mathscr{C}_n \text{ and } x \in \mathbb{C}^p.$$

By Hoeffding's inequality, we obtain that for any $t > 0$ and $\Omega \in \mathscr{C}_n$,

$$\mathsf{P}\left\{ R(\Omega) - \hat{R}_m(\Omega) \geq t \right\} \leq \mathrm{e}^{-2mt^2}.$$

By the union bound, we obtain that for any $t > 0$,

$$\mathsf{P}\left\{ R(\Omega) - \hat{R}_m(\Omega) < t \ \forall \Omega \in \mathscr{C}_n \right\} \geq 1 - \binom{p}{n} \mathrm{e}^{-2mt^2},$$

which implies the first inequality. The second inequality is proved similarly. □

Proof of Proposition 7.7. We write

$$\begin{aligned} R(\hat{\Omega}_m) - R(\Omega^\star) &= \left(R(\hat{\Omega}_m) - \hat{R}_m(\hat{\Omega}_m)\right) + \left(\hat{R}_m(\hat{\Omega}_m) - \hat{R}_m(\Omega^\star)\right) + \left(\hat{R}_m(\Omega^\star - R(\Omega^\star)\right) \\ &\leq \left(R(\hat{\Omega}_m) - \hat{R}_m(\hat{\Omega}_m)\right) + \left(\hat{R}_m(\Omega^\star - R(\Omega^\star)\right), \end{aligned}$$

where the inequality follows from the fact that $\hat{\Omega}_m$ is a minimizer of the empirical risk. By Lemma 7.8, we obtain that, for any $\delta > 0$,

$$\mathsf{P}\left\{ R(\hat{\Omega}_m) - R(\Omega^\star) \leq 2t_\delta \right\} > 1 - \delta,$$

where t_δ is defined in (7.6). Solving the equality $2t_\delta = \varepsilon$ proves the proposition. □

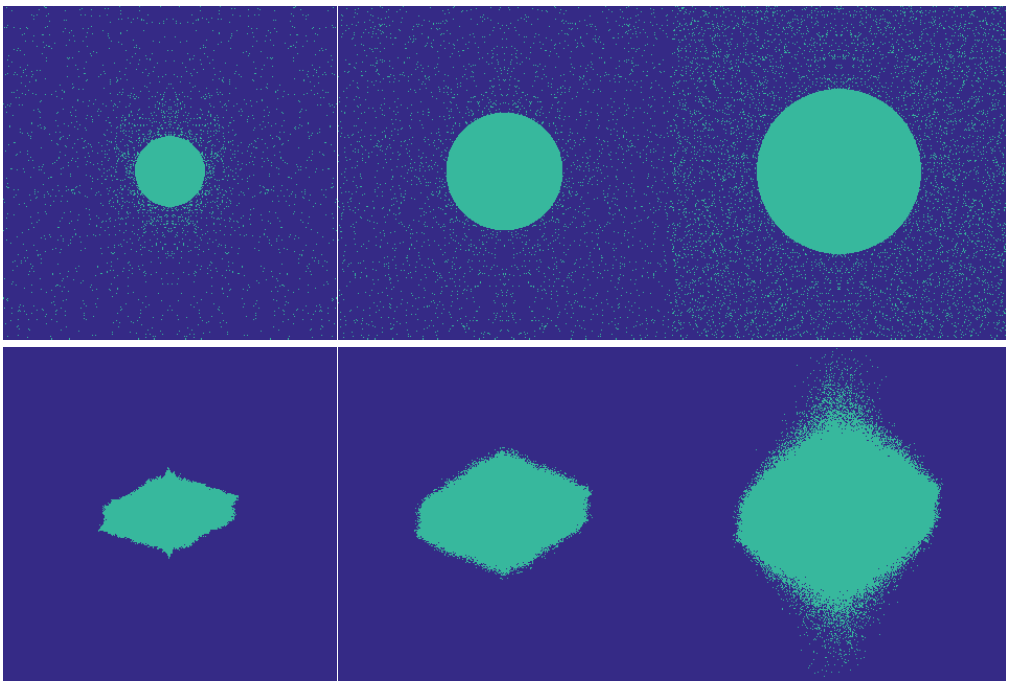
According to Proposition 7.7, it suffices to have $O(n \log p)$ training signals to ensure a small enough reconstruction error. Notice that this performance guarantee is a worst-case for all possible probability distributions Q; in practice, the required number of training signals can be significantly smaller, as the numerical results in the next section show.

7.5 Numerical Results

We use a three-dimensional data-set of raw knee-images data in the k-space[2]. We first take an inverse Fourier transform along the z-axis, and eliminate the $z-$slices of low energy that are close to the boundary of the datacube. These are noise-like slices that do not exhibit any knee feature, as they are close to the skin of the patient. We then investigate subsampling schemes in the 320×320 $x-y$ Fourier plane, which corresponds to compressive sampling for each z-slice.

We pick the images of the first ten patients in the given dataset for training, and test the learned sub-sampling patterns on the remaining ten patients—notice the number of training signals is much smaller than suggested by the worst-case guarantee. We compare our learning-based approach to the variable density sampling sheme proposed in [124], which is parametrized by the radius of the fully sampled region, r, and a polynomial degree, d. We use the values

[2]The data-set is available on http://mridata.org/fullysampled.



6.25% sampling 12.5% sampling 25% sampling

Figure 7.1 – First row: the subsampling maps of the tuned random variable sampling scheme [124]. Second row: the maps given by our learning-based approach.

of r and d that achieve the highest average peak signal-to-noise ratio (PSNR) on the training signals.

Figure 7.1 illustrates the best performing randomized indices and our learned set of indices in the $x-y$ plane of the k-space. Both the variable density approach [124] and our learning-based approach concentrates its sampling budget on the low frequencies, however the latter is endowed with the capability to adapt its frequency selection to the frequency content of the training signals instead of assuming a circularly symmetric selection.

Table 7.1 – Average PSNR on the test data

Indices	Sampling rate		
	6.25%	12.50%	25%
Best-n approx.	25.29 dB	26.36 dB	28.35 dB
Lustig *et al.*	24.51 dB	25.11 dB	26.05 dB
This work	24.66 dB	25.18 dB	26.12 dB

Table 7.1 shows the performance of both approaches on the test data, in addition to the error lower-bounds obtained by the best n-sample approximations with respect to the Fourier basis. It appears that the learning based approach slightly outperforms the randomized variable density based approach.

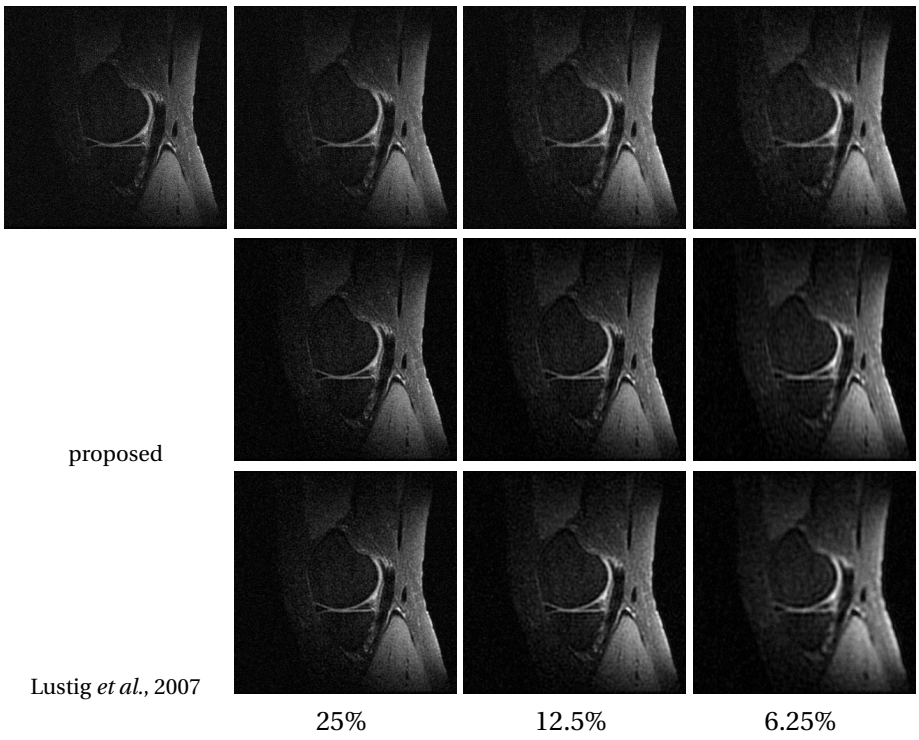


Figure 7.2 – MRI reconstructions of both schemes at different subsampling rates for a knee slice of patient #13, whose fully sampled reconstruction is shown on the top left.

However, the slight numerical improvements are actually accentuated when we look at the details of reconstructions, shown in Figure 7.2 for the test Patient #13. It is clear that the learning-based reconstructions provide more details especially for 6.25% and 12.5%.

7.6 Discussions

Rather than imposing subjective assumptions on the signal structure and taking the risk of overlooking other possible structures, our approach directly learns an optimal sub-sampling pattern for a given decoder from training signals. As the sub-sampling pattern is chosen to favor the given decoder, we have shown that even the very simple least squares estimator can be effective for compressive MRI. The empirical learned sub-sampling pattern complies with the practical wisdom that low-frequency samples are more important.

The trade-off is that our statistical error bound is weaker than those in standard compressive sampling literature. We cannot guarantee a small risk, but a small gap to the *unknown* optimal risk, due to the agnostic nature of our approach.

We discuss four possible research directions below.

7.6.1 Classification

In some application scenarios, the ultimate aim of MRI is not recovering the unknown image, but telling whether the unknown image possesses certain characteristics of interest. A doctor may be more interested in whether there is a tumor or not than the PSNR value. For such scenarios, one may replace the loss function in our MRI framework by the $0-1$ loss or any of its convex approximates, such as the logistic and hinge losses. We notice that standard PAC learning theory directly applies. The practical performance, however, needs to be checked via experiments on real data-sets.

7.6.2 Non-linear estimators

Our framework does not involve any subjective assumption on the signal structure. A proper introduction of assumptions on the signal structure can nevertheless be helpful. Empirical observations have shown that real-world signals posses structures such as sparsity, smoothness, and other more complicated ones; well-designed non-linear reconstruction methods, such as basis pursuit, the lasso, and neural networks match these structures and give better reconstruction performances. The theoretical framework presented in this chapter directly extends for non-linear estimators—one simply needs to replace the LS estimator by a non-linear one. The performance guarantee (Proposition 7.7) still holds. However, there may not exist efficient algorithms to compute the corresponding empirical risk minimizer.

In [82], we have developed an efficient greedy algorithm for any given estimator. The algorithm is not guaranteed to find the empirical risk minimizer, unless the ERM problem possesses some structure, e.g., sub-modularity; however, the empirical performance on real data-sets is already better than some state-of-the-art approaches.

It is reasonable to consider designing the sub-sampling pattern and the estimator jointly. The corresponding ERM formulation is immediate. Let $\mathscr{T}$ be the set of estimators—functions that map the measurement outcome to a vector in $\mathbb{C}^p$. For example, $\mathscr{T}$ can be the set of penalized LS estimators with penalization coefficients in a given interval, or the set of all possible realizations of a neural network. The ERM formulation is given by

$$(\hat{\tau}_m, \hat{\Omega}_m) \in \underset{(\tau,\Omega)}{\arg\min} \left\{ \frac{1}{m} \sum_{i=1}^{n} L(\tau, \Omega, x_i) \;\middle|\; (\tau, \Omega) \in \mathscr{T} \times \mathscr{C}_n \right\}, \tag{7.7}$$

where the loss function is still the normalized reconstruction error, i.e.,

$$L(\tau, \Omega, x) := \frac{\|x - \tau(P_\Omega \Phi x)\|_2^2}{\|x\|_2^2}.$$

Following standard arguments in statistical learning theory, the required number of training signals then depends on the *complexity* of $\mathscr{T}$; see, e.g., [5, 105, 163] for various complexity measures in statistical learning theory. Solving the optimization problem (7.7), however, is

computationally difficult in general.

7.6.3 Generalization error bound

It is not necessary, for the purpose of learning, to devise a rigorous optimization algorithm to solve the empirical risk minimization problem. If a heuristic algorithm yields a small empirical risk, though perhaps not the minimum empirical risk, and the number of training signals is sufficiently large, then a small risk can be guaranteed. This can be easily proved even for a general loss function and reconstruction algorithm.

Indeed, consider any loss function L that maps a given sub-sampling pattern Ω and signal x to a number in $[0,1]$. The following generalization error bound is standard in statistical learning theory.

Proposition 7.9. *For any $\varepsilon, \delta \in (0,1)$, if the number of training signals m satisfies*

$$m \geq \frac{\log\left[2\binom{p}{n}/\delta\right]}{\varepsilon^2},$$

it holds that

$$|R(\Omega) - \hat{R}_m(\Omega)| \leq \varepsilon, \quad \forall \Omega \in \mathscr{C}_n,$$

with probability at least $(1-\delta)$.

Proof. By Hoeffding's inequality and the union bound, we have

$$\mathsf{P}\left\{|R(\Omega) - \hat{R}_m(\Omega)| \geq t\right\} \leq 2\mathrm{e}^{-2mt^2}, \quad \forall\, t > 0 \text{ and } \Omega \in \mathscr{C}_n.$$

By the union bound, we have

$$\mathsf{P}\left\{|R(\Omega) - \hat{R}_m(\Omega)| \leq t \text{ for all } \Omega \in \mathscr{C}_n\right\} \geq 1 - 2\binom{p}{n}\mathrm{e}^{-2mt^2}.$$

The rest is similar to the proofs of Lemma 7.8 and Proposition 7.7. □

Therefore, a number of training signals of $O(n \log p)$ ensures that the empirical risk is a good estimate of the true risk. Notice that the proof does not assume the specific normalized reconstruction error loss and LS estimator; both can be chosen arbitrarily. Indeed, it is easily checked that Proposition 7.7 also admits such generality.

7.6.4 Effect of noise in training signals

In practice, the training signals are also obtained by measurements, and hence involve noise. Suppose that the noise is not negligible. Define the noisy training signals

$$z_i := x_i + w_i\,, \quad \forall\, i = 1,\dots,m\,,$$

where $w_i \in \mathbb{R}^p$ denotes some additive noise.

We clarify some notions first. Fix a unitary measurement matrix $\Phi \in \mathbb{C}^{p\times p}$. An estimator is a mapping τ that maps a measurement outcome and sub-sampling pattern, (y,Ω), to an estimate of the corresponding unknown signal. For example, the LS estimator corresponds to the choice

$$\tau(y,\Omega) = \Phi^{\mathrm{H}} P_\Omega^\top y\,.$$

We do not assume any specific estimator in this sub-section.

To emphasize the effect of noise, we write a loss function L in a slightly redundant manner. We write a loss function L as a mapping that maps a triple consisting of the sub-sampling pattern, unknown signal, and its estimate, $(\Omega, x, \hat{x})$, to a number in $[0,1]$. For example, the setup in the previous sections corresponds to the choice

$$L(\Omega, x, \tau(P_\Omega \Phi x)) \frac{\|x - \tau(P_\Omega\Phi x)\|_2^2}{\|x\|_2^2}\,.$$

We do not assume any specific loss function in this sub-section.

Now the issue is clear. We would like to minimize the risk

$$R(\Omega) := \mathsf{E}\, L(\Omega, x, \tau(P_\Omega \Phi x))\,, \quad \forall \Omega \in \mathscr{C}_n\,.$$

However, the empirical estimate of the risk we can compute, given the noisy training signals, is given by

$$\tilde{R}_m(\Omega) := \frac{1}{m}\sum_{i=1}^m L(\Omega, x_i + w_i, \tau(P_\Omega\Phi(x_i+w_i)))\,, \quad \forall \Omega \in \mathscr{C}_n\,.$$

We cannot expect that by the law of large numbers, $\tilde{R}_m$ is close to R when m is large.

Proposition 7.10. *Assume that* $\sup_i \{\|w_i\|_2\} \le W$ *for some* $W > 0$. *Suppose that the loss function* $L(\Omega, x, \tau(y))$ *is* L_1*-Lipshitz in* x *for any given* y, *and* L_2*-Lipschitz in* y *for any given* x. *For any* $\varepsilon, \delta \in (0,1)$, *if the number of training signals* m *satisfies*

$$m \ge \frac{\log\left[2\binom{p}{n}/\delta\right]}{\varepsilon^2}\,,$$

it holds that

$$|R(\Omega) - \tilde{R}_m(\Omega)| \leq (L_1 + L_2)W + \varepsilon\,, \quad \forall \Omega \in \mathscr{C}_n\,,$$

with probability at least $(1-\delta)$.

Proof. We write

$$\begin{aligned}
&|R(\Omega) - \tilde{R}_m(\Omega)| \\
&\quad = |R(\Omega) - \hat{R}_m(\Omega)| + \\
&\qquad \left| \hat{R}_m(\Omega) - \frac{1}{m}\sum_{i=1}^{m} L(\Omega, x_i, \tau(P_\Omega \Phi(x_i + w_i))) \right| + \\
&\qquad \left| \frac{1}{m}\sum_{i=1}^{m} L(\Omega, x_i, \tau(P_\Omega \Phi(x_i + w_i))) - \frac{1}{m}\sum_{i=1}^{m} L(\Omega, x_i + w_i, \tau(P_\Omega \Phi(x_i + w_i))) \right|.
\end{aligned}$$

The first term at the right-hand side is bounded above with high probability as in Proposition 7.9; the second term is bounded above by $L_2 W$ by the triangle inequality; similarly, the third term is bounded above by $L_1 W$. □

The Lipschitz parameter L_1 depends solely on the loss function. The Lipschitz parameter L_2 is closely related to the *robustness* of the estimator. Suppose that the output of the estimator is robust to measurement noise, in the sense that

$$\|\tau(y+\omega) - \tau(y)\|_2 \leq \gamma \|\omega\|\,, \quad \forall\, y, \omega \in \mathbb{R}^n\,.$$

Then L_2 exists if the loss function is Lipschitz continuous in its third argument.

In the generalization error bound, there is a non-zero gap between the risk and its empirical estimate based on noisy training signals. It is easily checked that the non-zero gap is unavoidable, even in the ideal case where $x_1 = \cdots = x_m = x$ for some $x \in \mathbb{R}^p$, and the estimator τ is a constant function that always outputs x. Proposition 7.10 shows that, however, one can improve on the generalization performance guarantee, by denoising the training signals before learning a sub-sampling pattern.

A result similar to Proposition 7.10 can be found in [82], where the additive Gaussian noise model is assumed.

8 Conclusions

We have presented rigorous solutions to address issues due to lack of smoothness and/or strong convexity.

- We have developed a unified framework to establish variable selection consistency of ℓ_1-penalized M-estimators, based on a novel local structured smoothness condition (LSSC). We have derived the sample complexity in the high-dimensional setting for several statistical learning problems.
- We have presented a sharp analysis of the estimation error of the lasso, based on a novel relaxed restricted strong convexity (RSC) condition. Our result establishes the minimax optimality of the lasso for estimating exactly or weakly sparse parameters.
- We have proved three convergence results for convex optimization.
  - Convergence of the Frank-Wolfe algorithm for objective functions involving the exp-linear loss.
  - Convergence of the mirror descent algorithm for locally relatively smooth objective functions.
  - Convergence of the exponentiated gradient method for convex differentiable objective functions.

  The first result shows that the Frank-Wolfe algorithm is a scalable approach to rigorous exp-linear optimization. Numerical results showed that the exponentiated gradient method with Armijo line search was the fastest guaranteed-to-converge algorithm for quantum state tomography, on real data-sets.
- We have proposed a novel framework for compressive MRI based on agnostic PAC learning, where the necessity of the famous restricted isometry property (RIP) vanishes. The framework leads to a computationally efficient compressive MRI system, with a rigorous statistical risk guarantee that does not require any *a priori* knowledge of the signal structure.

A key idea underlying most of our results above is *localization*. That is, we identify a specific set on which the smoothness/strong convexity condition is actually necessary, and only require the condition on the set. For example, the LSSC essentially requires the the Hessian of the loss function to be Lipschitz continuous, in a neighborhood of the parameter to be estimated; the relaxed RSC condition requires the RSC to hold, outside an ℓ_2-norm ball centered at the parameter to be estimated; the local relative smoothness condition requires the objective function to be relatively smooth, in a neighborhood of a limit point of the sequence of iterates.

Another key idea is *reformulation*. There can be several mathematical models corresponding to the same real-world application. Each model has its own pros and cons; in particular, a hard issue in a model may vanish in another model. Our approach to compressive MRI has demonstrated the power of a proper reformulation, showing the possibility of designing a compressive MRI system without any knowledge of the signal structure. The trade-off is that the performance guarantee is not on the statistical risk, but the gap to the unknown optimal risk.

8.1 Future research directions

It seems impossible to address all machine learning problems without smoothness and/or strong convexity in a unified framework. Focusing on the exp-linear loss, however, may lead to a deeper understanding of the necessity of smoothness and strong convexity conditions. There are already some theories for the exp-linear loss, yet many important applications involving the exp-linear loss currently lack complete solutions.

Recall that the exp-linear loss is defined as $f(x) := -\log\langle a, x\rangle$ for some vector a, or $f(X) := -\log\mathrm{Tr}(AX)$ for some matrix A. The exp-linear loss appears in several important applications; below are three examples.

- **Positron emission tomography (PET):** Maximum-likelihood (ML) PET requires solving the optimization problem [180]:

$$x^\star \in \operatorname*{arg\,min}_x \left\{ \frac{1}{n}\sum_{i=1}^{n} \left(\langle a_i, x\rangle - y_i \log\langle a_i, x\rangle \right) \,\middle|\, x \in \mathscr{P} \right\},$$

  for given $\{a_i\} \subset \mathbb{R}^p$ and $\{y_i\} \subset \mathbb{N}$, where $\mathscr{P}$ denotes the probability simplex.

- **Optimal portfolio selection (OPS):** The growth-optimal portfolio selection strategy requires solving the optimization problem [30]:

$$x^\star \in \operatorname*{arg\,min}_x \left\{ \mathsf{E}\left[-\log\langle a, x\rangle \right] \,\middle|\, x \in \mathscr{P} \right\},$$

  where the expectation is with respect to the random vector a. Each element of a denotes the *price relative* of an investment alternative. Notice that this strategy requires the

probability distribution of the price relatives.

A closely related problem is on-line portfolio selection. Fix some *time horizon* $T \in \mathbb{N}$. The aim of on-line portfolio selection is to generate portfolios $x_1, \ldots, x_T$ sequentially, to achieve a small *regret*. The regret is defined as

$$R_T := \sum_{i=1}^{T} \left(-\log\langle a_t, x_t\rangle\right) - \min_x \left\{ \sum_{i=1}^{T} \left(-\log\langle a_t, x\rangle\right) \,\middle|\, x \in \mathscr{P} \right\},$$

for sequentially incoming price relative vectors a_t.

- **Quantum state tomography (QST):** ML QST was discussed in Chapter 5. It requires solving the optimization problem [96]:

$$X^\star \in \arg\min_X \left\{ \frac{1}{n} \sum_{i=1}^{n} \left(-\log \mathrm{Tr}(A_i X)\right) \,\middle|\, X \in \mathscr{D} \right\},$$

for a given set of Hermitian matrices $\{A_i\}$, where $\mathscr{D}$ denotes the set of quantum density matrices (cf. (5.2)).

Other applications involve gamma regression (cf. Section 2.5.3), positive linear inverse problems [34], etc. Notice that the interior point method (IPM) for linear programming also involves minimizing exp-linear functions, which act as barrier functions for the linear constraints, as intermediate steps [136, 140].

We point out four interesting research directions below.

8.1.1 Compressive QST with guarantees

QST is essentially a matrix estimation problem, while we have only addressed the computational aspect of QST in Chapter 5. In a standard implementation of QST, when there are q qubits, one has to do about $3^q \cdot 100$ measurements; that is, the *sample complexity* grows exponentially with the number of qubits [85]. In many situations, it is already known that the matrix to be estimated is low-rank [85]. Can we exploit the theory of compressive sensing—low-rank matrix recovery in particular—to accelerate the measurement procedure?

Research Problem 1 *Characterize the sample complexity of QST, in terms of the rank of the unknown quantum density matrix.*

The possibility of compressive QST, when QST is approximated by a linear inverse problem, has been rigorously proved and attracted attention of the quantum computation community (see, e.g., [36, 75, 84, 107, 155, 167]). In practice, however, the ML estimator for the exact statistical model (cf. Example 5.1) was adopted, as it yields better statistical accuracy [85, 146]. The aim of Research Problem 1 is to provide a rigorous verification for the practice.

8.1.2 Provably faster PET/OPS/QST

In Chapter 5, we have rigorously proved that the exponentiated gradient method converges for QST, and observed that it was empirically the fastest on real data. As PET and OPS are also exp-linear optimization problems, it is easily checked that the exponentiated gradient method also applies to these two applications. A natural next step is the following.

Research Problem 2 *Characterize the convergence rate of the exponentiated gradient method applied to the exp-linear loss.*

Such a result is especially important for QST—none of the fast enough existing QST algorithms has a convergence speed guarantee. Notice that the standard approach to PET and OPS, based on expectation maximization, does not have a convergence speed guarantee either [58, 180].

It is already known that the IPM for linear programming converges slowly, when the number of linear constraints is large; addressing this issue has been an open problem in theoretical computer science for decades [115]. The rationale is similar to what we have discussed in Section 5.6: When there are many linear constraints, the step sizes have to be very small to ensure that none of the constraints will be violated. The standard IPM concerns with minimizing a sequence of exp-linear losses by Newton's method, and our results in Chapter 5 and 6 show that one may replace Newton's method by the exponentiated gradient method. Interestingly, as the iterates of the exponentiated gradient method are element-wise strictly positive, all constraints are automatically satisfied in Karmarkar's formulation of linear programs [102]. A good result for Research Problem 2 may help us attack this open problem.

8.1.3 On-line algorithms

On-line portfolio selection is a classical topic in on-line learning. It is known that, if there exists a constant strictly positive *market variability parameter (MVP)*, a fast rate ($R_T = O(\log T)$) can be achieved [2, 90, 91, 178]. Requiring existence of a constant MVP, however, is unrealistic—the MVP is a lower bound on the elements of the price relatives $a_1, \ldots, a_T$, which can be arbitrarily close to zero as $T \to \infty$. There are indeed algorithms for online portfolio selection without the MVP, but they are computationally very expensive [59, 60, 101].

PET and QST are very suitable for the on-line learning framework—the number of measurements is typically very large, so one would like to process the measurement outcomes sequentially, instead of in a batch, to reduce the computational burden. To the best of our knowledge, there does not exist any on-line algorithm for QST. There is one very recent result on "almost on-line" PET, which is sequential, but cannot run on the fly during the measurement process [70].

Research Problem 3 *Develop efficient on-line algorithms for on-line portfolio selection, PET, and QST, without unrealistic assumptions.*

There is one bonus. Given a "truly on-line" algorithm, the corresponding statistical per-

formance guarantee comes for free by online-to-batch conversion [47, 48]. Therefore, this research direction provides an alternative approach to addressing Research Problem 1.

8.1.4 OPS and optimal time series prediction

The special structure of the exp-linear loss had been exploited by information theorists to prove the optimality of growth-optimal portfolio selection [4, 30]. However, the growth-optimal portfolio selection strategy requires knowledge of the true probability distribution of the market, which is unrealistic. Can we solve the issue?

Research Problem 4 *Develop a probability-free version of the growth-optimal portfolio selection strategy.*

To be more precise, the goal is to show that the growth-optimal strategy is optimal under *Knightian uncertainty*—uncertainty for which we do not even have a probabilistic model. This is indeed the main focus of on-line learning and game-theoretic probability [49, 162]. The idea is to re-formulate and re-study the growth-optimal strategy under these frameworks.

The growth-optimal strategy was later generalized to derive the optimal strategy for time series prediction [3, 131], under some statistical assumptions. A good result for Research Problem 4 may lead to a similar result for time series prediction, *without any subjective, unverifiable statistical assumption*. Such a result would be of broad interest in the fields of machine learning and information theory.

A Mathematical Prerequisites

This chapter summarizes the mathematical facts necessary for the theories in this thesis.

A.1 Convex analysis

Let d be a positive integer. A set $\mathscr{X} \subseteq \mathbb{R}^d$ is said to be *convex*, if

$$x \in \mathscr{X}, y \in \mathscr{X} \Rightarrow \alpha x + (1-\alpha) y \in \mathscr{X}, \forall (x, y) \in \mathscr{X} \times \mathscr{X}, \alpha \in (0,1).$$

Let f be a function from $\mathbb{R}^d$ to $[-\infty, +\infty]$. Its *epigraph* is given by

$$\operatorname{epi} f := \left\{ (x, y) \in \mathbb{R}^d \times \mathbb{R} \;\middle|\; f(x) \leq y \right\}.$$

We say that f is *closed*, if $\operatorname{epi} f$ is a closed set. We say that f is *convex*, if $\operatorname{epi} f$ is a convex set. It is easily checked that f is convex, if and only if

$$f(\alpha x + (1-\alpha) y) \leq \alpha f(x) + (1-\alpha) f(y), \quad \forall (x, y) \in \mathbb{R}^d \times \mathbb{R}^d, \alpha \in (0,1).$$

The *(effective) domain* of f is the projection of $\operatorname{epi} f$ on $\mathbb{R}^d$, i.e.,

$$\operatorname{dom} f := \left\{ x \in \mathbb{R}^d \;\middle|\; f(x) < +\infty \right\}.$$

Consider the constrained optimization problem:

$$f^\star \in \min_x \left\{ f(x) \;\middle|\; x \in \mathscr{X} \right\},$$

where f is a closed convex function on $\mathbb{R}^d$, and $\mathscr{X}$ is a closed convex set in $\mathbb{R}^d$.

Theorem A.1. *Suppose that f is differentiable at a point $x^\star \in \mathscr{X}$. The point $x^\star$ minimizes f on*

$\mathscr{X}$, if and only if

$$\langle \nabla f(x^\star), x - x^\star \rangle \geq 0, \quad \forall x \in \mathscr{X}.$$

The notion of 2-smoothness has equivalent formulations (see, e.g., [136]). Recall that a function f is 2-smoothness, if its gradient is Lipschitz continuous.

Theorem A.2. *Let f be a continuously differentiable function on $\mathbb{R}^p$. The following three statements are equivalent.*

1. *The gradient of f is Lipschitz with parameter $L > 0$, i.e.,*

$$\|\nabla f(y) - \nabla f(x)\|_2 \leq L\|y - x\|_2, \quad \forall x, y \in \mathbb{R}^p.$$

2. *It holds that*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{L}{2}\|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^p.$$

3. *It holds that*

$$\alpha f(x) + (1-\alpha) f(y) \leq f(\alpha x + (1-\alpha) y) + \alpha(1-\alpha)\frac{L}{2}\|x - y\|_2^2, \quad \forall x, y \in \mathbb{R}^p \text{ and } \alpha \in [0,1].$$

The notion of strong convexity also has equivalent formulations (see, e.g., [136]).

Theorem A.3. *Let f be a function on $\mathbb{R}^p$. The following three statements are equivalent.*

1. *The function f is strongly convex with parameter $\mu > 0$, i.e.,*

$$(1-\alpha) f(x) + \alpha f(y) \geq f((1-\alpha)x + \alpha y) + \alpha(1-\alpha)\frac{\mu}{2}\|y - x\|_2^2, \quad \forall x, y \in \mathscr{X} \text{ and } \alpha \in [0,1].$$

2. *Suppose that f is continuously differentiable. It holds that*

$$\langle \nabla f(y) - \nabla f(x), y - x \rangle \geq \mu\|y - x\|_2^2, \quad \forall x, y \in \mathbb{R}^p.$$

3. *Suppose that f is twice continuously differentiable. It holds that*

$$\nabla^2 f(x) \geq \mu I, \quad \forall x \in \mathbb{R}^p.$$

A.2 Matrix analysis

Chapters 5 and 6 involve matrix functions. The functions $\exp(\cdot)$ and $\log(\cdot)$ in (6.1) denote matrix exponential and logarithm functions, respectively. In general, let $X \in \mathbb{C}^{d\times d}$ be Hermitian, and

$X = \sum_j \lambda_j P_j$ be its spectral decomposition. Let g be a real-valued function whose domain contains $\{\lambda_j\}$. Then $g(X) := \sum_j g(\lambda_j) P_j$.

The Peierls-Bogoliubov inequality says that the function

$$\varphi(t) := \log \mathrm{Tr} \exp(A + tB), \quad \forall\, t \in \mathbb{R},$$

is convex for any given Hermitian matrices A and B (see, e.g., [144]). Equivalently, we have

$$\varphi''(t) \geq 0, \quad \forall\, t \in \mathbb{R}.$$

The set of $d \times d$ *quantum density matrices* is given by

$$\mathscr{D} := \left\{ \rho \in \mathbb{C}^{d \times d} \;\middle|\; \rho \geq 0, \mathrm{Tr}\, \rho = 1 \right\}.$$

A quantum density matrix can be viewed as a matrix analogue of a probability distribution; in particular, it is easily checked that the diagonal of a quantum density matrix defines a probability distribution for a d-ary random variable (r.v.).

Let $\rho, \sigma \in \mathscr{D}$ be non-singular. The negative von Neumann entropy is defined as

$$h(\rho) := \mathrm{Tr}(\rho \log \rho) - \mathrm{Tr}(\rho).$$

The quantum relative entropy is defined as

$$D(\rho, \sigma) := \mathrm{Tr}(\rho \log \rho) - \mathrm{Tr}(\rho \log \sigma) - \mathrm{Tr}(\rho - \sigma).$$

The quantum relative entropy is jointly convex; that is (see, e.g., [46]), for every $\alpha \in [0, 1]$,

$$\alpha D(\rho_1, \sigma_1) + (1 - \alpha) D(\rho_2, \sigma_2) \geq D(\alpha \rho_1 + (1 - \alpha)\rho_2, \alpha \sigma_1 + (1 - \alpha)\sigma_2).$$

It is easily checked that the quantum relative entropy is the Bregman divergence induced by the negative von Neumann entropy; hence, it is always non-negative. Pinsker's inequality says that [92]

$$D(\rho, \sigma) \geq \frac{1}{2} \|\rho - \sigma\|_*^2,$$

where $\|\cdot\|_*$ denotes the nuclear norm. Therefore, $D(\rho, \sigma) = 0$ if and only if $\rho = \sigma$. Restricting the inequality above to diagonal matrices, we obtain the original form of Pinsker's inequality (see, e.g., [61]):

$$D(v, u) := \sum_{i=1}^{p} v_i \log\left(\frac{v_i}{u_i}\right) \geq \frac{1}{2} \|v - u\|_1^2, \quad \forall\, v, u \in \mathscr{P},$$

where $\|\cdot\|_1$ denotes the ℓ_1-norm, and $\mathscr{P}$ denotes the probability simplex in $\mathbb{R}^p$.

A.3 Concentration inequalities

The notion of subgaussian random variables (r.v.'s) is an extension of that of Gaussian r.v.'s.

Definition A.4. *A r.v. ξ is* subgaussian, *if there exists a constant $K > 0$ such that*

$$(\mathbb{E}|\xi|^p)^{1/p} \leq K\sqrt{p}, \quad \forall p \geq 1.$$

The subgaussian norm $\|\xi\|_{\psi_2}$ *of a subgaussian r.v. ξ is defined as the smallest K, i.e.,*

$$\|\xi\|_{\psi_2} := \sup\left\{ p^{-1/2}(\mathbb{E}|\xi|^p)^{1/p} \;\middle|\; p \geq 1 \right\}.$$

For example, Gaussian, Rademacher, and bounded r.v.'s are subgaussian.

Let $\xi_1, \ldots, \xi_n$ be independent and identically distributed Gaussian r.v.'s of zero mean and unit variance. Then their average $\bar{\xi}_n$ is a Gaussian r.v. of zero mean and variance $(1/n)$. It is easily verified, via Chernoff's bound, that

$$\begin{aligned} \mathsf{P}\left\{ \left|\bar{\xi}_n\right| \geq t \right\} &\leq \mathrm{e}^{-\lambda t} \mathsf{E}\, \mathrm{e}^{\lambda \bar{\xi}_n} \Big|_{\lambda = t} \\ &= \mathrm{e}^{-(1/2)nt^2}, \quad \forall t \geq 0. \end{aligned}$$

A similar result holds for subgaussian r.v.'s (see, e.g., [182]).

Theorem A.5 (Hoeffding-type inequality). *Let $\xi_1, \ldots, \xi_n$ be independent mean-zero subgaussian r.v.'s. Define*

$$K_{\max} := \max_i \left\{ \|\xi_i\|_{\psi_2} \;\middle|\; i = 1, \ldots, n \right\}.$$

There exists some universal constant $c > 0$, such that

$$\mathsf{P}\left\{ \left| \sum_{i=1}^n a_i \xi_i \right| \geq t \right\} \leq \mathrm{e} \cdot \exp\left(\frac{-ct^2}{K_{\max}^2 \|a\|_2^2} \right), \quad \forall a \in \mathbb{R}^n \text{ and } t \geq 0.$$

For bounded r.v.'s, a sharp inequality can be obtained via Hoeffding's lemma (see, e.g., [126]).

Theorem A.6 (Hoeffding's Inequality). *Let $\xi_1, \ldots, \xi_n$ be independent r.v.'s, such that ξ_i takes its value in $[a_i, b_i]$ almost surely for all $i \in \{1, \ldots, n\}$. Then*

$$\mathsf{P}\left\{ \left| \sum_{i=1}^n (\xi_i - \mathsf{E}\,\xi_i) \right| \geq t \right\} \leq 2\exp\left[-\frac{2t^2}{\sum_{i=1}^n (b_i - a_i)^2} \right], \quad \forall t \geq 0.$$

In the two inequalities above, the exponents of the probability bounds depends on t through t^2. In general, this may not be true. For r.v.'s that are not subgaussian, the following result is useful (see, e.g., [126]).

Theorem A.7 (Bernstein's Inequality). *Let $X_1, \ldots, X_n$ be independent real random variables. Suppose that there exist $v > 0$ and $c > 0$ such that $\sum_{i=1}^{n} \mathsf{E}\, X_i^2 \le v$, and*

$$\sum_{i=1}^{n} \mathsf{E}\, |X_i|^q \le \frac{q!}{2} v c^{q-2}$$

for all integers $q \ge 3$. Then

$$\mathsf{P}\left\{ \left| \sum_{i=1}^{n} (X_i - \mathsf{E}\, X_i) \right| \ge t \right\} \le 2 \exp\left[-\frac{t^2}{2(v + ct)} \right], \quad \forall\, t \ge 0\,.$$

Notice that the dependence of the exponent of the probability bound on t is different from that in Theorems A.5 and A.6.

Bibliography

[1] ADCOCK, B., HANSEN, A. C., POON, C., AND ROMAN, B. Breaking the coherence barrier: A new theory for compressed sensing. *Forum Math., Sigma 5*, e4 (2017).

[2] AGARWAL, A., HAZAN, E., KALE, S., AND SCHAPIRE, R. E. Algorithms for portfolio management based on the Newton method. In *Proc. 23rd Int. Conf. Machine Learning (ICML)* (2006), pp. 9 – 16.

[3] ALGOET, P. H. The strong law of large numbers for sequential decisions under uncertainty. *IEEE Trans. Inf. Theory 40*, 3 (1994), 609–633.

[4] ALGOET, P. H., AND COVER, T. M. Asymptotic optimality and asymptotic equipartition properties of log-optimum investment. *Ann. Probab. 16*, 2 (1988), 876–898.

[5] ANTHONY, M., AND BARTLETT, P. L. *Neural Network Learning: Theoretical Foundations*. Cambridge Univ. Press, Cambridge, UK, 1999.

[6] ARMIJO, L. Minimization of functions having Lipschitz continuous first partial derivatives. *Pac. J. Math. 16*, 1 (1966), 1–3.

[7] AUSLENDER, A., AND TEBOULLE, M. Interior gradient and proximal methods for convex and conic optimization. *SIAM J. Optim. 16*, 3 (2006), 697–725.

[8] BACH, F. Self-concordant analysis for logistic regression. *Electron. J. Stat. 4* (2010), 384–414.

[9] BACH, F. Learning with submodular functions: A convex optimization perspective. *Found. Trends Mach. Learn. 6*, 2–3 (2013), 145–373.

[10] BACH, F. Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression. *J. Mach. Learn. Res. 15* (2014), 595–627.

[11] BAHMANI, S., AND ROMBERG, J. Phase retrieval meets statistical learning theory: A flexible convex relaxation. In *Proc. 20th Int. Conf. Artificial Intelligence and Statistics* (2017), pp. 252–260.

[12] BALDASSARRE, L., LI, Y.-H., SCARLETT, J., GÖZCÜ, B., BOGUNOVIC, I., AND CEVHER, V. Learning-based compressive subsampling. *IEEE J. Sel. Topics Signal Process. 10*, 4 (2016), 809–822.

[13] BARANIUK, R. G., CEVHER, V., DUARTE, M. F., AND HEGDE, C. Model-based compressive sensing. *IEEE Trans. Inf. Theory 56*, 4 (Apr. 2010), 1982–2001.

[14] BARTLETT, P. L., JORDAN, M. I., AND MCAULIFFE, J. D. Convexity, classification, and risk bounds. *J. Am. Stat. Assoc. 101*, 473 (Mar. 2006), 138–156.

[15] BARTLETT, P. L., MENDELSON, S., AND NEEMAN, J. ℓ_1-regularized linear regression: persistence and oracle inequalities. *Probab. Theory Relat. Fields 154* (2012), 193–224.

[16] BAUSCHKE, H. H., BOLTE, J., AND TEBOULLE, M. A descent lemma beyond Lipschitz gradient continuity: first-order methods revisited and applications. *Math. Oper. Res. 42*, 2 (2017), 330–348.

[17] BAUSCHKE, H. H., BORWEIN, J. M., AND COMBETTES, P. L. Essential smoothness, essential strict convexity, and Legendre functions in Banach spaces. *Commun. Contemp. Math. 3*, 4 (2001), 615–647.

[18] BAUSCHKE, H. H., AND COMBETTES, P. L. *Convex analysis and monotone operator thoery in Hilbert spaces*. Springer, New York, NY, 2011.

[19] BECK, A., AND TEBOULLE, M. Mirror descent and nonlinear projected subgradient methods for convex optimization. *Oper. Res. Lett. 31* (2003), 167–175.

[20] BECK, A., AND TEBOULLE, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci. 2*, 1 (2009), 183–202.

[21] BECKER, S. R., CANDÈS, E. J., AND GRANT, M. C. Templates for convex cone problems with applications to sparse signal recovery. *Math. Program. Comput. 3* (2011), 165–218.

[22] BELLONI, A., CHERNOZHUKOV, V., AND WANG, L. Square-root lasso: pivotal recovery of sparse signals via conic programming. *Biometrika 98*, 4 (2011), 791–806.

[23] BERTSEKAS, D. P. On the Goldstein-Levitin-Polyak gradient projection method. *IEEE Trans. Automat. Contr. AC-21*, 2 (1976), 174–184.

[24] BERTSEKAS, D. P. *Nonlinear Programming*, 3rd ed. Athena Sci., Belmont, MA, 2016.

[25] BICKEL, P., RITOV, Y., AND TSYBAKOV, A. B. Simultaneous analysis of Lasso and Dantzig selector. *Ann. Stat. 37*, 4 (2009), 1705–1732.

[26] BLUME-KOHOUT, R. Hedged maximum likelihood quantum state estimation. *Phys. Rev. Lett. 105* (2010).

[27] BOBKOV, S. G., AND LEDOUX, M. On modified logarithmic Sobolev inequalities for Bernoulli and Poisson meausres. *J. Funct. Anal. 156* (1998), 347–365.

[28] BOLDUC, E., KNEE, G. C., GAUGER, E. M., AND LEACH, J. Projected gradient descent algorithms for quantum state tomography. *npj Quantum Inf. 3* (2017).

[29] BRAND, M. Fast low-rank modifications of the thin singular value decomposition. *Linear Algebra Appl. 415* (2006), 20–30.

[30] BREIMAN, L. Investment policies for expanding business optimal in a long-run sense. In *Stochastic Optimization Models in Finance*, W. T. Ziemba and R. G. Vickson, Eds. Academic Press, New York, NY, 1975, pp. 593–598.

[31] BUBECK, S. Convex optimization: Algorithms and complexity. *Found. Trends Mach. Learn. 8*, 3–4 (2015), 231–358.

[32] BÜHLMANN, P., AND VAN DE GEER, S. *Statistics for High-Dimensional Data*. Springer, Berlin, 2011.

[33] BUNEA, F. Honest variable selection in linear and logistic regression models via ℓ_1 and $\ell_1 + \ell_2$ penalization. *Electron. J. Stat. 2* (2008), 1153–1194.

[34] BYRNE, C., AND CENSOR, Y. Proximity function minimization using multiple Bregman projections, with application to split feasibility and Kullback-Leibler distance minimization. *Ann. Oper. Res. 105* (2001), 77–98.

[35] BYRNE, C. L. Alternating minimization as sequential unconstrained minimization: A survey. *J. Optim. Theory Appl. 156* (2013), 554–566.

[36] CAI, T. T., KIM, D., WANG, Y., YUAN, M., AND ZHOU, H. H. Optimal large-scale quantum state tomography with Pauli measurements. *Ann. Stat. 44*, 2 (2016), 682–712.

[37] CANDES, E., AND TAO, T. Decoding by linear programming. *IEEE Trans. Inf. Theory 51*, 12 (Dec. 2005), 4203–4215.

[38] CANDÈS, E. J. The restricted isometry property and its implications for compressed sensing. *C. R. Acad. Sci. Paris, Ser. I 346* (2008), 589–592.

[39] CANDÈS, E. J., ELDAR, Y. C., STROHMER, T., AND VORONINSKI, V. Phase retrieval via matrix completion. *SIAM Rev. 57*, 2 (2015), 225–251.

[40] CANDÈS, E. J., LI, X., AND SOLTANOLKOTABI, M. Phase retrieval from coded diffraction patterns. *Appl. Comput. Harmon. Anal. 39* (2015), 277–299.

[41] CANDÈS, E. J., LI, X., AND SOLTANOLKOTABI, M. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Trans. Inf. Theory 61*, 4 (2015), 1985–2007.

[42] CANDÈS, E. J., AND PLAN, Y. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Trans. Inf. Theory 57*, 4 (2011), 2342–2359.

[43] CANDÈS, E. J., ROMBERG, J., AND TAO, T. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory 52*, 2 (Feb. 2006), 489–509.

[44] CANDÈS, E. J., STROHMER, T., AND VORONINSKI, V. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Commun. Pure Appl. Math. LXVI* (2013), 1241–1274.

[45] CANDES, E. J., AND TAO, T. Near-optimal signal recovery from random projections: Universal encoding strategies? *IEEE Trans. Inf. Theory 52*, 12 (2006), 5406–5425.

[46] CARLEN, E. Trace inequalities and quantum entropy: An introductory course. In *Entropy and the Quantum*. Am. Math. Soc., Providence, RI, 2010, pp. 73–140.

[47] CESA-BIANCHI, N., CONCONI, A., AND GENTILE, C. On the generalization ability of on-line learning algorithms. *IEEE Trans. Inf. Theory 50*, 9 (2004), 2050–2057.

[48] CESA-BIANCHI, N., AND GENTILE, C. Improved risk tail bounds for on-line algorithms. *IEEE Trans. Inf. Theory 54*, 1 (2008), 386–390.

[49] CESA-BIANCHI, N., AND LUGOSI, G. *Prediction, Learning, and Games*. Cambridge Univ. Press, Cambridge, UK, 2006.

[50] CHANDRASEKARAN, V., RECHT, B., PARRILO, P. A., AND WILLSKY, A. S. The convex geometry of linear inverse problems. *Found. Comput. Math. 12* (2012), 805–849.

[51] CHAUFFERT, N., CIUCIU, P., KAHN, J., AND WEISS, P. Variable density sampling with continuous trajectories. *SIAM J. Imaging Sci. 7*, 4 (2014), 1962–1992.

[52] CHEN, C., AND HUANG, J. Compressive sensing MRI with wavelet tree sparsity. In *Adv. Neural Information Processing Systems 25* (2012).

[53] CHEN, Y., AND CANDÈS, E. J. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Adv. Neural Information Processing Systems 28* (2015).

[54] CHEN, Y., AND CANDÈS, E. J. Solving random quadratic systems of equations is nearly as easy as solving linear systems. *Commun. Pure Appl. Math. LXX* (2017), 0822–0883.

[55] COHEN, M. B., MĄDRY, A., TSIPRAS, D., AND VLADU, A. Matrix scaling and balancing via box constrained Newton's method and interior point methods. In *IEEE 58th Annu. Symp. Foundations of Computer Science* (2017), pp. 902–913.

[56] COMBETTES, P. L., AND WAJS, V. R. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul. 4*, 4 (2005), 1168–1200.

[57] CORMEN, T. H., LEISERSON, C. E., RIVEST, R. L., AND STEIN, C. *Introduction to Algorithms*, third ed. MIT Press, Cambridge, MA, 2009.

[58] COVER, T. M. An algorithm for maximizing expected log investment return. *IEEE Trans. Inf. Theory IT-30*, 2 (1984), 369–373.

[59] COVER, T. M. Universal portfolios. *Math. Finance 1*, 1 (1991), 1–29.

[60] COVER, T. M., AND ORDENTLICH, E. Universal portfolios with side information. *IEEE Trans. Inf. Theory 42*, 2 (1996), 348–363.

[61] CSISZÁR, I., AND KÖRNER, J. *Information Theory: Coding Theorems for Discrete Memoryless Systems*, second ed. Cambridge Univ. Press, Cambridge, UK, 2011.

[62] CSISZÁR, I., AND TUSNÁDY, G. Information geometry and alternating minimization procedures. *Stat. Decis.*, Supplement 1 (1984), 205–237.

[63] DAVENPORT, M. A., PLAN, Y., VAN DEN BERG, E., AND WOOTTERS, M. 1-bit matrix completion. *Inf. Inference 3* (2014), 189–223.

[64] DECARREAU, A., HILHORST, D., LEMARÉCHAL, C., AND NAVAZA, J. Dual methods in entropy maximization. application to some problems in crystallography. *SIAM J. Optim. 2*, 2 (1992), 173–197.

[65] DOLJANSKY, M., AND TEBOULLE, M. An interior proximal algorithm and the exponential multiplier method for semidefinite programming. *SIAM J. Optim. 9*, 1 (1998), 1–13.

[66] DONOHO, D. L. Compressed sensing. *IEEE Trans. Inf. Theory 52*, 4 (Apr. 2006), 1289–1306.

[67] DONOHO, D. L., AND TANNER, J. Neighborliness of randomly projected simplices in high dimensions. *Proc. Natl. Acad. Sci. 102*, 27 (2005), 9452–9457.

[68] DRUSVYATSKIY, D., IOFFE, A. D., AND LEWIS, A. S. Nonsmooth optimization using Taylor-like models: error bounds, convergence, and termination criteria. arXiv:1610.03446v1.

[69] ECKSTEIN, J., AND YAO, W. Understanding the convergence of the alternating direction method of multipliers: Theoretical and computational perspectives. 2015.

[70] EHRHARDT, M., MARKIEWICZ, P., CHAMBOLLE, A., RICHTÁRIK, P., SCHOTT, J., AND SCHÖNLIEB, C.-B. Faster PET reconstruction with a stochastic primal-dual hybrid gradient method. In *Proc. SPIE 10394, Wavelets and Sparsity XVII* (2017).

[71] EL HALABI, M., AND CEVHER, V. A totally unimodular view of structured sparsity. In *18th Int. Conf. Artificial Intelligence and Statistics* (2015).

[72] FAN, J., AND LV, J. Nonconcave penalized likelihood with NP-dimensionality. *IEEE Trans. Inf. Theory 57*, 8 (Aug. 2011), 5467–5484.

[73] FIENUP, J. R. Phase retrieval algorithms: a comparison. *Appl. Opt. 21*, 15 (1982), 2758–2769.

[74] FINO, B. J., AND ALGAZI, V. R. A unified treatment of discrete fast unitary transforms. *SIAM J. Comput. 6*, 4 (1977), 700–717.

[75] FLAMMIA, S. T., GROSS, D., LIU, Y.-K., AND EISERT, J. Quantum tomography via compressed sensing: Error bounds, sample complexity and efficient estimators. *New J. Phys. 14* (2012).

[76] FOUCART, S., AND RAUHUT, H. *A Mathematical Introduction to Compressive Sensing*. Birkhäuser, Basel, 2013.

[77] FREUND, R. M., AND GRIGAS, P. New analysis and results for the Frank-Wolfe method. *Math. Program., Ser. A 155* (2016), 199–230.

[78] GAFNI, E. M., AND BERTSEKAS, D. P. Convergence of a gradient projection method. LIDS-P-1201, Laboratory for Information and Decision Systems, Massachusetts Institute of Technology, 1982.

[79] GARBER, D., AND HAZAN, E. Faster rates for the Frank-Wolfe method over strongly-convex sets. In *Proc. 32nd Int. Conf. Machine Learning* (2015), pp. 541–549.

[80] GIRAUD, C. *Introduction to High-Dimensional Statistics*. CRC Press, Boca Raton, FL, 2015.

[81] GOLDSTEIN, T., AND STUDER, G. Phasemax: Convex phase retrieval via basis pursuit. *IEEE Trans. Inf. Theory 64*, 4 (2018), 2675–2689.

[82] GÖZCÜ, B., MAHABADI, K., LI, Y.-H., ILICAK, E., ÇUKUR, T., SCARLETT, J., AND CEVHER, V. Learning-based compressive MRI. *IEEE Trans. Med. Imag.* (2018). Early access.

[83] GREENSHTEIN, E., AND RITOV, Y. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli 6* (2004), 971–988.

[84] GROSS, D., LIU, Y.-K., FLAMMIA, S. T., BECKER, S., AND EISERT, J. Quantum state tomography via compressed sensing. *Phys. Rev. Lett. 105* (2010).

[85] HÄFFNER, H., HÄNSEL, W., ROOS, C. F., BENHELM, J., CHECK-AL-KAR, D., CHWALLA, M., KÖRBER, T., RAPOL, U. D., RIEBE, M., SCHMIDT, P. O., BECHER, C., GÜHNE, O., DÜR, W., AND BLATT, R. Scalable multiparticle entanglement of trapped ions. *Nature 438* (2005), 643–646.

[86] HARCHAOUI, Z., JUDITSKY, A., AND NEMIROVSKI, A. Conditional gradient algorithms for norm-regularized smooth convex optimization. *Math. Program., Ser. A 152*, 1 (2015), 75–112.

[87] HASTIE, T., TIBSHIRANI, R., AND WAINWRIGHT, M. *Statistical Learning with Sparsity: The Lasso and generalizations*. CRC Press, Boca Raton, FL, 2015.

[88] HAUSSLER, D. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inf. Comput. 100* (1992), 78–150.

[89] HAVIV, I., AND REGEV, O. The restricted isometry property of subsampled Fourier matrices. In *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2014–2016*, B. Klartag and E. Milman, Eds. Springer, Cham, 2017.

[90] HAZAN, E., AGARWAL, A., AND KALE, S. Logarithmic regret algorithms for online convex optimization. *Mach. Learn. 69* (2007), 169–192.

[91] HAZAN, E., AND KALE, S. An online portfolio selection algorithm with regret logarithmic in price variation. *Math. Finance 25*, 2 (2015), 288–310.

[92] HIAI, F., OHYA, M., AND TSUKADA, M. Sufficiency, KMS condition and relative entropy in von Neumann algebras. *Pac. J. Math. 96*, 1 (1981), 99–109.

[93] HIRIART-URRUTY, J.-B., STRODIOT, J.-J., AND NGUYEN, V. H. Generalized Hessian matrix and second-order optimality conditions for problem with $C^{1,1}$ data. *Appl. Math. Optim. 11* (1984), 43–56.

[94] HÖFFGEN, K.-U., AND SIMON, H.-U. Robust trainability of single neurons. *J. Comput. Syst. Sci. 50* (1995), 114–125.

[95] HORN, R. A., AND JOHNSON, C. R. *Matrix Analysis*. Cambridge Univ. Press, Cambridge, UK, 1985.

[96] HRADIL, Z. Quantum-state estimation. *Phys. Rev. A 55*, 3 (1997).

[97] IOFFE, A., AND MILOSZ, T. On a characterization of $C^{1,1}$ functions. *Cybern. Syst. Anal. 38*, 3 (2002), 313–322.

[98] JAGGI, M. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *Proc. 30th Int. Conf. Machine Learning* (2013).

[99] JI, S., XUE, Y., AND CARIN, L. Bayesian compressive sensing. *IEEE Trans. Sig. Process. 56*, 6 (2008), 2346–2356.

[100] JUDITSKY, A., AND NEMIROVSKI, A. First-order methods for nonsmooth convex large-scale optimization, I: General purpose methods. In *Optimization for Machine Learning*, S. Sra, S. Nowozin, and S. J. Wright, Eds. MIT Press, Cambridge, MA, 2012, ch. 5.

[101] KALAI, A., AND VEMPALA, S. Efficient algorithms for universal portfolios. *J. Mach. Learn. Res. 3* (2002), 423–440.

[102] KARMARKAR, N. A new polynomial-time algorithm for linear programming. In *Proc. 16th Annu. ACM Symp. Theory of Computing* (1984), pp. 302–311.

[103] KIVINEN, J., AND WARMUTH, M. K. Exponentiated gradient versus gradient descent for linear predictors. *Inf. Comput. 132* (1997), 1–63.

[104] KNEE, G. C., BOLDUC, E., LEACH, J., AND GAUGER, E. M. Maximum-likelihood quantum process tomography via projected gradient descent. arXiv:1803.10062v1.

[105] KOLTCHINSKII, V. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer-Verl., Berlin, 2011.

[106] KOLTCHINSKII, V. von Neumann entropy penalization and low-rank matrix estimation. *Ann. Stat. 39*, 6 (2011), 2936–2973.

[107] KOLTCHINSKII, V. A remark on low rank matrix recovery and noncommutative Bernstein type inequalities. In *From Probability to Statistics and Back: High-Dimensional Models and Processes*, M. Banerjee, F. Bunea, J. Huang, V. Koltchinskii, and M. H. Maathuis, Eds. Inst. Math. Stat., 2013, pp. 213–226.

[108] KOLTCHINSKII, V., AND MENDELSON, S. Bounding the smallest singular value of a random matrix without concentration. *Int. Math. Res. Not.* (2015).

[109] KONTOYIANNIS, I., AND MADIMAN, M. Measure concentration for compound Poisson distributions. *Electron. Commun. Probab. 11* (2006), 45–57.

[110] KULIS, B., SUSTIK, M. A., AND DHILLON, I. S. Low-rank kernel learning with Bregman matrix divergences. *J. Mach. Learn. Res. 10* (2009), 341–376.

[111] LACOSTE-JULIEN, S., AND JAGGI, M. On the global linear convergence of Frank-Wolfe optimization variants. In *Adv. Neural Information Processing Systems 28* (2015).

[112] LAM, C., AND FAN, J. Sparsistency and rates of convergence in large covariance matrix estimation. *Ann. Stat. 37* (2009), 4254–4278.

[113] LAURITZEN, S. L. *Graphical models*. Clarendon Press, Oxford, 1996.

[114] LEE, J. D., AND SUN, Y. On model selection consistency of regularized M-estimators. *Electron. J. Stat. 9* (2015), 608–642.

[115] LEE, Y. T., AND SIDFORD, A. Path finding I: Solving linear programs with $\tilde{O}(\sqrt{rank})$ linear system solves. In *IEEE 55th Annu. Symp. Foundations of Computer Science* (2015).

[116] LI, B., SAHOO, D., AND HOI, S. C. H. OLPS: A toolbox for on-line portfolio selection. *J. Mach. Learn. Res. 17* (2016), 1–5.

[117] LI, Y.-H., AND CEVHER, V. Consistency of ℓ_1-regularized maximum-likelihood for compressive Poisson regression. In *IEEE Int. Conf. Acoustics, Speech and Signal Processing* (2015), pp. 3606–3610.

[118] LI, Y.-H., AND CEVHER, V. Learning data triage: Linear decoding works for compressive MRI. In *IEEE Int. Conf. Acoustics, Speech and Signal Processing* (2016), pp. 4034–4038.

[119] LI, Y.-H., AND CEVHER, V. Convergence of the exponentiated gradient method with Armijo line search. arXiv:1712.08480.

[120] LI, Y.-H., HSIEH, Y.-P., ZERBIB, N., AND CEVHER, V. A geometric view on constrained M-estimators. arXiv:1506.08163.

[121] LI, Y.-H., RIOFRÍO, C. A., AND CEVHER, V. A general convergence result for mirror descent with Armijo line search. arXiv:1805.12232.

[122] LI, Y.-H., SCARLETT, J., RAVIKUMAR, P., AND CEVHER, V. Sparsistency of ℓ_1-regularized M-estimators. In *Proc. 18th Inf. Conf. Artificial Intelligence and Statistics* (2015), pp. 644–652.

[123] LU, H., FREUND, R. M., AND NESTEROV, Y. Relatively smooth convex optimization by first-order methods, and applications. *SIAM J. Optim. 28*, 1 (2018), 333–354.

[124] LUSTIG, M., DONOHO, D., AND PAULY, J. M. Sparse MRI: The application of compressed sensing for rapid MR imaging. *Magn. Reson. Med. 58* (2007), 1182–1195.

[125] MALLAT, S. *A Wavelet Tour of Signal Processing: The Sparse Way*, third ed. Academic Press, Burlington, MA, 2009.

[126] MASSART, P. *Concentration Inequalities and Model Selection*. Springer-Verl., Berlin, 2007.

[127] MCCULLAGH, P., AND NELDER, J. A. *Generalized Linear Models*, second ed. Chapman and Hall, London, 1989.

[128] MEINSHAUSEN, N., AND BÜHLMANN, P. High-dimensional graphs and variable selection with the Lasso. *Ann. Stat. 34*, 3 (2006), 1436–1462.

[129] MENDELSON, S. Learning without concentration. *J. ACM 62*, 3 (2015).

[130] MENDELSON, S., PAJOR, A., AND TOMCZAK-JAEGERMANN, N. Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal. 17* (2007), 1248–1282.

[131] MERHAV, N., AND FEDER, M. Universal prediction. *IEEE Trans. Inf. Theory 44*, 6 (1998), 2124–2147.

[132] MOHRI, M., ROSTAMIZADEH, A., AND TALWALKAR, A. *Foundations of Machine Learning*. MIT Press, Cambridge, MA, 2012.

[133] NATARAJAN, B. K. Sparse approximate solutions to linear systems. *SIAM J. Comput. 24*, 2 (1995), 227–234.

[134] NEGAHBAN, S. N., RAVIKUMAR, P., WAINWRIGHT, M. J., AND YU, B. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Stat. Sci. 27*, 4 (2012), 538–557.

[135] NEMIROVSKY, A. S., AND YUDIN, D. B. *Problem Complexity and Method Efficiency in Optimization*. John Wiley & Sons, Chichester, 1983.

[136] NESTEROV, Y. *Introductory Lectures on Convex Optimization*. Kluwer, Boston, MA, 2004.

[137] NESTEROV, Y. Gradient methods for minimizing composite functions. *Math. Program., Ser. B 140* (2013), 125–161.

[138] NESTEROV, Y. Complexity bounds for primal-dual methods minimizing the model of objective function. *Math. Program., Ser. A* (2017).

[139] NESTEROV, Y., AND NEMIROVSKI, A. On first-order algorithms for ℓ_1/nuclear norm minimization. *Acta Numer.* (2013), 509–575.

[140] NESTEROV, Y., AND NEMIROVSKII, A. *Interior-Point Polynomial Algorithms in Convex Programming*. SIAM, Philadelphia, PA, 1994.

[141] NETRAPALLI, P., JAIN, P., AND SANGHAVI, S. Phase retrieval using alternating minimization. In *Adv. Neural Information Processing Systems 26* (2013).

[142] OCHS, P., FADILI, J., AND BROX, T. Non-smooth non-convex Bregman minimization: Unification and new algorithms. arXiv:1707.02278v3.

[143] ODOR, G., LI, Y.-H., YURTSEVER, A., HSIEH, Y.-P., EL HALABI, M., TRAN-DINH, Q., AND CEVHER, V. Frank-Wolfe works for non-Lipschitz continuous gradient objectives: Scalable Poisson phase retrieval. In *IEEE Int. Conf. Acoustics, Speech and Signal Processing* (2016), pp. 6230–6234.

[144] OHYA, M., AND PETZ, D. *Quantum Entropy and Its Use*. Springer, Berlin, 1993.

[145] OYMAK, S., AND TROPP, J. A. Universality laws for randomized dimension reduction, with applications. *Inf. Inference* (2017).

[146] PARIS, M., AND ŘEHÁČEK, J., Eds. *Quantum State Estimation*. Springer, Berlin, 2004.

[147] PLAN, Y., AND VERSHYNIN, R. The generalized Lasso with non-linear observations. *IEEE Trans. Inf. Theory 62*, 3 (2016), 1528–1537.

[148] PLAN, Y., VERSHYNIN, R., AND YUDOVINA, E. High-dimensional estimation with geometric constraints. *Inf. Inference* (2017).

[149] POLYAK, B. T. *Introduction to Optimization*. Optimization Softw., Inc., New York, NY, 1987.

[150] RASKUTTI, G., WAINWRIGHT, M. J., AND YU, B. Restricted eigenvalue properties for correlated Gaussian designs. *J. Mach. Learn. Res. 11* (2010), 2241–2259.

[151] RASKUTTI, G., WAINWRIGHT, M. J., AND YU, B. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Trans. Inf. Theory 57*, 10 (Oct. 2011), 6976–6994.

[152] RAVIKUMAR, P., WAINWRIGHT, M. J., AND LAFFERTY, J. D. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Ann. Stat. 38*, 3 (2010), 1287–1319.

[153] RAVIKUMAR, P., WAINWRIGHT, M. J., RASKUTTI, G., AND YU, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. *Electron. J. Stat. 5* (2011), 935–980.

[154] RECHT, B., FAZEL, M., AND PARRILO, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Rev. 52*, 3 (2010), 471–501.

[155] RIOFRÍO, C. A., GROSS, D., FLAMMIA, S. T., MONZ, T., NIGG, D., BLATT, R., AND EISERT, J. Experimental quantum compressed sensing for a seven-qubit system. *Nature Commun.* (2017).

[156] ROCKAFELLAR, R. T. *Convex Analysis*. Princeton Univ. Press, Princeton, NJ, 1970.

[157] ROCKAFELLAR, R. T. Monotone operators and the proximal point algorithm. *SIAM J. Control Optim. 14*, 5 (1976), 877–898.

[158] ROCKAFELLAR, R. T., AND WETS, R. J. *Variational Analysis*. Springer, Berlin, 2009.

[159] ROTHMAN, A. J., BICKEL, P. J., LEVINA, E., AND ZHU, J. Sparse permutation invariant covariance estimation. *Electron. J. Stat. 2* (2008), 494–515.

[160] RUDELSON, M., AND VERSHYNIN, R. On sparse reconstruction from Fourier and Gaussian measurements. *Commun. Pure Appl. Math. LXI* (2008), 1025–1045.

[161] SALZO, S. The variable metric forward-backward splitting algorithms under mild differentiability assumptions. *SIAM J. Optim. 27*, 4 (2017), 2153–2181.

[162] SHAFER, G., AND VOVK, V. *Probability and Finance: It's Only a Game!* John Wiley & Sons, New York, NY, 2001.

[163] SHALEV-SHWARTZ, S., AND BEN-DAVID, S. *Understanding Machine Learning*. Cambridge Univ. Press, Cambridge, UK, 2014.

[164] SHANG, J., ZHANG, Z., AND NG, H. K. Superfast maximum likelihood reconstruction for quantum tomography. *Phys. Rev. A 95* (2017).

[165] SHECHTMAN, Y., ELDAR, Y. C., COHEN, O., CHAPMAN, H. N., MIAO, J., AND SEGEV, M. Phase retrieval with application to optical imaging. *IEEE Signal Process. Mag. 32*, 3 (2015), 87–109.

[166] SIVAKUMAR, V., BANERJEE, A., AND RAVIKUMAR, P. Beyond sub-Gaussian measurements: High-dimensional estimation with sub-exponential designs. In *Adv. Neural Information Processing Systems 28* (2015).

[167] SLAWSKI, M., LI, P., AND HEIN, M. Regularization-free estimation in trace regression with symmetric positive semidefinite matrices. In *Adv. Neural Information Processing Systems 28 (NIPS 2015)* (2015).

[168] SUN, T., AND TRAN-DINH, Q. Generalized self-concordant functions: a recipe for Newton-type methods. *Math. Program., Ser. A* (2018).

[169] TALAGRAND, M. *Upper and Lower Bounds for Stochastic Processes*. Springer, Berlin, 2014.

[170] TIBSHIRANI, R. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc., Ser. B 58*, 1 (1996), 267–288.

[171] TOŠIĆ, I., AND FROSSARD, P. Dictionary learning. *IEEE Signal Process. Mag. 28*, 2 (2011), 27–38.

[172] TRAN-DINH, Q., KYRILLIDIS, A., AND CEVHER, V. Composite self-concordant minimization. *J. Mach. Learn. Res. 16* (2015), 371–416.

[173] TRAN-DINH, Q., LI, Y.-H., AND CEVHER, V. Composite convex minimization involving self-concordant-like cost functions. In *Model. Comput. & Optim. in Inf. Syst. & Manage. Sci.* (Cham, 2015), Springer, pp. 155–168.

[174] TSUDA, K., RÄTSCH, G., AND WARMUTH, M. K. Matrix exponentiated gradient updates for on-line learning and Bregman projection. *J. Mach. Learn. Res. 6* (2005), 995–1018.

[175] TSYBAKOV, A. *Introduction to Nonparametric Estimation*. Springer, New York, 2009.

[176] VALIANT, L. G. A theory of the learnable. *Commun. ACM 27*, 11 (November 1984), 1134–1142.

[177] VAN DE GEER, S. The deterministic Lasso. Research report no. 140, Seminar für Statistik, Eidgenössische Technische Hochschule, 2007.

[178] VAN ERVEN, T., GRÜNWALD, P. D., MEHTA, N. A., REID, M. D., AND WILLIAMSON, R. C. Fast rates in statistical and online learning. *J. Mach. Learn. Res. 16* (2015), 1793–1861.

[179] VAPNIK, V. N. *The Nature of Statistical Learning Theory*. Springer-Verl., New York, NY, 2000.

[180] VARDI, Y., SHEPP, L. A., AND KAUFMAN, L. A statistical model for positron emission tomography. *J. Am. Stat. Assoc. 80*, 389 (1985), 8–20.

[181] VASANAWALA, S., MURPHY, M., ALLEY, M., LAI, P., KEUTZER, K., PAULY, J., AND LUSTIG, M. Practical parallel imaging compressed sensing MRI: Summary of two years of experience in accelerating body MRI of pediatric patients. In *IEEE Int. Symp. Biomedical Imaging* (2011), pp. 1039–1043.

[182] VERSHYNIN, R. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing: Theory and Applications*, Y. Eldar and G. Kutyniok, Eds. Cambridge Univ. Press, 2012, ch. 5, pp. 210–268.

[183] VERSHYNIN, R. Estimation in high dimensions: A geometric perspective. In *Sampling Theory, a Renaissance*, G. E. Pfander, Ed. Birkhäuser, Cham, 2015, ch. 1, pp. 3–66.

[184] ŘEHÁČEK, J., HRADIL, Z., KNILL, E., AND LVOVSKY, A. I. Diluted maximum-likelihood algorithm for quantum tomography. *Phys. Rev. A 75* (2007).

[185] WAINWRIGHT, M. J. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). *IEEE Trans. Inf. Theory 55*, 5 (2009), 2183–2202.

[186] YANG, J., ZHANG, Y., AND YIN, W. A fast alternating direction method for TVL1-L2 signal reconstruction from partial Fourier data. *IEEE J. Sel. Topics Signal Process. 4*, 2 (2010), 288–297.

[187] YURTSEVER, A., TRAN-DINH, Q., AND CEVHER, V. A universal primal-dual convex optimization framework. In *Adv. Neural Information Processing Systems 28* (2015).

[188] ZEIDLER, E. *Applied Functional Analysis: Main Principles and Their Applications*. Springer-Verl., New York, NY, 1995.

[189] ZERBIB, N., LI, Y.-H., HSIEH, Y.-P., AND CEVHER, V. Estimation error of the constrained lasso. In *54th Annu. Allerton Conf. Communication, Control, and computing* (2016).

[190] ZHANG, T. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Stat. 32*, 1 (2004), 56–134.

[191] ZHAO, P., AND YU, B. On model selection consistency of Lasso. *J. Mach. Learn. Res. 7* (2006), 2541–2563.