Mirror descent for high-dimensional statistical models



Fan Wu

St Peter's College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Michaelmas 2021

Acknowledgements

A number of wonderful people have made this thesis possible and the journey more enjoyable than it would otherwise have been. First and foremost, I would like to express my sincere gratitude to my supervisor Professor Patrick Rebeschini for his constant guidance, advice and encouragement. He has shown me how to conduct research and guided me through my doctoral studies with his wealth of knowledge and invaluable expertise. His guidance was especially important in times when I was unsure about my direction, and I feel priviliged to have him as my supervisor. I would also like to thank my funding body EPSRC & MRC for providing me with this great opportunity.

My time in Oxford has been made so much more enjoyable by many people, and I would like to give my thanks to Suzie Brown, Francesca Crucinio, James Hodgson, Leon Law, Jingyue Lu, Cian Naik, Marco Palma, Francesca Panero, Dominic Richards, Jean François Ton, Tomas Vaškevičius, Jin Xu, Xuewen Yu, Chao Zhang, Zihao Zhang, and everyone else who has been with me. A special thanks goes to my partner Sihan Zhou for her constant support and encouragement through all the ups and downs of the journey. Last but by no means least, I would like to thank my family, especially my parents, who have fostered my interest in maths, and my sister, who has always been there for me with advice when I needed it. They have always supported me in my education and goals in life, and none of this would have been possible without their unconditional love and support.

Abstract

As vast amounts of data are being produced and processed in various fields of science and engineering, optimizing high-dimensional statistical models has become a ubiquitous task in many modern applications. Often, it is crucial to exploit underlying low-dimensional structures to extract useful information from high-dimensional data and models. To this end, numerous optimization algorithms such as gradient descent and its variants have been extensively studied both empirically and theoretically, with a large part of the literature focusing on algorithms based on the Euclidean geometry.

In this thesis, we study mirror descent, an iterative first-order algorithm that can be adapted to possibly non-Euclidean geometries by choosing a suitable mirror map, applied to two canonical models in high-dimensional statistics with underlying low-dimensional structures, namely sparse phase retrieval and matrix sensing. In these two problems, we demonstrate that mirror descent, with different choices for the mirror map, can incorporate the low-dimensional structures of sparsity and low-rankness, respectively.

In sparse phase retrieval, we begin by proposing *Hadamard Wirtinger flow* (HWF), an algorithm that minimizes the *unregularized* empirical risk and can be obtained as a first-order approximation to mirror descent equipped with the hypentropy mirror map. In contrast to existing algorithms for sparse phase retrieval, HWF does not require any explicit regularization in the form of added regularization terms or thresholding steps to enforce sparsity of its estimates and is empirically seen to *adapt* to the sparsity level of the underlying signal. On the theoretical side, we adopt the mirror descent framework and consider the more general setting of *noisy* sparse phase retrieval, where early stopping becomes necessary to prevent the estimates of mirror descent from deteriorating after a certain number of iterations.

We show that early-stopped mirror descent achieves a nearly minimax-optimal rate of convergence in noisy sparse phase retrieval, which shows that, instead of explicit regularization in the form of added regularization terms or thresholding steps, the geometry defined by the hypentropy mirror map and early stopping can be utilized to reconstruct sparse signal vectors from noisy phaseless measurements.

In matrix sensing, we study mirror descent – with different choices for the mirror map when optimizing over general rectangular and positive semidefinite matrices – applied to the unregularized empirical risk. We characterize the limiting point of mirror descent via a quantity explicitly related to the nuclear norm, Frobenius norm, and the von Neumann entropy. Assuming either a restricted isometry property or the setting of matrix completion, we obtain recovery guarantees for mirror descent that improve upon existing theoretical results for implicit regularization-based algorithms, which do not rely on minimizing or regularizing the nuclear norm nor on a low-rank factorized parametrization to enforce low-rankness of their estimates, in terms of the sample complexity.

Contents

1 Introduction **1**

1.1 Background and literature review 4

1.1.1 Mirror descent 4

1.1.2 Sparse phase retrieval 7

1.1.3 Matrix sensing 9

1.2 Contributions 12

1.3 Thesis outline 17

2 Hadamard Wirtinger flow for sparse phase retrieval **18**

2.1 Introduction 18

2.1.1 Our contributions 20

2.2 Sparse phase retrieval 21

2.3 Hadamard Wirtinger flow 24

2.4 Support recovery 25

2.5 Parameter estimation 28

2.6 Conclusion 34

2.7 Chapter appendix 35

2.7.1 Understanding the dynamics of Hadamard Wirtinger flow 35

2.7.2 Proof of Lemma 2.1 38

3 Nearly minimax-optimal rates for noisy sparse phase retrieval via early-stopped mirror descent **45**

3.1 Introduction 46

3.2 Sparse phase retrieval 50

3.3 Mirror descent algorithms 52

3.4 Main results . . . 54
3.4.1 Proof ideas . . . 58
3.5 Numerical study . . . 62
3.5.1 Estimation error . . . 63
3.5.2 Warm-up time T_1 . . . 65
3.5.3 Convergence behavior of mirror descent . . . 65
3.6 Proofs . . . 66
3.6.1 Proof of Lemma 3.2 . . . 66
3.6.2 Proof of Theorem 3.3 . . . 67
3.7 Conclusion . . . 77
3.8 Chapter appendix . . . 78
3.8.1 Proof of Theorem 3.4 . . . 78
3.8.2 Proof of supporting lemmas . . . 85
3.8.3 Technical lemmas . . . 116

4 Implicit regularization in matrix sensing via mirror descent 137
4.1 Introduction . . . 137
4.1.1 Our contributions . . . 139
4.2 Background . . . 141
4.3 Algorithmic regularization of mirror descent . . . 142
4.4 The estimation problem . . . 146
4.4.1 Matrix sensing with restricted isometry property . . . 146
4.4.2 Matrix completion . . . 148
4.5 Connection with gradient descent . . . 150
4.6 Numerical simulations . . . 151
4.7 Conclusion . . . 152
4.8 Chapter appendix . . . 154
4.8.1 Mirror maps and gradients . . . 154
4.8.2 Proof of Theorem 4.1 . . . 155
4.8.3 Proof of Theorem 4.2 . . . 159
4.8.4 Proof of Theorem 4.3 . . . 162
4.8.5 Proof of Theorem 4.4 . . . 164

4.8.6 Proof of Proposition 4.5 167
4.8.7 Further claims 169
4.8.8 Additional experiments for matrix completion 170

5 Conclusion **173**

Bibliography **177**

List of Figures

2.1 Proportion of support variables $|\hat{\mathcal{S}} \cap \mathcal{S}|/|\mathcal{S}|$ correctly recovered plus/minus one standard deviation (vertical lines) by our method (red curve), SPARTA (blue curve) and SparseAltMinPhase (black curve), for different levels of $x^\star_{max}$. From left to right: (i) $x^\star_{max} = \frac{1}{\sqrt{k}}$, (ii) $x^\star_{max} = k^{-0.25}$, (iii) $x^\star_{max} = 0.7$, (iv) Gaussian signal $\mathbf{x}^\star$ (without restrictions on $x^\star_{max}$). 28

2.2 Empirical success rate for $n = 1000$ fixed against number of measurements m with sparsity level $k = 20$ fixed (left) and against sparsity level k with $m = 500$ measurements (right). 32

2.3 Empirical success rate against sparsity level k with $n = 1000, m = 500$ fixed. From left to right: (i) $x^\star_{max} = 1/\sqrt{k}$, (ii) $x^\star_{max} = k^{-0.25}$ and (iii) $x^\star_{max} = 0.7$. 32

2.4 Empirical success rate (red: high, blue: low) of HWF against sparsity level k and number of observations m, with $n = 1000$ fixed. Black line: $m = \frac{1}{3}k(x^\star_{max})^{-2} \log \frac{n}{k}$. 33

2.5 Empirical success rate for $n = 1000$ fixed against number of measurements m with sparsity level $k = 20$ fixed (left), and against sparsity level k with $m = 500$ measurements (right), for varying number of restarts $\bar{b} \in [1, 100]$. 33

2.6 Relative error (log-scale) of HWF for real/complex signals with $n = 1000$, $m = 500$ and $k = 10$ for 10,000 iterations (left) and zoom-in to 500 iterations (right). 34

2.7 Relative error (log-scale) of HWF with random initialization (blue) and our proposed initialization (2.4) (red) with $n = 1000$, $m = 700$ and $k = 10$. 35

3.1 Relative error $\mathrm{dist}(\mathbf{x}^\star, \mathbf{X}^{T_{stop}})/\|\mathbf{x}^\star\|_2$ plus/minus one standard deviation (shaded area) of early-stopped mirror descent, with the stopping time T_{stop} selected with oracle knowledge in red (circles), and using the hold-out method in blue (crosses). **Left:** relative error against noise-to-signal ratio $\sigma/\|\mathbf{x}^\star\|_2^2$, with $n = m = 2000$, $k = 10$. **Center:** log-log plot of relative error against sample size m, with $n = 2000$, $k = 10$, $\sigma/\|\mathbf{x}^\star\|_2^2 = 0.1$. **Right:** log-log plot of relative error against sparsity level k, with $n = 2000$, $m = 4000$, $\sigma/\|\mathbf{x}^\star\|_2^2 = 0.1$. 64

3.2 Length of warm-up period T_1 plus/minus one standard deviation (shaded area) of early-stopped mirror descent. **Left**: T_1 against $\log(\|\mathbf{x}^\star\|_2/\beta)$, with $n = 2000$, $m = 1500$, $k = 10$, $\sigma/\|\mathbf{x}^\star\|_2^2 = 0.1$. **Right**: T_1 against sparsity level k, with $n = 2000$, $m = 4000$, $\sigma/\|\mathbf{x}^\star\|_2^2 = 0.1$, $\beta/\|\mathbf{x}^\star\|_2 = 10^{-20}$. 65

3.3 Relative error $\mathrm{dist}(\mathbf{x}^\star, \mathbf{X}^t)/\|\mathbf{x}^\star\|_2$ (log-scale) of mirror descent for $n = 50000$, $m = 1000$, $k = 10$, and different values of β. **Left**: noiseless case with $\sigma/\|\mathbf{x}^\star\|_2^2 = 0$. **Right**: noisy case with $\sigma/\|\mathbf{x}^\star\|_2^2 = 0.5$. 66

4.1 Nuclear norm, effective rank [99] and reconstruction error against initialization size α for $n = 50$ and $r = 5$. Top row: $m = 3nr$ measurements. Bottom row: $m = nr$ measurements. 152

4.2 Nuclear norm, effective rank [99] and reconstruction error in matrix completion against initialization size α for $n = 50$ and $r = 5$. Top row: $m = 3nr$. Bottom row: $m = nr$. 170

Chapter 1

Introduction

Recent advances in technology have allowed the collection and processing of unprecedented amounts of data in many fields of science and engineering. In numerous modern applications including collaborative filtering [69], financial time series forecasting [111], genomic sequencing [35], image classification [78], optical imaging [85], speech recognition [52], and many more, using large datasets to conduct meaningful inference is often an essential task. This has resulted in an ever-increasing need for analyzing and understanding high-dimensional datasets as well as statistical models, which comprise a set of statistical assumptions that postulate a data generating process which connects a set of unknown parameters of interest with the collected data.

While massive datasets are being collected and high-dimensional statistical models are being deployed, they often contain irrelevant or redundant information. For instance, in image classification, higher resolution images lead to vector representations of higher dimensions, while the added pixels may not contain any additional information for tasks such as classifying handwritten digits. In collaborative filtering, where there can be a huge number of users and items, the user-item rating matrix often admits a low-rank representation, which is what allows the prediction of user-specific preferences from an incomplete set of user ratings. To summarize, while the collected data and statistical model may be high-dimensional, the relevant information can often be encoded via some underlying low-dimensional structure. The study of low-dimensional structures in high-dimensional statistical models is an active area of research, with an upcoming book [127] summarizing some of the recent developments.

To obtain a practicable estimate of the parameters or features of interest, the default approach is to formulate an optimization problem, where the goal is to minimize (or maximize) some objective or loss function. The growing need to optimize these high-dimensional models has spurred a flurry of research studying iterative first-order algorithms which only access the objective function via its gradient [12, 17]. Within this class of algorithms it is worth highlighting gradient descent, which minimizes a given objective function by iteratively moving in the direction of the negative gradient of the objective function, with the Euclidean distance of each step being controlled by a step size parameter that is to be chosen by the user. Gradient descent has proven to be a simple and computationally cheap, yet immensely successful algorithm for a variety of different optimization tasks, which has led to it serving as a basic building block for many other first-order algorithms [12, 17].

Another important family of algorithms within the class of iterative first-order methods is mirror descent [89], which differs from gradient descent and many of its variants in that it can also be adapted to non-Euclidean geometries by choosing a suitable strictly convex function, the so-called *mirror map*. In a nutshell, when taking a step in the direction of the negative gradient of the objective function, mirror descent can incorporate different notions of distance determined by the chosen mirror map, while gradient descent is tied to the Euclidean distance. In many settings in high-dimensional statistics, a non-Euclidean geometry has been shown to be better able to capture the specific structure of the problem at hand [117]. For instance, in convex settings with a simplex constraint, which occurs naturally in problems where the goal involves learning a probability distribution over a finite set, mirror descent equipped with the negative entropy as mirror map has been shown to improve upon gradient descent in terms of the dependence on the problem dimension [15].

When low-dimensional structures are present in the statistical model, they are typically addressed via some form of *explicit regularization* in a large part of the literature, which focuses the search to the region of the parameter space corresponding to the assumed low-dimensional structure [117]. For instance, augmenting the objective function with a norm-based penalty term and including projection steps in the optimization algorithm are two popular strategies that have led to plentiful theoretical developments across a

wide variety of tasks including sparse recovery [2, 108], sparse phase retrieval [20, 91], and matrix sensing [97, 98], just to name a few.

A more recent trend in the literature focuses on the study of *implicit regularization*, which is motivated by the observation that simple optimization algorithms often find solutions with desirable properties in high-dimensional models that do not have a unique minimizer, where it is a-priori unclear which minimizer an optimization algorithm will, if at all, converge to. Instead of using some form of explicit regularization to encourage the algorithm to stay within a certain region of the parameter space, a similar effect can often by achieved by a proper choice of the optimization algorithm and initialization [54, 80, 113, 126]. A widely used algorithmic principle that leverages this effect is *early stopping*, which has traditionally been studied in the context of iterative methods for ridge regression and kernel methods [95, 124, 134, 139], but more recently also for mirror descent [114], for instance.

In this thesis, we study mirror descent applied to the problems of sparse phase retrieval and matrix sensing. In sparse phase retrieval, the goal is to recover an unknown sparse signal vector from a collection of linear magnitude-only measurements, while the goal in matrix sensing is to recover an unknown low-rank matrix from a collection of linear measurements. In both problems, the assumptions of sparsity and low-rankness, respecitvely, allow the number of measurements to be (much) smaller than the number of parameters to be estimated, which would, in general, lead to an ill-posed problem without aforementioned structural assumptions. We derive theoretical guarantees for mirror descent, which can exploit the non-Euclidean geometries defined by suitable mirror maps to incorporate the restrictions to sparse vectors and low-rank matrices, respectively, without the need to explicitly enforce either low-dimensional structure.

The remainder of the introduction is structured as follows. In Section 1.1, we provide a general overview of the mirror descent algorithm and the problems of sparse phase retrieval and matrix sensing. The main contributions of the three papers that make up this thesis are summarized in Section 1.2. Section 1.3 outlines the structure of the remainder of this thesis.

1.1 Background and literature review

We begin by providing a general overview of the relevant concepts, existing algorithms, and the related literature for mirror descent, sparse phase retrieval and matrix sensing. Throughout this chapter, we will use boldface uppercase letters to denote matrices and boldface lowercase letters to denote vectors. Note that we will use a slightly different notation in Chapters 2 and 3 for ease of presentation.

1.1.1 Mirror descent

Mirror descent is a first-order optimization algorithm that was first introduced by Nemirovski and Yudin [89] to solve large-scale convex optimization problems, and it is becoming increasingly popular in various optimization and machine learning settings. In the following, we introduce mirror descent in $\mathbb{R}^n$, i.e. in the vector case. All definitions extend naturally to the matrix case (see Chapter 4), and a more comprehensive overview can be found in [17].

Let $f : \mathcal{X} \subseteq \mathbb{R}^n \to \mathbb{R}$ be a differentiable (possibly non-convex) function to be minimized, where $\mathcal{X} \subseteq \mathbb{R}^n$ is a convex set. We first introduce the key object that allows mirror descent to adapt to non-Euclidean geometries, the *mirror map*.

Definition 1.1. *Let $\mathcal{D} \subseteq \mathbb{R}^n$ be a convex open set that includes $\mathcal{X}$ in its closure, i.e. $\mathcal{X} \subseteq \overline{\mathcal{D}}$, and $\mathcal{X} \cap \mathcal{D} \neq \emptyset$. We say that $\Phi : \mathcal{D} \to \mathbb{R}$ is a mirror map if it is strictly convex and differentiable, and its gradient takes all values, i.e. $\{\nabla\Phi(\mathbf{x}) : \mathbf{x} \in \mathcal{D}\} = \mathbb{R}^n$.*

Given a mirror map Φ, the associated *Bregman divergence* is a measure of distance between two points $\mathbf{x}, \mathbf{y} \in \mathcal{D}$ defined by

$$D_\Phi(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}) - \Phi(\mathbf{y}) - \nabla\Phi(\mathbf{y})^\top(\mathbf{x} - \mathbf{y}).$$

We refer to the Bregman divergence as a measure of distance since $D_\Phi(\mathbf{x}, \mathbf{y}) \geq 0$, with equality if and only if $\mathbf{x} = \mathbf{y}$. It can be easier to analyze the convergence of mirror descent in terms of the Bregman divergence rather than the Euclidean distance, see e.g. [15, 17, 88, 142]. However, the Bregman divergence is neither symmetric nor does it satisfy the triangle inequality, hence D_Φ is, in general, not a metric.

The focus of this thesis will be *unconstrained* mirror descent, i.e. $\mathcal{X} = \mathcal{D}$, in which case projections onto $\mathcal{X}$ are not needed (cf. [17]) and the mirror descent update is given by

$$\nabla\Phi(\mathbf{x}^{t+1}) = \nabla\Phi(\mathbf{x}^t) - \eta_t \nabla f(\mathbf{x}^t),$$

where $\eta_t > 0$ is a sequence of step sizes.

The following proximal formulation of the mirror descent update illustrates how the mirror map Φ relates to the geometry of mirror descent. The mirror descent update can equivalently be written (see e.g. [15])

$$\mathbf{x}^{t+1} = \arg\min_{\mathbf{x}\in\mathcal{D}} \left\{ \nabla f(\mathbf{x}^t)^\top (\mathbf{x} - \mathbf{x}^t) + \frac{1}{\eta_t} D_\Phi(\mathbf{x}, \mathbf{x}^t) \right\},$$

that is in each step, mirror descent aims to minimize the first-order approximation of f around the current estimate $\mathbf{x}^t$ while not moving too far away from $\mathbf{x}^t$, with distances being measured in terms of the Bregman divergence D_Φ. A special case is the mirror map $\Phi(\mathbf{x}) = \|\mathbf{x}\|_2^2$, in which case the associated Bregman divergence is the Euclidean distance and we recover standard gradient descent.

Initially introduced to solve large-scale convex optimization problems, mirror descent and many of its variants based on above basic formulation have been studied extensively in settings in convex optimization [15, 17, 27, 88, 104] and online decision making [1, 7, 8, 38, 82, 86, 92, 103, 110]. One example where the flexibility of mirror descent leads to improved performance guarantees compared to gradient descent is convex optimization constrained to the simplex $\Delta_n = \{\mathbf{x} \in \mathbb{R}^n : \sum_{i=1}^n x_i = 1,\ x_i \geq 0\}$. It has been shown in [15] that mirror descent, equipped with the negative entropy as mirror map, achieves the nearly dimension-free rate of convergence $\sqrt{\frac{\log n}{t}}$, while gradient descent only achieves a rate of $\sqrt{\frac{n}{t}}$ in general. Mirror descent with various choices for the mirror map has since been applied to problems ranging from classification and regression [38, 92, 104] over bandits [1, 7, 8] to reinforcement learning [82, 86, 110].

Given the massive size of the literature on mirror descent, a full overview is outside the scope of the present thesis, and we only focus on three aspects most relevant to our works, namely non-convex optimization, implicit regularization and reparametrization.

Non-convex optimization. While the settings of convex optimization and online decision making have received the most attention in the mirror descent literature over

the past decades, many applications of interest involve non-convex optimization tasks, and a growing body of literature studies mirror descent algorithms in non-convex optimization settings, e.g. [37, 46, 47, 70, 142]. For instance, the mirror descent framework has been applied to the problem of online principal component analysis in [70], which includes an online version of phase retrieval (without sparsity) as a special case. More generally, the convergence of mirror descent has been established in a non-convex optimization setting assuming a so-called *variational coherence* property [142], which corresponds to the assumption that the inner product $\langle \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}' \rangle$ is non-negative for any global minimizer $\mathbf{x}'$ of f and any vector $\mathbf{x} \in \mathcal{X}$.

Implicit regularization. The phenomenon of implicit regularization has attracted much recent attention, with the analysis of Euclidean optimization algorithms such as gradient descent and its variants making up a large part of the literature. Recently, there has been a growing interest in studying the implicit regularization properties of mirror descent. For instance, linear regression and classification models have been considered in [53] and the implicit bias of mirror descent has been characterized in terms of the Bregman divergence, while the convergence of a stochastic version of mirror descent has been studied in the context of overparametrized nonlinear models [9, 10]. The implicit regularization of early-stopped mirror descent has been analyzed from a learning theoretical perspective in [114], which, in particular, established a connection with sparsity in the problem of sparse recovery.

Reparametrization. Non-Euclidean optimization algorithms have been studied extensively also outside the mirror descent framework. In particular, exponentiated gradient algorithms and reparametrized gradient descent are closely related to mirror descent in that certain exponentiated gradient algorithms, gradient descent with different parametrizations, and mirror descent with corresponding mirror maps, have been shown to be equivalent [3, 4, 48]. Exponentiated gradient algorithms [68] replace the additive updates in gradient descent with multiplicative updates and have been successfully applied to problems including linear regression [68], binary classification [76], and online allocation problems [44]. Gradient descent with different parametrizations has been studied in problems such as sparse recovery [56, 114, 140], matrix sensing [54, 74, 112], and neural network training [87, 126]. Specifically, gradient descent with

Hadamard parametrization has been shown to implicitly regularize and achieve the minimax-optimal rate of convergence in sparse recovery [113], for instance.

1.1.2 Sparse phase retrieval

Phase retrieval is the problem of reconstructing a signal vector $\mathbf{x}^\star \in \mathbb{R}^n$ from a set of quadratic measurements[1]

$$y_j = (\mathbf{a}_j^\top \mathbf{x}^\star)^2, \qquad j = 1, \dots, m,$$

where we consider the standard setup of i.i.d. $\mathbf{a}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ measurement vectors. Such a task arises naturally in applications such as crystallography [85], diffraction imaging [18], and optics [120], where detectors are typically only able to measure intensities, but not phases. While it has been shown that an arbitrary signal $\mathbf{x}^\star \in \mathbb{R}^n$ can be reconstructed with high probability from $m \geq 2n - 1$ Gaussian measurements [13], many applications of interest involve signals that are naturally sparse [57], which can be exploited to enable reconstruction of the signal from potentially far fewer measurements. In particular, it has been shown that $m \geq 4k - 1$ Gaussian measurements suffice for a k-sparse signal $\mathbf{x}^\star \in \mathbb{R}^n$ to be uniquely determined (up to a global sign) with high probability [73]. However, it remains an open question whether there exists an efficient algorithm that can provably recover general k-sparse signals with a sample complexity that scales linearly in k.

Many methods have been developed for phase retrieval, most of which have also been adapted to incorporate the common sparsity assumption. Most existing algorithms fall into one of three categories, namely alternating minimization-based methods, convex relaxation-based algorithms and non-convex optimization-based approaches, with the latter being the approach also taken in this thesis in Chapters 2 and 3.

Alternating minimization. The seminal error reduction algorithm by Gerchberg and Saxton [45] introduces a vector for the unobserved phases and alternates between optimizing the estimates for the signal and the phases, and a similar approach is taken in [90, 118]. This idea has been extended to the case of sparse signals by estimating

[1] We focus on the real case for the sake of clarity of the analytical results in this thesis. Nonetheless, mirror descent also extends to the complex case. Further, we only introduce the noiselesss case in this section. Noisy sparse phase retrieval is discussed in Chapter 3.

the support of the signal and restricting the search to the estimated support [90], or by incorporating a hard sparsity constraint [60], for instance. In both cases, reconstruction of a general signal vector requires $m \geq \mathcal{O}(k^2 \log n)$ Gaussian measurements.

Convex relaxation. Algorithms such as PhaseLift [21, 25] and PhaseCut [119] formulate a semidefinite programming problem by lifting the optimization variable into the space of $n \times n$ matrices, while PhaseMax [51] follows a non-lifting approach and proposes a convex optimization problem that can be solved via its dual formulation. Extending these approaches to sparse phase retrieval has led to algorithms such as compressive phase retrieval via lifting (CPRL) [91] and SparsePhaseMax [55], both of which rely on an ℓ_1-penalty term to promote sparsity and require $\mathcal{O}(k^2 \log n)$ measurements for successful reconstruction of the signal. While a recently proposed convex formulation using a mixed atomic matrix norm regularizer is exact with a sample complexity linear in the sparsity level k, there is no known efficient algorithm that provably solves the proposed convex program [84].

Non-convex optimization. A more recent line of research initiated by the Wirtinger flow algorithm [22] takes a non-convex optimization-based approach and aims to reconstruct the signal $\mathbf{x}^\star$ by minimizing the non-convex empirical risk

$$f(\mathbf{x}) = \frac{1}{m} \sum_{j=1}^{m} \Big((\mathbf{a}_j^\top \mathbf{x})^2 - y_j) \Big)^2.$$

Wirtinger flow addresses the difficulties posed by a non-convex objective function using the following two-stage procedure: (i) a carefully designed spectral initialization is employed that produces an initial guess which is then (ii) refined via gradient descent updates. A similar two-stage approach is adopted by many algorithms such as truncated Wirtinger flow [28], which considers the Poisson log-likelihood instead of the squared loss and adaptively drops terms that are too influential, reshaped Wirtinger flow [137], which considers an amplitude-based loss function, and truncated amplitude flow [121], which employs an orthogonality promoting initialization followed by truncated generalized gradient iterations. More recently, it has been shown that a special initialization scheme is not necessary and a simple random initialization suffices to solve the non-convex optimization problem [29, 106].

Most of these algorithms have been adapted to the case of sparse signals, leading to algorithms such as thresholded Wirtinger flow (TWF) [20], sparse truncated amplitude flow (SPARTA) [122], compressive reweighted amplitude flow (CRAF) [138] and sparse Wirtinger flow (SWF) [135]. To reconstruct a sparse signal, these algorithms all follow above two-stage recipe and (i) obtain an initial estimate by first estimating the full support of the signal $\mathbf{x}^\star$ and then running either a spectral, orthogonality promoting, or diagonal thresholding initialization scheme on the estimated support, followed by (ii) a refinement stage where thresholding steps are included after each gradient update in order to ensure that the estimates remain sparse. Theoretical recovery guarantees have been established assuming a sample complexity of $m \geq \mathcal{O}(k^2 \log n)$.

Other methods developed for sparse phase retrieval include GESPAR [100], which is based on greedily updating an estimate of the support, a generalized approximate message passing algorithm that uses a sparsity-inducing prior [102] and a majorization-minimization algorihtm that includes an ℓ_1-penalty term [94]. More recently, deep learning-based approaches to (sparse) phase retrieval have also been studied [65, 77].

1.1.3 Matrix sensing

Matrix sensing represents a cornerstone in modern statistics, with one of its best known applications being, popularized by the Netflix Prize competition, collaborative filtering [69]. In matrix sensing, the goal is to recover a rank-r matrix $\mathbf{X}^\star \in \mathbb{R}^{n \times n'}$ from a set of linear measurements

$$y_j = \langle \mathbf{A}_j, \mathbf{X}^\star \rangle, \qquad j = 1, \ldots, m,$$

where $\langle \mathbf{A}, \mathbf{B} \rangle = \mathrm{tr}(\mathbf{A}^\top \mathbf{B})$ denotes the Frobenius inner product and the sensing matrices $\mathbf{A}_j$'s are known. In particular, this formulation also includes the problem of matrix completion, where a random subset of the entries of $\mathbf{X}^\star$ are observed and which corresponds to sensing matrices with exactly one entry equal to one and all other entries equal to zero. Without loss of generality, we will assume $n \leq n'$ in what follows.

Matrix sensing is a well-studied topic and many algorithms have been developed to tackle this problem. We only summarize three approaches that are most relevant to this thesis, namely convex methods based on minimizing (or regularizing) the nuclear norm,

non-convex optimization-based algorithms utilizing a low-rank factorized parametrization, and algorithms based on implicit regularization. A more comprehensive overview focusing on non-convex optimization-based approaches can be found in [33].

Nuclear norm. Since low-rankness of a matrix corresponds to sparsity of its singular values, a natural idea is to extend strategies that are known to promote sparsity in the vector case to matrix sensing. Mirroring the relationship between the ℓ_1-norm and sparsity in the vector case, the nuclear norm can be seen as a convex relaxation of the rank of a matrix, and minimizing the nuclear norm has been shown to recover a given low-rank matrix $\mathbf{X}^\star$ from $r(n+n')$ (modulo logarithmic terms) observations in matrix completion [24, 97] or under the assumption that the sensing matrices $\mathbf{A}_j$'s satisfy a restricted isometry property (RIP) [98]. Since minimizing the nuclear norm under linear constraints is a convex optimization problem, plenty of algorithms such as interior point methods [105] and projected subgradient descent [17] can be used to solve matrix sensing. Another approach consists of using the nuclear norm to regularize the empirical risk [109]. However, since interior point methods use second order information and the first-order methods require computing the subgradient of the nuclear norm (see e.g. [123]) at every iteration, convex methods are typically computationally more expensive and less scalable than their non-convex optimization-based counterparts. On the other hand, non-convex optimization-based algorithms often require a separate analysis for matrix completion and matrix sensing with RIP.

Low-rank factorization. Another popular approach to matrix sensing is based on the low-rank factorized parametrization $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$ with $\mathbf{U} \in \mathbb{R}^{n\times r}$, $\mathbf{V} \in \mathbb{R}^{n'\times r}$ (or $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$ with $\mathbf{U} \in \mathbb{R}^{n\times r}$ when $\mathbf{X}^\star \in \mathbb{R}^{n\times n}$ is positive semidefinite), which directly confines the search to low-rank matrices. Then, matrix sensing can be cast as the non-convex optimization problem of minimizing the empirical risk

$$f(\mathbf{U},\mathbf{V}) = \frac{1}{4m}\sum_{j=1}^{m}\Big(\langle \mathbf{A}_j, \mathbf{U}\mathbf{V}^\top\rangle - y_j\Big)^2.$$

Keshavan et al. [66] proposed a two-stage non-convex procedure consisting of a spectral initialization followed by a refinement stage using (regularized) gradient descent on the Grassmann manifold and showed that their algorithm provably solves matrix completion with $\kappa^6 r^2(n+n')$ (modulo logarithmic terms) observations, where κ is the

condition number of $\mathbf{X}^\star$. Such a two-stage approach (without the Grassmann manifold) is also taken by many later low-rank factorization-based algorithms. For instance, Procrustes flow [112], which uses a projected gradient initialization scheme followed by gradient descent on the regularized empirical risk that penalizes $\|\mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top\|_F^2$, has been shown to recover a low-rank matrix under the RIP assumption, while gradient descent type algorithms that use a spectral initialization and include regularization terms based on $\|\mathbf{U}\|_F$ and $\|\mathbf{V}\|_F$ have been shown to solve matrix completion [107]. Another strategy that has been used to ensure that the algorithm stays within a region of benign local geometry during the refinement stage is the inclusion of projection steps [32, 141]. Recently, it has been shown that additional regularization during the refinement stage is in fact not necessary in matrix completion [62, 80] and in matrix sensing with RIP [79]. However, the low-rank factorized parametrization itself explicitly enforces a low-rankness constraint on the estimates of the algorithms.

Implicit regularization. Gunasekar et al. [54] empirically observed that vanilla gradient descent on the unregularized empirical risk using a random initialization close to zero and the parametrization $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$ (assuming $\mathbf{X}^\star \in \mathbb{R}^{n\times n}$ is positive semidefinite) converges to a low nuclear norm matrix and can recover $\mathbf{X}^\star$, even when $\mathbf{U} \in \mathbb{R}^{n\times n}$ and the parametrization does not pose any restrictions on the rank. This observation has inspired a line of research investigating the phenomenon of implicit regularization in matrix sensing. Theoretical convergence guarantees have been established in the special case where the sensing matrices $\mathbf{A}_j$'s commute [54], or when the sensing matrices satisfy an RIP assumption that leads to a sample complexity of $\mathcal{O}(\kappa^6 n r^2 \log^5 n)$ [74]. A low-rank initialization within a "capture neighborhood" has been studied in [39], which ensures that the iterates of gradient descent stay low-rank. The implicit regularization of gradient descent in matrix sensing has also been studied through a deep matrix factorization framework [6, 50, 75, 96], where the evolution of singular values has been analyzed and theoretical and empirical evidence is provided that suggests that the implicit bias of gradient descent in matrix sensing cannot be fully explained by the nuclear norm, but rather involves a notion of rank-minimization.

1.2 Contributions

We briefly summarize the main contributions of the three works [130, 132, 131] included in this integrated thesis. The first two works [130, 132] consider the problem of sparse phase retrieval, while the third work [131] considers the problem of matrix sensing.

Sparse phase retrieval. In the first work [130], we propose *Hadamard Wirtinger flow* (HWF), a simple two-stage procedure that follows the non-convex optimization-based approach outlined above, for the problem of sparse phase retrieval. In the first stage, we estimate a *single* coordinate on the support of the signal $\mathbf{x}^\star$ for a simple initialization corresponding to an indicator vector, which is then followed by gradient descent updates with the Hadamard parametrization $\mathbf{x} = \mathbf{u} \odot \mathbf{u} - \mathbf{v} \odot \mathbf{v}$ on the unregularized empirical risk, where $\odot$ denotes the elementwise Hadamard product. Note that gradient descent with Hadamard parametrization can be obtained as a first-order approximation to mirror descent equipped with the hypentropy mirror map [48, 114, 128], and the two algorithms have been observed to be indistiguishable in practice [128].

As discussed previously in Section 1.1.2, existing approaches to sparse phase retrieval typically rely on some form of explicit regularization such as ℓ_1-penalty terms, thresholding steps, sparsity-inducing priors or restrictions to low-dimensional subspaces to enforce or encourage sparsity of their estimates. On the other hand, our numerical experiments suggest that HWF *adapts* to the sparsity level of the signal $\mathbf{x}^\star$ and successfully recovers sparse signals from fewer measurements than existing non-convex optimization-based algorithms such as SPARTA and SWF, without requiring any of the aforementioned strategies to enforce sparsity. In particular, running HWF does not require knowledge of the sparsity level k nor the tuning of any parameters controlling the sparsity level of the estimates.

A partial theoretical analysis reveals that a single step of HWF can be used as support recovery method, which, when combined with existing algorithms such as SPARTA, yields an end-to-end procedure for sparse phase retrieval that provably recovers a k-sparse signal with a sample complexity that depends on the largest signal component and matches the sample requirements of the best known theoretical guarantees of

$\mathcal{O}(k^2 \log n)$ in the worst case, while the sample complexity of our procedure improves to $\mathcal{O}(k \log n)$ if the signal contains at least one large component.

In light of the close connection between HWF and mirror descent equipped with the hypentropy mirror map, we adopt the mirror descent framework in the second work [132] and provide a complete convergence analysis for the mirror descent algorithm. We generalize the setting considered in the first work [130] and consider noisy sparse phase retrieval, where we allow the measurements to be corrupted by additive sub-exponential noise with bounded sub-exponential norm $\sigma \geq 0$. The addition of noise to the model leads to early stopping being necessary to avoid deterioration of the estimates after a certain number of iterations.

Assuming that the sample size is at least of order k^2 (modulo logarithmic terms) and that the minimum non-zero signal component is of order $\|\mathbf{x}^\star\|_2/\sqrt{k}$, we show that early-stopped mirror descent applied to noisy sparse phase retrieval achieves a rate of convergence of $\frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k \log n}{m}}$, which matches the minimax-optimal rate up to a logarithmic term. While convergence guarantees have been established for various algorithms, many theoretical results are limited to the noiseless case. The noisy model has been studied in [31] and stability results have been established for a convex relaxation-based approach, while in [60] a non-convex optimization-based algorithm has been shown to be robust to noise. However, the aforementioned works did not consider the optimal dependence of the estimation error rate on the sample size, the sparsity level and the noise level. A formulation based on empirical risk minimization with a hard sparsity constraint has been shown to achieve a nearly minimax-optimal rate of convergence [71]. However, this formulation does not lead to a tractable algorithm due to the sparsity constraint. TWF [20] achieves the same rate of convergence as mirror descent and was the first tractable algorithm that has been shown to achieve a nearly minimax-optimal rate of convergence in noisy sparse phase retrieval.

A comparison with some of the existing results in sparse phase retrieval is summarized in Table 1.1. We do not report the sample complexity, since all listed algorithms require a sample complexity of k^2 (modulo logarithmic terms). Our mirror descent approach incurs a higher computational cost than existing non-convex optimization-based approaches, which we attribute, at least in part, to its simplicity. We expect it

Table 1.1: Comparison with prior theory for sparse phase retrieval

algorithm	type of regularization	initialization	computational complexity	rate of convergence
SparseAlt-MinPhase [90]	subspace restriction	spectral initialization	$k^3 n \log n$	n/a
CPRL [91]	ℓ_1-penalty	arbitrary	n^3	n/a
SPARTA [122]	thresholding steps	orthogonality promoting	$k^2 n \log n$	n/a
TWF [20]	thresholding steps	diagonal thresholding	$k^2 n \log n$	$\frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k \log n}{m}}$
Mirror descent (our theory)	implicit	indicator vector	$k^3 n \log^3 n$	$\frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k \log n}{m}}$

to be possible to reduce the computational complexity of mirror descent by including additional algorithmic principles: for instance, replacing our indicator vector initialization with a spectral initialization scheme, which is common in existing non-convex optimization-based approaches, we expect it to be possible to reduce the computational complexity by a factor $\sqrt{k}$, see Chapters 3 and 5 for more details.

The main distinguishing features between our and existing results are the following: (i) mirror descent does not require explicit regularization in the form of ℓ_1-penalty terms, thresholding steps, sparsity-inducing priors or restrictions to low-dimensional subspaces, but rather leverages the geometry defined by the hypentropy mirror map along with early stopping to ensure sparsity of its estimates, and (ii) compared to existing non-convex optimization-based approaches, our mirror descent approach does not involve a special initialization scheme for a good initial estimate (more precisely, an initial estimate $\mathbf{x}^0$ that correlates with the signal, $\langle \mathbf{x}^0, \mathbf{x}^\star \rangle = \Omega(\|\mathbf{x}^\star\|_2^2)$), but uses a simple indicator vector as initial estimate (in general only $\langle \mathbf{x}^0, \mathbf{x}^\star \rangle = \Omega(\|\mathbf{x}^\star\|_2^2/\sqrt{k})$).

More generally, the analysis leading to above results can be of interest to the mirror descent literature beyond the problem of sparse phase retrieval. Key to the results in [132] is a potential-based convergence analysis of mirror descent in terms of the Bregman divergence. This approach is not new and has been frequently applied in convex optimization settings, e.g. [15, 17, 88], as well as in a non-convex optimization settings where a variational coherence assumption has been introduced to establish

convergence of mirror descent [142]. Sparse phase retrieval, however, poses a non-convex optimization problem that does not satisfy variational coherence (not even the weaker versions proposed in [142]), and our analysis is based on establishing a quantitative control on the inner product $\langle \nabla f(\mathbf{x}^t), \mathbf{x}^t - \mathbf{x}^\star \rangle$ only along the path of mirror descent, up to a prescribed stopping time. While the analysis in [142] makes an a-priori assumotion of variational coherence and can only be used to show convergence to a global minimizer, we establish an adaptive lower bound for this inner product in our analysis, which allows us to derive convergence speed guarantees for mirror descent and is crucial to investigating early stopping.

Matrix Sensing. Unlike in sparse phase retrieval, plenty nuclear norm-based algorithms are known to provably solve matrix sensing with a sample complexity that scales linearly in the rank r, which is a corresponding notion of sparsity for matrices. However, the literature on implicit regularization-based algorithms for matrix sensing is more recent and less well developed. Theoretical guarantees have been established for gradient descent with full-rank factorized parametrization under the assumption of RIP with a constant which leads to a sample complexity of $\mathcal{O}(\kappa^6 n r^2 \log^5 n)$.

In the third work [131], we analyze mirror descent applied to the unregularized empirical risk in the problem of matrix sensing, with different choices for the mirror map when optimizing over general rectangular and positive semidefinite matrices, respectively. Again applying a potential-based analysis in terms of the Bregman divergence, we show that mirror descent converges to a global minimizer of the empirical risk that can be characterized by a quantity explicitly related to the nuclear norm, the Frobenius norm, and the von Neumann entropy. This characterization yields recovery guarantees for both matrix sensing with RIP and matrix completion with a sample complexity of $r(n + n')$ (modulo logarithmic terms), which matches the assumptions of nuclear norm-based approaches and improves upon the sample requirement of an existing implicit regularization-based algorithm [74].

A comparison with some of the existing results in matrix sensing is summarized in Table 1.2. The only existing recovery guarantee for an implicit regularization-based algorithm is limited to matrix sensing with RIP and positive semidefinite matrices, while our results also applies to matrix completion and general rectangular matrices.

Table 1.2: Comparison with prior theory for matrix sensing (logarithmic terms omitted)

algorithm	setting	type of regularization	initialization	sample complexity
Projected subgradient [97, 98]	completion and RIP	nuclear norm	arbitrary	$r(n+n')$
Procrustes flow [112]	RIP	low-rank factorization	projected gradient	$r(n+n')$
Projected gradient [141]	completion	low-rank factorization	spectral initialization	$\kappa^2 r^2(n+n')$
Gradient descent [74]	RIP, only psd $\mathbf{X}^\star$	implicit	close to zero	$\kappa^6 r^2 nr$
Mirror descent (our theory)	completion and RIP	implicit	close to zero	$r(n+n')$

Further, as in sparse phase retrieval, the empirical risk minimization-based algorithms typically require a special initialization scheme due to the non-convexity of the objective function, while a much simpler initialization suffices for mirror descent and gradient descent with full-rank factorized parametrization [74].

Note that we did not include the computational complexity since we consider both the settings of matrix sensing with RIP and matrix completion, which differ in that the computational cost of evaluating inner products of the form $\langle \mathbf{A}_j, \mathbf{X} \rangle$ is $\mathcal{O}(nn')$ in matrix sensing with RIP and $\mathcal{O}(1)$ in matrix completion. Especially when $r \ll \min\{n, n'\}$, we do not expect implicit regularization-based algorithms such as gradient descent with full-rank factorized parametrization and mirror descent, which optimize over $n \times n'$-matrices, to be able to match the computational efficiency of algorithms that use a low-rank factorized parametrization, which optimize over two smaller matrices of dimensions $n \times r$ and $n' \times r$, respectively.

Our analysis for mirror descent represents an alternative proof technique to the analysis in [53], where the limiting point of mirror descent has been characterized via KKT optimality conditions. Compared to [53], the main advantage of our potential-based approach is the fact that convergence of mirror descent does not need to be assumed a-priori, but follows as a by-product from the convexity of the empirical risk. Compared to our analysis for sparse phase retrieval [132], non-negativity of the inner product $\langle \nabla f(\mathbf{X}^t), \mathbf{X}^t - \mathbf{X}^\star \rangle$ is an immediate consequence of the convexity of the objective

function, which is what allows us to establish convergence guarantees for mirror descent assuming a sample complexity that scales only linearly in the rank r. However, since we do not establish a quantitative control on above inner product, our results do not include any convergence speed guarantees.

In the vector case, gradient descent with Hadamard parametrization can be obtained as a first-order approximation to mirror descent equipped with the hypentropy mirror map [48, 114, 128]. We show that this connection between reparametrized gradient descent and mirror descent extends to the matrix case in the special case where the sensing matrices $\mathbf{A}_j$'s are symmetric and commute, and present numerical experiments which suggest that, in some regimes, the estimates of mirror descent and gradient descent with full-rank factorized parametrization track each other closely, even when the sensing matrices $\mathbf{A}_j$'s do not commute. Hence, our analysis for mirror descent might present a partial explanation for the empirical success of gradient descent with full-rank factorized parametrization observed in [54].

1.3 Thesis outline

The remainder of this integrated thesis consists of three manuscripts, followed by conclusions and a discussion on possible future work. The following manuscripts are included in this thesis.

- **Chapter 2:** "Hadamard Wirtinger flow for sparse phase retrieval". **Fan Wu**, Patrick Rebeschini. In *International Conference on Artificial Intelligence and Statistics*, 2021 (oral presentation).
- **Chapter 3:** "Nearly minimax-optimal rates for noisy sparse phase retrieval via early-stopped mirror descent". **Fan Wu**, Patrick Rebeschini. *In submission* (a version containing preliminary results [128] appeared in *Advances in Neural Information Processing Systems*, 2020 as spotlight presentation).
- **Chapter 4:** "Implicit regularization in matrix sensing via mirror descent". **Fan Wu**, Patrick Rebeschini. In *Advances in Neural Information Processing Systems*, 2021.

Chapter 2

Hadamard Wirtinger flow for sparse phase retrieval

We consider the problem of reconstructing an n-dimensional k-sparse signal from a set of noiseless magnitude-only measurements. Formulating the problem as an unregularized empirical risk minimization task, we study the sample complexity performance of gradient descent with Hadamard parametrization, which we call Hadamard Wirtinger flow (HWF). Provided knowledge of the signal sparsity k, we prove that a single step of HWF is able to recover the support from $k(x^\star_{max})^{-2}$ (modulo logarithmic term) samples, where $x^\star_{max}$ is the largest component of the signal in magnitude. This support recovery procedure can be used to initialize existing reconstruction methods and yields algorithms with total runtime proportional to the cost of reading the data and improved sample complexity, which is linear in k when the signal contains at least one large component. We numerically investigate the performance of HWF at convergence and show that, while not requiring any explicit form of regularization nor knowledge of k, HWF adapts to the signal sparsity and reconstructs sparse signals with fewer measurements than existing gradient based methods.

2.1 Introduction

Phase retrieval, the problem of reconstructing a signal from the (squared) magnitude of its Fourier (or any linear) transform, arises in many fields of science and engineering. Such a task is naturally involved in applications such as crystallography [85] and diffraction imaging [18], where optical sensors are able to measure intensities, but not

phases of light waves. Due to the loss of phase information, the one-dimensional Fourier phase retrieval problem is ill-posed in general. Common approaches to overcome this ill-posedness include using prior information such as non-negativity, sparsity and the signal's magnitude [42, 57], or introducing redundancy into the measurements by oversampling random Gaussian measurements or coded diffraction patterns [22, 28].

In many applications, the underlying signal is naturally sparse [57]. A wide range of algorithms has been devised for phase retrieval with a sparse signal, including alternating minimization (SparseAltMinPhase) [90], non-convex optimization based approaches such as thresholded Wirtinger flow (TWF) [20], sparse truncated amplitude flow (SPARTA) [122], compressive reweighted amplitude flow (CRAF) [138] and sparse Wirtinger flow (SWF) [135], and convex relaxation based methods such as compressive phase retrieval via lifting (CPRL) [91] and SparsePhaseMax [55]. Other approaches to sparse phase retrieval include the greedy algorithm GESPAR [100], a generalized approximate message passing algorithm (PR-GAMP) [102] and majorization minimization algorithms [94].

A limitation of these algorithms is their sample complexity: the best known theoretical results require $\mathcal{O}(k^2 \log n)$ Gaussian measurements to guarantee successful reconstruction of a generic k-sparse signal $\mathbf{x}^\star \in \mathbb{R}^n$. On the other hand, it has been shown that reconstruction is possible from $\mathcal{O}(k \log n)$ phaseless measurements [40]; however, there is no known algorithm which provably achieves this in polynomial time. In fact, $\mathcal{O}(k^2 \log n)$ quadratic measurements are necessary for a certain class of convex relaxations, on which algorithms such as CPRL are based [73]. Other existing algorithms such as SPARTA and SparseAltMinPhase require a sample complexity of $\mathcal{O}(k^2 \log n)$ for the initial estimation of the support of the signal. With the knowledge of the support, these (and plenty other) algorithms require only $\mathcal{O}(k \log n)$ samples for the subsequent reconstruction of the signal. Hence, we identify the support recovery step as the bottleneck of the sample complexity of non-convex optimization based approaches to the sparse phase retrieval problem.

Additional structural assumptions have been considered to improve the sample complexity. It has been shown that a k-sparse signal $\mathbf{x}^\star$ can be reconstructed from $\mathcal{O}(k \log n)$ measurements if one is allowed to freely design the measurement vectors [58], or if the

signal $\mathbf{x}^\star$ is assumed to be block-sparse and the number of blocks containing non-zero entries is $\mathcal{O}(1)$ [59, 138]. However, exact knowledge of the additional structure as well as an algorithm designed to take advantage of it is necessary in both cases. Linear sample complexity has also been achieved assuming that the signal coefficients decay with power-law [60].

Another downside of the above algorithms is the fact that sparsity is enforced or promoted *explicitly*. For instance, CPRL and SparsePhaseMax augment the objective function with an ℓ_1-penalty term, which is known to promote sparsity. SWF and SPARTA include a thresholding step in their gradient updates, which projects the iterates onto the set of k-sparse vectors, and SparseAltMinPhase and GESPAR directly constrain the search to a k-dimensional subspace of $\mathbb{R}^n$, which needs to be carefully chosen and updated. In the case of CPRL and SparsePhaseMax, additional regularization parameters have to be tuned, while the thresholding step of SWF and SPARTA requires knowledge of the sparsity k.

2.1.1 Our contributions

In this work, we analyze gradient descent with Hadamard parametrization applied to the unregularized empirical risk for the problem of (noiseless) sparse phase retrieval and propose methods for support recovery and parameter estimation. The main contributions of this paper are stated below.

First, we propose a two-stage procedure for sparse phase retrieval, which we call *Hadamard Wirtinger flow* (HWF) (following the terminology used for Wirtinger flow [22], which considers gradient descent applied to the unregularized empirical risk under the natural parametrization to solve phase retrieval without the assumption on sparsity). In stage one, we estimate a *single* coordinate on the support to construct a simple initial estimate, without using a sophisticated initialization scheme typically required such as the spectral initialization used in WF and SWF or the orthogonality-promoting initialization used in SPARTA. For stage two, we consider the Hadamard parametrization, which has previously been applied to problems on sparse recovery [56, 113, 140] and matrix factorization [6, 54, 74], and apply gradient descent to the unregularized empirical risk under this parametrization.

Second, we prove that our proposed algorithm can be used to recover the support $\mathcal{S} = \{i : x_i^\star \neq 0\}$ with high probability by choosing the k largest components of the estimate obtained after one step of HWF, provided that $m \geq \mathcal{O}(\max\{k \log n, \log^3 n\}(x_{max}^\star)^{-2})$ and $x_{min}^\star = \Omega(1/\sqrt{k})$, where we write $x_{max}^\star := \max_i \frac{|x_i^\star|}{\|\mathbf{x}^\star\|_2}$ and $x_{min}^\star := \min_{i \in \mathcal{S}} \frac{|x_i^\star|}{\|\mathbf{x}^\star\|_2}$. Note that HWF does not require knowledge of the sparsity level k, while support recovery using one step of HWF does. With the knowledge of the support $\mathcal{S}$, plenty of algorithms provably recover the signal $\mathbf{x}^\star$ under linear sample complexity $\mathcal{O}(k \log n)$, see e.g. [22, 122]. Thus, provided knowledge of the sparsity level k, one step of HWF can be used as a support recovery tool and, combined with any of the aforementioned algorithms, it results in a procedure which provably recovers k-sparse signals from $\mathcal{O}(\max\{k \log n, \log^3 n\}(x_{max}^\star)^{-2})$ phaseless measurements.

If $x_{max}^\star = \Omega(1)$, then the sample complexity of this procedure reduces to $\mathcal{O}(k \log n)$, provided $k \geq \mathcal{O}(\log^2 n)$. Unlike previous results which leverage additional structural assumptions to achieve linear sample complexity, our procedure does not require knowledge of the value of $x_{max}^\star$, nor does it need to be modified in any way to accommodate for this additional structure; we run the exact same algorithm regardless of the value of $x_{max}^\star$.

Third, we present numerical experiments showing the low sample complexity of HWF. As a simple algorithm not requiring thresholding steps nor added regularization terms to promote sparsity, HWF is seen to *adapt* to the sparsity level of the underlying signal and to reconstruct signals from a similar number of Gaussian measurements as PR-GAMP, which has been empirically shown to achieve linear sample complexity in some regimes with Gaussian signals [102]. In particular, the numerical experiments suggest that the sample complexity required by HWF is lower than that of other gradient based methods such as SPARTA and SWF. Further, if the signal satisfies $x_{max}^\star = \Omega(1)$, the reconstruction performance of HWF is seen to be greatly improved, without any modifications to the algorithm being made.

2.2 Sparse phase retrieval

We denote vectors and matrices with boldface letters and real numbers with normal font, and, where appropriate, use uppercase letters for random and lowercase letters

for deterministic quantities. For vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ we write $\odot$ for the Hadamard product, $(\mathbf{u} \odot \mathbf{v})_i = u_i v_i$, and, for notational simplicity, $\mathbf{u}^2 = \mathbf{u} \odot \mathbf{u}$ for taking squares entry-wise. We use the common notation $[n] := \{1, ..., n\}$. Since it is impossible to distinguish $\mathbf{x}^\star$ from $-\mathbf{x}^\star$ using magnitude-only observations, we will often write $\mathbf{x}^\star$ for the solution set $\{\pm\mathbf{x}^\star\}$ and consider, for any $\mathbf{x} \in \mathbb{R}^n$, the distance $\text{dist}(\mathbf{x}, \mathbf{x}^\star) := \min\{\|\mathbf{x} - \mathbf{x}^\star\|_2, \|\mathbf{x} + \mathbf{x}^\star\|_2\}$. Further, we assume $\|\mathbf{x}^\star\|_2 = 1$ for notational simplicity; note that this assumption is not needed for our results.

The goal in phase retrieval is to reconstruct an unknown signal vector $\mathbf{x}^\star \in \mathbb{R}^n$ from a set of quadratic measurements $Y_j = (\mathbf{A}_j^T \mathbf{x}^\star)^2$, $j = 1, ..., m$, where we observe $\mathbf{A}_j \sim \mathcal{N}(0, \mathbf{I}_n)$ i.i.d.. For the sake of clarity, we focus on the real-valued model. Our proposed algorithm also works in the complex-valued Gaussian model, where $\mathbf{x}^\star \in \mathbb{C}^n$ and $\mathbf{A}_j \sim \mathcal{N}(0, \frac{1}{2}\mathbf{I}_n) + i\mathcal{N}(0, \frac{1}{2}\mathbf{I}_n)$.

Many methods have been devised to solve this problem. A popular class of algorithms performs alternating projections onto different constraint sets; these include the seminal error reduction algorithm proposed by [45] and alternating minimization (AltMinPhase) [90]. Another more recent approach is based on non-convex optimization: Wirtinger flow (WF) [22] and its variants [28, 137], truncated amplitude flow (TAF) [121] and the trust region method of [106] all minimize the empirical risk (which is non-convex due to the missing phase) based on different loss functions. The convex alternatives typically use matrix-lifting as in PhaseLift [21, 25] and PhaseCut [119], which allows phase retrieval to be formulated as a semidefinite programming problem, or consider a non-lifting convex relaxation and solve the dual problem as in PhaseMax [51].

Our approach for estimating the signal $\mathbf{x}^\star$ follows the established approach of empirical risk minimization. Writing $\mathbf{z} = (y, \mathbf{a}) \in \mathbb{R} \times \mathbb{R}^n$ for an observation, we consider the loss function $\ell(\mathbf{x}, \mathbf{z}) = \frac{1}{4}((\mathbf{a}^T\mathbf{x})^2 - y)^2$ and, given samples $\mathbf{Z}_1, ..., \mathbf{Z}_m$, the empirical risk

$$F(\mathbf{x}) = \frac{1}{4m} \sum_{j=1}^{m} \left((\mathbf{A}_j^T \mathbf{x})^2 - Y_j \right)^2 . \tag{2.1}$$

It is worth mentioning that in previous applications the amplitude-based loss function $\ell(\mathbf{x}, \mathbf{z}) = \frac{1}{2}(|\mathbf{a}^T\mathbf{x}| - \sqrt{y})^2$ has been numerically shown to be more effective in terms of sample complexity than the loss function based on squared magnitudes [121, 122, 137].

However, with our parametrization, we found the squared magnitude-based loss function to be more effective.

Without any restrictions on the signal $\mathbf{x}^\star \in \mathbb{R}^n$, $2n-1$ Gaussian measurements suffice for $\mathbf{x}^\star$ to be the unique (up to global sign) minimizer of $F(\mathbf{x})$ with high probability [13]. If $\mathbf{x}^\star$ is k-sparse, then it has been shown in [73] that

$$\{\pm\mathbf{x}^\star\} = \underset{\mathbf{x}:\|\mathbf{x}\|_0\leq k}{\operatorname{argmin}} F(\mathbf{x}) \tag{2.2}$$

holds with high probability if we have $m \geq 4k-1$ Gaussian measurements.

Solving (2.2) involves two main difficulties: (i) the objective function F is non-convex with potentially many local minima and saddle points, and (ii) due to the constraint $\|\mathbf{x}\|_0 \leq k$ the problem is of combinatorial nature and NP-hard in general.

Regarding the first difficulty, the non-convexity is typically addressed by using a spectral or orthogonality-promoting initialization, which produces an initial estimate close to $\mathbf{x}^\star$ and in a region where the objective function enjoys a benign local geometry, leading to linear convergence towards $\mathbf{x}^\star$ [22, 28, 121, 137]. Recently, it has been shown that such an initialization is not always necessary in the phase retrieval problem and that a random initialization can be used instead [29, 106].

Addressing the second difficulty, an approach replacing the constraint $\|\mathbf{x}\|_0 \leq k$ in (2.2) by adding a penalty term $\lambda\|\mathbf{x}\|_1$ was proposed in [133]. However, this procedure requires tuning of the regularization parameter λ to reach a desired sparsity level, and it is tailored for Fourier measurements only (in particular, it uses the fact that the DFT matrix is unitary). Recently, methods enforcing the constraint $\|\mathbf{x}\|_0 \leq k$ via a hard-thresholding step have received a lot of attention. These include SPARTA [122], CRAF [138] and SWF [135]. However, the implementation of such a thresholding step requires knowledge of k (or a suitable upper bound).

Our proposed method does not have to deal with these difficulties. We also approach the problem by minimizing the objective $F(\mathbf{x})$. However, unlike existing algorithms, we do not need to add any penalty term to the objective or to introduce a thresholding step to enforce the constraint $\|\mathbf{x}\|_0 \leq k$. Our simulations show that the iterates of gradient descent with Hadamard parametrization remain (approximately) in the low-dimensional

space of sparse vectors. Hence, we neither need to tune any regularization parameters, nor do we need knowledge of the underlying signal sparsity. Further, HWF does not need a sophisticated initialization scheme commonly used in non-convex optimization based approaches to (sparse) phase retrieval.

2.3 Hadamard Wirtinger flow

Consider the parametrization $\mathbf{x} = \mathbf{u} \odot \mathbf{u} - \mathbf{v} \odot \mathbf{v}$. Such a parametrization has previously been applied to problems such as sparse recovery [56, 113, 140] and matrix factorization [6, 54, 74]. Exploiting the restricted isometry property (RIP) assumed for these problems, the Hadamard parametrization has been shown to confine the gradient iterates to the low-dimensional spaces of sparse vectors and low-rank matrices, respectively. Overloading the notation, we write

$$F(\mathbf{u}, \mathbf{v}) = \frac{1}{4m} \sum_{j=1}^{m} \Big((\mathbf{A}_j^T(\mathbf{u}^2 - \mathbf{v}^2))^2 - Y_j \Big)^2,$$

for the empirical risk, with gradients

$$\begin{aligned} \nabla_{\mathbf{u}} F(\mathbf{u}, \mathbf{v}) &= \frac{2}{m} \sum_{j=1}^{m} \Big((\mathbf{A}_j^T(\mathbf{u}^2 - \mathbf{v}^2))^2 - (\mathbf{A}_j^T \mathbf{x}^\star)^2 \Big) (\mathbf{A}_j^T(\mathbf{u}^2 - \mathbf{v}^2)) \mathbf{A}_j \odot \mathbf{u} \\ &= 2 \nabla F(\mathbf{x}) \odot \mathbf{u}, \end{aligned}$$

and, similarly, $\nabla_{\mathbf{v}} F(\mathbf{u}, \mathbf{v}) = -2 \nabla F(\mathbf{x}) \odot \mathbf{v}$. We consider gradient descent in this parametrization,

$$\begin{aligned} \mathbf{X}^t &= \mathbf{U}^t \odot \mathbf{U}^t - \mathbf{V}^t \odot \mathbf{V}^t, \\ \mathbf{U}^{t+1} &= \mathbf{U}^t \odot \Big(\mathbf{1}_n - 2\eta \nabla F(\mathbf{X}^t) \Big), \\ \mathbf{V}^{t+1} &= \mathbf{V}^t \odot \Big(\mathbf{1}_n + 2\eta \nabla F(\mathbf{X}^t) \Big), \end{aligned} \tag{2.3}$$

where we denote by $\mathbf{1}_n \in \mathbb{R}^n$ the vector of all ones.

The reason why the Hadamard parametrization promotes sparsity is that this parametrization turns the additive updates of gradient descent into multiplicative updates. If we choose a small initialization, then, in the aforementioned problems with RIP assumptions, the multiplicative updates have been shown to lead to off-support variables staying negligibly small while support variables are being fitted. The variables grow exponentially, but at different (time-varying) rates, with off-support variables growing at a smaller rate than support variables. With additive updates, off-support

variables would not stay sufficiently small and the algorithm would typically converge towards non-sparse local minima. We provide a more detailed discussion on the role of the Hadamard parametrization in Appendix 2.7.1, which suggests the following initialization:

$$\mathbf{V}^0 = \alpha \mathbf{1}_n, \qquad U_i^0 = \begin{cases} \left(\frac{\hat{\theta}}{\sqrt{3}} + \alpha^2\right)^{\frac{1}{2}} & i = I_{max} \\ \alpha & i \neq I_{max} \end{cases} \tag{2.4}$$

where we write $\hat{\theta} = (\frac{1}{m}\sum_{j=1}^m Y_j)^{1/2}$ for the estimate of the signal size $\|\mathbf{x}^\star\|_2$ (see e.g. [22, 121]), $I_{max} = \text{argmax}_i \sum_{j=1}^m Y_j A_{ji}^2$ and $\alpha > 0$ is the initialization size.

We show in the next section that if $|x^\star_{I_{max}}| \geq \frac{1}{2} x^\star_{max}$ and m is sufficiently large, then, with high probability, we can recover the support by running one step of (2.3). The probability of finding a large coordinate I_{max} can be increased by allowing multiple restarts and considering not only the largest, but also a few more instances in $\{R_i\}_{i=1}^n := \{\frac{1}{m}\sum_{j=1}^m Y_j A_{ji}^2\}_{i=1}^n$. Specifically, if we allow $\bar{b}$ restarts, we consider different initialization as in (2.4) using each of the $\bar{b}$ largest instances in $\{R_i\}_{i=1}^n$. As pointed out in Section 2.2, $\mathbf{x}^\star$ is the sparsest minimizer of the objective F. Given the results from multiple runs, we can therefore choose the (approximately) sparsest solution, by which we mean the solution where the fewest coordinates make up most (e.g. 95%) of the norm $\|\mathbf{X}^{\bar{t}}\|_2$. This is summarized in Algorithm 1.

2.4 Support recovery

In this section we show, assuming $x^\star_{min} = \Omega(1/\sqrt{k})$, that one step of Algorithm 1 can be used to recover the support $\mathcal{S}$ from $\mathcal{O}(\max\{k \log n, \log^3 n\}(x^\star_{max})^{-2})$ Gaussian measurements. As pointed out in Section 2.2, the sample complexity bottleneck of reconstruction algorithms such as SPARTA and SparseAltMinPhase lies in the initial support recovery. In particular, both algorithms require $\mathcal{O}(k^2 \log n)$ measurements to guarantee successful support recovery; with the knowledge of the support $\mathcal{S}$, these and plenty other algorithms such as WF, TAF and PhaseLift only require a sample complexity of $\mathcal{O}(k \log n)$ to guarantee successful reconstruction of a k-sparse signal.

The main difference from previous work is that we only need a single coordinate i with $|x_i^\star| \geq \frac{1}{2} x^\star_{max}$ rather than the full support for our initialization. Therefore, our sample

Algorithm 1: Hadamard Wirtinger flow, $\bar{b}$ restarts

Input: observations $\{Y_j\}_{j=1}^m$, measurement vectors $\{\mathbf{A}_j\}_{j=1}^m$, step size η, iterations $\bar{t}$, initialization size α, number of restarts $\bar{b}$, sparsity tolerance κ

for $b = 1$ *to* $\bar{b}$ **do**

Set I_b to the b^{th} largest instance in $\{R_i\}_{i=1}^n$

Set $\mathbf{U}^{0,b} = \mathbf{V}^{0,b} = \alpha \mathbf{1}_n, \quad U_{I_b}^{0,b} = \left(\frac{\hat{\theta}}{\sqrt{3}} + \alpha^2\right)^{\frac{1}{2}}$

for $t = 0$ *to* $\bar{t}$ **do**

$$\mathbf{X}^{t,b} = \mathbf{U}^{t,b} \odot \mathbf{U}^{t,b} - \mathbf{V}^{t,b} \odot \mathbf{V}^{t,b}$$
$$\mathbf{U}^{t+1,b} = \mathbf{U}^{t,b} \odot \left(\mathbf{1}_n - 2\eta \nabla F(\mathbf{X}^{t,b})\right)$$
$$\mathbf{V}^{t+1,b} = \mathbf{V}^{t,b} \odot \left(\mathbf{1}_n + 2\eta \nabla F(\mathbf{X}^{t,b})\right)$$

Set B_{min} to be the index minimizing

$$\min_{\mathcal{C} \subset [n]} \left\{ |\mathcal{C}| : \sum_{i \in C} (\mathbf{X}_i^{\bar{t},b})^2 \geq (1-\kappa) \sum_{i=1}^n (\mathbf{X}_i^{\bar{t},b})^2 \right\}$$

return $\mathbf{X}^{\bar{t},B_{min}}$

complexity depends on $x^\star_{max}$, which is at least $1/\sqrt{k}$, rather than $x^\star_{min}$, which can be at most $1/\sqrt{k}$. This is made precise in the following Lemma.

Lemma 2.1 (Support recovery)**.** *Let $\mathbf{x}^\star \in \mathbb{R}^n$ be any k-sparse vector with $x^\star_{min} = \Omega(\frac{1}{\sqrt{k}})$, and assume that we are given measurements $\{Y_j = (\mathbf{A}_j^T \mathbf{x}^\star)^2\}_{j=1}^m$, where $\mathbf{A}_j \sim \mathcal{N}(0, \mathbf{I}_n)$, $j = 1, \ldots, m$, are i.i.d. Gaussian vectors. If $m \geq \mathcal{O}(\max\{k \log n, \log^3 n\}(x^\star_{max})^{-2})$, then, with probability at least $1 - \mathcal{O}(n^{-10})$, choosing the largest instance in $\{R_i\}_{i=1}^n := \{\frac{1}{m}\sum_{j=1}^m Y_j A_{ji}^2\}_{i=1}^n$ returns an index i with $|x_i^\star| \geq \frac{1}{2} x^\star_{max}$.*

Further, let $\mathbf{X}^1$ be the estimate obtained from running one step of Algorithm 1 with any $\eta, \alpha \in (0, \frac{1}{10})$ (and $\bar{b} = 1$, i.e. no multiple restarts). Then, with the same probability, we can recover the support $\mathcal{S} = \{i : x_i^\star \neq 0\}$ by choosing the k largest coordinates of $|\mathbf{X}^1|$ (where $|\cdot|$ denotes taking absolute values coordinate-wise).

The proof of Lemma 2.1 relies on standard concentration results and is deferred to Appendix 2.7.2.

Lemma 2.1 shows that, provided knowledge of the sparsity level k, Algorithm 1 can be used to recover the support of a k-sparse signal $\mathbf{x}^\star$ from $\mathcal{O}(\max\{k \log n, \log^3 n\}(x^\star_{max})^{-2})$

Gaussian measurements, which, provided $k \geq \mathcal{O}(\log^2 n)$, matches the best known bounds $\mathcal{O}(k^2 \log n)$ [90, 122] in the worst case, while it is an improvement if $\mathbf{x}^\star$ contains (at least) one large coordinate. For instance, if $x^\star_{max} = \Omega(1)$ and $k \geq \mathcal{O}(\log^2 n)$, only $\mathcal{O}(k \log n)$ samples are required for support recovery.

We validate this theoretical result in the following experiment. Let $\mathbf{x}^\star \in \mathbb{R}^{10000}$ be a k-sparse signal with randomly sampled support $\mathcal{S} = \{i_1, ..., i_k\}$ and normalized to $\|\mathbf{x}^\star\|_2 = 1$. We consider maximum signal values (i) $x^\star_{max} = \frac{1}{\sqrt{k}}$, (ii) $x^\star_{max} = k^{-0.25}$ and (iii) $x^\star_{max} = 0.7$. In case (i), we set $x^\star_{i_j} = \pm\frac{1}{\sqrt{k}}$ at random for all $j = 1, ..., k$. For the cases (ii) and (iii), we fix $x^\star_{i_1} = x^\star_{max}$, sample the other components from $x^\star_{i_j} \sim \mathcal{N}(0, 1)$ i.i.d., and then normalize them to satisfy $\|\mathbf{x}^\star\|_2 = 1$. We also consider a signal with $x^\star_{i_j} \sim \mathcal{N}(0, 1)$ i.i.d. normalized to $\|\mathbf{x}^\star\|_2 = 1$, without any restrictions on $x^\star_{max}$. We generate $m = 5000$ measurements $Y_j = (\mathbf{A}_j^T \mathbf{x}^\star)^2$ with $\mathbf{A}_j \sim \mathcal{N}(0, \mathbf{I}_n)$ i.i.d..

For our method (HWF) we run one step of Algorithm 1 and pick the k largest components of $|\mathbf{X}^1|$. We compare it with the support recovery methods used in SPARTA and SparseAltMinPhase (other algorithms like SWF and CRAF use the same support recovery method as SPARTA). Note that although correct identification of the full support is required for the theoretical guarantees of algorithms like SPARTA, it is not necessary in practice: it has been noted in [122] that, since the estimated support $\hat{\mathcal{S}}$ is only used for the orthogonality-promoting initialization, SPARTA can be successful as long as the initial estimate is sufficiently close to the underlying signal (more precisely, $\text{dist}(\mathbf{X}^0, \mathbf{x}^\star) \leq \frac{1}{10}\|\mathbf{x}^\star\|_2$), regardless of whether or not the full support has been correctly identified. Intuitively, the initialization produces an estimate sufficiently close to $\mathbf{x}^\star$ as long as the majority of the support is recovered. This intuition has been made rigorous for an alternative spectral initialization in [59].

We evaluate the proportion of correctly recovered support variables

$$\frac{|\hat{\mathcal{S}} \cap \mathcal{S}|}{|\mathcal{S}|}$$

obtained from 100 independent Monte Carlo trials, where $\hat{\mathcal{S}} \subset [n]$ denotes the estimated support.

Figure 2.1 confirms the predictions of Lemma 2.1. In the first case, where the signal only takes the values $x^\star_i \in \{-\frac{1}{\sqrt{k}}, 0, \frac{1}{\sqrt{k}}\}$, our theoretical bound reads $\mathcal{O}(k^2 \log n)$ (as

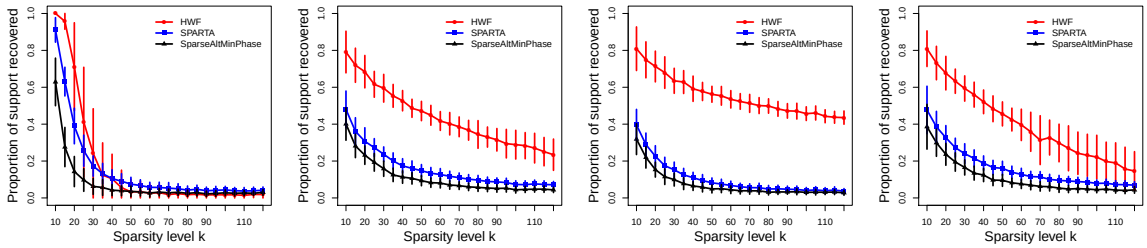


Figure 2.1: Proportion of support variables $|\hat{\mathcal{S}} \cap \mathcal{S}|/|\mathcal{S}|$ correctly recovered plus/minus one standard deviation (vertical lines) by our method (red curve), SPARTA (blue curve) and SparseAltMinPhase (black curve), for different levels of $x^\star_{max}$. From left to right: (i) $x^\star_{max} = \frac{1}{\sqrt{k}}$, (ii) $x^\star_{max} = k^{-0.25}$, (iii) $x^\star_{max} = 0.7$, (iv) Gaussian signal $\mathbf{x}^\star$ (without restrictions on $x^\star_{max}$).

we are considering a regime where $k \geq \mathcal{O}(\log^2 n)$), and we expect the support recovery performance of our method to be comparable to the other methods. For sparsity levels $k \geq 45$, our method is slightly worse than the other two, because it sometimes fails to identify a coordinate i with $x^\star_i \neq 0$ in the first step. This case is of little practical relevance, since if only such a small portion of the support is recovered, neither of the three algorithms is able to reconstruct $\mathbf{x}^\star$. As we increase $x^\star_{max}$, the performance of our method improves substantially, while the support recovery methods used in SPARTA and SparseAltMinPhase do not show any improvement (in fact, they get slightly worse, which can be attributed to the fact that, as we increase $x^\star_{max}$, the other coordinates become smaller since we keep $\|\mathbf{x}^\star\|_2 = 1$ fixed). Our method also shows better support recovery performance for Gaussian signals $\mathbf{x}^\star$, where we do not fix $x^\star_{max}$ (right figure).

2.5 Parameter estimation

In this section, we first demonstrate that our support recovery method, one step of HWF, can be combined with existing algorithms (such as SPARTA), which leads to a final procedure which provably recovers a k-sparse signal from $\mathcal{O}(\max\{k \log n, \log^3 n\}(x^\star_{max})^{-2})$ measurements; this is summarized in Algorithm 2.

Note that we can allow multiple restarts in this case as well ($\bar{b} > 1$) in order to further improve the probability of obtaining a good initialization. Since SPARTA only produces k-sparse solutions due to the thresholding step, we choose the final solution by selecting the one which produces the smallest gradient $\|\nabla F(\mathbf{X}^{\bar{t},b})\|_2$.

Algorithm 2: SPARTA-support, multiple restarts

Input: observations $\{Y_j\}_{j=1}^m$, measurement vectors $\{\mathbf{A}_j\}_{j=1}^m$, sparsity level k, step size η, iterations $\bar{t}$, initialization size α, number of restarts $\bar{b}$, parameters for SPARTA specified in [122]

for $b = 1$ *to* $\bar{b}$ **do**

Set I_b to the b^{th} largest instance in $\{R_i\}_{i=1}^n$

Set $\mathbf{U}^{0,b} = \mathbf{V}^{0,b} = \alpha \mathbf{1}_n, \quad U_{I_b}^{0,b} = \left(\frac{\hat{\theta}}{\sqrt{3}} + \alpha^2\right)^{\frac{1}{2}}$

Run one step of HWF (2.3) for $\mathbf{X}^{1,b}$

Set $\hat{\mathcal{S}}_b$ to the k largest coordinates of $|\mathbf{X}^{1,b}|$

Run $\bar{t}$ iterations of SPARTA using $\hat{\mathcal{S}}_b$ for $\mathbf{X}^{\bar{t},b}$

Set B_{min} to be the index minimizing $\|\nabla F(\mathbf{X}^{\bar{t},b})\|_2$

return $\mathbf{X}^{\bar{t},B_{min}}$

Our analysis from the previous section immediately leads to the following result.

Theorem 2.2. *Let $\mathbf{x}^\star \in \mathbb{R}^n$ be any k-sparse vector with $x^\star_{min} = \Omega(1/\sqrt{k})$, and assume that we are given measurements $\{Y_j = (\mathbf{A}_j^T \mathbf{x}^\star)^2\}_{j=1}^m$, where $\mathbf{A}_j \sim \mathcal{N}(0, \mathbf{I}_n)$, $j = 1, \ldots, m$, are i.i.d. Gaussian vectors. If $m \geq \mathcal{O}(\max\{k \log n, \log^3 n\}(x^\star_{max})^{-2})$, then, with the parameters specified in [122], successive estimates of SPARTA-support satisfy, with probability at least $1 - \mathcal{O}(m^{-1} + n^{-10})$ and for a universal constant $0 < v < 1$,*

$$\operatorname{dist}(\mathbf{X}^t, \mathbf{x}^\star) \leq \frac{1}{10}(1-\nu)^t \|\mathbf{x}^\star\|_2, \qquad t \geq 0.$$

Proof. By Lemma 2.1, one step of HWF recovers the true support with probability $1 - \mathcal{O}(n^{-10})$. The result then follows from Lemma 2 and 3 of [122]. □

Compared to Theorem 1 of [122], this result reduces the sample complexity from $\mathcal{O}(k^2 \log n)$ to $\mathcal{O}(k(x^\star_{max})^{-2} \log n)$, provided $k \geq \mathcal{O}(\log^2 n)$. The assumption $x^\star_{min} = \Omega(1/\sqrt{k})$ is likely an artifact of the proof method of [122] and not necessary. Intuitively, identifying the full support is not necessary, as a good initialization can also be obtained if only small coordinates with $x_i^\star \leq \mathcal{O}(1/\sqrt{k})$ are missed. This intuition has been made rigorous for an alternative spectral initialization [59]. However, the orthogonality-promoting initialization used in SPARTA has been experimentally found to produce an initial estimate closer to the signal $\mathbf{x}^\star$ than the spectral initialization [121, 138].

One step of Algorithm 1 requires $\mathcal{O}(nm)$ operations, and $\bar{t} = \mathcal{O}(\log(1/\epsilon))$ SPARTA iterations are sufficient to find an ϵ-accurate solution, so SPARTA-support incurs a

total computational cost of $\mathcal{O}(nm \log(1/\epsilon))$. This is proportional to the cost of reading the data modulo logarithmic terms.

However, SPARTA-support enforces sparsity of the estimates $\mathbf{X}^t$ explicitly via a hard-thresholding step, which requires knowledge of k (or an upper bound). Our simulations show that HWF *adapts* to the signal sparsity k: we neither need knowledge of k for thresholding steps, nor do we need to add a penalty term to the objective and tune regularization parameters to promote sparsity. Given enough samples, our algorithm automatically converges to the k-sparse signal $\mathbf{x}^\star$.

In the following, we present simulations evaluating the reconstruction performance of HWF and SPARTA-support relative to state-of-the-art methods for sparse phase retrieval. In particular, we will consider SPARTA, SWF and PR-GAMP.

Remark 2.1 (Comparison with PR-GAMP)**.** *Our numerical experiments show comparable sample complexities for PR-GAMP and HWF, with both being lower than the sample requirement of other gradient-based methods. PR-GAMP has been empirically shown to achieve linear sample complexity in some regimes with Gaussian signals [102]. However, PR-GAMP relies on the implementation and tuning of several algorithmic principles, such as damping, normalization, and expectation-maximization (EM) steps. On the one hand, the application of these algorithmic principles makes PR-GAMP difficult to analyze, as rigorous theoretical investigations are known to be challenging even for much simpler AMP-based algorithms [14]. On the other hand, running PR-GAMP requires tuning of several parameters, including the sparsity rate k/n via EM steps, and it requires choosing the prior distribution for the signal $\mathbf{x}^\star$. For our simulations we used the freely available GAMP package*[1] *that does automatic parameter tuning, using the Gauss-Bernoulli prior. HWF is a much simpler algorithm, as it is just vanilla gradient descent applied to the unregularized empirical risk with Hadamard parametrization. HWF does not rely on algorithmic principles to promote convergence to good solutions, and it is empirically seen to adapt to the sparsity level k. We leave it to future work to give a full theoretical account on the convergence guarantees of HWF and to consider more refined and fine-tuned formulations of HWF*

[1]We used the code available from `https://sourceforge.net/projects/gampmatlab/` for the implementation of PR-GAMP.

that can combine algorithmic principles typically used in the literature on sparsity (cf. Section 2.6, Conclusion).

In experiments where we do not fix $x^\star_{max}$, the true signal vector $\mathbf{x}^\star \in \mathbb{R}^{1000}$ was obtained by sampling $\mathbf{x}^\star \sim \mathcal{N}(0, \mathbf{I}_{1000})$, setting $(1000 - k)$ random entries of $\mathbf{x}^\star$ to 0 and normalizing $\|\mathbf{x}^\star\|_2 = 1$. Otherwise, $\mathbf{x}^\star$ is generated as described in Section 2.4. We obtain m noiseless measurements $Y_j = (\mathbf{A}_j^T \mathbf{x}^\star)^2$ with $\mathbf{A}_j \sim \mathcal{N}(0, \mathbf{I}_{1000})$ i.i.d..

For the parameters of SPARTA and SWF, we found the values suggested in the original papers to work best in our simulations and used these in all experiments. For HWF we found that a constant step size $\eta = 0.1$ works well (similar to WF [80]). For the other parameters, any small values work well without much difference and we set $\alpha = 0.001$, $\kappa = 0.05$ and allow $\bar{b} = 50$ restarts. We run all algorithms for a maximum of $\bar{t} = 100,000$ iterations or until $F(\mathbf{X}^t) \leq 10^{-7}$, and declare it a success if the relative error

$$\frac{\text{dist}(\mathbf{X}^{\bar{t}}, \mathbf{x}^\star)}{\|\mathbf{x}^\star\|_2}$$

is less than 0.01. We evaluate the empirical success rate obtained from 100 independent Monte Carlo trials. In all experiments, SPARTA, SWF and SPARTA-support were run with oracle knowledge of the true signal sparsity k, which is not needed for HWF.

In the first experiment, we fix the sparsity to $k = 20$ and vary m from 100 to 1000. Figure 2.2 (left) shows that HWF is able to reconstruct the signal reliably (with 95% success rate) from $m = 400$ measurements, which is slightly better than PR-GAMP ($m = 500$), while SPARTA and SWF both require almost twice as many observations ($m = 700$). Next, we fix $m = 500$ and vary the sparsity level k. Figure 2.2 (right) shows that HWF achieves a reconstruction rate of 95% for signals with up to 35 non-zero entries, while PR-GAMP achieves this success rate only for signals with up to 25 non-zero entries. PR-GAMP achieves slightly higher success rates than HWF for sparsity levels where neither algorithm is able to reliably reconstruct the signal.

In the previous section, we discussed the superior support recovery performance of our method as $x^\star_{max}$ increases. The next experiment examines whether this effect also translates into better reconstruction performance. To this end, we consider signals generated as in the experiments in Section 2.4, fix $m = 500$ and vary $k \in [10, 80]$.

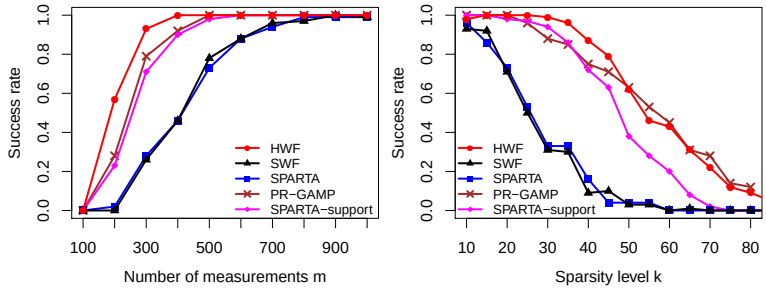


Figure 2.2: Empirical success rate for $n = 1000$ fixed against number of measurements m with sparsity level $k = 20$ fixed (left) and against sparsity level k with $m = 500$ measurements (right).

Figure 2.3 shows that, even when the signal only takes values $x_i^\star \in \left\{-\frac{1}{\sqrt{k}}, 0, \frac{1}{\sqrt{k}}\right\}$, HWF achieves higher success rates than SPARTA and SWF. SPARTA-support is comparable to them, as also the support recovery performance is similar for this $\mathbf{x}^\star$, and SPARTA-support subsequently applies the same steps as SPARTA. As $x^\star_{max}$ increases, the reconstruction performance of our methods improves, with HWF maintaining a higher success rate than SPARTA-support. As before, PR-GAMP achieves a 95% success rate up to slightly lower sparsity levels than HWF. PR-GAMP maintains success rates comparable to HWF as $x^\star_{max}$ increases, which might explain the linear sample complexity observed in [102] in some regimes for Gaussian signals. The maximum component of a Gaussian vector scales (in expectation) like $\sqrt{\log k}/\sqrt{k}$, which is, if k is not very large, noticeably larger than $1/\sqrt{k}$. Comparing the right plot of Figure 2.2 and the left plot of Figure 2.3, we see that PR-GAMP achieves higher success rates for Gaussian signals than for the signal with $x^\star_{max} = 1/\sqrt{k}$.

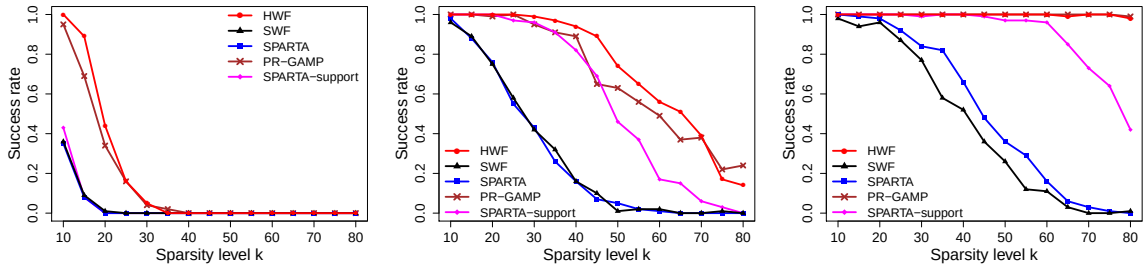


Figure 2.3: Empirical success rate against sparsity level k with $n = 1000, m = 500$ fixed. From left to right: (i) $x^\star_{max} = 1/\sqrt{k}$, (ii) $x^\star_{max} = k^{-0.25}$ and (iii) $x^\star_{max} = 0.7$.

Next, we examine how the sample complexity of HWF scales with the signal sparsity k. The success rate vs signal sparsity k and number of measurements m is shown in Figure 2.4, which suggests that the sample complexity scales as $\mathcal{O}(k(x^\star_{max})^{-2} \log \frac{n}{k})$, where we

obtain $x^\star_{max}$ as the average maximum coordinate of $100,000$ Gaussian k-sparse signals. We note that this scaling appears almost linear.

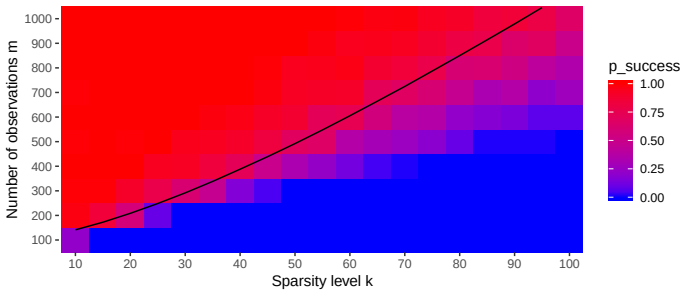


Figure 2.4: Empirical success rate (red: high, blue: low) of HWF against sparsity level k and number of observations m, with $n = 1000$ fixed. Black line: $m = \frac{1}{3}k(x^\star_{max})^{-2}\log\frac{n}{k}$.

One of the parameters in Algorithm 1 is the number of restarts $\bar{b}$. Increasing $\bar{b}$ also increases the probability of HWF finding the true signal, but this comes at the cost of an increase in computational time. For the next experiment, we run HWF in the same setting as the first two experiments and vary the number of allowed restarts $\bar{b}$ from 1 to 100. Figure 2.5 shows that increasing the number of allowed restarts indeed increases the probability of successful reconstruction, where the success rate barely increases further as we increase the number of restarts $\bar{b}$ beyond 50.

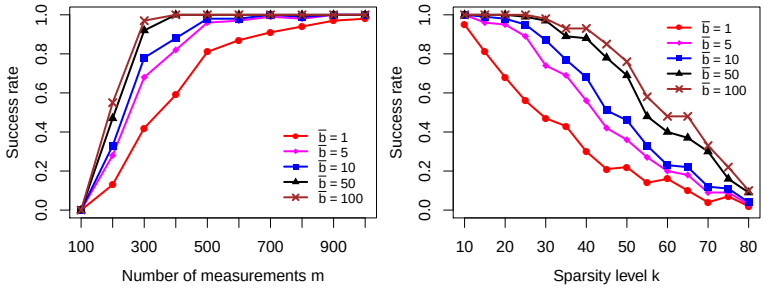


Figure 2.5: Empirical success rate for $n = 1000$ fixed against number of measurements m with sparsity level $k = 20$ fixed (left), and against sparsity level k with $m = 500$ measurements (right), for varying number of restarts $\bar{b} \in [1, 100]$.

Next, we examine the convergence behavior of HWF. We also test HWF in the complex-valued setting, where we generate vectors $\mathbf{x}^\star, \mathbf{A}_j \sim \mathcal{N}(0, \frac{1}{2}\mathbf{I}_{1000}) + i\mathcal{N}(0, \frac{1}{2}\mathbf{I}_{1000})$, set 990 random entries of $\mathbf{x}^\star$ to zero, normalize $\|\mathbf{x}^\star\|_2 = 1$ and generate $m = 500$ measurements $Y_j = |\mathbf{A}_j^{\mathrm{H}}\mathbf{x}^\star|^2$. Figure 2.6 shows that HWF is also able to reconstruct complex signals. While HWF converges faster in the real case, both cases exhibit sublinear convergence after a short "warm-up" period. This can be explained by our parametrization. Consider

the gradient $\nabla_{\mathbf{u}} F(\mathbf{U}^t, \mathbf{V}^t) = 2\nabla F(\mathbf{X}^t) \odot \mathbf{U}^t$: as the initialization size α is small, the gradient is small in the beginning due to the term $\mathbf{U}^t$. As $\mathbf{X}^t$ approaches $\mathbf{x}^\star$, the term $\nabla F(\mathbf{X}^t)$ converges to zero, which leads to linear convergence with a constant stepsize in the case of WF [80]. With our parametrization, $\nabla_{\mathbf{u}} F$ (or $\nabla_{\mathbf{v}} F$) converges to zero faster than $\nabla_{\mathbf{x}} F$, as we typically have $U_i^t \to 0$ or $V_i^t \to 0$ (or both, if $x_i^\star = 0$), leading to sublinear convergence.

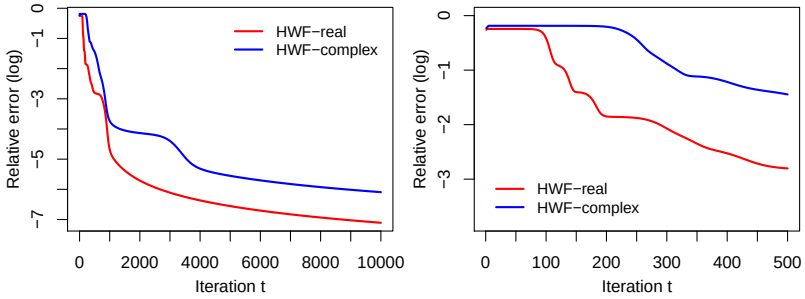


Figure 2.6: Relative error (log-scale) of HWF for real/complex signals with $n = 1000$, $m = 500$ and $k = 10$ for 10,000 iterations (left) and zoom-in to 500 iterations (right).

Finally, we present a numerical experiment that considers random initialization. In particular, we initialize U_i^0, V_i^0 to small Gaussian noise $\mathcal{N}(0, 0.01^2)$ for all $i = 1, \ldots, n$. In general, we find that more samples are required for HWF to successfully reconstruct the signal $\mathbf{x}^\star$ starting from a random initialization. Figure 2.7 shows that even in a setting with $n = 1000$, $m = 700$ and $k = 10$, where HWF with random initialization does converge to the signal $\mathbf{x}^\star$, the ℓ_2-error $\mathrm{dist}(\mathbf{X}^t, \mathbf{x}^\star)$ only decreases after an initial plateau, leading to slower convergence. This is in line with the intuition provided in Appendix 2.7.1, namely that (i) the signal can still be recovered as coordinates on the support of $\mathbf{x}^\star$ increase at a faster rate than coordinates not on the support, while (ii) all coordinates only change at a very slow rate initially, because the inner product $(\mathbf{X}^0)^T \mathbf{x}^\star$ is closer to zero with random initialization compared to our proposed initialization (2.4), which leads to the initial plateau; see Appendix 2.7.1 for details.

2.6 Conclusion

In this paper, we proposed HWF, which is a simple algorithm for sparse phase retrieval. We proved that one step of HWF can be used as a support recovery tool, which, combined with existing algorithms such as SPARTA, yields a computationally fast

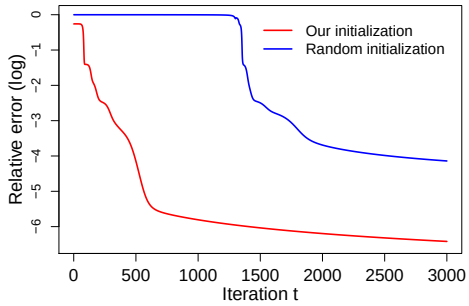


Figure 2.7: Relative error (log-scale) of HWF with random initialization (blue) and our proposed initialization (2.4) (red) with $n = 1000$, $m = 700$ and $k = 10$.

algorithm with improved sample complexity, which reads $\mathcal{O}(k \log n)$ if the signal contains at least one large component and $k \geq \mathcal{O}(\log^2 n)$. We have shown in numerical experiments that the sample complexity of HWF is lower than that of existing gradient based methods such as SPARTA and SWF, and comparable to PR-GAMP, which has been empirically shown to achieve linear sample complexity for Gaussian signals in some regimes [102]. While HWF does not require knowledge of the signal sparsity k, thresholding steps or any added regularization terms, this simplicity seems to come at the price of sublinear convergence and thus increased computational cost. We leave it to future work to investigate whether algorithmic principles such as the increasing step-size scheme considered in [113] or thresholding steps previously considered in the literature on sparse phase retrieval (e.g. [20, 122, 138]) can be used to accelerate the convergence speed of HWF or to further improve its sample complexity beyond the level of the empirical results observed for PR-GAMP, which already relies on a combination of many such algorithmic principles (damping, normalization, EM steps). Compared to PR-GAMP, the simplicity of HWF makes the algorithm potentially more amenable to a rigorous theoretical investigation that can support the high-level analysis presented in our work.

2.7 Chapter appendix

2.7.1 Understanding the dynamics of Hadamard Wirtinger flow

As discussed in Section 2.3, the Hadamard parametrization has previously been applied to problems such as sparse recovery [56, 113, 140] and matrix factorization [6, 54, 74],

where it turns the additive updates of gradient descent into multiplicative updates. The combination of multiplicative updates and a small initialization was shown to lead to sparsity in the aforementioned problems, under the assumption of the restricted isometry property (RIP).

The problem of sparse phase retrieval that we consider is known to satisfy RIP, cf. [116] for instance, and a similar explanation on why on-support variables and off-support variables can be made to grow at different speeds also holds in our setting. We now provide the main intuition behind the convergence properties of HWF by considering the evolution of the algorithm at the population level, i.e. in the case when $m = \infty$. While a rigorous convergence investigation of HWF is outside the scope of the present work, the analysis that we now provide is instrumental to construct a good initialization for Algorithm 1.

Consider the simplified setting where $\mathbf{x}^\star$ is non-negative, i.e. $x_i^\star \geq 0$ for all i. We can set $\mathbf{v} = \mathbf{0}$ in the parametrization, so that $\mathbf{x} = \mathbf{u}^2$. Further, assume that we have access to the population risk $f(\mathbf{x}) := \mathbb{E}[\ell(\mathbf{x}, \mathbf{Z})]$ (in other words, $m = \infty$), where $\mathbf{Z} = (Y, \mathbf{A})$ is defined by $Y = (\mathbf{A}^T\mathbf{x}^\star)^2$. Its gradient can be computed as

$$\nabla f(\mathbf{x}) = \Big(3\|\mathbf{x}\|_2^2 - 1\Big)\mathbf{x} - 2\Big(\mathbf{x}^T\mathbf{x}^\star\Big)\mathbf{x}^\star. \tag{2.5}$$

Under these two assumptions, first consider the initialization $\mathbf{x}^0 = \alpha^2\mathbf{1}_n$ for some small constant $\alpha > 0$. We can directly track the evolution of the estimates $\mathbf{x}^t$ (note that we use lowercase letters, since with $m = \infty$ the sequence is not random anymore) generated by Algorithm 1 via

$$x_i^{t+1} = x_i^t\Big(1 - 2\eta\Big[\Big(3\|\mathbf{x}^t\|_2^2 - 1\Big)x_i^t - 2\Big((\mathbf{x}^t)^T\mathbf{x}^\star\Big)x_i^\star\Big]\Big)^2.$$

This suggests that the evolution of $\mathbf{x}^t$ can be divided into two phases: if $\|\mathbf{x}^t\|_2^2 < \frac{1}{3}$, all coordinates grow ($x_i^{t+1} > x_i^t$), while coordinates $i \in \mathcal{S}$ on the support do so at a faster rate. If $\|\mathbf{x}^t\|_2^2 > \frac{1}{3}$, coordinates $i \notin \mathcal{S}$ decrease ($x_i^{t+1} < x_i^t$), while coordinates on the support increase if the product of the signal component $x_i^\star$ and the inner product $(\mathbf{x}^t)^T\mathbf{x}^\star$ is larger than the term $(3\|\mathbf{x}^t\|_2^2 - 1)x_i^t$.

If we choose $\alpha > 0$ small enough, we expect x_j^t to still be small (e.g. $< 1/n$) for $j \notin \mathcal{S}$ when $\|\mathbf{x}^t\|_2^2 \geq \frac{1}{3}$ first occurs, as x_i^t grows at a faster rate than x_j^t for $i \in \mathcal{S}$. Since

x_j^t decreases for $j \notin \mathcal{S}$ when $\|\mathbf{x}^t\|_2^2 \geq \frac{1}{3}$, we expect x_j^t to stay small throughout the algorithm for $j \notin \mathcal{S}$.

The smaller the step size η is, the more iterations are needed for the algorithm to converge. On the other hand, η cannot be too large; to illustrate this, consider the simplest case $n = 1$. The (scalar) gradient update becomes $x^{t+1} = x^t(1-6\eta[(x^t)^3 - x^t])$, and x^t diverges if η is too large. We found a constant step size $\eta = 0.1$ to work well in our simulations.

This recursion has three types of fixed points: $\mathbf{x}^{(1)} = \mathbf{0}$, any $\mathbf{x}^{(2)}$ satisfying $\|\mathbf{x}^{(2)}\|_2^2 = \frac{1}{3}$ and $(\mathbf{x}^{(2)})^T\mathbf{x}^\star = 0$, and $\mathbf{x}^{(3)} = \pm\mathbf{x}^\star$. The first fixed point $\mathbf{x}^{(1)}$ is repelling, as all coordinates grow if $\|\mathbf{x}^t\|_2^2 < \frac{1}{3}$. Similarly, the second fixed point $\mathbf{x}^{(2)}$ is repelling as x_i^t grows at a faster rate than x_j^t for $i \in \mathcal{S}, j \notin \mathcal{S}$. This leaves only $\mathbf{x}^{(3)}$, which is an attracting fixed point of the recursion. Thus, we expect Algorithm 1 to converge to $\mathbf{x}^\star$ if m is sufficiently large.

Guided by this intuition, we aim to construct an initialization $\mathbf{X}^0$ with $(\mathbf{X}^0)^T\mathbf{x}^\star$ large (more precisely, we will have $|(\mathbf{X}^0)^T\mathbf{x}^\star| \geq \frac{1}{4}x_{max}^\star$), while at the same time $\|\mathbf{X}^0\|_2^2$ should not be too large (e.g. fixed to $\|\mathbf{X}^0\|_2^2 = \frac{1}{3}\|\mathbf{x}^\star\|_2$; note that any other constant would also work, and that we use the estimate $\hat{\theta} = (\frac{1}{m}\sum_{j=1}^m Y_j)^{1/2}$ of the signal size $\|\mathbf{x}^\star\|_2$, see e.g. [22, 121]). In order to obtain such an initialization, it suffices to find a coordinate $i \in [n]$ with $|x_i^\star| \geq \frac{1}{2}x_{max}^\star$. Then, we can set $X_i^0 = \hat{\theta}/\sqrt{3}$ and $X_j^0 = 0$ for all $j \neq i$. Note that such an initialization is not necessary, and even with a random initialization (e.g. U_i^0, V_i^0 set to small random noise for all $i = 1, \ldots, n$), the above intuition that coordinates $i \in \mathcal{S}$ on the support grow at a faster rate than coordinates $i \notin \mathcal{S}$ not on the support, continues to hold. However, the initial inner product $(\mathbf{X}^0)^T\mathbf{x}^\star$ is closer to zero with random initialization compared to our proposed initialization, which leads to the population gradient $\nabla f(\mathbf{X}^t)_i$ initially being close to zero for all $i = 1, \ldots, n$, and therefore slow convergence.

Define the random variables $R_i = \frac{1}{m}\sum_{j=1}^m Y_j A_{ji}^2$ for $i = 1, \ldots, n$. These quantities were also used in [122] for support recovery, as one can compute $\mathbb{E}[R_i] = \|\mathbf{x}^\star\|_2^2 + 2x_i^2$ using the assumption $\mathbf{A}_j \sim \mathcal{N}(0, \mathbf{I}_n)$ i.i.d.. Hence, if the number of measurements m is large, the random variables $\{R_i\}_{i=1}^n$ will concentrate around their means, separating them for $i \in \mathcal{S}$ and $i \notin \mathcal{S}$. This intuition suggests the initialization proposed in Section 2.3.

2.7.2 Proof of Lemma 2.1

In the following, we assume, without loss of generality, that $\|\mathbf{x}^\star\|_2 = 1$; this assumption is made purely for notational simplicity, since we then have $x^\star_{max} = \max_i \frac{|x_i^\star|}{\|\mathbf{x}^\star\|_2} = \max_i |x_i^\star|$ and $x^\star_{min} = \min_{i:x_i^\star \neq 0} \frac{|x_i^\star|}{\|\mathbf{x}^\star\|_2} = \min_{i:x_i^\star \neq 0} |x_i^\star|$. If $\|\mathbf{x}^\star\|_2 \neq 1$ is unknown, then we only need to replace $x^\star_{max}$ and $x^\star_{min}$ with $\max_i |x_i^\star| = x^\star_{max}\|\mathbf{x}^\star\|_2$ and $\min_{i:x_i^\star \neq 0} |x_i^\star| = x^\star_{min}\|\mathbf{x}^\star\|_2$ respectively in the following proof. Further, note that knowledge of $\|\mathbf{x}^\star\|_2$ is not required for HWF.

The proof of Lemma 2.1 relies on the following result, which is a combination of Theorems 3.6 and 3.7 of [34].

Theorem 2.3. *[34] Let X_i be independent random variables satisfying $|X_i| \leq M$ for all $i \in [n]$. Let $X = \sum_{i=1}^n X_i$ and $\|X\| = \sqrt{\sum_{i=1}^n \mathbb{E}[X_i^2]}$. Then, we have*

$$\mathbb{P}[|X - \mathbb{E}[X]| > \lambda] \leq 2\exp\left(-\frac{\lambda^2}{2(\|X\|^2 + M\lambda/3)}\right).$$

Proof of the first claim.

We first show that by choosing the largest instance in $\{\frac{1}{m}\sum_{j=1}^m Y_j A_{ji}^2\}_{i=1}^n$, we obtain an index i with $|x_i^\star| \geq \frac{x^\star_{max}}{2}$ with high probability. Recall that we write $R_i = \frac{1}{m}\sum_{j=1}^m Y_j A_{ji}^2$. We can compute

$$\begin{aligned}
\mathbb{E}[R_i] &= \mathbb{E}[(\mathbf{A}_1^T\mathbf{x}^\star)^2 A_{1i}^2] \\
&= \mathbb{E}[A_{1i}^4 (x_i^\star)^2 + (\mathbf{A}_{1,-i}^T \mathbf{x}^\star_{-i})^2 A_{1i}^2] \\
&= 3(x_i^\star)^2 + \|\mathbf{x}^\star_{-i}\|_2^2 \\
&= \|\mathbf{x}^\star\|_2^2 + 2(x_i^\star)^2,
\end{aligned}$$

where we denote by $\mathbf{x}_{-i} \in \mathbb{R}^{n-1}$ the vector obtained by deleting the i-th entry from $\mathbf{x} \in \mathbb{R}^n$ and use the fact that $A_{ji} \sim \mathcal{N}(0,1)$ i.i.d. and hence $\mathbf{A}_{j,-i}^T \mathbf{x}^\star_{-i} \sim \mathcal{N}(0, \|\mathbf{x}^\star_{-i}\|_2^2)$, as $\mathbf{x}^\star \in \mathbb{R}^n$ is a fixed vector independent of the measurement vectors $\{\mathbf{A}_j\}_{j=1}^m$.

Let $I_{max} = \operatorname{argmax}_i R_i$. By definition, $R_{I_{max}} \geq R_i$ holds for all $i \in [n]$. If we can show

$|R_i - \mathbb{E}[R_i]| \le \frac{3}{4}(x^\star_{max})^2$ for all $i \in [n]$, then this would imply

$$\begin{aligned}\|\mathbf{x}^\star\|_2^2 + 2(x^\star_{I_{max}})^2 &= \mathbb{E}[R_{I_{max}}] \\ &= \mathbb{E}[R_i] + (R_i - \mathbb{E}[R_i]) + (\mathbb{E}[R_{I_{max}}] - R_{I_{max}}) + (R_{I_{max}} - R_i) \\ &\ge \mathbb{E}[R_i] - 2\max_j |R_j - \mathbb{E}[R_j]| \\ &\ge \|\mathbf{x}^\star\|_2^2 + 2(x^\star_i)^2 - \frac{3}{2}(x^\star_{max})^2,\end{aligned}$$

for any $i \in [n]$. In particular, if we choose $i = \mathrm{argmax}_j |x^\star_j|$, this implies $|x^\star_{I_{max}}| \ge \frac{1}{2}x^\star_{max}$, which concludes the proof of the first claim.

In order to show $|R_i - \mathbb{E}[R_i]| \le \frac{3}{4}(x^\star_{max})^2$, we use the following truncation argument: for any $i \in [n]$, we write

$$R_i = \frac{1}{m}\sum_{j=1}^m Y_j A_{ji}^2 = \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_j^T\mathbf{x}^\star)^2 A_{ji}^2 = \frac{1}{m}\sum_{j=1}^m (Z_{1,j} + Z_{2,j}),$$

where we write $Z_{1,j} = (\mathbf{A}_j^T\mathbf{x}^\star)^2 A_{ji}^2 \cdot \mathbf{1}(\max\{|\mathbf{A}_j^T\mathbf{x}^\star|, |A_{ji}|\} < \sqrt{44\log n})$, where $\mathbf{1}(\cdot)$ denotes the indicator function, and $Z_{2,j} = (\mathbf{A}_j^T\mathbf{x}^\star)^2 A_{ji}^2 - Z_{1,j}$. Since $Z_{1,j}$ is bounded, we can apply Theorem 2.3. To this end, compute the second moment

$$\sum_{j=1}^m \frac{1}{m^2}\mathbb{E}[Z_{1,j}^2] \le \sum_{j=1}^m \frac{1}{m^2}\mathbb{E}\Big[(\mathbf{A}_j^T\mathbf{x}^\star)^4 A_{ji}^4\Big] \le \sum_{j=1}^m \frac{1}{m^2}\sqrt{\mathbb{E}\Big[(\mathbf{A}_j^T\mathbf{x}^\star)^8\Big]\mathbb{E}\Big[A_{ji}^8\Big]} \le \frac{105}{m},$$

where we used the Cauchy-Schwarz inequality and the fact that $\mathbf{A}_j^T\mathbf{x}^\star \sim \mathcal{N}(0,1)$. With this, we have

$$\begin{aligned}\mathbb{P}\left[\left|\frac{1}{m}\sum_{j=1}^m Z_{1,j} - \mathbb{E}[Z_{1,j}]\right| > \frac{3}{8}(x^\star_{max})^2\right] &\le 2\exp\left(-\frac{\frac{9}{64}(x^\star_{max})^4}{2(\frac{105}{m} + \frac{44^2\log^2 n}{m}\cdot\frac{3}{8}(x^\star_{max})^2/3)}\right) \\ &\le \mathcal{O}(n^{-11})\end{aligned}$$

since $m \ge \mathcal{O}(\max\{k\log n, \log^3 n\}(x^\star_{max})^{-2})$.

For the second term $Z_{2,j}$, we can use the Chebyshev inequality: we have

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{m}\sum_{j=1}^m Z_{2,j}\right) &\le \frac{1}{m}\mathbb{E}\Big[(\mathbf{A}_1^T\mathbf{x}^\star)^4 A_{1i}^4 \cdot \mathbf{1}\Big(\max\{|\mathbf{A}_1^T\mathbf{x}^\star|, |A_{1i}|\} > \sqrt{44\log n}\Big)\Big] \\ &\le \frac{1}{m}\sqrt{\mathbb{E}\Big[(\mathbf{A}_1^T\mathbf{x}^\star)^8 A_{1i}^8\Big]\cdot\mathbb{P}\Big[\max\{|\mathbf{A}_1^T\mathbf{x}^\star|, |A_{1i}|\} > \sqrt{44\log n}\Big]} \\ &\le \frac{45\sqrt{1001}}{m}\cdot 2n^{-11},\end{aligned}$$

and hence, by the Chebyshev inequality,

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{j=1}^{m} Z_{2,j} - \mathbb{E}[Z_{2,j}]\right| > \frac{3}{8}(x^\star_{max})^2\right] \leq \frac{\frac{45\sqrt{1001}}{m}\cdot 2n^{-11}}{\frac{9}{64}(x^\star_{max})^4} \leq \mathcal{O}(n^{-11}).$$

Put together, this implies that

$$\mathbb{P}\left[|R_i - \mathbb{E}[R_i]| > \frac{3}{4}(x^\star_{max})^2\right] \leq \mathcal{O}(n^{-11}).$$

Taking the union bound over all $i \in [n]$ implies that $|R_i - E[R_i]| \leq \frac{3}{4}(x^\star_{max})^2$ holds for all $i \in [n]$ with probability at least $1 - O(n^{-10})$. This concludes the proof of the first claim of Lemma 2.1.

Proof of the second claim.

First, note that the gradient of the empirical risk $F(\mathbf{x})$ is given by

$$\nabla F(\mathbf{x}) = \frac{1}{m}\sum_{j=1}^{m}\Big((\mathbf{A}_j^T\mathbf{x})^2 - (\mathbf{A}_j^T\mathbf{x}^\star)^2\Big)(\mathbf{A}_j^T\mathbf{x})\mathbf{A}_j.$$

By the dominated convergence theorem, the gradient of the population risk $f(\mathbf{x})$ can then be computed as

$$\begin{aligned}\nabla f(\mathbf{x}) = \mathbb{E}[\nabla F(\mathbf{x})] &= \mathbb{E}\Big[\Big((\mathbf{A}_1^T\mathbf{x})^2 - (\mathbf{A}_1^T\mathbf{x}^\star)^2\Big)(\mathbf{A}_1^T\mathbf{x})\mathbf{A}_1\Big] \\ &= \Big(3\|\mathbf{x}\|_2^2 - 1\Big)\mathbf{x} - 2\Big(\mathbf{x}^T\mathbf{x}^\star\Big)\mathbf{x}^\star\end{aligned}$$

for any fixed vector $\mathbf{x} \in \mathbb{R}^n$. Further, we have the initialization

$$\begin{aligned}U_i^0 &= \begin{cases}\left(\frac{\hat{\theta}}{\sqrt{3}} + \alpha^2\right)^{\frac{1}{2}} & i = I_{max} \\ \alpha & i \neq I_{max}\end{cases} \\ V_i^0 &= \alpha\end{aligned}$$

which leads to

$$X_i^0 = \begin{cases}\frac{\hat{\theta}}{\sqrt{3}} & i = I_{max} \\ 0 & i \neq I_{max}\end{cases}$$

Hence, we have

$$\nabla f(\mathbf{X}^0)_i = \Big(\hat{\theta}^2 - 1\Big)X_i^0 - \frac{2\hat{\theta}}{\sqrt{3}}x^\star_{I_{max}}x^\star_i.$$

In particular, we have $\nabla f(\mathbf{X}^0)_j = 0$ for $j \notin \mathcal{S}$. Using standard concetration bounds for sub-exponential random variables (see e.g. Proposition 5.16 of [115]), we can bound

with probability $1 - \mathcal{O}(n^{-10})$ (recall that we have assumed $\|\mathbf{x}^\star\|_2 = 1$ for notational simplicity),

$$\left|\hat{\theta}^2 - 1\right| = \left|\frac{1}{m}\sum_{j=1}^{m}(\mathbf{A}_j^T\mathbf{x}^\star)^2 - 1\right| \leq 9\sqrt{\frac{\log n}{m}}.$$

This bound implies $2\hat{\theta} \geq \sqrt{3}$, where we used $m \geq \mathcal{O}(\max\{k\log n, \log^3 n\}(x^\star_{max})^{-2})$.

For the second claim we need to show that $|X_i^1| > |X_j^1|$ holds whenever $i \in \mathcal{S}$ and $j \notin \mathcal{S}$. We can assume without loss of generality that $x^\star_{I_{max}} > 0$. First, consider the case $i \neq I_{max}$. Since $|X_i^1| = |(U_i^1)^2 - (V_i^1)^2|$, it suffices to show, assuming $x_i^\star > 0$, that

$$U_i^1 > \max\{U_j^1, V_j^1\} \tag{2.6}$$

$$V_i^1 < \min\{U_j^1, V_j^1\} \tag{2.7}$$

holds simultaneously. The case $x_i^\star < 0$ can be dealt with the same way, exchanging the roles of U_i^1 and V_i^1. We can bound

$$\begin{aligned} U_i^1 &= \alpha\Big(1 - 2\eta\nabla F(\mathbf{X}^0)_i\Big) \\ &\geq \alpha\Big(1 - 2\eta\nabla f(\mathbf{X}^0)_i - 2\eta|\nabla F(\mathbf{X}^0)_i - \nabla f(\mathbf{X}^0)_i|\Big), \end{aligned}$$

and

$$U_j^1 \leq \alpha\Big(1 + 2\eta|\nabla F(\mathbf{X}^0)_j - \nabla f(\mathbf{X}^0)_j|\Big).$$

We have shown above that (recall that $X_i^0 = 0$ for $i \neq I_{max}$)

$$-\nabla f(\mathbf{X}^0)_i = \frac{2\hat{\theta}}{\sqrt{3}}x^\star_{I_{max}}x_i^\star \geq \frac{1}{2}x^\star_{max}x^\star_{min},$$

since from the first part we know that $x^\star_{I_{max}} \geq \frac{1}{2}x^\star_{max}$ and we assumed $x_i^\star > 0$. Hence, if we can show

$$\max_i |\nabla F(\mathbf{X}^0)_i - \nabla f(\mathbf{X}^0)_i| \leq \frac{1}{4}x^\star_{max}x^\star_{min}, \tag{2.8}$$

then $U_i^1 \geq U_j^1$ follows. We also have

$$V_j^1 \leq \alpha\Big(1 + 2\eta|\nabla F(\mathbf{X}^0)_j - \nabla f(\mathbf{X}^0)_j|\Big),$$

which then implies $U_i^1 \geq V_j^1$, completing the proof of (2.6); (2.7) can be shown the same way.

The case $i = I_{max}$ also follows from the bound (2.8). Since $m \geq \mathcal{O}(k(x^\star_{max})^{-2} \log n)$, we can bound

$$\begin{aligned}|\nabla F(\mathbf{X}^0)_i| &\leq |\nabla f(\mathbf{X}^0)_i| + |\nabla F(\mathbf{X}^0)_i - \nabla f(\mathbf{X}^0)_i| \\ &\leq 9\sqrt{\frac{\log n}{m}}\frac{\hat{\theta}}{\sqrt{3}} + \frac{2\hat{\theta}}{\sqrt{3}}(x^\star_{max})^2 + \frac{1}{2\sqrt{3}}x^\star_{max}x^\star_{min} \\ &\leq 2,\end{aligned}$$

where we used that $x^\star_{max} \leq 1$. Since we assume $\eta \leq 0.1$, we can bound

$$|2\eta\nabla F(\mathbf{X}^0)_i| \leq 0.4,$$

which, since also $\alpha \leq 0.1$, implies

$$U_i^1 \geq \left(\frac{\hat{\theta}}{\sqrt{3}} + \alpha^2\right)^{\frac{1}{2}}(1 - 0.4) \geq 2\max\{U_j^1, V_j^1, V_i^1\},$$

and hence

$$|X_i^1| = (U_i^1)^2 - (V_i^1)^2 \geq \max\left\{(U_j^1)^2, (V_j^1)^2\right\} \geq |X_j^1|.$$

What is left to show is (2.8). Since $\mathbf{X}^0$ is not independent from $\{\mathbf{A}_j\}_{j=1}^m$, we cannot immediately apply the truncation argument from the proof of the first claim. Therefore, define the (deterministic) vectors $\mathbf{x}^{(l)} \in \mathbb{R}^n$ for $l = 1, ..., n$ by

$$x_i^{(l)} = \begin{cases} \frac{1}{\sqrt{3}} & i = l \\ 0 & i \neq l \end{cases}$$

Now, we need to show that the empirical gradient

$$\nabla F(\mathbf{x}^{(l)})_i = \frac{1}{m}\sum_{j=1}^m ((\mathbf{A}_j^T\mathbf{x}^{(l)})^2 - (\mathbf{A}_j^T\mathbf{x}^\star)^2)(\mathbf{A}_j^T\mathbf{x}^{(l)})A_{ji}$$

is close to its expectation $\nabla f(\mathbf{x}^{(l)})_i$. Using the same truncation argument as in the proof of the first claim, we can show

$$\mathbb{P}\left[\left|\nabla f(\mathbf{x}^{(l)})_i - \nabla F(\mathbf{x}^{(l)})_i\right| \geq \frac{1}{8}x^\star_{max}x^\star_{min}\right] \leq \mathcal{O}(n^{-12}),$$

Taking the union bound over all i and l implies that

$$\max_l \max_i |\nabla F(\mathbf{x}^{(l)})_i - \nabla f(\mathbf{x}^{(l)})_i| \leq \frac{1}{8}x^\star_{max}x^\star_{min}$$

holds with probability $1 - \mathcal{O}\left(n^{-10}\right)$. The bound (2.8) now follows since $\mathbf{X}^0$ is close to $\mathbf{x}^{(I_{max})}$. We can write

$$\begin{aligned}\left|\nabla F(\mathbf{X}^0) - \nabla F(\mathbf{x}^{(I_{max})})\right| \leq &\left|\frac{1}{m}\sum_{j=1}^{m} A_{ji}\Big((\mathbf{A}_j^T\mathbf{X}^0)^3\Big) - \Big(\mathbf{A}_j^T\mathbf{x}^{(I_{max})}\Big)^3\right| \\ &+ \left|\frac{1}{m}\sum_{j=1}^{m} A_{ji}\Big(\mathbf{A}_j^T\mathbf{x}^\star\Big)^2\Big(\mathbf{A}_j^T\Big(\mathbf{X}^0 - \mathbf{x}^{(I_{max})}\Big)\Big)\right|.\end{aligned}$$

As both terms can be bounded the same way, we only demonstrate the following computations for the first term. Using the definitions and Hölder's inequality, we can bound

$$\begin{aligned}\left|\frac{1}{m}\sum_{j=1}^{m} A_{ji}\Big((\mathbf{A}_j^T\mathbf{X}^0)^3\Big) - \Big(\mathbf{A}_j^T\mathbf{x}^{(I_{max})}\Big)^3\right| &= \left|\frac{1}{m}\sum_{j=1}^{m} A_{ji}A_{jI_{max}}^3\frac{\hat{\theta}^3 - 1}{3\sqrt{3}}\right| \\ &\leq \frac{1}{m}\sum_{j=1}^{m}\left|A_{ji}A_{jI_{max}}^3\right| \cdot \left|\frac{\hat{\theta}^3 - 1}{3\sqrt{3}}\right| \\ &\leq \left(\frac{1}{m}\sum_{j=1}^{m} A_{ji}^4\right)^{1/4}\left(\frac{1}{m}\sum_{j=1}^{m} A_{jI_{max}}^4\right)^{3/4}\left|\frac{\hat{\theta}^3 - 1}{3\sqrt{3}}\right|.\end{aligned}$$

It follows from standard Gaussian concentration that the first two sums are bounded by $\mathcal{O}(1)$ with high probability. As shown above, we can bound

$$\left|\frac{\hat{\theta}^3 - 1}{3\sqrt{3}}\right| \leq \mathcal{O}\left(\sqrt{\frac{\log n}{m}}\right) \leq \mathcal{O}(x_{max}^\star x_{min}^\star),$$

where we used the assumption $x_{min}^\star = \Omega(1/\sqrt{k})$. Repeating the same computation for the second term, we can show that

$$\left|\nabla F(\mathbf{X}^0) - \nabla F(\mathbf{x}^{(I_{max})})\right| \leq \frac{1}{16}x_{max}^\star x_{min}^\star.$$

Recalling the definition of the population gradient ∇f, we can also bound

$$\left|\nabla f(\mathbf{X}^0) - \nabla f(\mathbf{x}^{(I_{max})})\right| \leq \frac{1}{16}x_{max}^\star x_{min}^\star,$$

which completes the proof of (2.8) and therefore also completes the proof of Lemma 2.1. □

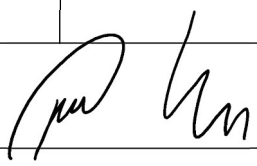
Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

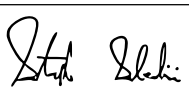
Title of Paper	Hadamard Wirtinger flow for sparse phase retrieval
Publication Status	□Published
Publication Details	"Hadamard Wirtinger flow for sparse phase retrieval", Fan Wu, Patrick Rebeschini. In International Conference on Aritificial Intelligence and Statistics, 2021

Student Confirmation

Student Name:	Fan Wu		
Contribution to the Paper	Derived technical contributions of the manuscript and coded the numerical experiments. Wrote first draft of the manuscript, and wrote later versions of the paper alongside Patrick Rebeschini. Wrote first draft of the response to the reviewers, and implemented feedback from the reviewers.		
Signature		Date	29/12/2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Patrick Rebeschini, Associate Professor		
Supervisor comments			
Signature		Date	30/12/2021

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 3

Nearly minimax-optimal rates for noisy sparse phase retrieval via early-stopped mirror descent

This paper studies early-stopped mirror descent applied to noisy sparse phase retrieval, which is the problem of recovering a k-sparse signal $\mathbf{x}^\star \in \mathbb{R}^n$ from a set of quadratic Gaussian measurements corrupted by sub-exponential noise. We consider the (non-convex) unregularized empirical risk minimization problem and show that early-stopped mirror descent, when equipped with the hypentropy mirror map and proper initialization, achieves a nearly minimax-optimal rate of convergence, provided the sample size is at least of order k^2 (modulo logarithmic term) and the minimum (in modulus) non-zero entry of the signal is on the order of $\|\mathbf{x}^\star\|_2/\sqrt{k}$. Our theory leads to a simple algorithm that does not rely on explicit regularization or thresholding steps to promote sparsity. More generally, our results establish a connection between mirror descent and sparsity in the non-convex problem of noisy sparse phase retrieval, adding to the literature on early stopping that has mostly focused on non-sparse, Euclidean, and convex settings via gradient descent. Our proof combines a potential-based analysis of mirror descent with a quantitative control on a variational coherence property that we establish along the path of mirror descent, up to a prescribed stopping time.

3.1 Introduction

In many fields of science and engineering, one is tasked with phase retrieval, the problem of recovering an n-dimensional signal $\mathbf{x}^\star$ from phaseless measurements

$$Y_j = (\mathbf{A}_j^\top \mathbf{x}^\star)^2 + \varepsilon_j, \qquad j = 1, ..., m, \tag{3.1}$$

where the sensing vectors $\mathbf{A}_1, ..., \mathbf{A}_m$ are observed, and the random variables $\varepsilon_1, ..., \varepsilon_m$ model a possible contamination of the measurements with noise. This problem arises naturally in applications such as optics, X-ray crystallography and astronomy, where it is often easier to build detectors which measure intensities and not phases [101]. In many applications of interest, the underlying signal $\mathbf{x}^\star$ is naturally sparse [36, 57, 85]. (Noisy) sparse phase retrieval represents a cornerstone in high-dimensional statistics. It is more challenging than sparse linear regression due to the missing phase information, which often leads to non-convex problems. A geometric analysis of a natural least-squares formulation of phase retrieval (without sparsity and noise) shows that, with high probability, there are no spurious local minimizers, provided the number of Gaussian measurements m is at least of order n modulo logarithmic terms [106]. However, recovery of a sparse signal is possible from potentially far fewer measurements $m \ll n$, in which case the least-squares formulation studied in [106] has potentially exponentially many local minimizers.

Over the years, a variety of methods have been proposed to tackle the problem of sparse phase retrieval. These methods typically enforce sparsity either by adding explicit regularization via penalty terms [31, 91], sparsity-inducing priors [102] or sparsity constraints [60, 90, 100], or by using algorithmic principles such as thresholding steps [20, 122, 138]. The noisy model has been studied and stability results have been established in [31] for a convex relaxation-based formulation, which includes an ℓ_1-penalty term to promote sparsity, and in [60] for an alternating minimization-based approach that uses an explicit sparsity constraint to enforce sparsity. However, the optimal dependence of the estimation error rate on the sample size, the sparsity of the signal and the noise level was not considered in aforementioned works. Empirical risk minimization with sparsity constraint has been shown to achieve a nearly minimax-optimal rate of convergence in [71], though it does not lead to a tractable algorithm

due to the non-convex sparsity constraint. Thresholded Wirtinger flow (TWF) [20] is a non-convex optimization-based algorithm that relies on thresholding steps to enforce sparsity, and has been shown to achieve a nearly minimax-optimal rate of convergence for noisy sparse phase retrieval, provided the number of measurements m is of order k^2 modulo logarithmic terms.

Mirror descent was first introduced by Nemirovski and Yudin [89] for solving large-scale convex optimization problems, and has since gained in popularity in various optimization and machine learning settings. An appealing property of mirror descent is the fact that it can be adapted to the (possibly non-Euclidean) geometry of the problem at hand by choosing a suitable strictly convex function, the so-called mirror map. In convex optimization, the algorithm often admits a potential-based convergence analysis in terms of the Bregman divergence associated to the mirror map, see e.g. [15, 17, 88]. There is a large body of literature that analyzes mirror descent in optimization (both convex and non-convex) and online learning settings, see e.g. [8, 46, 103]. In particular, an algorithm based on online mirror descent has been studied for an online version of phase retrieval (without sparsity) [70]. In the case of *noiseless* sparse phase retrieval, an approach based on continuous-time mirror descent has been recently shown to lead to recovery of the signal up to an arbitrary precision, without requiring explicit regularization terms or thresholding steps [128]. These results, however, do not apply to noisy measurements where perfect recovery of the signal is no longer possible and where the iterates of mirror descent may cease to be (approximately) sparse as time increases.

In the present work, we build upon the framework considered in [128] and we present a full analysis of *early-stopped* mirror descent (both continuous-time and discrete-time), equipped with the hypentropy mirror map, applied to minimize the (non-convex) unregularized empirical risk to solve *noisy* sparse phase retrieval. We establish the lower bound $\frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{k/m}$ for the minimax-optimal rate of convergence for noisy sparse phase retrieval under sub-exponential noise for signals $\mathbf{x}^\star$ with minimum non-zero entry (in modulus) on the order of $\|\mathbf{x}^\star\|_2/\sqrt{k}$, where σ describes the noise level. We prove that early-stopped mirror descent achieves this rate up to a factor $\sqrt{\log n}$, provided the sample size is on the order of k^2 (modulo logarithmic term) and the sensing vectors

$\mathbf{A}_1, ..., \mathbf{A}_m$ are independent standard Gaussian vectors. Up to logarithmic terms, the sample size requirement matches that of existing results in the literature on (noisy) sparse phase retrieval [20, 31].

The proposed approach yields a simple algorithm for noisy sparse phase retrieval, which, unlike most existing algorithms, does not rely on added regularization terms or algorithmic principles such as thresholding steps to enforce sparsity of the estimates. Running mirror descent involves tuning at most two parameters, namely the step size η (only for the discrete-time version of the algorithm) and the mirror map parameter β. Our analysis shows that η should be chosen smaller than a quantity depending on the signal size $\|\mathbf{x}^\star\|_2$, while the choice of β depends on $\|\mathbf{x}^\star\|_2$ and the (known) ambient dimension n. As the signal size $\|\mathbf{x}^\star\|_2$ can be easily estimated by considering the average size of the observations [22, 121], the tuning of η and β only involves known or easily estimated quantities, and, in particular, does not require knowledge (or estimation) of the sparsity k or the noise level σ. The optimal stopping time, whose knowledge is not required a priori to run mirror descent, is the only parameter that depends on the unknown quantities k and σ. Our numerical simulations attest that a commonly-used data-dependent stopping rule based on cross-validation, where no knowledge of k or σ is used, yields results that validate our theoretical findings.

The main idea behind the proof of our results is to combine a generic potential-based analysis of mirror descent in terms of the Bregman divergence with a quantitative control of a variational coherence property that we establish along the path of mirror descent, when properly initialized, up to a prescribed stopping time. Variational coherence has been previously used as an *a-priori* assumption to establish convergence results for mirror descent in non-convex optimization. As defined in [142], variational coherence corresponds to the requirement that the inner product $\langle \nabla F(\mathbf{x}), \mathbf{x} - \mathbf{x}' \rangle$ is non-negative for any vector $\mathbf{x}$ in a certain region, where F is the objective function to be minimized and $\mathbf{x}'$ is a minimizer of F. Two versions of variational coherence have been considered in [142]: a global property that holds for any vector $\mathbf{x}$ and any global minimizer $\mathbf{x}'$, and a local property that only needs to be satisfied for any vector $\mathbf{x}$ in an open neighborhood of a local minimizer $\mathbf{x}'$. The global property precludes the existence of local minimizers (that are not also global minimizers) and saddle points,

and is not satisfied in the problem of sparse phase retrieval when F is taken to be the unregularized empirical risk. Further, both the global and local versions of variational coherence are formulated to yield convergence towards local minimizers of the objective function, which is not the goal in noisy sparse phase retrieval as, almost surely, the signal $\mathbf{x}^\star$ does not coincide with a local minimizer $\mathbf{x}'$ of the empirical risk. Moreover, in these formulations, variational coherence only yields convergence results but it does not yield a direct control on the speed of convergence. In contrast, our analysis unveils and exploits a *quantitative* control on variational coherence within a region where the trajectory of mirror descent lies up to the stopping time, upon proper initialization. The object of our analysis is given by the inner product $\langle \nabla F(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \rangle$, where $\mathbf{X}(t)$ represents the iterate of mirror descent at time t and where the signal $\mathbf{x}^\star$ now replaces the local minimizer $\mathbf{x}'$ of F. In our analysis, we establish a lower bound for this inner product that is strictly positive and directly depends on the iterates of the algorithm. This adaptive lower bound yields a control on the rate of convergence of mirror descent, which is key to investigate early stopping and establish our results. In existing gradient-based approaches to sparse phase retrieval, the inclusion of thresholding steps confines the algorithm to the low-dimensional subspace of sparse vectors, see e.g. [20, 122, 135]. On the other hand, our analysis shows that, when properly initialized, the iterates of mirror descent have negligibly small off-support coordinates until a certain stopping time, allowing us to focus on the subspace of (approximately) sparse vectors.

Our work adds to the literature on early stopping for iterative methods, which has been primarily developed in the context of ridge regression and kernel methods [95, 124, 134, 139], i.e. convex problems based on the Euclidean geometry which, in particular, do not involve sparsity. An exception is the work [113], where a connection between early stopping and sparsity has been established for gradient descent with Hadamard parametrization applied to the (convex) problem of sparse recovery. This result was later recovered in [114] using mirror descent with the hypentropy mirror map. Our results extend this line of research by establishing a connection between early stopping and sparsity for mirror descent applied to the non-convex problem of noisy sparse phase retrieval. Our approach shows how to apply the generic potential-based analysis of mirror descent to non-convex problems where a quantitative control on variational coherence is exhibited along the path traced by the algorithm.

The remainder of the paper is organized as follows. In Section 3.2, we give the required background and a brief literature review on sparse phase retrieval. In Section 3.3, we describe our approach to solving sparse phase retrieval using mirror descent. In Section 3.4, we present our main results for mirror descent in continuous time and in discrete time. In Section 3.5, we perform a numerical study to verify that the dependence of the estimation error achieved by mirror descent with respect to the noise level σ, sample size m and sparsity k can match the behavior prescribed by our upper bounds. In Section 3.6, we present the proof of the results for the continuous-time algorithm. The proof of the results for the discrete-time algorithm involves further technicalities, and we defer it to the Appendix. Code to reproduce our numerical study can be found at `https://github.com/fawuuu/esmd_nspr`.

Notation. We use boldface letters for vectors and matrices, normal font for real numbers, and, generally, uppercase letters for random and lowercase letters for deterministic quantities. For any number $n \in \mathbb{N}$, we write $[n] = \{1, ..., n\}$. For any vector $\mathbf{x} \in \mathbb{R}^n$ and $p \geq 0$, we write $\|\mathbf{x}\|_p = (\sum_{i=1}^n |x_i|^p)^{1/p}$, which is the standard ℓ_p-norm for $p \geq 1$. For a random variable X, we write $\|X\|_{\psi_1} = \sup_{p \geq 1} p^{-1}(\mathbb{E}[|X|^p])^{1/p}$ for its sub-exponential norm. We denote by $\mathcal{S} = \{i \in [n] : x_i^\star \neq 0\}$ the support of $\mathbf{x}^\star$. For any $\mathbf{x} \in \mathbb{R}^n$ and subset of coordinates $\mathcal{C} \subseteq [n]$, we write $\mathbf{x}_\mathcal{C} = (x_i)_{i \in \mathcal{C}} \in \mathbb{R}^{|\mathcal{C}|}$, and, similarly for $i \in [n]$, we write $\mathbf{x}_{-i} = (x_j)_{j \in [n] \setminus \{i\}} \in \mathbb{R}^{n-1}$. The notation $f(n) \lesssim g(n)$ (resp. $f(n) \gtrsim g(n)$, $f(n) \asymp g(n)$) means that there are constants $c_1, c_2 > 0$ such that $f(n) \leq c_1 g(n)$ (resp. $f(n) \geq c_2 g(n)$, $c_2 g(n) \leq f(n) \leq c_1 g(n)$).

3.2 Sparse phase retrieval

The goal in sparse phase retrieval is to reconstruct an unknown k-sparse signal vector $\mathbf{x}^\star \in \mathbb{R}^n$ from a set of (possibly noisy) quadratic measurements $Y_j = (\mathbf{A}_j^\top \mathbf{x}^\star)^2 + \varepsilon_j$, $j = 1, ..., m$. We consider the standard setting where the sensing vectors $\mathbf{A}_j \sim \mathcal{N}(0, \mathbf{I}_n)$ are independent standard Gaussian vectors, and the noise terms ε_j are independent centered sub-exponential random variables with maximum sub-exponential norm $\sigma = \max_{j \in [m]} \|\varepsilon_j\|_{\psi_1} \geq 0$.[1]

[1]For the clarity of the analytical results, we focus on noisy sparse phase retrieval with real signal and real Gaussian measurement vectors. Our results extend naturally to the complex case.

We follow the well-established approach to estimating the signal $\mathbf{x}^\star$ based on non-convex optimization [20, 122, 135, 138]. Given observations $\{\mathbf{A}_j, Y_j\}_{j=1}^m$, we consider the following (non-convex) unregularized empirical risk minimization problem:

$$F(\mathbf{x}) = \frac{1}{4m} \sum_{j=1}^{m} \Big((\mathbf{A}_j^\top \mathbf{x})^2 - Y_j \Big)^2. \tag{3.2}$$

It is worth mentioning that a risk function based on amplitude measurements $|\mathbf{A}_j^\top \mathbf{x}^\star|$ has also been considered [122, 138]. However, the objective function becomes non-smooth in that case, and an analysis via the mirror descent framework appears more challenging.

Without any restrictions, the non-convex function F in (3.2) could potentially have exponentially many local minimizers and saddle points if $m \lesssim n$, cf. [106]. In the noiseless case, it has been shown in [73] that $m \geq 4k - 1$ Gaussian measurements suffice for $\{\pm\mathbf{x}^\star\}$ to be the sparsest minimizer of F with high probability, that is

$$\{\pm\mathbf{x}^\star\} = \operatorname{argmin}_{\mathbf{x}:\|\mathbf{x}\|_0 \leq k} F(\mathbf{x}). \tag{3.3}$$

However, solving (3.3) is challenging due to the non-convexity of the objective F and the combinatorial nature of the sparsity constraint $\|\mathbf{x}\|_0 \leq k$. Existing gradient-based methods such as thresholded Wirtinger flow (TWF) [20], sparse truncated amplitude flow (SPARTA) [122], compressive reweighted amplitude flow (CRAF) [138] and sparse Wirtinger flow (SWF) [135] minimize the non-convex empirical risk by employing a non-trivial diagonal thresholding, orthogonality-promoting or spectral initialization scheme, which produces an initial estimate close to the signal (up to a global sign) $\{\pm\mathbf{x}^\star\}$, inside the so-called basin of attraction. This is followed by a refinement procedure via thresholded gradient descent iterations, where the thresholding steps are necessary to enforce sparsity of the iterates.

Numerous other algorithms have been developed to solve sparse phase retrieval, which rely on a strategy other than thresholding steps to exploit sparsity. One such strategy is to include an ℓ_1-regularization term to the objective function, which is commonly found in convex relaxation-based approaches such as compressive phase retrieval via lifting (CPRL) [91] and SparsePhaseMax [55], but also in the generalized message passing algorithm PR-GAMP [102], which includes a sparsity-promoting prior. Alternatively,

one can directly restrict the search to k-sparse vectors, either with a preliminary support recovery step as in the alternating minimization algorithm SparseAltMinPhase [90], or by updating the estimated support in every iteration of the search in a greedy fashion as in GESPAR [100].

Hadamard Wirtinger flow (HWF) [130] is an algorithm that performs gradient descent on the unregularized empirical risk (3.2) using the Hadamard parametrization, and it can be recovered as a first-order approximation to mirror descent equipped with the hypentropy mirror map [114, 128]. HWF has been empirically studied in [130], and in some settings it has been shown to exhibit favorable sample complexities compared to other gradient-based approaches to sparse phase retrieval in numerical simulations.

3.3 Mirror descent algorithms

Our proposed approach to solving noisy sparse phase retrieval consists of running unconstrained mirror descent equipped with the hypentropy mirror map to minimize the empirical risk (3.2). In particular, when initialized as described below, we require neither an added regularization term nor thresholding steps to enforce sparsity of the estimates. We begin by giving a brief overview of the main quantities and definitions for mirror descent. The key object defining the geometry of mirror descent is the *mirror map*.

Definition 3.1. *Let $\mathcal{D} \subseteq \mathbb{R}^n$ be a convex open set. We say that $\Phi : \mathcal{D} \to \mathbb{R}$ is a mirror map if it is strictly convex, differentiable and $\{\nabla\Phi(\mathbf{x}) : \mathbf{x} \in \mathcal{D}\} = \mathbb{R}^n$.*

A central quantity in the analysis of mirror descent is the *Bregman divergence* associated to a mirror map Φ, which is a measure of distance between two points $\mathbf{x}, \mathbf{y} \in \mathcal{D}$ given by

$$D_\Phi(\mathbf{x}, \mathbf{y}) = \Phi(\mathbf{x}) - \Phi(\mathbf{y}) - \nabla\Phi(\mathbf{y})^\top(\mathbf{x} - \mathbf{y}).$$

As a special case, we obtain the squared Euclidean distance by choosing the mirror map $\Phi(\mathbf{x}) = \|\mathbf{x}\|_2^2$, in which case mirror descent coincides with standard gradient descent.

As we consider unconstrained mirror descent, we have $\mathcal{D} = \mathbb{R}^n$. Let $F : \mathbb{R}^n \to \mathbb{R}$ be the objective function we seek to minimize. The trajectory of mirror descent is

characterized by the mirror map Φ and an initial point $\mathbf{X}(0)$, and is, in continuous time, defined by the ordinary differential equation [89]

$$\frac{d}{dt}\mathbf{X}(t) = -\Big(\nabla^2\Phi(\mathbf{X}(t))\Big)^{-1}\nabla F(\mathbf{X}(t)). \tag{3.4}$$

In discrete time, the mirror descent update is given by

$$\nabla\Phi(\mathbf{X}^{t+1}) = \nabla\Phi(\mathbf{X}^t) - \eta\nabla F(\mathbf{X}^t), \tag{3.5}$$

where $\eta > 0$ is the step-size. Throughout, we write $\mathbf{X}(t)$ and $\mathbf{X}^t$ for the iterates of mirror descent in continuous and discrete time, respectively.

For the mirror map, we choose the hypentropy mirror map [48] given by

$$\Phi(\mathbf{x}) = \sum_{i=1}^{n}\bigg(x_i \operatorname{arcsinh}\Big(\frac{x_i}{\beta}\Big) - \sqrt{x_i^2+\beta^2}\bigg), \tag{3.6}$$

for some parameter $\beta > 0$. We provide a discussion on the choice of the parameter β in Section 3.4. This choice of mirror map is motivated by the fact that, equipped with this mirror map, mirror descent is approximated by gradient descent with Hadamard parametrization [48, 114], which has been studied in the recovery of low-rank structures in sparse recovery [56, 113, 140] and matrix factorization [6, 54, 74].

For the initialization, we follow the approach outlined in [130], which estimates a single coordinate on the support of the signal $\mathbf{x}^\star$ and initializes all other coordinates to zero,

$$X_i(0) \equiv X_i^0 = \begin{cases} \frac{1}{\sqrt{3}}\sqrt{\frac{1}{m}\sum_{j=1}^m Y_j} & i = I_0 \\ 0 & i \neq I_0 \end{cases}, \qquad I_0 \in \arg\max_i \bigg\{\frac{1}{m}\sum_{j=1}^m Y_j A_{ji}^2\bigg\}. \tag{3.7}$$

The term $(\frac{1}{m}\sum_{j=1}^m Y_j)^{1/2}$ is an estimate of the magnitude $\|\mathbf{x}^\star\|_2$, see e.g. [22, 121], and Lemma 1 of [130] guarantees in the noiseless case that $x^\star_{I_0} \neq 0$ with high probability, provided $m \gtrsim (x^\star_{max})^{-2}\max\{k\log n, \log^3 n\}$, where $x^\star_{max} = \max_i \frac{|x_i^\star|}{\|\mathbf{x}^\star\|_2}$. The following is an immediate extension of Lemma 1 of [130] to the noisy observation model.

Lemma 3.1. *Let $\mathbf{x}^\star \in \mathbb{R}^n$ be any k-sparse vector, and let the observations $\{Y_j, \mathbf{A}_j\}_{j=1}^m$ be given as in (3.1), where $\mathbf{A}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ i.i.d. and $\{\varepsilon_j\}_{j=1}^m$ are independent centered sub-exponential random variables with maximum sub-exponential norm $\sigma = \max_j \|\varepsilon_j\|_{\psi_1}$. Let $I_0 \in \arg\max_i\{\frac{1}{m}\sum_{j=1}^m Y_j A_{ji}^2\}$. There exist universal constants $c_s, c_p > 0$ such that the following holds. If $m \geq c_s(1 + \frac{\sigma^2}{\|\mathbf{x}^\star\|_2^4})(x^\star_{max})^{-2}\max\{k\log n, \log^3 n\}$, then, with probability at least $1 - c_p n^{-10}$ we have $|x^\star_{I_0}| \geq \frac{1}{2}x^\star_{max}\|\mathbf{x}^\star\|_2$.*

Proof. By Lemma 3.13 in Appendix 3.8.3, $|\frac{1}{m}\sum_{j=1}^{m}\varepsilon_j A_{ji}^2| \le c\sigma\sqrt{\frac{\log n}{m}}$ holds with probability $1 - c_p n^{-10}$, where $c > 0$ is an absolute constant. The rest of the proof follows the same steps as the proof of Lemma 1 in [130], and we omit the details. □

3.4 Main results

We show that, with high probability, early-stopped mirror descent recovers any k-sparse signal $\mathbf{x}^\star \in \mathbb{R}^n$ with $x^\star_{min} \gtrsim 1/\sqrt{k}$ from $(1 + \frac{\sigma^2}{\|\mathbf{x}^\star\|_2^4})k^2$ (modulo logarithmic term) Gaussian measurements, where we write $x^\star_{min} = \min_{i:x_i^\star \neq 0} \frac{|x_i^\star|}{\|\mathbf{x}^\star\|_2}$.

We begin by characterizing the relationship between the Bregman divergence $D_\Phi(\mathbf{x}^\star, \mathbf{x})$ associated to the hypentropy mirror map Φ in (3.6) and the ℓ_2-norm $\|\mathbf{x} - \mathbf{x}^\star\|_2$.

Lemma 3.2 ([128]). *Let $\mathbf{x}^\star \in \mathbb{R}^n$ be any k-sparse vector with $x^\star_{min} \ge c_\star/\sqrt{k}$ for a constant $c_\star > 0$. Let $\mathcal{S} = \{i \in [n] : x_i^\star \neq 0\}$ be its support, and let Φ be as in (3.6) with parameter $\beta > 0$.*

- *For any vector $\mathbf{x} \in \mathbb{R}^n$, we have*

$$\|\mathbf{x} - \mathbf{x}^\star\|_2^2 \le 2\sqrt{\max\{\|\mathbf{x}\|_\infty^2, \|\mathbf{x}^\star\|_\infty^2\} + \beta^2}\, D_\Phi(\mathbf{x}^\star, \mathbf{x}). \tag{3.8}$$

- *Let $\mathbf{x} \in \mathbb{R}^n$ be any vector with $x_i x_i^\star \ge 0$ (no mismatched sign) and $|x_i| \ge \frac{1}{2}|x_i^\star|$ for all $i = 1, ..., n$. Then, we have*

$$D_\Phi(\mathbf{x}^\star, \mathbf{x}) \le \frac{\sqrt{k}}{c_\star\|\mathbf{x}^\star\|_2}\|\mathbf{x}_\mathcal{S} - \mathbf{x}^\star_\mathcal{S}\|_2^2 + \|\mathbf{x}_{\mathcal{S}^c}\|_1. \tag{3.9}$$

A proof of Lemma 3.2 appears in the appendix of [129]. For completeness, we include a proof of Lemma 3.2 in Section 3.6.1 below. Lemma 3.2 implies that, if we are interested in convergence with respect to the ℓ_2-norm, we can consider the Bregman divergence D_Φ as a proxy for it. As it is impossible to distinguish $\mathbf{x}^\star$ from $-\mathbf{x}^\star$ using phaseless measurements, we consider the ℓ_2-distance and Bregman divergence from the solution set $\{\pm\mathbf{x}^\star\}$ given by $\mathrm{dist}(\mathbf{x}^\star, \mathbf{x}) = \min\{\|\mathbf{x} - \mathbf{x}^\star\|_2, \|\mathbf{x} + \mathbf{x}^\star\|_2\}$ and $\mathrm{dist}_\Phi(\mathbf{x}^\star, \mathbf{x}) = \min\{D_\Phi(\mathbf{x}^\star, \mathbf{x}), D_\Phi(-\mathbf{x}^\star, \mathbf{x})\}$, respectively. Applying the bound in (3.8) to both $\mathbf{x}^\star$ and $-\mathbf{x}^\star$, we find the upper bound

$$\mathrm{dist}(\mathbf{x}^\star, \mathbf{x})^2 \le 2\sqrt{\max\{\|\mathbf{x}\|_\infty^2, \|\mathbf{x}^\star\|_\infty^2\} + \beta^2}\, \mathrm{dist}_\Phi(\mathbf{x}^\star, \mathbf{x}).$$

The setting for our convergence results is summarized in the following assumption.

Assumption I. *For some universal constants $c_\star, c_s > 0$, the following holds. The signal $\mathbf{x}^\star \in \mathbb{R}^n$ is k-sparse with $x^\star_{min} \geq c_\star/\sqrt{k}$. The observations $\{Y_j, \mathbf{A}_j\}_{j=1}^m$ are i.i.d. given as in (3.1), with $\mathbf{A}_j \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ and ε_j centered sub-exponential with maximum sub-exponential norm $\sigma = \max_j \|\varepsilon_j\|_{\psi_1}$. The sample size m is at least $m \geq c_s(1 + \frac{\sigma^2}{\|\mathbf{x}^\star\|_2^4}) \max\{k^2 \log^2 n,\ \log^5 n\}$.*

We now formulate our results for continuous-time mirror descent in terms of the Bregman divergence D_Φ. Lemma 3.2 allows to translate the upper bound on $\mathrm{dist}_\Phi(\mathbf{x}^\star, \mathbf{X}(t))$ into an upper bound on the ℓ_2-distance $\mathrm{dist}(\mathbf{x}^\star, \mathbf{X}(t))$. Recall that we denote the support of the signal $\mathbf{x}^\star$ by $\mathcal{S} = \{i \in [n] : x_i^\star \neq 0\}$.

Theorem 3.3. *Let Assumption I hold. Then, there exist universal constants $c_p, c, c_1, c_2 > 0$ such that the following holds. Let $\mathbf{X}(t)$ be defined by the continuous-time mirror descent algorithm given by (3.4) with mirror map (3.6) and initialization (3.7) with $\beta \leq c_1\|\mathbf{x}^\star\|_2/n^3$. Let $\delta = \sqrt{n\beta/\|\mathbf{x}^\star\|_2}$, and define*

$$T_2 = \inf\left\{t > 0 : \frac{\mathrm{dist}(\mathbf{x}^\star, \mathbf{X}(t))}{\|\mathbf{x}^\star\|_2} \leq c \max\left\{\sqrt{c_\star\sqrt{k}\delta},\ \frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k\log n}{m}}\right\}\right\}.$$

Then, there exists $T_1 \leq \frac{c_2}{\|\mathbf{x}^\star\|_2^3} k \log(\frac{\|\mathbf{x}^\star\|_2}{\beta}) \log(k \log \frac{\|\mathbf{x}^\star\|_2}{\beta})$ such that

$$\frac{\mathrm{dist}_\Phi(\mathbf{x}^\star, \mathbf{X}(t))}{\|\mathbf{x}^\star\|_2} \leq \frac{6\sqrt{k}}{c_\star} \exp\left(-\frac{c_\star\|\mathbf{x}^\star\|_2^3}{4\sqrt{k}}(t - T_1)\right) \quad \textit{for all } T_1 \leq t \leq T_2, \tag{3.10}$$

with probability at least $1 - c_p n^{-10}$. Furthermore, for all $t \leq T_2$, we have

$$\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1 \leq \delta\|\mathbf{x}^\star\|_2. \tag{3.11}$$

Analogous results also hold in discrete time.

Theorem 3.4. *Let Assumption I hold. Then, there exist universal constants $c_p, c, c_1, c_2, c_3 > 0$ such that the following holds. Let $\mathbf{X}^t$ be defined by the discrete-time mirror descent algorithm given by (3.5) with mirror map (3.6) and initialization (3.7) with $\beta \leq c_1\|\mathbf{x}^\star\|_2/n^3$ and $\eta \leq c_3/\|\mathbf{x}^\star\|_2^3$. Let $\delta = \sqrt{n\beta/\|\mathbf{x}^\star\|_2}$, and define*

$$T_2 = \inf\left\{t > 0 : \frac{\mathrm{dist}(\mathbf{x}^\star, \mathbf{X}^t)}{\|\mathbf{x}^\star\|_2} \leq c \max\left\{\sqrt{c_\star\sqrt{k}\delta},\ \frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k\log n}{m}}\right\}\right\}.$$

Then, there exists $T_1 \leq \frac{c_2}{\eta\|\mathbf{x}^\star\|_2^3} k \log(\frac{\|\mathbf{x}^\star\|_2}{\beta}) \log(k \log \frac{\|\mathbf{x}^\star\|_2}{\beta})$ *such that*

$$\frac{\mathrm{dist}_\Phi(\mathbf{x}^\star, \mathbf{X}^t)}{\|\mathbf{x}^\star\|_2} \leq \frac{6\sqrt{k}}{c_\star}\left(1 - \frac{c_\star \eta \|\mathbf{x}^\star\|_2^3}{8\sqrt{k}}\right)^{t-T_1} \quad \text{for all } T_1 \leq t \leq T_2, \tag{3.12}$$

with probability at least $1 - c_p n^{-10}$. *Furthermore, for all* $t \leq T_2$, *we have*

$$\|\mathbf{X}^t_{\mathcal{S}^c}\|_1 \leq \delta \|\mathbf{x}^\star\|_2. \tag{3.13}$$

In the noiseless case, when $\sigma = 0$, Theorem 3.3 recovers the results in [128] and Theorem 3.4 shows that any desired relative error $\varepsilon > 0$ can be achieved by running mirror descent with the choice $\beta \asymp \varepsilon^2 \|\mathbf{x}^\star\|_2 / n^3$ for $k \log \frac{\|\mathbf{x}^\star\|_2}{\beta}$ (modulo double-logarithmic term) iterations.

We now discuss some of the main implications of Theorem 3.3 and Theorem 3.4. For notational simplicity, we focus on discrete-time mirror descent throughout the remainder of this section, but the same considerations also apply to the continuous-time algorithm.

Nearly minimax-optimal rate. Theorem 3.4 implies that mirror descent achieves a relative mean-squared error $\frac{\mathrm{dist}(\mathbf{X}^{T_2}, \mathbf{x}^\star)}{\|\mathbf{x}^\star\|_2}$ on the order of $\frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k \log n}{m}}$, provided the mirror map parameter β is chosen smaller than $\sigma^4 k \log^2 n / (m^2 n \|\mathbf{x}^\star\|_2^7)$. This is identical to the estimation error rate achieved by thresholded Wirtinger flow [20], and nearly matches the lower bound on the minimax-optimal rate which we derive below in Theorem 3.5. As we only consider k-sparse signals with minimum non-zero entry on the order of $\|\mathbf{x}^\star\|_2/\sqrt{k}$, our lower bound on the minimax rate of convergence differs by a factor $\sqrt{\log(en/k)}$ from the bound of $\frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k \log(en/k)}{m}}$ for k-sparse vectors, which has been stated as Theorem 3.2 in [20].

Theorem 3.5. *Let* $\Theta(k, n, r) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = r, \|\mathbf{x}\|_0 = k, \min_{i:x_i \neq 0} |x_i| \geq c_\star r/\sqrt{k}\}$ *for some constant* $c_\star > 0$. *Let* $\{\mathbf{A}_j\}_{j=1}^m$ *be a collection of i.i.d.* $\mathcal{N}(0, \mathbf{I}_n)$ *random vectors,* $\{\varepsilon_j\}_{j=1}^m$ *be a collection of i.i.d.* $\mathcal{N}(0, \sigma^2)$ *random variables, and let* $\{Y_j\}_{j=1}^m$ *be given as in (3.1). There exist universal constants* $c_s, c > 0$ *such that if* $m \geq c_s(1 + \frac{\sigma^2}{r^4}) k \log(en/k)$, *then*

$$\inf_{\widehat{\mathbf{x}}} \sup_{\mathbf{x}^\star \in \Theta(k,n,r)} \mathbb{P}_{(\mathbf{A}, \mathbf{Y}|\mathbf{x}^\star)}\left[\frac{\mathrm{dist}(\mathbf{x}^\star, \widehat{\mathbf{x}})}{r} \geq \frac{c\sigma}{r^2}\sqrt{\frac{k}{m}}\right] \geq \frac{1}{5}, \tag{3.14}$$

where the infimum is taken over all procedures $\widehat{\mathbf{x}}(\mathbf{A}, \mathbf{Y})$.

Proof. Theorem 3.5 follows from an application of Theorem C in [71]. We first introduce some notation used for Theorem C. For any vector $\mathbf{x} \in \mathbb{R}^n$ and $r > 0$, we write $\mathcal{B}_2(\mathbf{x}, \rho) \subseteq \mathbb{R}^n$ for the ℓ_2-ball with center $\mathbf{x}$ and radius ρ. For any set $\mathcal{X} \in \mathbb{R}^n$ and $\rho > 0$, we denote the packing number with respect to the ℓ_2-norm by

$$M(\mathcal{X}, \rho) = \sup\{n : \exists \mathbf{x}_1, ..., \mathbf{x}_n \in \mathcal{X} \text{ s.t } \forall i \neq j, \|\mathbf{x}_i - \mathbf{x}_j\|_2 > \rho\}.$$

To apply Theorem C, we need to bound the packing number of $\Theta(k, n, r) \cap \mathcal{B}_2(\mathbf{x}_0, c_0\rho)$, where $\mathbf{x}_0 \in \Theta(k, n, r)$ is an arbitrary vector. More precisely, we will show that, for all $\rho > 0$ and $c_0 \geq 2$,

$$\sup_{\mathbf{x}_0 \in \Theta(k,n,r)} \log^{1/2} M(\Theta(k, n, r) \cap \mathcal{B}_2(\mathbf{x}_0, c_0\rho), \rho) \asymp \sqrt{k}. \tag{3.15}$$

Theorem 3.5 then follows from an application of Theorem C of [71] (c.f. Corollary 6.2 in [71]).

In the minimax rate given in Corollary 6.2 in [71] and Theorem 3.2 in [20], the set $\Theta(n, k, r)$ is defined without the restriction $\min_{i:x_i \neq 0} |x_i| > c_\star r/\sqrt{k}$. In that case, standard estimates on the packing number yield

$$\log^{1/2} M(\widetilde{\Theta}(k, n, r) \cap \mathcal{B}_2(\mathbf{x}_0, c_0\rho), \rho) \asymp \sqrt{k \log(en/k)},$$

where we write $\widetilde{\Theta}(k, n, r) = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 = r, \|\mathbf{x}\|_0 = k\}$, see e.g. Lemma 1.4.2 and Lemma 2.2.17 in [26]). On the other hand, with the restriction $\min_{i:x_i \neq 0} |x_i| > c_\star r/\sqrt{k}$, when we fix any $\mathbf{x}_0 \in \Theta(k, n, r)$, then any vector $\mathbf{x} \in \Theta(k, n, r) \cap \mathcal{B}(\mathbf{x}_0, c_0\rho)$ must be supported on the support of $\mathbf{x}_0$ for $\rho < \sqrt{2}c_\star r/(c_0\sqrt{k})$, because any non-zero coordinate of $\mathbf{x}_0$ and $\mathbf{x}$ has absolute value at least $c_\star r/\sqrt{k}$, and therefore $\|\mathbf{x} - \mathbf{x}_0\|_2 \geq \sqrt{2}c_\star r/\sqrt{k}$ if $\mathbf{x}$ and $\mathbf{x}_0$ have different supports. This leads to a packing number that scales as c_0^k (as opposed to a scaling of $\binom{n}{k}$ in the case of $\widetilde{\Theta}(n, k, r)$), and therefore the estimate in (3.15). □

On early stopping and sparsity. The bound (3.13) in Theorem 3.4 controls the magnitude off-support coordinates can attain while mirror descent runs, up to a time T_2. To show the bound (3.13), we crucially use the assumption $x^\star_{min} \gtrsim 1/\sqrt{k}$, which is needed to show that all support coordinates grow faster than any off-support coordinate, which then leads to $\|\mathbf{X}^t_{\mathcal{S}^c}\|_1 < \delta\|\mathbf{x}^\star\|_2$ for a sufficiently small δ; see Section 3.4.1 for

an outline of the proof idea, and Stage (i), part (b) in Section 3.6.2 for details. The fact that mirror descent iterates stay approximately k-sparse allows us to show bound (3.12): after an initial "warm-up" period of length T_1, mirror descent converges linearly to the solution set $\{\pm \mathbf{x}^\star\}$ up to a time T_2, at which a precision determined by the mirror map parameter β and the quantity $\frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k \log n}{m}}$ is reached. The necessity for early stopping is explained by the fact that, as elaborated in the proof sketch below, we need to control the size of off-support variables $\|\mathbf{X}^t_{\mathcal{S}^c}\|_1$, which is guaranteed to be sufficiently small up to the time T_2.

Using the fact that the Bregman divergence decreases linearly for $T_1 \leq t \leq T_2$, a quick calculations shows that $T_2 - T_1 \lesssim \sqrt{k} \log \frac{k\|\mathbf{x}^\star\|_2}{\beta}$, see Stage (ii), part (a) in Section 3.6.2 for details. Therefore, an upper bound for T_2 is given by $k \log \frac{\|\mathbf{x}^\star\|_2}{\beta}$ (modulo double-logarithmic term). In applications, we can use a data-dependent stopping rule such as the hold-out method outlined in [95] to estimate T_2: we can run mirror descent using a fraction $\alpha \in (0,1)$ of the data and return the iterate $\mathbf{X}^t$ which minimizes the empirical risk F evaluated on the remaining $(1-\alpha)$ of the data. Section 3.5 presents a numerical study that also investigates this data-dependent stopping rule.[1]

Role of the mirror map parameter. In Theorem 3.4, choosing a small parameter β is needed to guarantee that off-support variables stay sufficiently small in the bound (3.13). The quantity δ in the bound (3.13) depends polynomially on β, while choosing a small β leads to the length of the initial warm-up period T_1 increasing as $\log \frac{\|\mathbf{x}^\star\|_2}{\beta}$. In practice, we can therefore simply choose a very small β (e.g. 10^{-20}), so that convergence is linear up to a precision on the order of $\frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k \log n}{m}}$. A similar trade-off between computational cost and statistical accuracy with respect to the size of the initialization has previously been observed in [113] in the setting of noisy sparse linear regression.

3.4.1 Proof ideas

We give a high-level overview of the main ideas behind the proofs of Theorem 3.3 and Theorem 3.4. We begin with the proof of Theorem 3.3. Without loss of generality, let us assume that the initialization in (3.7) satisfies $x^\star_{I_0} > 0$, so that $\mathbf{X}(0)^\top \mathbf{x}^\star > 0$. In this

[1] Establishing theoretical guarantees for data-dependent stopping rules is outside the scope of our contribution.

case, we will show that $\mathbf{X}(t)$ converges to $\mathbf{x}^\star$. Otherwise, we can show convergence to $-\mathbf{x}^\star$.

Employing a potential-based analysis in terms of the Bregman divergence, we will bound the Bregman divergence $D_\Phi(\mathbf{x}^\star, \mathbf{X}(t))$, which then also yields a bound on the ℓ_2-distance $\|\mathbf{X}(t) - \mathbf{x}^\star\|_2$ via Lemma 3.2. A quick calculation yields

$$\frac{d}{dt} D_\Phi(\mathbf{x}', \mathbf{X}(t)) = -\Big\langle \nabla F(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}' \Big\rangle, \tag{3.16}$$

where $\mathbf{x}' \in \mathbb{R}^n$ is any reference point. If F were to be convex, then choosing $\mathbf{x}'$ in equation (3.16) to be any global minimizer of F would imply

$$\Big\langle \nabla F(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}' \Big\rangle \geq F(\mathbf{X}(t)) - F(\mathbf{x}'),$$

which shows that continuous-time mirror descent strictly decreases the Bregman divergence as long as $F(\mathbf{X}(t)) > F(\mathbf{x}')$. As the Bregman divergence is bounded from below by zero by construction due to the convexity of the mirror map, this implies that $F(\mathbf{X}(t))$ must converge to $\min_{\mathbf{x}} F(\mathbf{x})$, which in turn implies that $\mathbf{X}(t)$ converges to a global minimizer of F, provided the function F is continuous. For non-convex objective functions, the inner product in (3.16) has been used to define the notion of *variational coherence* [142], namely the assumption that

$$\Big\langle \nabla F(\mathbf{x}), \mathbf{x} - \mathbf{x}' \Big\rangle \geq 0 \qquad \text{for all } \mathbf{x} \in \mathbb{R}^n,\ \mathbf{x}' \in \mathcal{X}^\star, \tag{3.17}$$

where $\mathcal{X}^\star = \arg\min_{\mathbf{z}} F(\mathbf{z})$, and that there exists an $\mathbf{x}' \in \mathcal{X}^\star$ such that $\langle \nabla F(\mathbf{x}), \mathbf{x} - \mathbf{x}' \rangle = 0$ implies that also also $\mathbf{x} \in \mathcal{X}^\star$. Under this assumption, convergence of a stochastic version of mirror descent towards the set of minimizers of F has been shown in [142]. In the batch setting, convergence of mirror descent towards a global minimizer of F can be shown as in the convex case under the assumption of variational coherence. A local version of variational coherence, only assuming that inequality (3.17) holds for a local minimizer $\mathbf{x}'$ and for $\mathbf{x}$ belonging to an open neighborhood of $\mathbf{x}'$, was also considered in [142], yielding a local version of the convergence results.

In order to study convergence rates, a more refined notion, namely quantitative control on a strictly positive lower bound in (3.17), would be necessary. For example, if the function F were to satisfy a lower bound of the form

$$\Big\langle \nabla F(X(t)), \mathbf{X}(t) - \mathbf{x}' \Big\rangle \geq c D_\Phi(\mathbf{x}', \mathbf{X}(t))$$

for some constant $c > 0$ and minimizer $\mathbf{x}' \in \mathcal{X}^\star$, then this would imply linear convergence towards $\mathbf{x}'$, since we have

$$\frac{d}{dt} D_\Phi(\mathbf{x}', \mathbf{X}(t)) \leq -c D_\Phi(\mathbf{x}', \mathbf{X}(t)) \quad \Rightarrow \quad D_\Phi(\mathbf{x}', \mathbf{X}(t)) \leq e^{-ct} D_\Phi(\mathbf{x}', \mathbf{X}(t)).$$

By design, both the global and local version of variational coherence can only be used to establish convergence towards local minimizers of the objective function. As in our case we are interested in establishing convergence towards the statistical signal $\mathbf{x}^\star$, we consider the inner product $\langle \nabla F(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \rangle$ and bound this quantity from below within a region where the trajectory of mirror descent is confined to stay up to the prescribed stopping time, upon proper initialization. In order to obtain results on the rate of convergence, it is crucial to quantify the lower bound in (3.17). In particular, we derive a lower bound of the form

$$\Big\langle \nabla F(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \Big\rangle \geq \begin{cases} \frac{c_1}{\sqrt{k}} & \text{if } \|\mathbf{X}(t)\|_2^2 \leq \frac{2}{5}\|\mathbf{x}^\star\|_2^2 \\ c_2 \|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2 & \text{else} \end{cases} \tag{3.18}$$

for some constants $c_1, c_2 > 0$. The analysis is divided into two stages: (i) an initial warm-up period of length T_1, where the Bregman divergence decreases first at a constant rate $1/\sqrt{k}$, after which we can relate the quantity $\|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2$ in the second bound in (3.18) to the Bregman divergence $D_\Phi(\mathbf{x}^\star, \mathbf{X}(t))$ by counting the number of "small" coordinates for which $|X_i(t)| < \frac{1}{2}|x_i^\star|$, and (ii) a subsequent stage where we show linear convergence by relating the Bregman divergence to the quantity $\|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2$ in the lower bound (3.18) via Lemma 3.2.

The proof of the lower bound (3.18) involves three main steps. In the first step, writing $f(\mathbf{x}) = \mathbb{E}[F(\mathbf{x})]$ for the population risk, we assume that we have access to the population gradient, which, by the dominated convergence theorem, can be computed as

$$\nabla f(\mathbf{x}) = \mathbb{E}[\nabla F(\mathbf{x})] = \Big(3\|\mathbf{x}\|_2^2 - \|\mathbf{x}^\star\|_2\Big)\mathbf{x} - 2\Big(\mathbf{x}^\top \mathbf{x}^\star\Big)\mathbf{x}^\star.$$

In this setting, the lower bounds in (3.18) on the inner product $\langle \nabla f(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \rangle$ can be obtained via algebraic manipulations: the first bound in (3.18) relies on Lemma 3.8 and the assumption $x_{min}^\star \geq c_\star/\sqrt{k}$ to bound the inner product $\mathbf{X}(t)^\top \mathbf{x}^\star$ from below by $c_1/\sqrt{k}$, while the second bound only uses the Cauchy-Schwarz inequality.

In the second step, we need to control the deviation of the empirical quantity

$$\langle \nabla F(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \rangle = \frac{1}{m} \sum_{j=1}^{m} \Big((\mathbf{A}_j^\top \mathbf{X}(t))^2 - Y_j \Big) \Big(\mathbf{A}_j^\top \mathbf{X}(t) \Big) \Big(\mathbf{A}_j^\top (\mathbf{X}(t) - \mathbf{x}^\star) \Big) \quad (3.19)$$

from the population quantity $\langle \nabla f(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \rangle$. The deviation is bounded in Lemma 3.7, and we summarize the main proof ideas in the following. If the iterate $\mathbf{X}(t)$ were to be independent from the measurement vectors $\{\mathbf{A}_j\}_{j=1}^m$ and noise terms $\{\varepsilon_j\}_{j=1}^m$, then, conditionally on $\mathbf{X}(t)$, the term in (3.19) would be a fourth order Gaussian polynomial, which can be controlled using standard concentration bounds. More precisely, we could use a truncation argument and split it into a bounded and an unbounded part, where the unbounded part is non-zero only with small probability, and use Lipschitz concentration and Hoeffding's bounds to control the bounded part, and Chebyshev's inequality for the unbounded part. However, the assumption of independence is not reasonable. Instead, we use an ϵ-net argument followed by above concentration argument for each fixed vector $\mathbf{x}$ in the ϵ-net. If we considered an ϵ-net in an unit ball in $\mathbb{R}^n$, our argument would require a sample size of $m \gtrsim nk$, with which plenty of algorithms provably recover the signal $\mathbf{x}^\star$ regardless of the assumption on sparsity, e.g. [22, 31, 90]. To obtain our sample requirement of k^2 (modulo logarithmic terms), we show that mirror descent exhibits algorithmic regularization in the sense that $\mathbf{X}(t)$ stays "almost k-sparse" (more precisely, we show that $\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1 < \delta \|\mathbf{x}^\star\|_2$), and apply a union bound to control the deviation of the term in (3.19) from its expectation for all bounded vectors $\mathbf{x}$ supported on $\mathcal{S}$.

In the third step, in order to show that off-support coordinates stay negligibly small, we consider the trajectory of a coordinate $X_i(t)$, which is defined by (cf. (3.4))

$$\frac{d}{dt} X_i(t) = -\sqrt{X_i(t)^2 + \beta^2}\, \nabla F(\mathbf{X}(t))_i.$$

If we imagined that $\beta = 0$ and $\nabla F(\mathbf{X}(t))_i = -c$ was a constant, then the trajectory would simply be given by $X_i(t) = e^{ct}$. Using the assumption $x^\star_{min} \geq c_\star / \sqrt{k}$ along with the concentration argument outlined above, we can show that, for any $i \in \mathcal{S}$ and $j \notin \mathcal{S}$,

$$|\nabla F(\mathbf{X}(t))_j| \leq \gamma |\nabla F(\mathbf{X}(t))_i|$$

holds with high probability for some $\gamma < 1$. Informally, the behavior of support and off-support coordinates is comparable to $\beta(1+c)^t$ and $\beta(1+\gamma c)^t$, respectively, and the

gap can be made large by choosing β small and t large. A detailed proof can be found in Section 3.6.2.

The main ideas of the proof of Theorem 3.4 are similar to those of Theorem 3.3. To show that the Bregman divergence is decreasing in discrete time, we need to bound the difference

$$D_\Phi(\mathbf{x}^\star, \mathbf{X}^{t+1}) - D_\Phi(\mathbf{x}^\star, \mathbf{X}^t) = -\eta\Big\langle \nabla F(\mathbf{X}^t), \mathbf{X}^t - \mathbf{x}^\star \Big\rangle + D_\Phi(\mathbf{X}^t, \mathbf{X}^{t+1}). \tag{3.20}$$

The first term in (3.20) can be bounded as in continuous time, while the second term $D_\Phi(\mathbf{X}^t, \mathbf{X}^{t+1})$ measures the distance between consecutive iterates in terms of the Bregman divergence and is due to the discretization and can be bounded using the strong convexity of the mirror map Φ in (3.6). Unlike in continuous time, we need to establish coordinate-wise upper and lower bounds for how much each coordinate X_i^t changes in one iteration in the discrete-time case in order to characterize the region to which the trajectory of mirror descent is confined. This is not required in continuous time, since in that case we can use the continuity of $\mathbf{X}_i(t)$. The details of the proof can be found in Appendix 3.8.1.

3.5 Numerical study

In this section, we provide numerical experiments demonstrating how the minimum relative estimation error achieved after t_{max} iterations, that is

$$\min_{1 \le t \le t_{max}} \frac{\operatorname{dist}(\mathbf{x}^\star, \mathbf{X}^t)}{\|\mathbf{x}^\star\|_2}, \tag{3.21}$$

depends on the noise-to-signal ratio $\sigma/\|\mathbf{x}^\star\|_2^2$, the sample size m and the sparsity level k, and how the length of the warm-up period as defined in Section 3.6.2, that is

$$T_1 = \inf\left\{ t > 0 : \min_{i \in \mathcal{S}} \frac{|X_i^t|}{|x_i^\star|} > \frac{1}{2} \right\},$$

depends on the mirror map parameter β and the sparsity level k. The main purpose of this section is to validate how the theoretical bounds in Theorem 3.4 on the estimation error and T_1, as well as the convergence behavior suggested by Theorem 3.4, can be representative of the actual performance of mirror descent.

The exponentiated gradient algorithm EG± without normalization [68] is an equivalent formulation of mirror descent equipped with the hypentropy map (Theorem 24 [48]), which is given by

$$\begin{aligned} \mathbf{X}^t &= \mathbf{U}^t - \mathbf{V}^t, \\ \mathbf{U}^{t+1} = \mathbf{U}^t \odot \exp\Big(-\eta \nabla F(\mathbf{X}^t)\Big), &\quad \mathbf{V}^{t+1} = \mathbf{V}^t \odot \exp\Big(\eta \nabla F(\mathbf{X}^t)\Big), \end{aligned} \tag{3.22}$$

where $\odot$ denotes the elementwise Hadamard product, and the initialization (3.7) becomes

$$U_i(0) = \begin{cases} \frac{\hat{\theta}}{2\sqrt{3}} + \sqrt{\frac{\hat{\theta}^2}{12} + \frac{\beta^2}{4}} & i = I_0 \\ \frac{\beta}{2} & i \neq I_0 \end{cases}, \qquad V_i(0) = \begin{cases} -\frac{\hat{\theta}}{2\sqrt{3}} + \sqrt{\frac{\hat{\theta}^2}{12} + \frac{\beta^2}{4}} & i = I_0 \\ \frac{\beta}{2} & i \neq I_0 \end{cases},$$

where we write $\hat{\theta} = (\frac{1}{m}\sum_{j=1}^m Y_j)^{1/2}$ for the estimate of the signal size $\|\mathbf{x}^\star\|_2$. In our simulations, we found EG± to be numerically more stable than the update (3.5), especially for small values of β due to the division by β in (3.5).

In all experiments, we generate a k-sparse vector $\mathbf{x}^\star \in \mathbb{R}^n$ by sampling $x_i^\star$ uniformly from $[-1, -0.15] \cup [0.15, 1]$ for $i = 1, \ldots, n$ and setting $(n-k)$ random entries of $\mathbf{x}^\star$ to zero. We sample m i.i.d. measurement vectors $\mathbf{A}_j \sim \mathcal{N}(0, \mathbf{I}_n)$ and noise terms $\varepsilon_j \sim \mathcal{N}(0, \sigma^2)$ (note that we have $\|\varepsilon_j\|_{\psi_1} = \sqrt{2/\pi}\sigma$), and generate observations $Y_j = (\mathbf{A}_j^\top \mathbf{x}^\star)^2 + \varepsilon_j$. For mirror descent, we find that the step size $\eta = 0.3/(\sum_{j=1}^m Y_j/m)^{3/2}$ works well, which is consistent with the value prescribed by Theorem 3.4.

3.5.1 Estimation error

First, we examine how the relative estimation error of early-stopped mirror descent depends on the noise-to-signal ratio $\sigma/\|\mathbf{x}^\star\|_2^2$, the sample size m and the sparsity level k. Theorem 3.4 provides an upper bound for the relative estimation error of order $\frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k \log n}{m}}$, provided β is chosen small enough so that $\sqrt{c_\star \sqrt{k}\delta} \leq \frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k \log n}{m}}$. In the following experiments, we fix $n = 2000$ and $\beta = 10^{-20}$. We run mirror descent for $t_{max} = 5000$ iterations and consider the minimum relative estimation error defined in (3.21), averaged across 100 independent Monte Carlo trials.

Evaluating the minimum relative error defined in (3.21) requires oracle knowledge of the underlying signal $\|\mathbf{x}^\star\|_2$. For the sake of completeness, in our simulations we also consider the hold-out method to implement early stopping, even if this data-dependent

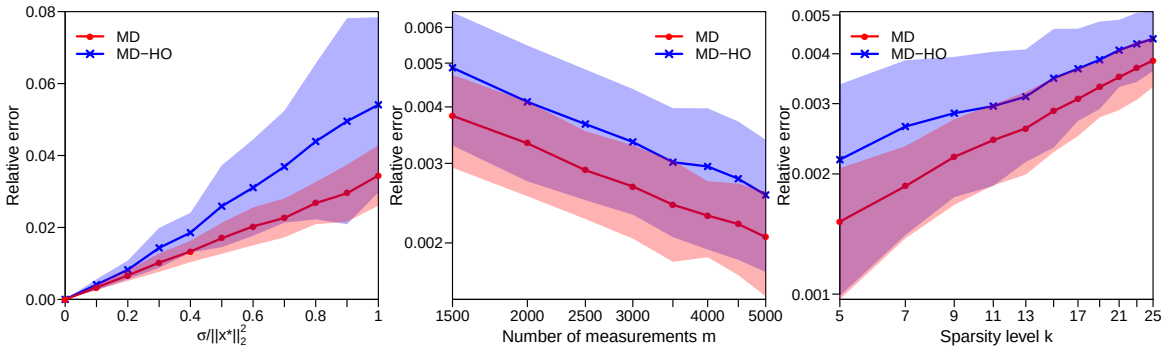


Figure 3.1: Relative error $\text{dist}(\mathbf{x}^\star, \mathbf{X}^{T_{stop}})/\|\mathbf{x}^\star\|_2$ plus/minus one standard deviation (shaded area) of early-stopped mirror descent, with the stopping time T_{stop} selected with oracle knowledge in red (circles), and using the hold-out method in blue (crosses). **Left:** relative error against noise-to-signal ratio $\sigma/\|\mathbf{x}^\star\|_2^2$, with $n = m = 2000$, $k = 10$. **Center:** log-log plot of relative error against sample size m, with $n = 2000$, $k = 10$, $\sigma/\|\mathbf{x}^\star\|_2^2 = 0.1$. **Right:** log-log plot of relative error against sparsity level k, with $n = 2000$, $m = 4000$, $\sigma/\|\mathbf{x}^\star\|_2^2 = 0.1$.

rule goes beyond the reach of the theoretical results that we have developed (in this, we follow the same approach considered in prior related literature [95]). The hold-out method only uses observed data to select a stopping time. More precisely, we run mirror descent using 90% of the data, and select the stopping time $T_{stop} \in \{1, ..., t_{max}\}$ for which the iterate $\mathbf{X}^{T_{stop}}$ minimizes the risk F evaluated on the remaining 10% of the data.

In the first experiment, we fix $k = 10$, $m = 2000$, and vary the noise-to-signal ratio $\sigma/\|\mathbf{x}^\star\|_2^2$. The left plot in Figure 3.1 verifies the linear relationship between the relative estimation error and $\sigma/\|\mathbf{x}^\star\|_2^2$ suggested by the upper bound in Theorem 3.4.

Next, we fix $k = 10$, $\sigma/\|\mathbf{x}^\star\|_2^2 = 0.1$, and increase m from 1500 to 5000. The log-log plot (Figure 3.1, center) shows a slope of -0.5137, which validates the prediction of Theorem 3.4 that the estimation error decreases as $1/\sqrt{m}$ as the sample size increases.

Finally, we fix $m = 4000$, $\sigma/\|\mathbf{x}^\star\|_2^2 = 0.1$, and vary the sparsity k. The log-log plot (Figure 3.1, right) shows a slope of 0.5775, which, considering the standard deviation, is in line with the upper bound of Theorem 3.4 which increases as $\sqrt{k}$ as the sparsity level increases.

In all experiments, the relative error and its standard deviation are larger when the stopping time is selected using the hold-out method compared to a stopping time

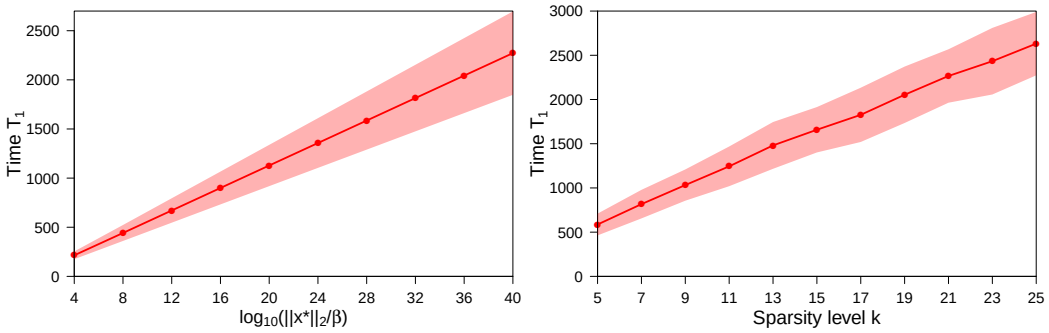


Figure 3.2: Length of warm-up period T_1 plus/minus one standard deviation (shaded area) of early-stopped mirror descent. **Left**: T_1 against $\log(\|\mathbf{x}^\star\|_2/\beta)$, with $n = 2000$, $m = 1500$, $k = 10$, $\sigma/\|\mathbf{x}^\star\|_2^2 = 0.1$. **Right**: T_1 against sparsity level k, with $n = 2000$, $m = 4000$, $\sigma/\|\mathbf{x}^\star\|_2^2 = 0.1$, $\beta/\|\mathbf{x}^\star\|_2 = 10^{-20}$.

selected using oracle knowledge of the signal $\mathbf{x}^\star$, but the same trends are exhibited.

3.5.2 Warm-up time T_1

We note that the asymptotic computational complexity of the algorithm is controlled by the growth of T_1, since $T_2 - T_1 \lesssim \sqrt{k} \log \frac{k\|\mathbf{x}^\star\|_2}{\beta} \lesssim T_1$, see Stage (ii) part (a) in Section 3.6.2. As a result, we focus on T_1 in the following simulations, where we examine how the length of the warm-up period T_1 depends on the mirror map parameter β and the sparsity level k. Theorem 3.4 provides an upper bound of order $k \log \frac{\|\mathbf{x}^\star\|_2}{\beta}$ (modulo double-logarithmic terms).

We first fix $n = 2000$, $m = 1500$, $k = 10$, $\sigma/\|\mathbf{x}^\star\|_2^2 = 0.1$ and decrease $\beta/\|\mathbf{x}^\star\|_2$ from 10^{-4} to 10^{-40}. The left plot in Figure 3.2 validates the upper bound in Theorem 3.4, the length of the warm-up period T_1 increases linearly with $\log \frac{\|\mathbf{x}^\star\|_2}{\beta}$.

In the second experiment, we set $n = 2000$, $m = 4000$, $\sigma/\|\mathbf{x}^\star\|_2^2 = 0.1$, $\beta = 10^{-20}$ and increase k from 5 to 25. Again, the right plot in Figure 3.2 shows that T_1 increases linearly with the sparsity level k, validating the upper bound in Theorem 3.4.

3.5.3 Convergence behavior of mirror descent

The next experiment examines the convergence behavior of mirror descent and its dependence on the parameter β. We only present the ℓ_2-error $\mathrm{dist}(\mathbf{x}^\star, \mathbf{X}^t)$. The Bregman divergence term displays the same qualitative behavior. We set $n = 50000$, $m = 1000$, $k = 10$ and run mirror descent for different values of β. In both the noiseless

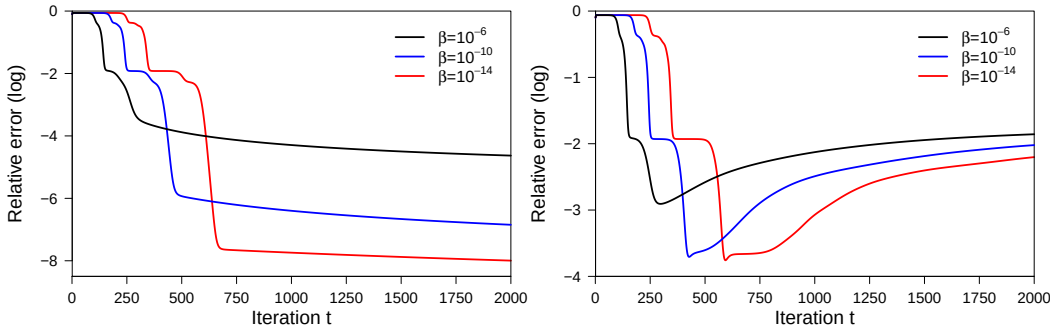


Figure 3.3: Relative error $\mathrm{dist}(\mathbf{x}^\star, \mathbf{X}^t)/\|\mathbf{x}^\star\|_2$ (log-scale) of mirror descent for $n = 50000$, $m = 1000$, $k = 10$, and different values of β. **Left**: noiseless case with $\sigma/\|\mathbf{x}^\star\|_2^2 = 0$. **Right**: noisy case with $\sigma/\|\mathbf{x}^\star\|_2^2 = 0.5$.

$(\sigma/\|\mathbf{x}^\star\|_2^2 = 0)$ and noisy $(\sigma/\|\mathbf{x}^\star\|_2^2 = 0.5)$ case, we observe the expected dependence on β: as β decreases from 10^{-6} to 10^{-14}, the length of the initial warm-up period increases as $\log \frac{1}{\beta}$, while we have linear convergence up to an improved accuracy that depends polynomially on β. In the noisy case, the precision up to which we have linear convergence barely improves as we decrease β from 10^{-10} to 10^{-14}, which we attribute to the noise term $\frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k \log n}{m}}$ dominating the achievable accuracy. Figure 3.3 suggests that early stopping is indeed necessary in the presence of noise.

3.6 Proofs

In this section, we prove the results stated in Section 3.4. Since the proof of Theorem 3.4 largely follows the same steps as the proof of Theorem 3.3, with a few technical difficulties arising due to the discretization, we defer the proof of Theorem 3.4 to Appendix 3.8.1.

3.6.1 Proof of Lemma 3.2

The proof of Lemma 3.2 relies on the fact that the Bregman divergence associated to the hypentropy mirror map can be decomposed into a sum of functions, each of which depends only on a single coordinate x_i and is (locally) bounded from above and below by quadratic functions.

Proof of Lemma 3.2. Using the identiy $\mathrm{arcsinh}(x) = \log(x + \sqrt{1 + x^2})$, we can write

$$D_\Phi(\mathbf{x}^\star, \mathbf{x}) = \sum_{i=1}^{n} g_i(x_i),$$

where

$$g_i(x_i) = \sqrt{x_i^2 + \beta^2} - \sqrt{(x_i^\star)^2 + \beta^2} - x_i^\star \log \frac{x_i + \sqrt{x_i^2 + \beta^2}}{x_i^\star + \sqrt{(x_i^\star)^2 + \beta^2}}.$$

We begin by showing the bound (3.9). For any $i \in [n]$, we have $g_i(x_i^\star) = 0$ and

$$g_i'(x_i) = \frac{x_i}{\sqrt{x_i^2 + \beta^2}} - x_i^\star \frac{1 + x_i/\sqrt{x_i^2 + \beta^2}}{x_i + \sqrt{x_i^2 + \beta^2}} = \frac{x_i - x_i^\star}{\sqrt{x_i^2 + \beta^2}}. \tag{3.23}$$

As we assume $x_i x_i^\star \geq 0$ and $|x_i| \geq \frac{1}{2}|x_i^\star|$, we have $\sqrt{z^2 + \beta^2} \geq \frac{x_{min}^\star \|\mathbf{x}^\star\|_2}{2} \geq \frac{c_\star \|\mathbf{x}^\star\|_2}{2\sqrt{k}}$ for any $z \in (x_i, x_i^\star)$ (or for any $z \in (x_i^\star, x_i)$ if $x_i^\star < x_i$). Hence, using the convention $\int_a^b f(x)dx = -\int_b^a f(x)dx$ when $a > b$, we can bound

$$g_i(x_i) = g_i(x_i^\star) + \int_{x_i^\star}^{x_i} g'(z)dz < \int_{x_i^\star}^{x_i} \frac{2\sqrt{k}}{c_\star \|\mathbf{x}^\star\|_2}(z - x_i^\star)dz = \frac{\sqrt{k}}{c_\star \|\mathbf{x}^\star\|_2}(x_i - x_i^\star)^2.$$

Summing over all $i \in \mathcal{S}$, this gives

$$\sum_{i \in \mathcal{S}} g_i(x_i) \leq \frac{\sqrt{k}}{c_\star \|\mathbf{x}^\star\|_2} \|\mathbf{x}_\mathcal{S} - \mathbf{x}_\mathcal{S}^\star\|_2^2$$

as claimed. On the other hand, if $x_i^\star = 0$, we have

$$g_i(x_i) = \sqrt{x_i^2 + \beta^2} - \beta \leq |x_i|,$$

which gives

$$\sum_{i \notin \mathcal{S}} g_i(x_i) \leq \|x_{\mathcal{S}^c}\|_1,$$

and completes the proof of the bound (3.9) in Lemma 3.2. The other bound (3.8) can be shown similarly using (3.23). We have $\sqrt{z^2 + \beta^2} \leq \sqrt{\max\{\|\mathbf{x}\|_\infty^2, \|\mathbf{x}^\star\|_\infty^2\} + \beta^2}$ for all $z \in (x_i, x_i^\star)$ (or $z \in (x_i^\star, x_i)$ if $x_i^\star < x_i$), which gives

$$\begin{aligned} g(x_i) = g(x_i^\star) + \int_{x_i}^{x_i^\star} g'(z)dz &> \int_{x_i}^{x_i^\star} \Big(\max\{\|\mathbf{x}\|_\infty^2,\, \|\mathbf{x}^\star\|_\infty^2\} + \beta^2\Big)^{-\frac{1}{2}} (z - x_i^\star)dz \\ &= \frac{1}{2}\Big(\max\{\|\mathbf{x}\|_\infty^2,\, \|\mathbf{x}^\star\|_\infty^2\} + \beta^2\Big)^{-\frac{1}{2}} (x_i - x_i^\star)^2. \end{aligned}$$

Summing over all $i \in [n]$ gives the bound (3.8) of Lemma 3.2. □

3.6.2 Proof of Theorem 3.3

In this section, we will make the following assumptions, which are purely for notational simplicity.

- We will assume $\|\mathbf{x}^\star\|_2 = 1$. The general case $\|\mathbf{x}^\star\|_2 \neq 1$ immediately follows by writing $\mathbf{x}^\star = \frac{\mathbf{x}^\star}{\|\mathbf{x}^\star\|_2}\|\mathbf{x}^\star\|_2$, $\mathbf{x} = \frac{\mathbf{x}}{\|\mathbf{x}^\star\|_2}\|\mathbf{x}^\star\|_2$ and $\varepsilon_j = \frac{\varepsilon_j}{\|\mathbf{x}^\star\|_2}\|\mathbf{x}^\star\|_2$ in what follows.
- We will assume $\mathbf{X}(0)^\top \mathbf{x}^\star \geq 0$, in which case we show convergence towards $\mathbf{x}^\star$. Otherwise, we can show convergence towards $-\mathbf{x}^\star$.

The gradient of the objective function (3.2) is given by

$$\nabla F(\mathbf{x}) = \frac{1}{m}\sum_{j=1}^{m}\Big((\mathbf{A}_j^\top \mathbf{x})^2 - (\mathbf{A}_j^\top \mathbf{x}^\star)^2 - \varepsilon_j\Big)\Big(\mathbf{A}_j^\top \mathbf{x}\Big)\mathbf{A}_j.$$

A brief calculation gives the following expression for its expectation, the population gradient.

$$\nabla f(\mathbf{x}) = \mathbb{E}[\nabla F(\mathbf{x})] = \Big(3\|\mathbf{x}\|_2^2 - 1\Big)\mathbf{x} - 2\Big(\mathbf{x}^\top \mathbf{x}^\star\Big)\mathbf{x}^\star.$$

First, we provide three supporting lemmas characterizing the behavior of mirror descent. We defer the proofs of the following lemmas to Appendix 3.8.2. Lemmas 3.6 and 3.7 show that the gradient $\nabla F(\mathbf{x})$ is close to its expectation $\nabla f(\mathbf{x})$.

Lemma 3.6. *Let $\mathbf{x}^\star \in \mathbb{R}^n$ be a k-sparse vector, and let $\mathcal{S} = \{i \in [n] : x_i^\star \neq 0\}$ be its support. Let $\{\mathbf{A}_j\}_{j=1}^m$ be a collection of i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ random vectors and $\{\varepsilon_j\}_{j=1}^m$ a collection of independent centered sub-exponential random variables with maximum sub-exponential norm $\sigma = \max_j \|\varepsilon_j\|_{\psi_1}$. There exist universal constants $c, c_p > 0$ such that for any constant $\gamma > 0$, there is a constant $c_s(\gamma) > 0$ such that the following holds. For any $0 \leq \delta \leq \frac{1}{4}$, let*

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 \leq \|\mathbf{x}^\star\|_2, \|\mathbf{x}_{\mathcal{S}^c}\|_1 \leq \delta\|\mathbf{x}^\star\|_2\}.$$

If $m \geq c_s(\gamma)(1 + \frac{\sigma^2}{\|\mathbf{x}^\star\|_2^4})\max\{k^2 \log^2 n,\ \log^5 n\}$, then, with probability at least $1 - c_p n^{-10}$,

$$\begin{aligned}\frac{\|\nabla F(\mathbf{x}) - \nabla f(\mathbf{x})\|_\infty}{\|\mathbf{x}^\star\|_2^2} &\leq \gamma \min\left\{\frac{\|\mathbf{x}_{\mathcal{S}}\|_1}{k},\ \frac{\|\mathbf{x}_{\mathcal{S}} - \mathbf{x}_{\mathcal{S}}^\star\|_2}{\sqrt{k}}\right\} + c\delta\|\mathbf{x}^\star\|_2 \\ &\quad + \frac{c\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{\log n}{m}}\min\Big\{\|\mathbf{x}\|_1,\ (1+\delta)\|\mathbf{x}^\star\|_2 + \sqrt{k}\|\mathbf{x}_{\mathcal{S}} - \mathbf{x}_{\mathcal{S}}^\star\|_2\Big\}\end{aligned}$$

holds for all $\mathbf{x} \in \mathcal{X}$.

In particular, if $\frac{\|\mathbf{x}-\mathbf{x}^\star\|_2}{\|\mathbf{x}^\star\|_2} \geq \max\Big\{\frac{c\sqrt{k}\delta}{\gamma}, \frac{c(1+\delta)\sigma}{\gamma\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k\log n}{m}}\Big\}$, $\|\mathbf{x}_{\mathcal{S}}\|_1 \geq \frac{1}{2}\|\mathbf{x}^\star\|_2$, $\delta \leq \frac{\gamma}{2ck}$ and $c_s(\gamma) \geq \frac{c^2}{\gamma^2}$, then we have the simplified bound

$$\frac{\|\nabla F(\mathbf{x}) - \nabla f(\mathbf{x})\|_\infty}{\|\mathbf{x}^\star\|_2^2} \leq 4\gamma \min\left\{\frac{\|\mathbf{x}_{\mathcal{S}}\|_1}{k},\ \frac{\|\mathbf{x} - \mathbf{x}^\star\|_2}{\sqrt{k}}\right\} \quad \textit{for all } \mathbf{x} \in \mathcal{X}.$$

Lemma 3.7. *Let $\mathbf{x}^\star \in \mathbb{R}^n$ be a k-sparse vector, and let $\mathcal{S} = \{i \in [n] : x_i^\star \neq 0\}$ be its support. Let $\{\mathbf{A}_j\}_{j=1}^m$ be a collection of i.i.d. $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ random vectors and $\{\varepsilon_j\}_{j=1}^m$ a collection of independent centered sub-exponential random variables with maximum sub-exponential norm $\sigma = \max_j \|\varepsilon_j\|_{\psi_1}$. There exist universal constants $c, c_p > 0$ such that for any constant $\gamma > 0$, there is a constant $c_s(\gamma) > 0$ such that the following holds. For any $0 \le \delta \le \frac{1}{4}\min\{\frac{\gamma}{c}, 1\}$, let*

$$\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2 \le \|\mathbf{x}^\star\|_2, \|\mathbf{x}_{\mathcal{S}^c}\|_1 \le \delta\|\mathbf{x}^\star\|_2\}.$$

If $m \ge c_s(\gamma)(1 + \frac{\sigma^2}{\|\mathbf{x}^\star\|_2^4})\max\{k^2\log^2 n,\ \log^5 n\}$, then, with probability at least $1 - c_p n^{-10}$,

$$\frac{\big|\langle \nabla F(\mathbf{x}) - \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star\rangle\big|}{\|\mathbf{x}^\star\|_2^3} \le \gamma \min\left\{\frac{\|\mathbf{x}_S\|_1}{\sqrt{k}} + \delta\|\mathbf{x}^\star\|_2,\ \frac{\|\mathbf{x}_{\mathcal{S}} - \mathbf{x}_{\mathcal{S}}^\star\|_2^2}{\|\mathbf{x}^\star\|_2} + \delta^2\|\mathbf{x}^\star\|_2\right\} + \frac{c\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{\log n}{m}}\Big(\delta\|\mathbf{x}^\star\|_2 + \min\Big\{\|\mathbf{x}_{\mathcal{S}}\|_1, \sqrt{k}\|\mathbf{x}_{\mathcal{S}} - \mathbf{x}_{\mathcal{S}}^\star\|_2\Big\}\Big)$$

holds for all $\mathbf{x} \in \mathcal{X}$.

In particular, if $\frac{\|\mathbf{x}-\mathbf{x}^\star\|_2}{\|\mathbf{x}^\star\|_2} \ge \max\Big\{\delta, \frac{c\sigma}{\gamma\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k\log n}{m}}\Big\}$, $\|\mathbf{x}_{\mathcal{S}}\|_1 \ge \frac{1}{2}\|\mathbf{x}^\star\|_2$, $\delta \le \frac{1}{2\sqrt{k}}$ and $c_s(\gamma) \ge \frac{c^2}{\gamma^2}$, then we have the simplified bound

$$\frac{\big|\langle \nabla F(\mathbf{x}) - \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star\rangle\big|}{\|\mathbf{x}^\star\|_2^3} \le 4\gamma \min\left\{\frac{\|\mathbf{x}_S\|_1}{\sqrt{k}},\ \frac{\|\mathbf{x} - \mathbf{x}^\star\|_2^2}{\|\mathbf{x}^\star\|_2}\right\} \quad \text{for all } \mathbf{x} \in \mathcal{X}.$$

The following lemma characterizes the region to which the trajectory of early-stopped mirror descent (both in continuous and discrete time) is confined.

Lemma 3.8. *Let Assumption I hold. There exist universal constants $c_p, c_1, c_2, c_3 > 0$ such that the following holds. Let $\delta \le c_1/n$ and let $\mathbf{X}(t)$ be defined by the continuous-time mirror descent algorithm given by (3.4) with mirror map (3.6) and initialization (3.7) with $\beta \le \delta$. Assume that there is a $T > 0$ such that*

$$\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1 \le \delta\|\mathbf{x}^\star\|_2 \qquad \text{and} \qquad \frac{\|\mathbf{X}(t) - \mathbf{x}^\star\|_2}{\|\mathbf{x}^\star\|_2} \ge c_2 \max\left\{\sqrt{k}\delta,\ \frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k\log n}{m}}\right\}$$

are satisfied for all $t \le T$. Then, there is a $\xi \in \{-1, +1\}$, such that, with probability at least $1 - c_p n^{-10}$, the following holds for all $t \le T$:

$$\xi X_i(t) x_i^\star \ge 0 \quad \text{for all } i \in [n], \tag{3.24}$$

$$\frac{1}{3} - (3 + 9\sigma)\sqrt{\frac{\log n}{m}} \le \frac{\|\mathbf{X}(t)\|_2^2}{\|\mathbf{x}^\star\|_2^2} \le 2, \tag{3.25}$$

$$\sqrt{3} \cdot 2\Big|\mathbf{X}(t)^\top \mathbf{x}^\star\Big| \ge 3\|\mathbf{X}(t)\|_2^2 - \|\mathbf{x}^\star\|_2^2. \tag{3.26}$$

The same result holds for $\mathbf{X}^t$ defined by the discrete-time mirror descent algorithm given by (3.5) with $\eta \leq c_3/\|\mathbf{x}^\star\|_2^3$.

The following inequalities will be useful throughout the following proofs. Under the assumptions of Lemma 3.8, we can use (3.25) to bound

$$\|\mathbf{X}_{\mathcal{S}}(t)\|_1 = \|\mathbf{X}(t)\|_1 - \|\mathbf{X}_{\mathcal{S}^c}(t)\|_1 \geq \left(\frac{1}{3} - (3+9\sigma)\sqrt{\frac{\log n}{m}}\right)^{-\frac{1}{2}} - \delta \geq \frac{1}{2}, \tag{3.27}$$

and, using (3.24) of Lemma 3.8, we can bound the inner product

$$\mathbf{X}(t)^\top \mathbf{x}^\star \geq \|\mathbf{X}_{\mathcal{S}}(t)\|_1 x^\star_{min} \geq \frac{c_\star}{2\sqrt{k}}. \tag{3.28}$$

Proof of Theorem 3.3. We first consider the population dynamics to illustrate the main ideas of the proof of Theorem 3.3. That is, we first assume that we had access to the population gradient ∇f, or in other words that $m = \infty$. The proof of Theorem 3.3 relies on the identity

$$\frac{d}{dt} D_\Phi(\mathbf{x}^\star, \mathbf{X}(t)) = -\Big\langle \nabla F(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \Big\rangle, \tag{3.29}$$

and we bound the inner product on the right hand side to show that the Bregman divergence $D_\Phi(\mathbf{x}^\star, \mathbf{X}(t))$ decreases as claimed.

Initial Bregman divergence

Writing $\hat{\theta} = (\frac{1}{m}\sum_{j=1}^m Y_j)^{1/2}$ for the estimate of the signal size $\|\mathbf{x}^\star\|_2$, we can use standard concentration bounds for sub-exponential random variables (see e.g. Prop. 5.16 of [115]) to bound, with probability $1 - 4n^{10}$,

$$\hat{\theta}^2 = \frac{1}{m}\sum_{j=1}^m Y_j = \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_j^\top \mathbf{x}^\star)^2 + \frac{1}{m}\sum_{j=1}^m \varepsilon_j < \left(1 + (9+25\sigma)\sqrt{\frac{\log n}{m}}\right)\|\mathbf{x}^\star\|_2^2,$$

Using the definition of the initialization (3.7), we can bound the initial Bregman divergence

$$\begin{aligned}
D_\Phi(\mathbf{x}^\star, \mathbf{X}(0)) &= \sum_{i \neq I_0} \beta - \sqrt{(x_i^\star)^2 + \beta^2} - x_i^\star \log \frac{\beta}{x_i^\star + \sqrt{(x_i^\star)^2 + \beta^2}} \\
&\quad + \sqrt{\hat{\theta}^2/3 + \beta^2} - \sqrt{(x^\star_{I_0})^2 + \beta^2} - x^\star_{I_0} \log \frac{\hat{\theta}/\sqrt{3} + \sqrt{\hat{\theta}^2/3 + \beta^2}}{x^\star_{I_0} + \sqrt{(x^\star_{I_0})^2 + \beta^2}} \\
&\leq \sum_{i \neq I_0} |x_i^\star| \log \frac{1}{\beta} + \underbrace{\beta - \sqrt{(x_i^\star)^2 + \beta^2} + |x_i^\star| \log\left(|x_i^\star| + \sqrt{(x_i^\star)^2 + \beta^2}\right)}_{\leq 0} + 1 \\
&\leq \|\mathbf{x}^\star\|_1 \log \frac{1}{\beta} + 1,
\end{aligned} \tag{3.30}$$

where we used the assumption $\|\mathbf{x}^\star\|_2 = 1$ and the fact that the function $x \log \frac{\beta}{x+\sqrt{x^2+\beta^2}}$ is symmetric, that is $x \log \frac{\beta}{x+\sqrt{x^2+\beta^2}} = -x \log \frac{\beta}{-x+\sqrt{x^2+\beta^2}}$.

Bounding $\langle \nabla f(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \rangle$

We can compute

$$\begin{aligned}\Big\langle \nabla f(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \Big\rangle = -\Big[3\Big(\|\mathbf{X}(t)\|_2^2 - 1\Big) + 2\Big(\mathbf{X}(t)^\top \mathbf{x}^\star\Big)\Big]\Big(\mathbf{X}(t)^\top \mathbf{x}^\star\Big) \\ + \Big(3\|\mathbf{X}(t)\|_2^2 - 1\Big)\|\mathbf{X}(t)\|_2^2.\end{aligned}$$

To bound this quantity, we distinguish two cases.

- **Case 1:** $\|\mathbf{X}(t)\|_2^2 \le \frac{2}{5}$

  In this case, we use the fact that $3\|\mathbf{X}(t)\|_2^2 - 1 \ge -(9 + 27\sigma)\sqrt{\frac{\log n}{m}}$ by Lemma 3.8 to bound

  $$\begin{aligned}\Big\langle \nabla f(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \Big\rangle &\ge \left(\frac{9}{5} - 2\sqrt{\frac{2}{5}}\right)\Big(\mathbf{X}(t)^\top \mathbf{x}^\star\Big) - \frac{18 + 54\sigma}{5}\sqrt{\frac{\log n}{m}} \\ &\ge \frac{1}{2}\Big(\mathbf{X}(t)^\top \mathbf{x}^\star\Big), \end{aligned} \tag{3.31}$$

  where for the first inequality we used that $\mathbf{X}(t)^\top \mathbf{x}^\star \le \|\mathbf{X}(t)\|_2 \|\mathbf{x}^\star\|_2$ by the Cauchy-Schwarz inequality, and the second inequality holds by (3.28) since $m \ge c_s k^2 \log^2 n$.

- **Case 2:** $\|\mathbf{X}(t)\|_2^2 > \frac{2}{5}$

  In this case, we can write $2(\mathbf{X}(t)^\top \mathbf{x}^\star) = \|\mathbf{X}(t)\|_2^2 + \|\mathbf{x}^\star\|_2^2 - \|\mathbf{X}(t) - \mathbf{x}\|_2^2$ to compute

  $$\begin{aligned}\Big\langle \nabla f(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \Big\rangle = \Big(2 - 4\|\mathbf{X}(t)\|_2^2 + \Delta\Big)\Big(\mathbf{X}(t)^\top \mathbf{x}^\star\Big) \\ + \Big(3\|\mathbf{X}(t)\|_2^2 - 1\Big)\|\mathbf{X}(t)\|_2^2,\end{aligned}$$

  where we denote $\Delta = \|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2$. If $2 - 4\|\mathbf{X}(t)\|_2^2 + \Delta \ge 0$, this is lower bounded by

  $$\Big\langle \nabla f(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \Big\rangle \ge \frac{2}{25}. \tag{3.32}$$

  Otherwise, again using $\mathbf{X}(t)^\top \mathbf{x}^\star \le \|\mathbf{X}(t)\|_2 \|\mathbf{x}^\star\|_2$, we have

  $$\begin{aligned}\Big\langle \nabla f(\mathbf{X}(t)), \mathbf{x}^\star - \mathbf{X}(t) \Big\rangle &\ge \underbrace{3\|\mathbf{X}(t)\|_2^4 - 4\|\mathbf{X}(t)\|_2^3 - \|\mathbf{X}(t)\|_2^2 + 2\|\mathbf{X}(t)\|_2}_{\ge 0} + \Delta\|\mathbf{X}(t)\|_2 \\ &\ge \sqrt{\frac{2}{5}}\Delta. \end{aligned} \tag{3.33}$$

The inequalities (3.31)–(3.33) can be used to show that the Bregman divergence $D_\Phi(\mathbf{x}^\star, \mathbf{X}(t))$ decreases as claimed. In particular, we can use Lemma 3.2 to bound $\|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2$ in terms of $D_\Phi(\mathbf{x}^\star, \mathbf{X}(t))$, and inequality (3.33) yields a bound leading to linear convergence as long as $\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1$ is sufficiently small compared to $\|\mathbf{X}_{\mathcal{S}}(t) - \mathbf{x}^\star_{\mathcal{S}}\|_2^2$. In order to make the proof of Theorem 3.3 rigorous, we need to replace the population gradient ∇f by the empirical gradient ∇F in the outline we provided above. To this end, we divide the analysis of the convergence of mirror descent into two stages bounded by

$$T_1 = \inf\left\{t > 0 : \min_{i \in \mathcal{S}} \frac{|X_i(t)|}{|x_i^\star|} > \frac{1}{2}\right\}, \text{ and}$$
$$T_2 = \inf\left\{t > 0 : \frac{\|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2}{\|\mathbf{x}^\star\|_2^2} \leq c^2 \max\left\{c_\star \sqrt{k}\delta,\ \frac{\sigma^2}{\|\mathbf{x}^\star\|_2^4} \frac{k \log n}{m}\right\}\right\},$$

respectively, where $\delta = \sqrt{n\beta}$. We consider the stages (i) $t \leq T_1$ and (ii) $T_1 < t \leq T_2$. Note that we have $T_2 > T_1$, because if there is an index $i \in \mathcal{S}$ with $|X_i(t)| < \frac{1}{2}|x_i^\star|$, then we also have $\|\mathbf{X}(t) - \mathbf{x}^\star\|_2 > \frac{c_\star}{2\sqrt{k}}$. In both stages, we will (a) bound the length of the stage by using (3.29) to bound T_i, and (b) show that off-support coordinates satisfy $\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1 \leq \delta$. Throughout the proof we will assume that the inequalities in Lemmas 3.6–3.8 are satisfied, which happens with probability at least $1 - c_p n^{-10}$. Note that as $\|\mathbf{X}_{\mathcal{S}}(t)\|_1 \geq \|\mathbf{X}(t)\|_2 - \delta \geq \frac{1}{2}$ by the lower bound in (3.25) of Lemma 3.8 and since $\delta \lesssim n^{-1}$, the conditions for the simplified bounds in Lemmas 3.6 and 3.7 are satisfied for $t \leq T_2$.

Stage (i), part (a): $t \leq T_1$, bound T_1

Assume for now that we have already shown $\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1 < \delta_1 = n\beta^{3/4}$ for all $t \leq T_1$. Note that we have $\delta_1 \leq \delta$ since $\beta \leq n^{-2}$. We have already computed the rate (3.29) at which the Bregman divergence $D_\Phi(\mathbf{x}^\star, \mathbf{X}(t))$ decreases if we had access to the population gradient ∇f in (3.31) and (3.33). Using these bounds, we can bound the inner product in (3.29) with the empirical gradient ∇F.

- **Case 1: $\|\mathbf{X}(t)\|_2^2 \leq \frac{2}{5}$**

  By Lemma 3.7 and the definitions of δ_1 and T_2, we have

  $$\left|\left\langle \nabla F(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \right\rangle - \left\langle \nabla f(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \right\rangle\right| \leq \frac{c_\star}{4\sqrt{k}} \|\mathbf{X}_{\mathcal{S}}(t)\|_1 \leq \frac{1}{4}\left(\mathbf{X}(t)^\top \mathbf{x}^\star\right).$$

Together with (3.31), this bound leads to

$$\begin{aligned}\frac{d}{dt}D_\Phi(\mathbf{x}^\star, \mathbf{X}(t)) &\le -\Big\langle \nabla f(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \Big\rangle \\ &\quad + \Big|\Big\langle \nabla F(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \Big\rangle - \Big\langle \nabla f(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \Big\rangle\Big| \\ &\le -\frac{1}{4}\Big(\mathbf{X}(t)^\top \mathbf{x}^\star\Big) \\ &\le -\frac{c_\star}{8\sqrt{k}}, \end{aligned} \tag{3.34}$$

where for the last line we used (3.28).

- **Case 2:** $\|\mathbf{X}(t)\|_2^2 \ge \frac{2}{5}$

  As in the previous case, we can use Lemma 3.7 to bound the difference $\langle \nabla F(\mathbf{X}(t)) - \nabla f(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star\rangle$. Recalling (3.32) and (3.33), we obtain

$$\frac{d}{dt}D_\Phi(\mathbf{x}^\star, \mathbf{X}(t)) = -\Big\langle \nabla F(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \Big\rangle \le -\frac{1}{25} \tag{3.35}$$

  if $2 - 4\|\mathbf{X}(t)\|_2^2 + \|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2 \ge 0$, and

$$\frac{d}{dt}D_\Phi(\mathbf{x}^\star, \mathbf{X}(t)) = -\Big\langle \nabla F(\mathbf{X}(t)), \mathbf{X}(t) - \mathbf{x}^\star \Big\rangle \le -\frac{1}{2}\|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2 \tag{3.36}$$

  if $2 - 4\|\mathbf{X}(t)\|_2^2 + \|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2 < 0$, where we used that

$$\Big|\Big\langle \nabla F(\mathbf{X}(t)) - \nabla f(\mathbf{X}(t),\ \mathbf{X}(t) - \mathbf{x}^\star \Big\rangle\Big| \le \left(\sqrt{\frac{2}{5}} - \frac{1}{2}\right)\|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2,$$

  for all $t \le T_2$ by the simplified bound in Lemma 3.7.

We can now bound T_1. Define

$$T_0 = \inf\left\{t > 0 : \|\mathbf{X}(t)\|_2^2 > \frac{2}{5}\right\},$$

as the time until which Case 1 holds. Then, the bound (3.34) from Case 1 implies

$$\min\{T_1, T_0\} \le \frac{8\sqrt{k}D_\Phi(\mathbf{x}^\star, \mathbf{X}(0))}{c_\star} \le \frac{8}{c_\star}k\log\frac{1}{\beta} + \frac{8\sqrt{k}}{c_\star} \le \frac{c_2}{2}k\log\frac{1}{\beta},$$

provided that $c_2 \ge \frac{16}{c_\star} + \frac{16}{c_\star\sqrt{k}\log\frac{1}{\beta}}$, where we used (3.30) to bound the initial Bregman divergence $D_\Phi(\mathbf{x}^\star, \mathbf{X}(0))$. If $T_1 < T_0$, then we have bounded T_1 as desired. Otherwise, we can use the bounds (3.35) and (3.36) from Case 2 to control $T_1 - T_0$. The first bound (3.35) can apply at most for $t \le 25D_\Phi(\mathbf{x}^\star, \mathbf{X}(0)) \le 25\sqrt{k}\log\frac{1}{\beta} + 25$, where we again used the bound (3.30). As the bound (3.36) depends on the ℓ_2-distance

$\|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2$, the idea is to show that this quantity is sufficiently large as long as the Bregman divergence $D_\Phi(\mathbf{x}^\star, \mathbf{X}(t))$ is large. To this end, define

$$S(t) = \left\{ i \in \mathcal{S} : \frac{|X_i(t)|}{|x_i^\star|} < \frac{1}{2} \right\}.$$

With this, we can bound

$$\begin{aligned} \|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2 &\geq \sum_{i \in S(t)} \left(\frac{1}{2} x_i^\star\right)^2 + \sum_{i \notin S(t)} \left(X_i(t) - x_i^\star\right)^2 \\ &\geq \frac{c_\star^2}{4k} |S(t)| + \sum_{i \notin S(t)} \left(X_i(t) - x_i^\star\right)^2. \end{aligned} \tag{3.37}$$

As in the computation for (3.30) and the proof of Lemma 3.2, we can bound

$$D_\Phi(\mathbf{x}^\star, \mathbf{X}(t)) \leq \sum_{i \in S(t)} |x_i^\star| \log \frac{1}{\beta} + \frac{\sqrt{k}}{c_\star} \sum_{i \notin S(t)} \left(X_i(t) - x_i^\star\right)^2 + \|\mathbf{X}_{\mathcal{S}^c}(t)\|_1. \tag{3.38}$$

Because $\|\mathbf{x}^\star\|_\infty \leq \|\mathbf{x}^\star\|_2 = 1$, we have $\sum_{i \in S(t)} |x_i^\star| \leq |S(t)|$. Further, since $t < T_1$, we also have $\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1 \leq \delta_1 \leq \frac{1}{4}(x_{min}^\star)^2 \leq \sum_{i \notin S(t)} (X_i(t) - x_i^\star)^2$. With this, we can combine (3.36), (3.37) and (3.38) to bound

$$\frac{d}{dt} D_\Phi(\mathbf{x}^\star, \mathbf{X}(t)) \leq -\frac{1}{2} \|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2 \leq -\frac{c_\star^2}{8k \log \frac{1}{\beta}} D_\Phi(\mathbf{x}^\star, \mathbf{X}(t)),$$

which shows that the Bregman divergence $D_\Phi(\mathbf{x}^\star, \mathbf{X}(t))$ decreases linearly at the rate $c_\star^2/(8k \log \frac{1}{\beta})$ for $T_0 < t < T_1$. We have

$$D_\Phi(\mathbf{x}^\star, \mathbf{X}(T_0)) \leq D_\Phi(\mathbf{x}^\star, \mathbf{X}(0)) \leq \sqrt{k} \log \frac{1}{\beta} + 1,$$

and, for $t < T_1$,

$$\frac{c_\star^2}{4k} \leq \|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2 \leq 2\sqrt{2 + \beta^2}\, D_\Phi(\mathbf{x}^\star, \mathbf{X}(t)) \leq 3 D_\Phi(\mathbf{x}^\star, \mathbf{X}(t)),$$

where for the second inequality we used that $\|\mathbf{X}(t)\|_\infty \leq \sqrt{2}$ by Lemma 3.8 and the bound (3.8) of Lemma 3.2. This implies

$$T_1 - T_0 \leq \frac{8k \log \frac{1}{\beta}}{c_\star^2} \log\left(\frac{12}{c_\star^2} k^{\frac{3}{2}} \log \frac{1}{\beta} + \frac{12}{c_\star^2} k\right) \leq \frac{c_2}{2} k \log\left(\frac{1}{\beta}\right) \log\left(k \log \frac{1}{\beta}\right),$$

provided that $c_2 \geq \frac{24}{c_\star^2} + \frac{16}{c_\star^2} \log \frac{24}{c_\star^2}$.

Stage (i)**, part** (b)**:** $t \leq T_1$**, bound** $\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1$

The idea to controlling $\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1$ is as follows: we will show that for coordinates $j \notin \mathcal{S}$

and $i \in \mathcal{S}$, $X_j(t)$ can only grow at a comparatively slower rate than $X_i(t)$. We will show that the growth of both coordinates is bounded by exponentials, and use the fact that, for any fixed $\epsilon > 0$, the gap between $\beta(1+2\epsilon)^t$ and $\beta(1+\epsilon)^t$ can be made arbitrarily large by choosing t large and β small enough. Recall that

$$\frac{d}{dt}X_i(t) = -\sqrt{X_i(t)^2 + \beta^2}\,\nabla F(\mathbf{X}(t))_i.$$

By Lemma 3.8, we have $\sqrt{3} \cdot 2(\mathbf{X}(t)^\top \mathbf{x}^\star) \geq 3\|\mathbf{X}(t)\|_2^2 - 1$. For any $i \in \mathcal{S}$ with $x_i^\star > 0$ and $X_i(t) \leq \frac{1}{2}x_i^\star$, we have

$$\nabla f(\mathbf{X}(t))_i \leq \left(\frac{\sqrt{3}}{2} - 1\right) \cdot 2\Big(\mathbf{X}(t)^\top \mathbf{x}^\star\Big)x_i^\star \leq -\frac{c_\star^2 \|\mathbf{X}_{\mathcal{S}}(t)\|_1}{4k}.$$

As before, Lemma 3.6 gives

$$\Big|\nabla F(\mathbf{X}(t))_i - \nabla f(\mathbf{X}(t))_i\Big| \leq \frac{c_\star^2 \|\mathbf{X}_{\mathcal{S}}(t)\|_1}{8k},$$

for $t \leq T_2$. This allows us to bound

$$\nabla F(\mathbf{X}(t))_i \leq -\frac{c_\star^2 \|\mathbf{X}_{\mathcal{S}}(t)\|_1}{8k},$$

from which it follows that $\frac{d}{dt}X_i(t) > 0$. The analogous result holds for coordinates $i \in \mathcal{S}$ with $x_i^\star < 0$. In other words, once we have $|X_i(t_0)| \geq \frac{1}{2}|x_i^\star|$ for some $t_0 > 0$, then $|X_i(t)| \geq \frac{1}{2}|x_i^\star|$ continues to hold for $t \geq t_0$. With this, we can define $I_1 \in \mathcal{S}$ to be the last coordinate which crosses this threshold, that is for which $\frac{|X_i(t)|}{|x_i^\star|} \geq \frac{1}{2}$, and assume without loss of generality that $x_{I_1}^\star > 0$. By definition, we have $|X_{I_1}(t)| \leq \frac{1}{2}|x_{I_1}^\star|$ for all $t \leq T_1$. For any $j \notin \mathcal{S}$ with $X_j(t) \geq 0$, we have

$$\nabla f(\mathbf{X}(t))_j \geq -(9 + 27\sigma)\sqrt{\frac{\log n}{m}}X_j(t).$$

As before, we can use Lemma 3.6 to obtain

$$\nabla F(\mathbf{X}(t))_j \geq -\frac{1}{4\sqrt{2}}\frac{c_\star^2 \|\mathbf{X}_{\mathcal{S}}(t)\|_1}{8k}.$$

The analogous result holds for $X_j(t) < 0$, which shows that, for any $j \notin \mathcal{S}$,

$$|\nabla F(\mathbf{X}(t))_j| \leq \frac{1}{4\sqrt{2}}|\nabla F(\mathbf{X}(t))_{I_1}| \tag{3.39}$$

holds for all $t \leq T_1$. The idea is that $X_{I_1}(t)$ and $X_j(t)$ both grow (approximately) exponentially, but at different (time-varying) rates. By the definition of I_1, we have

$X_{I_1}(T_1) = \frac{1}{2}x^\star_{I_1}$, and $X_j(T_1)$ can be made arbitrarily small by choosing a sufficiently small parameter β.

To make this rigorous, let $T_\beta = \inf\{t > 0 : \mathbf{X}_{I_1}(t) \geq \beta\}$ be the time when $\mathbf{X}_{I_1}(t)$ first reaches β. Then, since $\sqrt{X_{I_1}(t)^2 + \beta^2} \geq X_{I_1}(t)$, we can bound

$$X_{I_1}(t) \geq \beta \exp\left(-\int_{T_\beta}^t \nabla F(\mathbf{X}(s))ds\right) \quad \Rightarrow \quad \exp\left(-\int_{T_\beta}^t \nabla F(\mathbf{X}(s))ds\right) \leq \frac{X_{I_1}(t)}{\beta}.$$

Recalling the bound (3.39), we have for $j \notin \mathcal{S}$,

$$|X_j(T_\beta)| \leq \beta, \qquad \left|\frac{d}{dt}X_j(t)\right| \leq \frac{1}{4\sqrt{2}}\sqrt{X_j(t)^2 + \beta^2}\,|\nabla F(\mathbf{X}(t))_{I_1}|.$$

Then, since $\sqrt{X_j(t)^2 + \beta^2} \leq \sqrt{2}|X_j(t)|$ for $X_j(t) \geq \beta$, we can bound

$$X_j(t) \leq \beta \exp\left(-\frac{1}{4}\int_{T_\beta}^t \nabla F(\mathbf{X}(s))ds\right) \leq \beta\left(\frac{X_{I_1}(t)}{\beta}\right)^{1/4} \leq \beta^{3/4},$$

for all $t \leq T_1$, where we used the fact that $X_{I_1}(t) \leq \frac{1}{2}x^\star_{I_1} \leq 1$ for $t \leq T_1$. As this holds for every $j \notin \mathcal{S}$, we have, for all $t \leq T_1$,

$$\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1 \leq n\beta^{3/4} = \delta_1$$

Stage (ii)**, part** (a)**:** $T_1 < t \leq T_2$**, bound** $D_\Phi(\mathbf{x}^\star, \mathbf{X}(t))$

As in Stage (i), we first assume that we have already shown $\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1 \leq \delta$ for all $t \leq T_2$. As before, we can use Lemma 3.7 to derive the bound (3.36). By the definition of T_1, we have $|X_i(t)| \geq \frac{1}{2}|x^\star_i|$ for all $i \in [n]$. By Lemma 3.8 we also have $X_i(t)x^\star_i \geq 0$ (since we assumed $x^\star_{I_0} > 0$), so the assumptions for inequality (3.9) of Lemma 3.2 are satisfied, and we can bound

$$D_\Phi(\mathbf{x}^\star, \mathbf{X}(t)) \leq \frac{\sqrt{k}}{c_\star}\|\mathbf{X}_{\mathcal{S}}(t) - \mathbf{x}^\star_{\mathcal{S}}\|_2^2 + \|\mathbf{X}_{\mathcal{S}^c}(t)\|_1 \leq \frac{2\sqrt{k}}{c_\star}\|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2,$$

where we used that $\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1 \leq \delta \leq \frac{\sqrt{k}}{c_\star}\|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2$. With this, inequality (3.36) reads

$$\frac{d}{dt}D_\Phi(\mathbf{x}^\star, \mathbf{X}(t)) \leq -\frac{c_\star}{4\sqrt{k}}D_\Phi(\mathbf{x}^\star, \mathbf{X}(t)).$$

Using the fact that $\|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2 \leq 3$ by Lemma 3.8, we can bound $D_\Phi(\mathbf{x}^\star, \mathbf{X}(T_1)) \leq \frac{6\sqrt{k}}{c_\star}$. Hence, we have for $T_1 \leq t \leq T_2$,

$$D_\Phi(\mathbf{x}^\star, \mathbf{X}(t)) \leq D_\Phi(\mathbf{x}^\star, \mathbf{X}(T_1)) \exp\left(-\frac{c_\star}{4\sqrt{k}}(t - T_1)\right) \leq \frac{6\sqrt{k}}{c_\star}\exp\left(-\frac{c_\star}{4\sqrt{k}}(t - T_1)\right).$$

Together with the fact that $\|\mathbf{X}(t)\|_\infty^2 \le 2$ by Lemma 3.8, bound (3.8) of Lemma 3.2 implies

$$D_\Phi(\mathbf{x}^\star, \mathbf{X}(t)) > \frac{1}{3}\|\mathbf{X}(t) - \mathbf{x}^\star\|_2^2 \ge \frac{c^2 c_\star \sqrt{k}\delta}{3}$$

for $t \le T_2$, so we have $T_2 - T_1 \le \frac{4\sqrt{k}}{c_\star} \log \frac{18}{c^2 c_\star^2 \delta}$.

Stage (ii)**, part** (b)**:** $T_1 < t \le T_2$**, bound** $\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1$

Recall that $|X_j(t)| \le \beta^{3/4}$ for all $j \notin \mathcal{S}$ and $t \le T_1$. Further, as in the previous stage we can use Lemma 3.6 to show

$$|\nabla F(\mathbf{X}(t))_j| \le \frac{c_\star \|\mathbf{X}_{\mathcal{S}}(t)\|_1}{32k} \le \frac{c_\star}{16\sqrt{2k}},$$

where we used that $\|\mathbf{X}_{\mathcal{S}}(t)\|_1 \le \sqrt{k}\|\mathbf{X}_{\mathcal{S}}(t)\|_2 \le \sqrt{2k}$. Noting $\sqrt{x^2+\beta^2} \le \sqrt{2}x$ for $x \ge \beta$, we can bound, for $t \le T_2$,

$$\begin{aligned}
|X_j(t)| &\le \beta^{3/4} \exp\left(\frac{c_\star}{16\sqrt{k}}(T_2 - T_1)\right) \\
&\le \beta^{3/4} \exp\left(\frac{c_\star}{16\sqrt{k}} \frac{4\sqrt{k}}{c_\star} \log \frac{18}{c^2 c_\star^2 \delta}\right) \\
&\le \beta^{3/4} \left(\frac{18}{c^2 c_\star^2 \delta}\right)^{1/4} \\
&\le \frac{\delta}{n},
\end{aligned}$$

provided that $\beta \le (c^2 c_\star^2/18)^2/n^3$, where we used the definition $\delta = \sqrt{n\beta}$. This completes the proof that $\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1 \le \delta$ for all $t \le T_2$. □

3.7 Conclusion

In this paper, we establish a general theory of mirror descent with early stopping for the problem of noisy sparse phase retrieval. We provide a full convergence analysis in both continuous and discrete time and we show that, when equipped with the hypentropy mirror map and with proper initialization, early-stopped mirror descent achieves a nearly minimax-optimal rate of convergence, provided the number of measurements m is sufficiently large and the minimum (in modulus) non-zero signal component is on the order of $\|\mathbf{x}^\star\|_2/\sqrt{k}$. These conditions have been previously considered to establish theoretical results for sparse phase retrieval [90, 122, 138].

Previous procedures achieving a nearly minimax-optimal rate of convergence include empirical risk minimization with sparsity constraint [71], which does not lead to a tractable algorithm, and thresholded Wirtinger flow [20], which relies on thresholding steps to promote sparsity. In contrast, unlike most existing algorithms designed to solve sparse phase retrieval, mirror descent does not rely on added regularization terms or thresholding steps to promote sparsity, and no a-priori knowledge of the sparsity k or the noise level σ is needed to *run* the algorithm. Our numerical simulations attest that a standard data-dependent stopping rule that does not require knowledge of k or σ yields results that validate our theoretical findings.

Most of the literature on early stopping for iterative gradient-based methodologies has focused on the convex setting of ridge regression and kernel methods via Euclidean gradient descent and boosting. Our results establish connections between early stopping and sparsity in the non-convex setting of sparse phase retrieval via the general framework of mirror descent. The potential-based analysis that we present unveils and exploits a quantitative version of variational coherence that is satisfied along the path traced by the iterates of mirror descent, and it might inspire similar approaches to investigate implicit regularization via early stopping in other non-convex settings in statistical inference.

3.8 Chapter appendix

The appendix is organized as follows. We prove Theorem 3.4 in Appendix 3.8.1. In Appendix 3.8.2, we prove Lemmas 3.6–3.8 from Section 3.6.2. Further technical lemmas are provided in Appendix 3.8.3.

3.8.1 Proof of Theorem 3.4

As in the proof of Theorem 3.3, we assume $\|\mathbf{x}^\star\|_2 = 1$ and $(\mathbf{X}^0)^\top \mathbf{x}^\star \geq 0$ for notational simplicity. The proof of Theorem 3.4 largely follows the same steps as the proof of Theorem 3.3, and we can reuse many of the bounds shown before. Before proving the theorem, we derive some generic bounds on how much X_i^t can increase or decrease in

one iteration. The update (3.22) can be written as (see e.g. Theorem 24 [48])

$$X_i^{t+1} - X_i^t = \frac{\sqrt{(X_i^t)^2 + \beta^2} + X_i^t}{2}\Big(\exp(-\eta \nabla F(\mathbf{X}^t)_i) - 1\Big) - \frac{\sqrt{(X_i^t)^2 + \beta^2} - X_i^t}{2}\Big(\exp(\eta \nabla F(\mathbf{X}^t)_i) - 1\Big). \tag{3.40}$$

Assuming that the conditions for the simplified bound in Lemma 3.6 are satisfied, we can bound

$$\Big|\nabla F(\mathbf{X}^t)_i - \nabla f(\mathbf{X}^t)_i\Big| \leq \frac{0.1 c_\star\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big)}{\sqrt{k}} \leq 0.1\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big)|x_i^\star|, \tag{3.41}$$

with probability $1 - c_p n^{-10}$, where we used (3.28) to bound $\|\mathbf{X}_{\mathcal{S}}^t\|_1 \leq \frac{\sqrt{k}}{c_\star}((\mathbf{X}^t)^\top \mathbf{x}^\star)$ in the first inequality, and for the second inequality we used the assumption $x_{min}^\star \geq \frac{c_\star}{\sqrt{k}}$. Recalling the definition of the population gradient

$$\nabla f(\mathbf{x}) = \Big(3\|\mathbf{x}\|_2^2 - 1\Big)\mathbf{x} - 2\Big(\mathbf{x}^\top \mathbf{x}^\star\Big)\mathbf{x}^\star,$$

an application of the triangle inequality yields

$$|\nabla F(\mathbf{X}^t)_i| \leq \Big|\Big(3\|\mathbf{X}^t\|_2^2 - 1\Big)X_i^t\Big| + \Big|2.1\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big)x_i^\star\Big| \leq 7.1\sqrt{2}, \tag{3.42}$$

provided that $\|\mathbf{X}^t\|_2 \leq \sqrt{2}$, where we used that $(\mathbf{X}^t)^\top \mathbf{x}^\star \leq \|\mathbf{X}^t\|_2 \|\mathbf{x}^\star\|_2$ by the Cauchy-Schwarz inequality. Hence, we have $|\eta \nabla F(\mathbf{X}^t)_i| \leq \frac{1}{2}$, provided that $\eta \leq \frac{1}{14.2\sqrt{2}}$.

To bound the exponentials in (3.40), we use that $e^x \leq 1 + \frac{3}{2}x$ for $0 \leq x \leq \frac{1}{2}$, and $1 + x \leq e^x$ for $-\frac{1}{2} \leq x \leq 0$. Without loss of generality, we assume $X_i^t \geq 0$; analogous bounds can be derived for $X_i^t < 0$. If $\nabla F(\mathbf{X}^t)_i < 0$ (i.e. $X_i^{t+1} > X_i^t$), then we can bound

$$\begin{aligned} X_i^{t+1} - X_i^t &\leq \frac{\sqrt{(X_i^t)^2 + \beta^2} + X_i^t}{2}\Big(-\frac{3}{2}\eta \nabla F(\mathbf{X}^t)_i\Big) - \frac{\sqrt{(X_i^t)^2 + \beta^2} - X_i^t}{2}\eta \nabla F(\mathbf{X}^t)_i \\ &\leq -\frac{3}{2}\eta \nabla F(\mathbf{X}^t)_i \sqrt{(X_i^t)^2 + \beta^2}. \end{aligned}$$

If $\nabla F(\mathbf{X}^t)_i > 0$, we obtain the analogous lower bound

$$X_i^{t+1} - X_i^t \geq -\frac{5}{4}\eta \nabla F(\mathbf{X}^t)_i \sqrt{(X_i^t)^2 + \beta^2}.$$

Combining both bounds yields an upper bound on how much X_i^t can increase or decrease in one iteration:

$$|X_i^{t+1} - X_i^t| \leq \frac{3}{2}\eta |\nabla F(\mathbf{X}^t)_i| \sqrt{(X_i^t)^2 + \beta^2}. \tag{3.43}$$

Similarly, we can use the bounds $1 + x \le e^x$ for $0 \le x \le \frac{1}{2}$ and $e^x \le 1 + \frac{x}{2}$ for $-\frac{1}{2} \le x \le 0$ to obtain the lower bound

$$|X_i^{t+1} - X_i^t| \ge \frac{1}{2}\eta |\nabla F(\mathbf{X}^t)_i| |X_i^t|. \tag{3.44}$$

Proof of Theorem 3.4. The proof of Theorem 3.4 relies on the following identity, which follows from the definition of the Bregman divergence and the mirror descent update (3.5), to show that $D_\Phi(\mathbf{x}^\star, \mathbf{X}^t)$ decreases as claimed:

$$D_\Phi(\mathbf{x}^\star, \mathbf{X}^{t+1}) - D_\Phi(\mathbf{x}^\star, \mathbf{X}^t) = -\eta \Big\langle \nabla F(\mathbf{X}^t), \mathbf{X}^t - \mathbf{x}^\star \Big\rangle + D_\Phi(\mathbf{X}^t, \mathbf{X}^{t+1}). \tag{3.45}$$

The first term in (3.45) has been bounded in the continuous-time case, and we can reuse those bounds. The second term measures the distance between the iterates $\mathbf{X}^t$ and $\mathbf{X}^{t+1}$ in terms of the Bregman divergence due to the discretization, and we will derive bounds for the second term similar to those shown in the continuous-time case for the first term. The function Φ is $\frac{1}{2}$-strongly convex with respect to the ℓ_2-norm on the ℓ_2-ball $\{\mathbf{x} \in \mathbb{R}^n : \|\mathbf{x}\|_2^2 \le 2\}$ provided $\beta \le \sqrt{2}$, since the Hessian $\nabla^2\Phi(\mathbf{x})$ is a diagonal matrix with entries

$$\nabla^2\Phi(\mathbf{x})_{ii} = \frac{1}{\sqrt{x_i^2 + \beta^2}} \ge \frac{1}{\sqrt{2 + \beta^2}} \ge \frac{1}{2}.$$

Hence, we can bound

$$\begin{aligned}
D_\Phi(\mathbf{X}^t, \mathbf{X}^{t+1}) &= \Phi(\mathbf{X}^t) - \Phi(\mathbf{X}^{t+1}) - \langle \nabla\Phi(\mathbf{X}^{t+1}), \mathbf{X}^t - \mathbf{X}^{t+1} \rangle \\
&\le \langle \nabla\Phi(\mathbf{X}^t) - \nabla\Phi(\mathbf{X}^{t+1}), \mathbf{X}^t - \mathbf{X}^{t+1} \rangle - \frac{1}{4}\|\mathbf{X}^t - \mathbf{X}^{t+1}\|_2^2 \\
&= \langle \eta \nabla F(\mathbf{X}^t), \mathbf{X}^t - \mathbf{X}^{t+1} \rangle - \frac{1}{4}\|\mathbf{X}^t - \mathbf{X}^{t+1}\|_2^2 \\
&\le \eta \|\nabla F(\mathbf{X}^t)\|_2 \|\mathbf{X}^t - \mathbf{X}^{t+1}\|_2 - \frac{1}{4}\|\mathbf{X}^t - \mathbf{X}^{t+1}\|_2^2 \\
&\le \eta^2 \|\nabla F(\mathbf{X}^t)\|_2^2,
\end{aligned} \tag{3.46}$$

where the second line follows from strong convexity of Φ, the third line from the definition of the mirror descent update, the fourth line from the Cauchy-Schwarz inequality, and for the last line we optimized the quadratic function in $\|\mathbf{X}^t - \mathbf{X}^{t+1}\|_2$.

As in continuous time, we divide the proof of Theorem 3.4 into two stages bounded by

$$\begin{aligned}
T_1 &= \inf\left\{ t > 0 : \min_{i \in \mathcal{S}} \frac{|X_i^t|}{|x_i^\star|} > \frac{1}{2} \right\}, \text{ and} \\
T_2 &= \inf\left\{ t > 0 : \frac{\|\mathbf{X}^t - \mathbf{x}^\star\|_2^2}{\|\mathbf{x}^\star\|_2^2} \le c^2 \max\left\{ c_\star \sqrt{k}\delta,\ \frac{\sigma^2}{\|\mathbf{x}^\star\|_2^4} \frac{k \log n}{m} \right\} \right\},
\end{aligned}$$

respectively, where $\delta = \sqrt{n\beta}$, and (a) bound the length of the stage by using (3.45) to bound T_i, and (b) show that off-support coordinates $\|\mathbf{X}^t_{\mathcal{S}^c}\|_1$ stay sufficiently small. Throughout the proof we will assume that the inequalities in Lemmas 3.6–3.8 are satisfied, which happens with probability at least $1 - c_p n^{-10}$. As in the continuous-time case, the conditions for the simplified bounds in Lemmas 3.6 and 3.7 are satisfied for $t \le T_2$.

Stage (i), part (a): $t \le T_1$, bound T_1

Assume for now that we have already shown $\|\mathbf{X}^t_{\mathcal{S}^c}\|_1 < \delta_1 = 2n\beta^{3/4}$ for all $t \le T_1$. As in continuous time, we consider the following two cases:

- **Case 1:** $\|\mathbf{X}^t\|_2^2 \le \frac{2}{5}$

  In this case, we have shown in (3.34) that

  $$-\eta\Big\langle \nabla F(\mathbf{X}^t), \mathbf{X}^t - \mathbf{x}^\star \Big\rangle \le -\frac{\eta}{4}\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big).$$

  It remains to bound the term $D_\Phi(\mathbf{X}^t, \mathbf{X}^{t+1})$. Using the bounds (3.42), (3.46) and the inequality $(a+b)^2 \le 2a^2 + 2b^2$, we can bound

  $$\begin{aligned} D_\Phi(\mathbf{X}^t, \mathbf{X}^{t+1}) &\le 2\eta^2 \sum_{i=1}^n \Big(3\|\mathbf{X}^t\|_2^2 - 1\Big)^2 (X_i^t)^2 + 2.1^2 \Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big)^2 (x_i^\star)^2 \\ &\le 2\eta^2 \Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big)^2 \sum_{i=1}^n \Big(12 (X_i^t)^2 + 2.1^2 (x_i^\star)^2\Big) \\ &\le \frac{\eta}{8}\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big) \end{aligned}$$

  for $\eta \le \frac{1}{94}$, where for the second inequality we used (3.26) of Lemma 3.8 to bound $3\|\mathbf{X}^t\|_2^2 - 1 \le \sqrt{3} \cdot 2((\mathbf{X}^t)^\top \mathbf{x}^\star)$, and for the last inequality we used the assumptions $\|\mathbf{X}^t\|_2^2 \le 2/5$ and $\|\mathbf{x}^\star\|_2 = 1$, and the Cauchy-Schwarz inequality to bound $(\mathbf{X}^t)^\top \mathbf{x}^\star \le \sqrt{2/5}$. Plugging this bound into (3.45), we have

  $$D_\Phi(\mathbf{x}^\star, \mathbf{X}^{t+1}) - D_\Phi(\mathbf{x}^\star, \mathbf{X}^t) \le -\frac{\eta}{8}\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big) \le -\frac{c_\star \eta}{16\sqrt{k}}, \tag{3.47}$$

  where for the second inequality we used (3.28).

- **Case 2:** $\|\mathbf{X}^t\|_2^2 \ge \frac{2}{5}$

  In this case, we have already shown that

  $$-\eta\Big\langle \nabla F(\mathbf{X}^t), \mathbf{X}^t - \mathbf{x}^\star \Big\rangle \le -\frac{\eta}{25} \tag{3.48}$$

if $4\|\mathbf{X}^t\|_2^2 - 2 - \|\mathbf{X}^t - \mathbf{x}^\star\|_2^2 \leq 0$, and

$$-\eta\Big\langle \nabla F(\mathbf{X}^t), \mathbf{X}^t - \mathbf{x}^\star \Big\rangle \leq -\frac{\eta}{2}\|\mathbf{X}^t - \mathbf{x}^\star\|_2^2 \tag{3.49}$$

if $4\|\mathbf{X}^t\|_2^2 - 2 - \|\mathbf{X}^t - \mathbf{x}^\star\|_2^2 > 0$. It remains to bound $D_\Phi(\mathbf{X}^t, \mathbf{X}^{t+1})$ in terms of the expressions in (3.48) and (3.49). As the bound corresponding to (3.48) can be obtained the same (but easier) way as for (3.49), we only show the computation for the latter case.

Substituting $\mathbf{Z} = \mathbf{X}^t - \mathbf{x}^\star$, we can write

$$\begin{aligned}
|\nabla f(\mathbf{X}(t))_i| &= \Big|\Big(3\|\mathbf{X}^t\|_2^2 - 1\Big)Z_i + \Big(3\|\mathbf{Z} + \mathbf{x}^\star\|_2^2 - 1 - 2(\mathbf{Z} + \mathbf{x}^\star)^\top \mathbf{x}^\star\Big)x_i^\star\Big| \\
&= \Big|\Big(3\|\mathbf{X}^t\|_2^2 - 1\Big)Z_i + \mathbf{Z}^\top(3\mathbf{X}^t + \mathbf{x}^\star)x_i^\star\Big| \\
&\leq 5|Z_i| + (1 + 3\sqrt{2})\|\mathbf{Z}\|_2|x_i^\star|,
\end{aligned}$$

where we used $\|\mathbf{X}^t\|_2 \leq \sqrt{2}$ by (3.25) of Lemma 3.8. By Lemma 3.6, we have

$$|\nabla F(\mathbf{X}^t)_i - \nabla f(\mathbf{X}^t)_i| \leq \frac{0.1c_\star\|\mathbf{X}^t - \mathbf{x}^\star\|_2}{\sqrt{k}} \leq 0.1\|\mathbf{X}^t - \mathbf{x}^\star\|_2|x_i^\star|,$$

for $t \leq T_2$, so that we can bound

$$|\nabla F(\mathbf{X}^t)_i| \leq 5.4\|\mathbf{X}^t - \mathbf{x}^\star\|_2|x_i^\star| + 5|X_i^t - x_i^\star|.$$

As in the previous case, this bound together with (3.42) and (3.46) implies

$$\begin{aligned}
D_\Phi(\mathbf{X}^t, \mathbf{X}^{t+1}) &\leq 2\eta^2 \sum_{i=1}^n \Big(5.4^2\|\mathbf{X}^t - \mathbf{x}^\star\|_2^2(x_i^\star)^2 + 5^2(\mathbf{X}_i^t - x_i^\star)^2\Big) \\
&\leq 2\eta^2\|\mathbf{X}^t - \mathbf{x}^\star\|_2^2\left(5^2 + \sum_{i=1}^n 5.4^2(x_i^\star)^2\right) \\
&\leq \frac{\eta}{4}\|\mathbf{X}^t - \mathbf{x}^\star\|_2^2
\end{aligned}$$

for $\eta \leq \frac{1}{434}$, where we used that $\|\mathbf{x}^\star\|_2 = 1$. Plugging this bound into (3.45) yields

$$D_\Phi(\mathbf{x}^\star, \mathbf{X}^{t+1}) - D_\Phi(\mathbf{x}^\star, \mathbf{X}^t) \leq -\frac{\eta}{4}\|\mathbf{X}^t - \mathbf{x}^\star\|_2^2. \tag{3.50}$$

We can now bound T_1. Define

$$T_0 = \inf\left\{t > 0 : \|\mathbf{X}^t\|_2^2 > \frac{2}{5}\right\},$$

as the time until which Case 1 holds. Then, the bound (3.47) from Case 1 shows that

$$\min\{T_1, T_0\} \le \frac{16\sqrt{k} D_\Phi(\mathbf{x}^\star, \mathbf{X}^0)}{c_\star \eta} \le \frac{16}{c_\star \eta} k \log \frac{1}{\beta} + \frac{16\sqrt{k}}{c_\star \eta} \le \frac{c_2}{2\eta} k \log \frac{1}{\beta},$$

provided that $c_2 \ge \frac{32}{c_\star} + \frac{32}{c_\star \sqrt{k} \log \frac{1}{\beta}}$, where we used (3.30) to bound the initial Bregman divergence $D_\Phi(\mathbf{x}^\star, \mathbf{X}^0)$. If $T_1 < T_0$, then we have bounded T_1 as desired. Otherwise, we can use (3.50) and the analogous bound corresponding to (3.48) to control $T_1 - T_0$. As in the continuous-time case, the bound corresponding to (3.48) can apply for at most $t \le 50\sqrt{k} \log \frac{1}{\beta} + 50$ iterations. Following the same steps as in the continuous-time case, we can derive the bounds (3.37) and (3.38), which allow us to bound the ℓ_2-distance $\|\mathbf{X}^t - \mathbf{x}^\star\|_2^2$ in terms of the Bregman divergence $D_\Phi(\mathbf{x}^\star, \mathbf{X}^t)$. Combined with (3.50), this gives

$$D_\Phi(\mathbf{x}^\star, \mathbf{X}^{t+1}) - D_\Phi(\mathbf{x}^\star, \mathbf{X}^t) \le -\frac{\eta}{4} \|\mathbf{X}^t - \mathbf{x}^\star\|_2^2 \le -\frac{c_\star^2 \eta}{16 k \log \frac{1}{\beta}} D_\Phi(\mathbf{x}^\star, \mathbf{X}^t),$$

which shows that $D_\Phi(\mathbf{x}^\star, \mathbf{X}^t)$ decreases at least linearly at the rate $(1 - c_\star^2 \eta / (16 k \log \frac{1}{\beta}))$ for $T_0 \le t < T_1$. We have

$$D_\Phi(\mathbf{x}^\star, \mathbf{X}^{T_0}) \le D_\Phi(\mathbf{x}^\star, \mathbf{X}^0) \le \sqrt{k} \log \frac{1}{\beta} + 1,$$

and, for $t \le T_1$,

$$\frac{c_\star^2}{4k} \le \|\mathbf{X}^t - \mathbf{x}^\star\|_2^2 \le 2\sqrt{2 + \beta^2}\, D_\Phi(\mathbf{x}^\star, \mathbf{X}^t) \le 3 D_\Phi(\mathbf{x}^\star, \mathbf{X}^t),$$

where for the second inequality we used the bound (3.8) of Lemma 3.2 and $\|\mathbf{X}^t\|_\infty \le \sqrt{2}$ by Lemma 3.8. This implies

$$T_1 - T_0 \le \frac{16 k \log \frac{1}{\beta}}{c_\star^2 \eta} \log\left(\frac{12}{c_\star^2} k^{\frac{3}{2}} \log \frac{1}{\beta} + \frac{12}{c_\star^2} k \right) \le \frac{c_2}{2\eta} k \log\left(\frac{1}{\beta}\right) \log\left(k \log \frac{1}{\beta} \right)$$

provided that $c_2 \ge \frac{48}{c_\star^2} + \frac{32}{c_\star^2} \log \frac{24}{c_\star^2}$.

Stage (i), part (b): $t \le T_1$, bound $\|\mathbf{X}^t_{\mathcal{S}^c}\|_1$

As in the continuous-time case, let $I_1 \in \mathcal{S}$ be the last coordinate which satisfies $|X_i^t| \ge \frac{1}{2}|x_i^\star|$ for some $t > 0$, and let $j \notin \mathcal{S}$. To simplify notation, assume $X_{I_1}^t, X_j^t > 0$, in which case we can show $\nabla F(\mathbf{X}^t)_{I_1} < 0$ for all $t \le T_1$ as in the continuous-time case.

As shown in (3.39), we have (using different constants in Lemma 3.6)

$$|\nabla F(\mathbf{X}^t)_j| \le \frac{\sqrt{2}}{30} |\nabla F(\mathbf{X}^t)_{I_1}| \tag{3.51}$$

for all $t \le T_1$. The idea of the proof is the same as in continuous time: the growth of both $X^t_{I_1}$ and X^t_j is bounded by exponentials with different (time-varying) rates. Using (3.44), we have

$$X^{t+1}_{I_1} \ge X^t_{I_1} - \frac{1}{2}\eta \nabla F(\mathbf{X}^t)_{I_1} X^t_{I_1} \ge \Big(1 - \frac{1}{2}\eta \nabla F(\mathbf{X}^t)_{I_1}\Big) X^t_{I_1}, \tag{3.52}$$

and similarly, using (3.43), we have, for $X^t_j \ge \beta$,

$$X^{t+1}_j \le \Big(1 + \frac{3}{2}\sqrt{2}\eta |\nabla F(\mathbf{X}^t)_j|\Big) X^t_j. \tag{3.53}$$

Let $G_t = -\nabla F(\mathbf{X}^t)_{I_1}$, and let $T_\beta = \min\{t : |X^t_{I_1}| \ge \beta\}$. Then, we use (3.52) to bound

$$X^{T_1}_{I_1} \ge X^{T_\beta}_{I_1} \prod_{t=T_\beta}^{T_1} \Big(1 + \frac{1}{2}\eta G_t\Big),$$

and, since $|X^{T_\beta}_j| \le 2\beta$ and $\sqrt{(X^{T_\beta}_j)^2 + \beta^2} \le \sqrt{2}|X^{T_\beta}_j|$ for $|X^{T_\beta}_j| \ge \beta$, we can use (3.51) and (3.53) to bound

$$X^{T_1}_j \le 2\beta \prod_{t=T_\beta}^{T_1} \Big(1 + \frac{1}{10}\eta G_t\Big) \le 2\beta \left(\prod_{t=T_\beta}^{T_1} 1 + \frac{1}{2}\eta G_t\right)^{\frac{1}{4}} \le 2\beta \left(\frac{X^{T_1}_{i_1}}{X^{T_\beta}_{i_1}}\right)^{\frac{1}{4}} \le 2\beta \left(\frac{1}{2\beta}\right)^{\frac{1}{4}} \le 2\beta^{\frac{3}{4}},$$

where for the second inequality we used the inequality $\frac{x}{x+1} \le \log(1+x) \le x$ for $x \ge 0$, with which we have, for $\eta G_t \le \frac{1}{2}$,

$$\log\Big(1 + \frac{1}{10}\eta G_t\Big) \le \frac{1}{10}\eta G_t \le \frac{1}{4}\frac{\eta G_t/2}{1 + \eta G_t/2} \le \frac{1}{4}\log\Big(1 + \frac{1}{2}\eta G_t\Big).$$

Stage (ii)**, part** (a)**:** $T_1 < t \le T_2$**, bound** $D_\Phi(\mathbf{x}^\star, \mathbf{X}^t)$

As in the continuous-time case, we can use Lemma 3.2 to bound

$$D_\Phi(\mathbf{x}^\star, \mathbf{X}^t) \le \frac{2\sqrt{k}}{c_\star} \|\mathbf{X}^t - \mathbf{x}^\star\|_2^2,$$

and Lemma 3.8 to bound $D_\Phi(\mathbf{x}^\star, \mathbf{X}^{T_1}) \le \frac{6\sqrt{k}}{c_\star}$. With this, inequality (3.50) reads

$$D_\Phi(\mathbf{x}^\star, \mathbf{X}^{t+1}) - D_\Phi(\mathbf{x}^\star, \mathbf{X}^t) \le -\frac{c_\star \eta}{8\sqrt{k}} D_\Phi(\mathbf{x}^\star, \mathbf{X}^t).$$

Hence, we have for $T_1 \le t \le T_2$,

$$D_\Phi(\mathbf{x}^\star, \mathbf{X}^t) \le D_\Phi(\mathbf{x}^\star, \mathbf{X}^{T_1}) \Big(1 - \frac{c_\star \eta}{8\sqrt{k}}\Big)^{t - T_1} \le \frac{6\sqrt{k}}{c_\star} \Big(1 - \frac{c_\star \eta}{8\sqrt{k}}\Big)^{t - T_1}.$$

Together with the fact that $\|\mathbf{X}^t\|_\infty^2 \leq 2$, bound (3.8) of Lemma 3.2 implies

$$D_\Phi(\mathbf{x}^\star, \mathbf{X}^t) \geq \frac{1}{3}\|\mathbf{X}^t - \mathbf{x}^\star\|_2^2 \geq \frac{c^2 c_\star \sqrt{k}\delta}{3}$$

for $t \leq T_2$, so we can bound

$$T_2 - T_1 \leq \frac{8\sqrt{k}}{c_\star \eta} \log \frac{18}{c^2 c_\star^2 \delta}.$$

Stage (ii)**, part** (b)**:** $T_1 < t \leq T_2$**, bound** $\|\mathbf{X}^t_{\mathcal{S}^c}\|_1$

Recall that we have shown $|X_j^t| \leq 2\beta^{\frac{3}{4}}$ for all $j \notin \mathcal{S}$ and $t \leq T_1$. As in the continuous-time case, we can use Lemma 3.6 to show

$$|\nabla F(\mathbf{X}^t)_j| \leq \frac{c_\star \|\mathbf{X}^t_{\mathcal{S}}\|_1}{64k} \leq \frac{c_\star}{32\sqrt{2k}},$$

where we used that $\|\mathbf{X}^t_{\mathcal{S}}\|_1 \leq \sqrt{k}\|\mathbf{X}^t_{\mathcal{S}}\|_2 \leq \sqrt{2k}$. As in Stage (i), we can bound, for $t \leq T_2$ (as $\sqrt{x^2+\beta^2} \leq \sqrt{2}x$ for $x \geq \beta$),

$$\begin{aligned}
|X_j(t)| &\leq 2\beta^{\frac{3}{4}}\left(1 + \frac{3}{\sqrt{2}}\eta\frac{c_\star}{32\sqrt{k}}\right)^{T_2 - T_1} \\
&\leq 2\beta^{\frac{3}{4}} \exp\left(\frac{\eta c_\star}{32\sqrt{k}} \frac{8\sqrt{k}}{c_\star \eta} \log \frac{18}{c^2 c_\star^2 \delta}\right) \\
&\leq 2\beta^{\frac{3}{4}}\left(\frac{18}{c^2 c_\star^2 \delta}\right)^{\frac{1}{4}} \\
&\leq \frac{\delta}{n},
\end{aligned}$$

provided that $\beta \leq (c^2 c_\star^2/288)^2/n^3$, where we used the definition $\delta = \sqrt{n\beta}$. This completes the proof that $\|\mathbf{X}^t_{\mathcal{S}^c}\|_1 \leq \delta$ for all $t \leq T_2$. □

3.8.2 Proof of supporting lemmas

In this section, we provide the proofs of the supporting lemmas stated in Section 3.6.2. Throughout this section, we will assume $\|\mathbf{x}^\star\|_2 = 1$ for notational simplicity's sake. The general case $\|\mathbf{x}^\star\|_2 \neq 1$ follows by writing $\mathbf{x}^\star = \frac{\mathbf{x}^\star}{\|\mathbf{x}^\star\|_2}\|\mathbf{x}^\star\|_2$, $\mathbf{x} = \frac{\mathbf{x}}{\|\mathbf{x}^\star\|_2}\|\mathbf{x}^\star\|_2$ and $\varepsilon_j = \frac{\varepsilon_j}{\|\mathbf{x}^\star\|_2}\|\mathbf{x}^\star\|_2$ in what follows.

Proof of Lemma 3.6. We will bound the difference $|\nabla F(\mathbf{x})_i - \nabla f(\mathbf{x})_i|$ for any $i \in [n]$. Lemma 3.6 then follows by taking a union bound. First, we write $\mathbf{w} \in \mathbb{R}^n$ for the

vector $\mathbf{x}_{\mathcal{S}}$ padded with zeros, that is $w_i = x_i$ for $i \in \mathcal{S}$ and $w_i = 0$ otherwise. For convenience, we denote by

$$\nabla \widetilde{F}(\mathbf{x}) = \frac{1}{m}\sum_{j=1}^{m}\Big((\mathbf{A}_j^\top \mathbf{x})^2 - (\mathbf{A}_j^\top \mathbf{x}^\star)^2\Big)\Big(\mathbf{A}_j^\top \mathbf{x}\Big)\mathbf{A}_j$$

the gradient if the measurements were noiseless. Then, since the random variables $\{\varepsilon_j\}_{j=1}^m$ are centered and independent of the measurement vectors $\{\mathbf{A}_j\}_{j=1}^m$, we can compute $\nabla f(\mathbf{x}) = \mathbb{E}[\nabla F(\mathbf{x})] = \mathbb{E}[\nabla \widetilde{F}(\mathbf{x})]$ and decompose

$$\begin{aligned}\Big|\nabla F(\mathbf{x})_i - \nabla f(\mathbf{x})_i\Big| \le \Big|\nabla F(\mathbf{x})_i - \nabla \widetilde{F}(\mathbf{x})_i\Big| + \Big|\nabla \widetilde{F}(\mathbf{x})_i - \nabla \widetilde{F}(\mathbf{w})_i\Big| \\ + \Big|\nabla \widetilde{F}(\mathbf{w})_i - \mathbb{E}[\nabla \widetilde{F}(\mathbf{w})_i]\Big| + \Big|\mathbb{E}[\nabla \widetilde{F}(\mathbf{w})_i] - \mathbb{E}[\nabla \widetilde{F}(\mathbf{x})_i]\Big|, \quad (3.54)\end{aligned}$$

and we will bound the four terms separately.

Step 1: Bound the term $|\nabla F(\mathbf{x})_i - \nabla \widetilde{F}(\mathbf{x})_i|$

We begin by bounding the first term in (3.54) by

$$c\sigma\sqrt{\frac{\log n}{m}}\min\Big\{\|\mathbf{x}\|_1,\ 1 + \delta + \sqrt{k}\|\mathbf{x}_{\mathcal{S}} - \mathbf{x}^\star_{\mathcal{S}}\|_2\Big\}.$$

To obtain the second bound in the minimum, we write $\mathbf{z} = \mathbf{x} - \mathbf{x}^\star$ and bound

$$\begin{aligned}\Big|\nabla F(\mathbf{x})_i - \nabla \widetilde{F}(\mathbf{x})_i\Big| &= \left|\frac{1}{m}\sum_{j=1}^{m}\varepsilon_j(\mathbf{A}_j^\top \mathbf{x})A_{ji}\right| \\ &\le \left|\frac{1}{m}\sum_{j=1}^{m}\varepsilon_j(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})A_{ji}\right| + \left|\frac{1}{m}\sum_{j=1}^{m}\varepsilon_j(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})A_{ji}\right| \\ &\le \left|\frac{1}{m}\sum_{j=1}^{m}\varepsilon_j\Big(\mathbf{A}_{j,\mathcal{S}}^\top(\mathbf{z}_{\mathcal{S}} + \mathbf{x}^\star_{\mathcal{S}})\Big)A_{ji}\right| + \left|\sum_{l\notin\mathcal{S}} x_l \frac{1}{m}\sum_{j=1}^{m}\varepsilon_j A_{jl}A_{ji}\right| \\ &\le \Big(\|\mathbf{z}_{\mathcal{S}}\|_1 + \|\mathbf{x}_{\mathcal{S}^c}\|_1\Big)\max_{l\in[n]}\left|\frac{1}{m}\sum_{j=1}^{m}\varepsilon_j A_{jl}A_{ji}\right| + \left|\frac{1}{m}\sum_{j=1}^{m}\varepsilon_j(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}^\star_{\mathcal{S}})A_{ji}\right| \\ &\le c\sigma\sqrt{\frac{\log n}{m}}\Big(\|\mathbf{z}_{\mathcal{S}}\|_1 + 1 + \delta\Big),\end{aligned}$$

with probability $1 - \frac{c_p}{3}n^{-11}$ for any constant $c > 0$ that is at least the universal constant from Lemma 3.13, where we used Hölder's inequality in the penultimate line. The second bound in the minimum follows as $\|\mathbf{z}_{\mathcal{S}}\|_1 \le \sqrt{k}\|\mathbf{z}_{\mathcal{S}}\|_2$. The first bound in the minimum follows by directly bounding $\frac{1}{m}\sum_{j=1}^m \varepsilon_j(\mathbf{A}_j^\top \mathbf{x})A_{ji}$ using Hölder's inequality and Lemma 3.13 as above, without substituting $\mathbf{z} = \mathbf{x} - \mathbf{x}^\star$.

Step 2: Bound the term $|\nabla \widetilde{F}(\mathbf{x})_i - \nabla \widetilde{F}(\mathbf{w})_i|$

Next, we bound the second term in (3.54) by $\frac{c}{2}\delta$. We can write

$$\nabla \widetilde{F}(\mathbf{x})_i = \frac{1}{m}\sum_{j=1}^{m}\Big((\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}} + \mathbf{A}_{j,\mathcal{S}^c}\mathbf{x}_{\mathcal{S}^c})^3 - (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}} + \mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})(\mathbf{A}_j^\top \mathbf{x}^\star)^2\Big)A_{ji}.$$

Then, we have

$$\begin{aligned}\nabla \widetilde{F}(\mathbf{x})_i - \nabla \widetilde{F}(\mathbf{w})_i = \frac{3}{m}\sum_{j=1}^{m} A_{ji}(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})^2(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c}) + \frac{3}{m}\sum_{j=1}^{m} A_{ji}(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})^2 \\ + \frac{1}{m}\sum_{j=1}^{m} A_{ji}(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})^3 - \frac{1}{m}\sum_{j=1}^{m} A_{ji}(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})(\mathbf{A}_j^\top \mathbf{x}^\star)^2. \quad (3.55)\end{aligned}$$

These four terms can be bounded as follows: for the first term, we have

$$\begin{aligned}\left|\frac{1}{m}\sum_{j=1}^{m} A_{ji}(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})^2(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})\right| &= \left|\sum_{l\notin\mathcal{S}} x_l \frac{1}{m}\sum_{j=1}^{m} A_{ji}A_{jl}(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})^2\right| \\ &\le \|\mathbf{x}_{\mathcal{S}^c}\|_1 \max_{l\notin\mathcal{S}}\left|\frac{1}{m}\sum_{j=1}^{m} A_{ji}A_{jl}(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})^2\right| \\ &\le \|\mathbf{x}_{\mathcal{S}^c}\|_1 \max_{l\notin\mathcal{S}}\sqrt{\frac{1}{m}\sum_{j=1}^{m} A_{ji}^2A_{jl}^2}\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})^4},\end{aligned}$$

where we used Hölder's inequality to obtain both inequalities. The first sum is bounded by Lemma 3.13: recalling $m \ge c_s(\gamma)k^2\log^2 n$, we have with probability at least $1 - c_2 n^{-13}$, where c_2 is the universal constant from Lemma 3.13,

$$\max_{l\notin\mathcal{S}}\left|\frac{1}{m}\sum_{j=1}^{m} A_{ji}^2A_{jl}^2\right| \le 1 + \frac{1}{k} \le 2.$$

By Lemma 3.11 with $t = 5\sqrt{\log n}$, we can bound, with probability $1 - 4n^{-12.5}$,

$$\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})^4} \le \frac{1}{\sqrt{m}}\Big((3m)^{\frac{1}{4}} + \sqrt{k} + 5\sqrt{\log n}\Big)^2 \le 11$$

for all $\mathbf{x} \in \mathcal{X}$, where we used the fact that $5\sqrt{\log n} \le m^{\frac{1}{4}}$ holds, provided we have $c_s(\gamma) \ge 5^4 \min\{k^{-2},\ \log^{-3} n\}$. Put together, this gives

$$\frac{1}{m}\sum_{j=1}^{m} A_{ji}(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})^2(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c}) \le c'\delta,$$

where $c' = 11\sqrt{2}$. The other terms in (3.55) can be bounded the same way. For instance, we can write

$$\left|\frac{1}{m}\sum_{j=1}^{m} A_{ji}(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})^2\right| = \left|\sum_{l\notin\mathcal{S}} x_l \sum_{r\notin\mathcal{S}} x_r \frac{1}{m}\sum_{j=1}^{m} A_{ji}A_{jl}A_{js}(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})\right|,$$

and, following the same steps as above, we obtain the bounds

$$\left|\frac{1}{m}\sum_{j=1}^{m} A_{ji}(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})(\mathbf{A}_j^\top \mathbf{x}^\star)^2\right| \le c'\delta,$$
$$\left|\frac{1}{m}\sum_{j=1}^{m} A_{ji}(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_S)(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})^2\right| \le c''\delta^2,$$
$$\left|\frac{1}{m}\sum_{j=1}^{m} A_{ji}(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})^3\right| \le c'''\delta^3,$$

for all $\mathbf{x} \in \mathcal{X}$ with probability $1 - 3(c_2+4)n^{-12.5}$. Recalling that $\delta \le \frac{1}{4}$, we have for any constant $c > 0$ satisfying $4c'\delta + 3c''\delta^2 + c'''\delta^3 \le \frac{c}{2}\delta$,

$$\left|\nabla\widetilde{F}(\mathbf{x})_i - \nabla\widetilde{F}(\mathbf{w})_i\right| \le \frac{c}{2}\delta \quad \text{for all } \mathbf{x} \in \mathcal{X}$$

with probability at least $1 - \frac{c_p}{3}n^{-11}$.

Step 3: Bound the term $|\mathbb{E}[\nabla\widetilde{F}(\mathbf{x})_i] - \mathbb{E}[\nabla\widetilde{F}(\mathbf{w})_i]|$

We use the Cauchy-Schwarz inequality to bound each of the four terms in (3.55) in expectation. The first term in (3.55) can be bounded by

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} A_{ji}(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})^2(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})\right] &\le \mathbb{E}\Big[A_{1i}^2\Big]^{\frac{1}{2}}\mathbb{E}\Big[(\mathbf{A}_{1,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})^4(\mathbf{A}_{1,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})^2\Big]^{\frac{1}{2}} \\
&= \Big(\mathbb{E}\Big[(\mathbf{A}_{1,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})^4\Big]\mathbb{E}\Big[(\mathbf{A}_{1,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})^2\Big]\Big)^{\frac{1}{2}} \\
&\le \sqrt{3\|\mathbf{x}_{\mathcal{S}^c}\|_2^2} \\
&\le \sqrt{3}\delta
\end{aligned}$$

for all $\mathbf{x} \in \mathcal{X}$, where we used that $\mathbf{A}_{1,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}} \sim \mathcal{N}(0, \|\mathbf{x}_{\mathcal{S}}\|_2^2)$ and $\mathbf{A}_{1,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c} \sim \mathcal{N}(0, \|\mathbf{x}_{\mathcal{S}^c}\|_2^2)$ are independent, and that $\|\mathbf{x}_{\mathcal{S}}\|_2 \le 1$ and $\|\mathbf{x}_{\mathcal{S}^c}\|_2 \le \delta$ by the definition of $\mathcal{X}$. Similarly, we can bound the other terms:

$$\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} A_{ji}(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})(\mathbf{A}_j^\top \mathbf{x}^\star)^2\right] \le \sqrt{3}\delta,$$
$$\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} A_{ji}(\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_{\mathcal{S}})(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})^2\right] \le \sqrt{3}\delta^2,$$
$$\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m} A_{ji}(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})^3\right] \le \sqrt{15}\delta^3.$$

This completes the proof that

$$\Big|\mathbb{E}[\nabla\widetilde{F}(\mathbf{x})_i] - \mathbb{E}\Big[\nabla\widetilde{F}(\mathbf{w})_i\Big]\Big| \leq \frac{c}{2}\delta.$$

Step 4, part (a)**: Bound the term** $|\nabla\widetilde{F}(\mathbf{w})_i - \mathbb{E}[\nabla\widetilde{F}(\mathbf{w})_i]|$ **by** $\gamma\frac{\|\mathbf{x}_S\|_1}{k}$

Finally, we need to bound the term $|\nabla\widetilde{F}(\mathbf{w})_i - \mathbb{E}[\nabla\widetilde{F}(\mathbf{w})_i]|$ in (3.54) for all $\mathbf{x} \in \mathcal{X}$ and $i \in [n]$ with probability $1 - \frac{c_p}{3}n^{-10}$, which then completes the proof of Lemma 3.6. We begin by showing the bound $\gamma\frac{\|\mathbf{x}_S\|_1}{k}$.

We decompose $\nabla\widetilde{F}(\mathbf{w})_i$ in a straightforward, albeit somewhat lengthy manner. We have

$$\begin{aligned}
\nabla\widetilde{F}(\mathbf{w})_i &= \frac{1}{m}\sum_{j=1}^{m}\Big((\mathbf{A}_j^\top\mathbf{w})^3 - (\mathbf{A}_j^\top\mathbf{w})(\mathbf{A}_j^\top\mathbf{x}^\star)^2\Big)A_{ji} \\
&= \Big(w_i^3 - w_i(x_i^\star)^2\Big)\frac{1}{m}\sum_{j=1}^{m}A_{ji}^4 + \Big(3w_i^2 - (x_i^\star)^2\Big)\frac{1}{m}\sum_{j=1}^{m}A_{ji}^3(\mathbf{A}_{j,-i}^\top\mathbf{w}_{-i}) \\
&\quad - 2w_ix_i^\star\frac{1}{m}\sum_{j=1}^{m}A_{ji}^3(\mathbf{A}_{j,-i}^\top\mathbf{x}_{-i}^\star) + 3w_i\frac{1}{m}\sum_{j=1}^{m}A_{ji}^2(\mathbf{A}_{j,-i}^\top\mathbf{w}_{-i})^2 \\
&\quad - 2x_i^\star\frac{1}{m}\sum_{j=1}^{m}A_{ji}^2(\mathbf{A}_{j,-i}^\top\mathbf{w}_{-i})(\mathbf{A}_{j,-i}^\top\mathbf{x}_{-i}^\star) - w_i\frac{1}{m}\sum_{j=1}^{m}A_{ji}^2(\mathbf{A}_{j,-i}^\top\mathbf{x}_{-i}^\star)^2 \\
&\quad + \frac{1}{m}\sum_{j=1}^{m}A_{ji}(\mathbf{A}_{j,-i}^\top\mathbf{w}_{-i})^3 - \frac{1}{m}\sum_{j=1}^{m}A_{ji}(\mathbf{A}_{j,-i}^\top\mathbf{w}_{-i})(\mathbf{A}_{j,-i}^\top\mathbf{x}_{-i}^\star)^2 \\
&=: B_1 + B_2 + B_3 + B_4 + B_5 + B_6 + B_7 + B_8,
\end{aligned}$$

and we will show that $|B_l - \mathbb{E}[B_l]|$ is small for all $l = 1, ..., 8$. By Lemmas 3.13 and 3.17, all the following statements hold with probability $1 - \frac{c_p}{3}n^{-10}$ for all $\mathbf{w} \in \mathbb{R}^n$ with $\|\mathbf{w}\|_2 \leq 1$ and $\mathbf{w}_{S^c} = \mathbf{0}$, and for all $i = 1, ..., n$. We write the bounds in terms of k instead of m using the assumption $m \geq c_s(\gamma)k^2\log^2 n$.

- For the first term, we have $\mathbb{E}[B_1] = 3(w_i^3 - w_i(x_i^\star)^2)$, and

  $$\Big|B_1 - \mathbb{E}[B_1]\Big| = \left|\Big(w_i^3 - w_i(x_i^\star)^2\Big)\left(\frac{1}{m}\sum_{j=1}^{m}A_{ji}^4 - 3\right)\right| \leq \frac{\gamma\|\mathbf{w}\|_1}{8k}$$

  by (3.79) of Lemma 3.13, where we used $|w_i^3 - w_i(x_i^\star)^2| \leq \|\mathbf{w}\|_1$.

- For the second term, we have $\mathbb{E}[B_2] = 0$, and

  $$|B_2| \leq \Big|3w_i^2 - (x_i^\star)^2\Big|\|\mathbf{w}\|_1\max_{l\neq i}\left|\frac{1}{m}\sum_{j=1}^{m}A_{ji}^3A_{jl}\right| \leq \frac{\gamma\|\mathbf{w}\|_1}{8k}$$

  by (3.79) of Lemma 3.13.

- For the third term, we have $\mathbb{E}[B_3] = 0$, and

$$|B_3| \leq |2w_i x_i^\star| \frac{\gamma}{16k} \leq \frac{\gamma \|\mathbf{w}\|_1}{8k}$$

by (3.80) of Lemma 3.13.

- For the fourth term, we have $\mathbb{E}[B_4] = 3w_i \|\mathbf{w}_{-i}\|_2^2$, and

$$\begin{aligned}\Big|B_4 - \mathbb{E}[B_4]\Big| &\leq 3|w_i| \left| \sum_{l \neq i} w_l^2 \left(\frac{1}{m} \sum_{j=1}^m A_{ji}^2 A_{jl}^2 - 1 \right) + \sum_{l \neq i} w_l \sum_{r \neq l,i} w_r \frac{1}{m} \sum_{j=1}^m A_{ji}^2 A_{jl} A_{jr} \right| \\ &\leq \frac{\gamma \|\mathbf{w}\|_1}{16k} + \frac{\gamma \|\mathbf{w}\|_1}{16k} \\ &= \frac{\gamma \|\mathbf{w}\|_1}{8k},\end{aligned}$$

where we used Lemma (3.79) of 3.13 to bound the first and (3.90) of Lemma 3.17 to bound the second term.

- For the fifth term, we have $\mathbb{E}[B_5] = -2x_i^\star \sum_{l \neq i} w_l x_l^\star$, and

$$\Big|B_5 - \mathbb{E}[B_5]\Big| \leq |2x_i^\star| \left| \sum_{l \neq i} w_l \left(\frac{1}{m} \sum_{j=1}^m A_{ji}^2 A_{jl} (\mathbf{A}_{j,-i}^\top \mathbf{x}_{-i}^\star) - x_l^\star \right) \right| \leq \frac{\gamma \|\mathbf{w}\|_1}{8k}$$

where we used Hölder's inequality and (3.80) of Lemma 3.13.

- For the sixth term, we have $\mathbb{E}[B_6] = -w_i \|\mathbf{x}_{-i}^\star\|_2^2$, and

$$\Big|B_6 - \mathbb{E}[B_6]\Big| \leq |w_i| \frac{\gamma}{8k} \leq \frac{\gamma \|\mathbf{w}\|_1}{8k}$$

by (3.81) of Lemma 3.13.

- For the seventh term, we have $\mathbb{E}[B_7] = 0$, and

$$|B_7| \leq \frac{\gamma \|\mathbf{w}\|_1}{8k}$$

by (3.89) of Lemma 3.17.

- Finally, for the eighth term, we have $\mathbb{E}[B_8] = 0$, and

$$|B_8| \leq \left| \sum_{l \neq i} w_l \frac{1}{m} \sum_{j=1}^m A_{ji} A_{jl} (\mathbf{A}_{j,-i}^\top \mathbf{x}_{-i}^\star)^2 \right| \leq \frac{\gamma \|\mathbf{w}\|_1}{8k}$$

where we used Hölder's inequality and (3.81) of Lemma 3.13.

Step 4, part (b): Bound the term $|\nabla\widetilde{F}(\mathbf{w})_i - \mathbb{E}[\nabla\widetilde{F}(\mathbf{w})_i]|$ by $\gamma\frac{\|\mathbf{x}_\mathcal{S}-\mathbf{x}^\star_\mathcal{S}\|_2}{\sqrt{k}}$
To show this bound, we parametrize $\mathbf{z} = \mathbf{w} - \mathbf{x}^\star$ and write

$$\begin{aligned}\nabla\widetilde{F}(\mathbf{w})_i &= \frac{1}{m}\sum_{j=1}^{m}\Big((\mathbf{A}_j^\top(\mathbf{z}+\mathbf{x}^\star))^3 - (\mathbf{A}_j^\top(\mathbf{z}+\mathbf{x}^\star))(\mathbf{A}_j^\top\mathbf{x}^\star)^2\Big)A_{ji}\\ &= \frac{1}{m}\sum_{j=1}^{m}\Big((\mathbf{A}_j^\top\mathbf{z})^3 + 3(\mathbf{A}_j^\top\mathbf{z})^2(\mathbf{A}_j^\top\mathbf{x}^\star) + 2(\mathbf{A}_j^\top\mathbf{z})(\mathbf{A}_j^\top\mathbf{x}^\star)^2\Big)A_{ji}.\end{aligned}$$

We can decompose this expression as before and observe that each term depends at least linearly on $\|\mathbf{z}\|_2$. Further, because $\mathbf{z}_{\mathcal{S}^c} = \mathbf{w}_{\mathcal{S}^c} = \mathbf{0}$, we have $\|\mathbf{z}\|_1 \le \sqrt{k}\|\mathbf{z}\|_2$. Hence, we obtain the bound $\gamma\frac{\|\mathbf{z}\|_1}{k} \le \gamma\frac{\|\mathbf{z}\|_2}{\sqrt{k}}$ the same way as above using the bounds in Lemmas 3.13 and 3.17. We omit the details to avoid repetition.

All in all, putting everything together we have, with probability $1 - \frac{c_p}{3}n^{-10}$,

$$\Big|\nabla\widetilde{F}(\mathbf{w})_i - \mathbb{E}[\nabla\widetilde{F}(\mathbf{w})_i]\Big| \le \gamma\min\left\{\frac{\|\mathbf{w}\|_1}{k},\ \frac{\|\mathbf{z}\|_2}{\sqrt{k}}\right\} = \gamma\min\left\{\frac{\|\mathbf{x}_\mathcal{S}\|_1}{k},\ \frac{\|\mathbf{x}_\mathcal{S}-\mathbf{x}^\star_\mathcal{S}\|_2}{\sqrt{k}}\right\},$$

for all $\mathbf{w}\in\mathbb{R}^n$ with $\|\mathbf{w}\|_2 \le 1$ and $\mathbf{w}_{\mathcal{S}^c} = \mathbf{0}$, and all $i = 1, ..., n$, which completes the proof of Lemma 3.6. Finally, the simplified bound can be obtained directly by plugging in the additional assumptions. □

Proof of Lemma 3.7. As in the proof of Lemma 3.6, we assume $\|\mathbf{x}^\star\|_2 = 1$ for notational simplicity, and denote the gradient corresponding to noiseless measurements by $\nabla\widetilde{F}$. Then,

$$\Big|\langle\nabla F(\mathbf{x}) - \nabla f(\mathbf{x}), \mathbf{x}-\mathbf{x}^\star\rangle\Big| \le \Big|\langle\nabla F(\mathbf{x}) - \nabla\widetilde{F}(\mathbf{x}), \mathbf{x}-\mathbf{x}^\star\rangle\Big| + \Big|\langle\nabla\widetilde{F}(\mathbf{x}) - \nabla f(\mathbf{x}), \mathbf{x}-\mathbf{x}^\star\rangle\Big|,$$

and we begin by bounding the first of the two terms.

Step 1: Bound the term $\Big|\langle\nabla F(\mathbf{x}) - \nabla\widetilde{F}(\mathbf{x}), \mathbf{x}-\mathbf{x}^\star\rangle\Big|$
We begin by bounding the first term $\Big|\langle\nabla F(\mathbf{x}) - \nabla\widetilde{F}(\mathbf{x}), \mathbf{x}-\mathbf{x}^\star\rangle\Big|$ by

$$c\sigma\sqrt{\frac{\log n}{m}}\Big(\delta + \min\{\|\mathbf{x}_\mathcal{S}\|_1,\ \sqrt{k}\|\mathbf{x}_\mathcal{S}-\mathbf{x}^\star_\mathcal{S}\|_2\}\Big).$$

To obtain the second bound in the minimum, we can substitute $\mathbf{z} = \mathbf{x} - \mathbf{x}^\star$ and write

$$\begin{aligned}\Big|\langle\nabla F(\mathbf{x}) - \nabla\widetilde{F}(\mathbf{x}), \mathbf{x}-\mathbf{x}^\star\rangle\Big| &= \left|\frac{1}{m}\sum_{j=1}^{m}\varepsilon_j(\mathbf{A}_j^\top\mathbf{x})(\mathbf{A}_j^\top(\mathbf{x}-\mathbf{x}^\star))\right|\\ &\le \left|\frac{1}{m}\sum_{j=1}^{m}\varepsilon_j(\mathbf{A}_j^\top\mathbf{z})^2\right| + \left|\frac{1}{m}\sum_{j=1}^{m}\varepsilon_j(\mathbf{A}_j^\top\mathbf{z})(\mathbf{A}_j^\top\mathbf{x}^\star)\right|.\end{aligned}$$

Using Lemmas 3.13 and 3.17 we can bound the first term with probability $1 - \frac{c_p}{3} n^{10}$,

$$\begin{aligned}
\left| \frac{1}{m} \sum_{j=1}^{m} \varepsilon_j (\mathbf{A}_j^\top \mathbf{z})^2 \right| &\leq \left| \frac{1}{m} \sum_{j=1}^{m} \varepsilon_j (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{z}_\mathcal{S})^2 \right| + 2 \left| \frac{1}{m} \sum_{j=1}^{m} \varepsilon_j (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{z}_\mathcal{S})(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{z}_{\mathcal{S}^c}) \right| \\
&\quad + \left| \frac{1}{m} \sum_{j=1}^{m} \varepsilon_j (\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{z}_{\mathcal{S}^c})^2 \right| \\
&\leq \frac{c}{4} \sigma \|\mathbf{z}_\mathcal{S}\|_1 \sqrt{\frac{\log n}{m}} + 2 \|\mathbf{z}_{\mathcal{S}^c}\|_1 \max_{l \notin \mathcal{S}} \left| \sum_{s \in \mathcal{S}} z_s \frac{1}{m} \sum_{j=1}^{m} \varepsilon_j A_{js} A_{jl} \right| \\
&\quad + \|\mathbf{z}_{\mathcal{S}^c}\|_1^2 \max_{l,s \notin \mathcal{S}} \left| \frac{1}{m} \sum_{j=1}^{m} \varepsilon_j A_{jl} A_{js} \right| \\
&\leq \frac{c}{4} \sigma \|\mathbf{z}_\mathcal{S}\|_1 \sqrt{\frac{\log n}{m}} + 2 \frac{c}{4} \sigma \|\mathbf{z}_\mathcal{S}\|_1 \sqrt{\frac{\log n}{m}} \delta + \frac{c}{4} \sigma \sqrt{\frac{\log n}{m}} \delta^2,
\end{aligned}$$

provided that $\frac{c}{4}$ is at least as large as the universal constants from Lemmas 3.13 and 3.17, where we used

$$\frac{1}{m} \sum_{j=1}^{m} \varepsilon_j (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{z}_\mathcal{S})^2 = \sum_{l \in \mathcal{S}} z_l^2 \frac{1}{m} \sum_{j=1}^{m} \varepsilon_j A_{jl}^2 + \sum_{\substack{l,s \in \mathcal{S} \\ l \neq s}} z_l z_s \frac{1}{m} \sum_{j=1}^{m} \varepsilon_j A_{jl} A_{js}$$

and Lemmas 3.13 and 3.17 for the second inequality, and Lemma 3.13 to bound the other terms. Similarly, we can bound

$$\left| \frac{1}{m} \sum_{j=1}^{m} \varepsilon_j (\mathbf{A}_j^\top \mathbf{z})(\mathbf{A}_j^\top \mathbf{x}^\star) \right| = \left| \sum_{l=1}^{n} z_l \frac{1}{m} \sum_{j=1}^{m} \varepsilon_j A_{jl} (\mathbf{A}_j^\top \mathbf{x}^\star) \right| \leq \frac{c}{2} \sigma \Big(\|\mathbf{z}_\mathcal{S}\|_1 + \delta \Big) \sqrt{\frac{\log n}{m}},$$

where we used Lemma 3.13. All in all, this yields with probability $1 - \frac{c_p}{3} n^{-10}$, recalling that $\|\mathbf{z}_\mathcal{S}\|_1 \leq \sqrt{k} \|\mathbf{x}_\mathcal{S} - \mathbf{x}_\mathcal{S}^\star\|_2$,

$$\left| \langle \nabla F(\mathbf{x}) - \nabla \widetilde{F}(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle \right| \leq c\sigma \sqrt{\frac{\log n}{m}} \Big(\delta + \sqrt{k} \|\mathbf{x}_\mathcal{S} - \mathbf{x}_\mathcal{S}^\star\|_2 \Big).$$

The first bound in the minimum can be obtained following exactly the same steps as above without substituting $\mathbf{z} = \mathbf{x} - \mathbf{x}^\star$.

Step 2: Bound the term $\left| \langle \nabla \widetilde{F}(\mathbf{x}) - \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle \right|$ by $\gamma(\frac{\|\mathbf{x}_\mathcal{S}\|_1}{\sqrt{k}} + \delta)$

Writing $\mathbf{w} \in \mathbb{R}^n$ for the vector $\mathbf{x}_\mathcal{S}$ padded with zeros, i.e. $w_i = x_i$ for $i \in \mathcal{S}$ and $w_i = 0$ otherwise, we can bound

$$\begin{aligned}
\left| \langle \nabla \widetilde{F}(\mathbf{x}) - \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle \right| &\leq \left| \langle \nabla \widetilde{F}(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle - \langle \nabla \widetilde{F}(\mathbf{w}), \mathbf{w} - \mathbf{x}^\star \rangle \right| \\
&\quad + \left| \langle \nabla \widetilde{F}(\mathbf{w}), \mathbf{w} - \mathbf{x}^\star \rangle - \mathbb{E}[\langle \nabla \widetilde{F}(\mathbf{w}), \mathbf{w} - \mathbf{x}^\star \rangle] \right| \\
&\quad + \left| \mathbb{E}[\langle \nabla \widetilde{F}(\mathbf{w}), \mathbf{w} - \mathbf{x}^\star \rangle] - \mathbb{E}[\langle \nabla \widetilde{F}(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle] \right|, \quad (3.56)
\end{aligned}$$

where we used $\nabla f(\mathbf{x}) = \mathbb{E}[\nabla F(\mathbf{x})] = \mathbb{E}[\nabla \widetilde{F}(\mathbf{x})]$. To bound these three terms, we write

$$\begin{aligned}\langle \nabla \widetilde{F}(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle &= \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_j^\top \mathbf{x})^4 - \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_j^\top \mathbf{x})^3 (\mathbf{A}_j^\top \mathbf{x}^\star) \\ &\quad - \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_j^\top \mathbf{x})^2 (\mathbf{A}_j^\top \mathbf{x}^\star)^2 + \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_j^\top \mathbf{x})(\mathbf{A}_j^\top \mathbf{x}^\star)^3 \\ &=: g_{1,\mathbf{x}}(\mathbf{A}) - g_{2,\mathbf{x}}(\mathbf{A}) - g_{3,\mathbf{x}}(\mathbf{A}) + g_{4,\mathbf{x}}(\mathbf{A}),\end{aligned}$$

and note that we have analogous expressions for $\langle \widetilde{F}(\mathbf{w}), \mathbf{w} - \mathbf{x}^\star \rangle$, $\mathbb{E}[\langle \widetilde{F}(\mathbf{w}), \mathbf{w} - \mathbf{x}^\star \rangle]$ and $\mathbb{E}[\langle \widetilde{F}(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle]$. We will show that, with probability $1 - \frac{c_p}{3} n^{-10}$, the four terms $g_{i,\mathbf{x}}(\mathbf{A})$ deviate by at most $\frac{\gamma}{4}(\frac{\|\mathbf{x}\|_1}{\sqrt{k}} + \delta)$ from their means by bounding each difference as in (3.56).

Step 2, part (a): Split $g_{i,\mathbf{x}}(\mathbf{A})$ into a term depending on $\mathbf{x}_\mathcal{S}$ and a residual
We first show that the first and last terms in (3.56) are both bounded by $\frac{\gamma}{2}\delta$. To this end, we split each of the four terms $g_{i,\mathbf{x}}(\mathbf{A})$ into a term that only depends on $\mathbf{x}_\mathcal{S}$ (which corresponds to $\langle \nabla \widetilde{F}(\mathbf{w}), \mathbf{w} - \mathbf{x}^\star \rangle$) and a residual term that also depends on $\mathbf{x}_{\mathcal{S}^c}$ (which corresponds to the difference $\langle \nabla \widetilde{F}(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle - \langle \nabla \widetilde{F}(\mathbf{w}), \mathbf{w} - \mathbf{x}^\star \rangle$). We only present the computation for $g_{1,\mathbf{x}}(\mathbf{A})$, since the other three terms can be bounded following the same steps. We have

$$\begin{aligned}g_{1,\mathbf{x}}(\mathbf{A}) &= \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_\mathcal{S} + \mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})^4 \\ &= \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_\mathcal{S})^4 + \frac{4}{m}\sum_{j=1}^m (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_\mathcal{S})^3 (\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c}) + \frac{6}{m}\sum_{j=1}^m (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_\mathcal{S})^2 (\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})^2 \\ &\quad + \frac{4}{m}\sum_{j=1}^m (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_\mathcal{S})(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})^3 + \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c})^4.\end{aligned}$$

The first term only depends on $\mathbf{x}_\mathcal{S} = \mathbf{w}_\mathcal{S}$, and we write $h_{1,\mathbf{w}}(\mathbf{A}) = \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{w}_\mathcal{S})^4$ (and analogously $h_{i,\mathbf{w}}(\mathbf{A}) = \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{w}_\mathcal{S})^{5-i} (\mathbf{A}_j^\top \mathbf{x}^\star)^{i-1}$ for $i = 2, 3, 4$). The other terms can be bounded as follows:

$$\begin{aligned}\frac{4}{m}\sum_{j=1}^m (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_\mathcal{S})^3 (\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{x}_{\mathcal{S}^c}) &= 4\sum_{l \notin \mathcal{S}} x_l \frac{1}{m}\sum_{j=1}^m A_{jl} (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_\mathcal{S})^3 \\ &\leq 4\|\mathbf{x}_{\mathcal{S}^c}\|_1 \max_{l \notin \mathcal{S}} \left| \frac{1}{m}\sum_{j=1}^m A_{jl} (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{x}_\mathcal{S})^3 \right| \\ &\leq \frac{\gamma}{11}\delta,\end{aligned}$$

provided that $m \geq (\frac{44\tilde{c}}{\gamma})^2 k \log^2 n$, where we used that $\|\mathbf{x}_{\mathcal{S}}\|_1 \leq \sqrt{k}$ and $\tilde{c}$ is the universal constant from Lemma 3.17. For the first inequality we used Hölder's inequality, and the second inequality holds by (3.89) of Lemma 3.17 with probability $1 - \frac{c_p}{9} n^{-10}$. Similarly, we can bound

$$\begin{aligned}\frac{6}{m}\sum_{j=1}^{m}(\mathbf{A}_{j,\mathcal{S}}^{\top}\mathbf{x}_{\mathcal{S}})^2(\mathbf{A}_{j,\mathcal{S}^c}^{\top}\mathbf{x}_{\mathcal{S}^c})^2 &= 6\sum_{l\notin\mathcal{S}} x_l \sum_{r\notin\mathcal{S}} x_r \frac{1}{m}\sum_{j=1}^{m} A_{jl}A_{jr}(\mathbf{A}_{j,\mathcal{S}}^{\top}\mathbf{x}_{\mathcal{S}})^2 \\ &\leq 6\|\mathbf{x}_{\mathcal{S}^c}\|_1^2 \max_{l,r\notin\mathcal{S}} \sqrt{\frac{1}{m}\sum_{j=1}^{m} A_{jl}^2 A_{jr}^2}\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\mathbf{A}_{j,\mathcal{S}}^{\top}\mathbf{x}_{\mathcal{S}})^4} \\ &\leq \frac{c}{11}\delta^2,\end{aligned}$$

provided that $c > 0$ is large enough, where we used Hölder's inequality in the second line and the last inequality holds by Lemmas 3.11 and 3.13 with probability $1 - \frac{2c_p}{9} n^{-10}$. The same computation yields

$$\frac{4}{m}\sum_{j=1}^{m}(\mathbf{A}_{j,\mathcal{S}}^{\top}\mathbf{x}_{\mathcal{S}})(\mathbf{A}_{j,\mathcal{S}^c}^{\top}\mathbf{x}_{\mathcal{S}^c})^3 \leq \frac{c}{11}\delta^3,$$

$$\frac{1}{m}\sum_{j=1}^{m}(\mathbf{A}_{j,\mathcal{S}^c}^{\top}\mathbf{x}_{\mathcal{S}^c})^4 \leq \frac{c}{11}\delta^4.$$

Putting everything together, we can bound the residual term by $\frac{c}{8}\delta$ since, by assumption, $\delta \leq \frac{1}{4}\min\{\frac{\gamma}{c}, 1\}$. Bounding $g_{2,\mathbf{x}}(\mathbf{A})$, $g_{3,\mathbf{x}}(\mathbf{A})$ and $g_{4,\mathbf{x}}(\mathbf{A})$ the same way, we have, with probability $1 - \frac{c_p}{3} n^{-10}$,

$$\left|\langle \nabla \widetilde{F}(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star\rangle - \langle \nabla \widetilde{F}(\mathbf{w}), \mathbf{w} - \mathbf{x}^\star\rangle\right| \leq \frac{\gamma}{2}\delta.$$

For the difference in expectation, we bound, using $\|\mathbf{x}\|_2 \leq 1$ and $\|\mathbf{x}_{\mathcal{S}^c}\|_2 \leq \|\mathbf{x}_{\mathcal{S}^c}\|_1 \leq \delta$,

$$\begin{aligned}\mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m}(\mathbf{A}_{j,\mathcal{S}}^{\top}\mathbf{x}_{\mathcal{S}})^3(\mathbf{A}_{j,\mathcal{S}^c}^{\top}\mathbf{x}_{\mathcal{S}^c})\right] &= \mathbb{E}\Big[(\mathbf{A}_{1,\mathcal{S}}^{\top}\mathbf{x}_{\mathcal{S}})^3\Big]\mathbb{E}\Big[\mathbf{A}_{j,\mathcal{S}^c}^{\top}\mathbf{x}_{\mathcal{S}^c}\Big] = 0, \\ \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m}(\mathbf{A}_{j,\mathcal{S}}^{\top}\mathbf{x}_{\mathcal{S}})^2(\mathbf{A}_{j,\mathcal{S}^c}^{\top}\mathbf{x}_{\mathcal{S}^c})^2\right] &= \mathbb{E}\Big[(\mathbf{A}_{1,\mathcal{S}}^{\top}\mathbf{x}_{\mathcal{S}})^2\Big]\mathbb{E}\Big[(\mathbf{A}_{j,\mathcal{S}^c}^{\top}\mathbf{x}_{\mathcal{S}^c})^2\Big] \leq \delta^2, \\ \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m}(\mathbf{A}_{j,\mathcal{S}}^{\top}\mathbf{x}_{\mathcal{S}})(\mathbf{A}_{j,\mathcal{S}^c}^{\top}\mathbf{x}_{\mathcal{S}^c})^3\right] &= \mathbb{E}\Big[\mathbf{A}_{1,\mathcal{S}}^{\top}\mathbf{x}_{\mathcal{S}}\Big]\mathbb{E}\Big[(\mathbf{A}_{j,\mathcal{S}^c}^{\top}\mathbf{x}_{\mathcal{S}^c})^3\Big] = 0, \\ \mathbb{E}\left[\frac{1}{m}\sum_{j=1}^{m}(\mathbf{A}_{j,\mathcal{S}^c}^{\top}\mathbf{x}_{\mathcal{S}^c})^4\right] &= \mathbb{E}\Big[(\mathbf{A}_{j,\mathcal{S}^c}^{\top}\mathbf{x}_{\mathcal{S}^c})^4\Big] \leq 3\delta^4.\end{aligned}$$

Repeating this for $g_{2,\mathbf{x}}(\mathbf{A})$, $g_{3,\mathbf{x}}(\mathbf{A})$ and $g_{4,\mathbf{x}}(\mathbf{A})$ shows

$$\left|\mathbb{E}\Big[\langle \nabla F(\mathbf{x}), \mathbf{x} - \mathbf{x}^*\rangle\Big] - \mathbb{E}\Big[\langle \nabla F(\mathbf{w}), \mathbf{w} - \mathbf{x}^*\rangle\Big]\right| \leq \frac{\gamma}{2}\delta.$$

Step 2, part (b)**: Bound the term** $|\langle \nabla F(\mathbf{w}, \mathbf{w} - \mathbf{x}^\star\rangle - \mathbb{E}[\langle \nabla F(\mathbf{w}), \mathbf{w} - \mathbf{x}^\star\rangle]|$
Finally, we need to bound the second term in (3.56), i.e. we need to show that the terms $h_{i,\mathbf{w}}(\mathbf{A})$ concentrate around their respective expectations. Let N_ϵ be a smallest ϵ-net of the set $\mathcal{X} = \{\mathbf{w} \in \mathbb{R}^n : \mathbf{w}_{\mathcal{S}^c} = \mathbf{0}, \|\mathbf{w}\|_2 = 1\}$ (which corresponds to the unit sphere in $\mathbb{R}^k$), where $\epsilon = c_1 \gamma n^{-3}$ for some constant $c_1 > 0$. We will first show that we can bound the difference $|h_{i,\mathbf{w}}(\mathbf{A}) - \mathbb{E}[h_{i,\mathbf{w}}(\mathbf{A})]| \leq \frac{\gamma}{8}\frac{\|\mathbf{w}\|_1}{\sqrt{k}}$ for every $\mathbf{w} \in N_\epsilon$ via concentration of Lipschitz functions for Gaussian random variables. Then, we will extend this bound to every $\mathbf{w} \in \mathcal{X}$.

The idea is to apply concentration of Lipschitz functions of Gaussian random variables to show that the terms $h_{i,\mathbf{w}}(\mathbf{A})$ are close to their respective expectations. However, as $h_{i,\mathbf{w}}$ is not globally Lipschitz continuous, we cannot directly apply Theorem 3.9. Similar to the proof of Theorem 3 in [43], we will first restrict $h_{i,\mathbf{w}}$ to the set $\mathcal{A}$ defined as the intersection of the sets defined in Lemma 3.12 and Lemma 3.14, where we can bound the Lipschitz constant of $h_{i,\mathbf{w}}$. Then, we extend this restricted function to a function $\tilde{h}_{i,\mathbf{w}}$ on the entire space in a way such that $\tilde{h}_{i,\mathbf{w}}$ is globally Lipschitz continuous, and apply Theorem 3.9 to $\tilde{h}_{i,\mathbf{w}}$. This also provides a high probability bound for $h_{i,\mathbf{w}}(\mathbf{A})$, since, by construction, $\tilde{h}_{i,\mathbf{w}}(\mathbf{A}) = h_{i,\mathbf{w}}(\mathbf{A})$ with high probability.

Step 2, part $(b.1)$**: Bound the Lipschitz constant of** $h_{i,\mathbf{w}}$ **restricted to** $\mathcal{A}$
Let the subset $\mathcal{A} \subset \mathbb{R}^{m\times n}$ be defined as the intersection of the sets defined in Lemma 3.12 with $t = 5\sqrt{\log n}$ and Lemma 3.14. More precisely, we require the projection of $\mathcal{A}$ onto $\mathbb{R}^{m\times \mathcal{S}}$ to be as in Lemma 3.12. By the aforementioned lemmas, we have $\mathbb{P}[\mathcal{A}^c] \leq c_2 n^{-12}$ for a universal constant $c_2 > 0$. Since $\mathcal{A}$ is convex, it follows from the mean-value theorem that the Lipschitz constant of $h_{i,\mathbf{w}}$ restricted to $\mathcal{A}$ is bounded by the norm of its gradient $\|\nabla h_{i,\mathbf{w}}\|_2$. For any $\{\mathbf{a}_j\}_{j=1}^m \in \mathcal{A}$, we can compute the following.

- $\frac{\partial}{\partial a_{jl}} h_{1,\mathbf{w}}(\mathbf{a}) = \frac{4}{m} w_l (\mathbf{a}_j^\top \mathbf{w})^3$. Hence, we can bound the Lipschitz constant by the squareroot of

$$\begin{aligned}\|\nabla h_{1,\mathbf{w}}(\mathbf{a})\|_2^2 &= \frac{16}{m^2}\sum_{l=1}^n w_l^2 \sum_{j=1}^m (\mathbf{a}_j^\top \mathbf{w})^6 \\ &\leq \frac{16}{m^2}\Big((15m)^{\frac{1}{6}} + \sqrt{8\log(2k)}\|\mathbf{w}\|_1 + 5\sqrt{\log n}\Big)^6 \\ &\leq c_3 \frac{\|\mathbf{w}\|_1^2 \log k}{m},\end{aligned}$$

where we used $\|\mathbf{w}\|_2 = 1$, $\|\mathbf{w}\|_1 \leq \sqrt{k}$, $m \geq c_s(\gamma)\max\{k^2\log^2 n, \log^5 n\}$ and Lemma 3.12.

- $\frac{\partial}{\partial a_{jl}} h_{2,\mathbf{w}}(\mathbf{a}) = \frac{1}{m}(x_l^\star(\mathbf{a}_j^\top \mathbf{w})^3 + 3w_l(\mathbf{a}_j^\top \mathbf{w})^2(\mathbf{a}_j^\top \mathbf{x}^\star))$. Hence, we can bound the Lipschitz constant by the squareroot of

$$\begin{aligned}
\|\nabla h_{2,\mathbf{w}}(\mathbf{a})\|_2^2 &= \frac{2}{m^2}\sum_{l=1}^{n}\sum_{j=1}^{m}(x_l^\star)^2(\mathbf{a}_j^\top \mathbf{w})^6 + 9w_l^2(\mathbf{a}_j^\top \mathbf{w})^4(\mathbf{a}_j^\top \mathbf{x}^\star)^2 \\
&\leq \frac{2}{m^2}\Big((15m)^{\frac{1}{6}} + 2\sqrt{2\log(2k)}\|\mathbf{w}\|_1 + 5\sqrt{\log n}\Big)^6 \\
&\quad + \frac{18}{m}\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\mathbf{a}_j^\top \mathbf{w})^8}\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\mathbf{a}_j^\top \mathbf{x}^\star)^4} \\
&\leq c_4\frac{\|\mathbf{w}\|_1^2\log k}{m},
\end{aligned}$$

where we use Lemma 3.12 to bound the sum $\frac{1}{m}\sum_{j=1}^{m}(\mathbf{a}_j^\top \mathbf{w})^8$ and Lemma 3.14 to bound $\frac{1}{m}\sum_{j=1}^{m}(\mathbf{a}_j^\top \mathbf{x}^*)^4$.

- $\frac{\partial}{\partial a_{ji}} h_{3,\mathbf{w}}(\mathbf{a}) = \frac{2}{m}(x_l^\star(\mathbf{a}_j^\top \mathbf{w})^2(\mathbf{a}_j^\top \mathbf{x}^\star) + w_l(\mathbf{a}_j^\top \mathbf{w})(\mathbf{a}_j^\top \mathbf{x}^\star)^2)$. Hence, we can bound the Lipschitz constant by the squareroot of

$$\begin{aligned}
\|\nabla h_{3,\mathbf{w}}(\mathbf{a})\|_2^2 &= \frac{4}{m^2}\sum_{l=1}^{n}\sum_{j=1}^{m}2(x_l^\star)^2(\mathbf{a}_j^\top \mathbf{w})^4(\mathbf{a}_j^\top \mathbf{x}^\star)^2 + 2w_l^2(\mathbf{a}_j^\top \mathbf{w})^2(\mathbf{a}_j^\top \mathbf{x}^\star)^4 \\
&\leq \frac{8}{m}\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\mathbf{a}_j^\top \mathbf{w})^8}\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\mathbf{a}_j^\top \mathbf{x}^\star)^4} \\
&\quad + \frac{8}{m}\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\mathbf{a}_j^\top \mathbf{w})^4}\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\mathbf{a}_j^\top \mathbf{x}^\star)^8} \\
&\leq c_5\frac{\|\mathbf{w}\|_1^2\log k}{m},
\end{aligned}$$

again by Lemmas 3.12 and 3.14.

- $\frac{\partial}{\partial a_{ji}} h_{4,\mathbf{w}}(\mathbf{a}) = \frac{1}{m}(3x_l^\star(\mathbf{a}_j^\top \mathbf{w})(\mathbf{a}_j^\top \mathbf{x}^\star)^2 + w_l(\mathbf{a}_j^\top \mathbf{x}^\star)^3)$. Hence, we can bound the Lipschitz constant by the squareroot of

$$\begin{aligned}
\|\nabla h_{4,\mathbf{w}}(\mathbf{a})\|_2^2 &= \frac{2}{m^2}\sum_{l=1}^{n}\sum_{j=1}^{m}9(x_l^\star)^2(\mathbf{a}_j^\top \mathbf{w})^2(\mathbf{a}_j^\top \mathbf{x}^\star)^4 + w_l^2(\mathbf{a}_j^\top \mathbf{x}^\star)^6 \\
&\leq \frac{18}{m}\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\mathbf{a}_j^\top \mathbf{w})^4}\sqrt{\frac{1}{m}\sum_{j=1}^{m}(\mathbf{a}_j^\top \mathbf{x}^\star)^8} + \frac{2}{m}\frac{1}{m}\sum_{j=1}^{m}(\mathbf{a}_j^\top \mathbf{x}^\star)^6 \\
&\leq c_6\frac{\|\mathbf{w}\|_1^2\log k}{m},
\end{aligned}$$

again by Lemmas 3.12 and 3.14.

Step 2, part $(b.2)$**: Construct a globally Lipschitz continuous extension**

We only present the following steps for the term $h_{1,\mathbf{w}}$, as the proofs for the other three terms follow the exact same steps. Consider the following Lipschitz extension of $h_{1,\mathbf{w}}$:

$$\tilde{h}_{1,\mathbf{w}}(\mathbf{a}) = \inf_{\mathbf{a}' \in \mathcal{A}} \Big(h_{1,\mathbf{w}}(\mathbf{a}') + \mathrm{Lip}(h_{1,\mathbf{w}}) \|\mathbf{a} - \mathbf{a}'\|_2 \Big),$$

where we write $\mathrm{Lip}(h_{1,\mathbf{w}}) = \sqrt{c_3 \frac{\|\mathbf{w}\|_1^2 \log k}{m}}$. By definition, we have $\tilde{h} = h$ on $\mathcal{A}$, and it follows from an application of the triangle inequality that $\tilde{h}$ is globally Lipschitz continuous with Lipschitz constant $\mathrm{Lip}(h)$ (see e.g. Theorem 7.2 of [83]).

We will show that $\tilde{h}_{1,\mathbf{w}}$ concentrates around its mean, which can potentially differ from the mean of $h_{1,\mathbf{w}}$. Since $h_{1,\mathbf{w}}$ and $\tilde{h}_{1,\mathbf{w}}$ differ only on $\mathcal{A}^c$ (which has probability less than $c_2 n^{-12}$), we can bound, using the Cauchy-Schwarz inequality and denoting by $\mathbf{1}(\cdot)$ the indicator function,

$$\mathbb{E}\Big[|h_{1,\mathbf{w}}(\mathbf{A})| \mathbf{1}_{\mathcal{A}^c}(\mathbf{A})\Big] \leq \sqrt{\mathbb{E}\Big[h_{1,\mathbf{w}}(\mathbf{A})^2\Big]} \sqrt{\mathbb{E}\Big[\mathbf{1}_{\mathcal{A}^c}(\mathbf{A})\Big]} \leq \frac{c'}{n^6},$$

where we used $\mathbb{E}[h_{1,\mathbf{w}}(\mathbf{A})^2] \leq 3 + 105/m$. Similarly, since $\tilde{h}_{1,\mathbf{w}}(\mathbf{a}) \leq \mathrm{Lip}(h_{1,\mathbf{w}}) \|\mathbf{a}\|_2$,

$$\mathbb{E}\Big[|\tilde{h}_{1,\mathbf{w}}(\mathbf{A})| \mathbf{1}_{\mathcal{A}^c}(\mathbf{A})\Big] \leq \mathrm{Lip}(h_{1,\mathbf{w}}) \sqrt{\mathbb{E}\Big[\|\mathbf{A}\|_2^2\Big]} \sqrt{\mathbb{E}\Big[\mathbf{1}_{\mathcal{A}^c}(\mathbf{A})\Big]} \leq \frac{c''}{n^5}.$$

All in all, this shows that

$$\Big| \mathbb{E}[h_{1,\mathbf{w}}(\mathbf{A})] - \mathbb{E}[\tilde{h}_{1,\mathbf{w}}(\mathbf{A})] \Big| \leq \frac{c_7}{n^5}$$

for a constant $c_7 > 0$. Finally, using the triangle inequality and Theorem 3.9, we have

$$\mathbb{P}\left[\Big| \tilde{h}_{1,\mathbf{w}}(\mathbf{A}) - \mathbb{E}[h_{1,\mathbf{w}}(\mathbf{A})] \Big| > \frac{\gamma}{8} \frac{\|\mathbf{w}\|_1}{\sqrt{k}} \right] \leq 2 \exp\left(- \frac{(\frac{\gamma}{8} \frac{\|\mathbf{w}\|_1}{\sqrt{k}} - \frac{c_7}{n^5})^2}{2 c_3 \frac{\|\mathbf{w}\|_1^2 \log k}{m}} \right) \leq 2 \exp(-c_8 k \log n)$$

for a constant $c_8 \leq \frac{\gamma^2 c_s(\gamma)}{128 c_3} - \frac{\gamma c_s(\gamma) c_7}{8 c_3 n^5}$.

Step 2, part $(b.3)$**: Union bound over** $\mathbf{w} \in N_\epsilon$

A union bound over all $\mathbf{w} \in N_\epsilon$, which has cardinality bounded by $(3/\epsilon)^k$, yields

$$\mathbb{P}\left[\Big| \tilde{h}_{1,\mathbf{w}}(\mathbf{A}) - \mathbb{E}[h_{1,\mathbf{w}}(\mathbf{A})] \Big| > \frac{\gamma}{8} \frac{\|\mathbf{w}\|_1}{\sqrt{k}} \text{ for some } \mathbf{w} \in N_\epsilon \right] \leq 2 \exp\left(-c_8 k \log n + k \log \frac{3}{\epsilon} \right).$$

Since $h_{1,\mathbf{w}}(\mathbf{a}) = \tilde{h}_{1,\mathbf{w}}(\mathbf{a})$ for all $\mathbf{w}$ provided that $\mathbf{a} \in \mathcal{A}$, this implies

$$\mathbb{P}\left[\Big| h_{1,\mathbf{w}}(\mathbf{A}) - \mathbb{E}[h_{1,\mathbf{w}}(\mathbf{A})] \Big| > \frac{\gamma}{8} \frac{\|\mathbf{w}\|_1}{\sqrt{k}} \text{ for some } \mathbf{w} \in N_\epsilon \right] \leq 2 \exp(-c_9 k \log n) + c_2 n^{-12},$$

for a constant $c_9 \leq c_8 - 3 - \frac{1}{\log n} \log \frac{3}{c_1 \gamma}$, as we have $\epsilon = c_1 \gamma n^{-3}$.

Step 2, part $(b.4)$**: From** ϵ**-net to the full sphere**

Next, we show that $|h_{1,\mathbf{w}}(\mathbf{A}) - \mathbb{E}[h_{1,\mathbf{w}}(\mathbf{A})]| \leq \frac{\gamma}{4} \frac{\|\mathbf{w}\|_1}{\sqrt{k}}$ for any $\mathbf{w} \in \mathcal{X}$. The case $\|\mathbf{w}\|_2 < 1$ follows by considering the vector $\mathbf{w}/\|\mathbf{w}\|_2$ and rescaling. For any $\mathbf{w} \in \mathcal{X}$, let $\mathbf{w}' \in N_\epsilon$ with $\|\mathbf{w} - \mathbf{w}'\|_2 \leq \epsilon$. Then,

$$\begin{aligned}\Big|h_{1,\mathbf{w}}(\mathbf{A}) - \mathbb{E}[h_{1,\mathbf{w}}(\mathbf{A})]\Big| \leq \Big|h_{1,\mathbf{w}}(\mathbf{A}) - h_{1,\mathbf{w}'}(\mathbf{A})\Big| + \Big|h_{1,\mathbf{w}'}(\mathbf{A}) - \mathbb{E}[h_{1,\mathbf{w}'}(\mathbf{A})]\Big| \\ + \Big|\mathbb{E}[h_{1,\mathbf{w}'}(\mathbf{A})] - \mathbb{E}[h_{1,\mathbf{w}}(\mathbf{A})]\Big|.\end{aligned}$$

The first and third term can be bounded using $a^4 - b^4 = (a^2 + b^2)(a + b)(a - b)$:

$$\begin{aligned}\Big|h_{1,\mathbf{w}}(\mathbf{A}) - h_{1,\mathbf{w}'}(\mathbf{A})\Big| &= \left|\frac{1}{m}\sum_{j=1}^{m}\Big((\mathbf{A}_j^\top \mathbf{w})^2 + (\mathbf{A}_j^\top \mathbf{w}')^2\Big)\Big(\mathbf{A}_j^\top(\mathbf{w} + \mathbf{w}')\Big)\Big(\mathbf{A}_j^\top(\mathbf{w} - \mathbf{w}')\Big)\right| \\ &\leq \max_j 2\|\mathbf{A}_j\|_2^2 \cdot 2\|\mathbf{A}_j\|_2\|\mathbf{A}_j\|_2\|\mathbf{w} - \mathbf{w}'\|_2 \\ &\leq 4\Big(\sqrt{k} + 5\sqrt{\log n}\Big)^4 \|\mathbf{w} - \mathbf{w}'\|_2 \\ &\leq \frac{\gamma}{16}\frac{\|\mathbf{w}\|_1}{\sqrt{k}},\end{aligned}$$

with probability $1 - mn^{-12.5}$, provided c_1 is sufficiently large, where we used the fact that the norm $\|\cdot\|_2$ is 1-Lipschitz continuous and applied Theorem 3.9 to bound the term $\|\mathbf{A}_j\|_2$, and used $\|\mathbf{w} - \mathbf{w}'\|_2 \leq \epsilon$ for the last inequality. For the expectation, the same argument yields

$$\begin{aligned}\Big|\mathbb{E}[h_{1,\mathbf{w}}(\mathbf{A})] - \mathbb{E}[h_{1,\mathbf{w}'}(\mathbf{A})]\Big| &\leq \mathbb{E}\Big[2\|\mathbf{A}_j\|_2^2 \cdot 2\|\mathbf{A}_j\|_2\|\mathbf{A}_j\|_2\|\mathbf{w} - \mathbf{w}'\|_2\Big] \\ &= 4\Big(3k + k(k-1)\Big)\|\mathbf{w} - \mathbf{w}'\|_2 \\ &\leq \frac{\gamma}{16}\frac{\|\mathbf{w}\|_1}{\sqrt{k}}.\end{aligned}$$

This completes the proof that

$$\mathbb{P}\left[\Big|h_{1,\mathbf{w}}(\mathbf{A}) - \mathbb{E}[h_{1,\mathbf{w}}(\mathbf{A})]\Big| < \frac{\gamma}{4}\frac{\|\mathbf{w}\|_1}{\sqrt{k}} \text{ for all } \mathbf{w} \in \mathcal{X}\right] \geq 1 - \frac{c_p}{24}n^{-10},$$

provided that $c_p \geq 24 + 24c_2 n^{-2}$, and if $m \leq n^{2.5}$. The other case $m > n^{2.5}$ is simpler and can be shown following the same steps, writing the probabilities in terms of m instead of n. We omit the details to avoid repetition. Repeating the same steps for the terms $h_{2,\mathbf{w}}$, $h_{3,\mathbf{w}}$ and $h_{4,\mathbf{w}}$ shows that

$$\mathbb{P}\left[\Big|\langle \nabla \widetilde{F}(\mathbf{w}), \mathbf{w} - \mathbf{x}^\star\rangle - \mathbb{E}\Big[\langle \nabla \widetilde{F}(\mathbf{w}), \mathbf{w} - \mathbf{x}^\star\rangle\Big]\Big| \leq \gamma\frac{\|\mathbf{w}\|_1}{\sqrt{k}} \text{ for all } \mathbf{w} \in \mathcal{X}\right] \geq \ 1 - \frac{c_p}{6}n^{-10}.$$

Finally, the bound also holds for any vector $\mathbf{w} \in \mathbb{R}^n$ with $\mathbf{w}_{\mathcal{S}^c} = \mathbf{0}$ and $\|\mathbf{w}\|_2 < 1$ by considering $\mathbf{w}/\|\mathbf{w}\|_2$, and noting that each of the four terms which make up $\langle \nabla \widetilde{F}(\mathbf{w}), \mathbf{w} - \mathbf{x}^\star \rangle$ scale at least linearly in $\|\mathbf{w}\|_2$.

Step 3: Bound the term $\left|\langle \nabla \widetilde{F}(\mathbf{x}) - \nabla f(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle\right|$ by $\gamma(\|\mathbf{x}_{\mathcal{S}} - \mathbf{x}^\star_{\mathcal{S}}\|_2^2 + \delta^2)$

Substituting $\mathbf{z} = \mathbf{x} - \mathbf{x}^\star$, we have

$$\langle \nabla \widetilde{F}(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle = \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_j^\top \mathbf{z})^4 + \frac{3}{m}\sum_{j=1}^m (\mathbf{A}_j^\top \mathbf{z})^3 (\mathbf{A}_j^\top \mathbf{x}^\star) + \frac{2}{m}\sum_{j=1}^m (\mathbf{A}_j^\top \mathbf{z})^2 (\mathbf{A}_j^\top \mathbf{x}^\star)^2. \tag{3.57}$$

The proof follows the same steps as the proof of the bound $\gamma(\frac{\|\mathbf{x}_{\mathcal{S}}\|_1}{\sqrt{k}} + \delta)$, writing $\mathbf{z}_{\mathcal{S}} = \|\mathbf{z}_{\mathcal{S}}\|_2 \frac{\mathbf{z}_{\mathcal{S}}}{\|\mathbf{z}_{\mathcal{S}}\|_2}$ and using the fact that $\|\mathbf{z}_{\mathcal{S}^c}\|_1 = \|\mathbf{x}_{\mathcal{S}^c}\|_1 \le \delta$ and $\frac{\|\mathbf{z}_{\mathcal{S}}\|_1}{\sqrt{k}} \le \|\mathbf{z}_{\mathcal{S}}\|_2$. We then obtain the desired bound since each of the three terms depends at least quadratically on $\mathbf{z}$.

We illustrate this argument for the last term in (3.57). The other two terms can be controlled similarly (they are easier to control because of the higher order dependence on $\mathbf{z}$). We have

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^m (\mathbf{A}_j^\top \mathbf{z})^2 (\mathbf{A}_j^\top \mathbf{x}^\star)^2 &= \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{z}_{\mathcal{S}})^2 (\mathbf{A}_j^\top \mathbf{x}^\star)^2 + \frac{2}{m}\sum_{j=1}^m (\mathbf{A}_{j,\mathcal{S}}^\top \mathbf{z}_{\mathcal{S}})(\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{z}_{\mathcal{S}^c})(\mathbf{A}_j^\top \mathbf{x}^\star)^2 \\ &\quad + \frac{1}{m}\sum_{j=1}^m (\mathbf{A}_{j,\mathcal{S}^c}^\top \mathbf{z}_{\mathcal{S}^c})^2 (\mathbf{A}_j^\top \mathbf{x}^\star)^2 \\ &=: B_1 + B_2 + B_3.\end{aligned}$$

The same computation as in Step 2, part (b) shows that, with probability $1 - \frac{c_p}{27} n^{-10}$,

$$\left|B_1 - \mathbb{E}[B_1]\right| \le \frac{\gamma}{12}\|\mathbf{z}_{\mathcal{S}}\|_2^2.$$

As in Step 2, part (a), we can bound with probability $1 - \frac{c_p}{27} n^{-10}$,

$$\begin{aligned}\left|B_2 - \mathbb{E}[B_2]\right| &= 2\left|\sum_{i \notin \mathcal{S}} z_i \sum_{l \in \mathcal{S}} z_l \frac{1}{m}\sum_{j=1}^m A_{ji} A_{jl} (\mathbf{A}_j^\top \mathbf{x}^\star)^2\right| \\ &\le 2\|\mathbf{z}_{\mathcal{S}^c}\|_1 \|\mathbf{z}_{\mathcal{S}}\|_1 \max_{i \notin \mathcal{S}, l \in \mathcal{S}} \left|\frac{1}{m}\sum_{j=1}^m A_{ji} A_{jl} (\mathbf{A}_j^\top \mathbf{x}^\star)^2\right| \\ &\le \frac{\gamma}{6}\|\mathbf{z}_{\mathcal{S}}\|_2 \delta,\end{aligned}$$

where we used Hölder's inequality, (3.81) of Lemma 3.13 together with the assumption $m \ge c_s(\gamma) k^2 \log^2 n$, and the fact that $\|\mathbf{z}_{\mathcal{S}}\|_1 \le \sqrt{k}\|\mathbf{z}_{\mathcal{S}}\|_2$. The same argument gives

$$|B_3 - \mathbb{E}[B_3]| \le \frac{\gamma}{12}\delta^2$$

with probability $1 - \frac{c_p}{27} n^{-10}$. Combining these bounds, we have, with probability $1 - \frac{c_p}{9} n^{-10}$,

$$\left| \frac{2}{m} \sum_{j=1}^{m} (\mathbf{A}_j^\top \mathbf{z})^2 (\mathbf{A}_j^\top \mathbf{x}^\star)^2 - \mathbb{E}\left[\frac{2}{m} \sum_{j=1}^{m} (\mathbf{A}_j^\top \mathbf{z})^2 (\mathbf{A}_j^\top \mathbf{x}^\star)^2 \right] \right| \leq \frac{\gamma}{3} \Big(\|\mathbf{z}_{\mathcal{S}}\|_2^2 + \delta^2 \Big),$$

where we used the inequality $2ab \leq a^2 + b^2$. Repeating these steps for the other two terms in (3.57) completes the proof that, with probability $1 - \frac{c_p}{3} n^{-10}$, we have

$$\Big| \langle \nabla \widetilde{F}(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle - \mathbb{E}[\langle \nabla \widetilde{F}(\mathbf{x}), \mathbf{x} - \mathbf{x}^\star \rangle] \Big| \leq \gamma \Big(\|\mathbf{x}_{\mathcal{S}} - \mathbf{x}_{\mathcal{S}}^\star\|_2^2 + \delta^2 \Big) \quad \text{for all } \mathbf{x} \in \mathcal{X}.$$

Finally, the simplified bound can be obtained directly by plugging in the additional assumptions. □

We present the proof of Lemma 3.8 separately for the continuous-time and discrete-time cases, since the discrete-time case involves additional technical challenges.

Proof of Lemma 3.8 in the continuous-time case. We will show that the three inequalities (3.24), (3.25) and (3.26) are satisfied by showing that, as long as all inequalities are satisfied, neither can be violated first.

Let I_0 be the index for which we have the non-zero initialization $X_{I_0}(0) > 0$. As we can only recover the signal $\mathbf{x}^\star$ up to a global sign from phaseless measurements, we can assume without loss of generality that $x_{I_0}^\star > 0$, since we can otherwise replace $\mathbf{x}^\star$ by $-\mathbf{x}^\star$ in the proof below. That is, we need to show (3.24) with $\xi = +1$.

At $t = 0$, (3.24) is satisfied by the definition of the initialization. Using standard concentration bounds for sub-exponential random variables (see e.g. Prop. 5.16 of [115]), we can bound with probability $1 - 4n^{-10}$,

$$\frac{1}{m} \sum_{j=1}^{m} Y_j = \frac{1}{m} \sum_{j=1}^{m} (\mathbf{A}_j^\top \mathbf{x}^\star)^2 + \frac{1}{m} \sum_{j=1}^{m} \varepsilon_j > 1 - (9 + 25\sigma) \sqrt{\frac{\log n}{m}}.$$

Hence, the initialization (3.7) satisfies $\|\mathbf{X}(0)\|_2^2 \geq \frac{1}{3} - (3 + 9\sigma)\sqrt{\frac{\log n}{m}}$. Similarly, we can show $\|\mathbf{X}(0)\|_2^2 \leq 2$ and therefore (3.25). Finally, inequality (3.26) is satisfied at $t = 0$ since we have

$$\mathbf{X}(0)^\top \mathbf{x}^\star \geq \|\mathbf{X}_{\mathcal{S}}(0)\|_1 x_{min}^\star \geq \frac{c_\star}{2\sqrt{k}} \geq (9 + 25\sigma) \sqrt{\frac{\log n}{m}} \geq 3\|\mathbf{X}(0)\|_2^2 - 1.$$

Step 1: (3.24) continues to hold as long as (3.25) holds

We prove this inequality by contradiction. Let $T_1 = \inf\{t \geq 0 : X_i(t)x_i^\star < 0 \text{ for some } i\}$ be the first time inequality (3.24) is violated. Assume that $T_1 < T$ and that (3.25) holds for all $t \leq T_1$. Let i be the index for which $X_i(t)x_i^\star < 0$ first occurs. This is only possible for a coordinate $i \in \mathcal{S}$, and by continuity we must have $X_i(T_1) = 0$. Without loss of generality, assume that $x_i^\star > 0$. We will show that

$$\frac{d}{dt}X_i(T_1) = -\sqrt{X_i(T_1)^2 + \beta^2}\,\nabla F(\mathbf{X}(T_1))_i > 0,$$

that is $X_i(t)$ must become positive for t close enough to T_1, which is a contradiction to the definition of T_1, and we hence must have $T_1 \geq T$. We can bound, since both (3.24) and (3.25) hold at T_1,

$$\mathbf{X}(T_1)^\top \mathbf{x}^\star \geq \|\mathbf{X}_{\mathcal{S}}(T_1)\|_1 x_{min}^\star \geq \|\mathbf{X}_{\mathcal{S}}(T_1)\|_1 \frac{c_\star}{\sqrt{k}}, \tag{3.58}$$

and hence

$$\nabla f(\mathbf{X}(T_1))_i = -2\Big(\mathbf{X}(T_1)^\top \mathbf{x}^\star\Big)x_i^\star \leq -2\|\mathbf{X}_{\mathcal{S}}(T_1)\|_1 \frac{c_\star^2}{k},$$

where we used $x_i^\star \geq x_{min}^\star \geq \frac{c_\star}{\sqrt{k}}$. As we assume $T_1 < T$, we can use the simplified bound in Lemma 3.6 to bound

$$|\nabla F(\mathbf{X}(T_1))_i - \nabla f(\mathbf{X}(T_1))_i| \leq 0.1\frac{c_\star^2\|\mathbf{X}_{\mathcal{S}}(T_1)\|_1}{k} \leq \frac{1}{20}|\nabla f(\mathbf{X}(T_1))_i|$$

with probability $1 - c_p n^{-10}$ if c_s is sufficiently large, since we have, by assumption, $\|\mathbf{X}_{\mathcal{S}^c}(T_1)\|_1 \leq \delta \leq c_1/n$. Hence, we have $\nabla F(\mathbf{X}(T_1))_i < 0$, which implies $\frac{d}{dt}X_i(T_1) > 0$ and contradicts the definition of T_1. Therefore, we must have $T_1 \geq T$.

Step 2: (3.25) continues to hold as long as (3.24) and (3.26) hold

We begin by showing the lower bound in (3.25).

Define $T_2 = \inf\{t \geq 0 : \|\mathbf{X}(t)\|_2^2 < \frac{1}{3} - (3 + 9\sigma)\sqrt{\frac{\log n}{m}}\}$, and assume that $T_2 < T$ and that inequalities (3.24) and (3.26) are satisfied for all $t \leq T_2$. By continuity, we must have $\|\mathbf{X}(T_2)\|_2^2 = \frac{1}{3} - (3 + 9\sigma)\sqrt{\frac{\log n}{m}}$, and we will show that

$$\frac{d}{dt}\|\mathbf{X}(T_2)\|_2^2 = -2\sum_{i=1}^n X_i(T_2)\frac{d}{dt}X_i(T_2) = -2\sum_{i=1}^n X_i(T_2)\sqrt{X_i(T_2)^2 + \beta^2}\,\nabla F(\mathbf{X}(T_2))_i$$

is positive, which implies $\|\mathbf{X}(t)\|_2^2 > \frac{1}{3} - (3+9\sigma)\sqrt{\frac{\log n}{m}}$ for t close enough to T_2 and contradicts the definition of T_2. Hence, we must have $T_2 \geq T$. Since $3\|\mathbf{X}(T_2)\|_2^2 - 1 < 0$, we can bound the population gradient using (3.58) and the assumption $\|\mathbf{X}_{\mathcal{S}^c}(t)\|_1 \leq \delta$:

$$\nabla f(\mathbf{X}(T_2))_i \begin{cases} \leq -2\|\mathbf{X}_{\mathcal{S}}(T_2)\|_1 \frac{c_\star^2}{k} & x_i^\star > 0 \\ \in (-\delta, \delta) & x_i^\star = 0 \\ \geq 2\|\mathbf{X}_{\mathcal{S}}(T_2)\|_1 \frac{c_\star^2}{k} & x_i^\star < 0 \end{cases} \tag{3.59}$$

Recall from the previous step that, with probability $1 - c_p n^{-10}$ and for all $i \in [n]$,

$$|\nabla F(\mathbf{X}(T_2))_i - \nabla f(\mathbf{X}(T_2))_i| \leq 0.1 \frac{c_\star^2 \|\mathbf{X}_{\mathcal{S}}(T_2)\|_1}{k}, \tag{3.60}$$

which results in bounds analogous to (3.59) for the gradient $\nabla F(\mathbf{X}(T_2))_i$. In order to bound $\frac{d}{dt}\|\mathbf{X}(T_2)\|_2^2$, we write

$$\left| \sum_{i \notin \mathcal{S}} X_i(T_2) \sqrt{X_i(T_2)^2 + \beta^2}\, \nabla F(\mathbf{X}(T_2))_i \right| \leq \|\mathbf{X}_{\mathcal{S}^c}(T_2)\|_1 \leq \delta,$$

where we used $|\sqrt{X_i(T_2)^2 + \beta^2}\, \nabla F(\mathbf{X}(T_2))_i| \leq 1$. Using $\|\mathbf{X}_{\mathcal{S}}(T_2)\|_1 \geq \frac{1}{2}$, we have

$$-\sum_{i \in \mathcal{S}} X_i(T_2) \sqrt{X_i(T_2)^2 + \beta^2}\, \nabla F(\mathbf{X}(T_2))_i \geq \sum_{i \in \mathcal{S}} X_i(T_2)^2 \frac{c_\star^2}{2k} \geq 0.15 \frac{c_\star^2}{k},$$

since $\|\mathbf{X}_{\mathcal{S}}(T_2)\|_2^2 = \|\mathbf{X}(T_2)\|_2^2 - \|\mathbf{X}_{\mathcal{S}^c}(T_2)\|_2^2 \geq 0.3$. This implies $\frac{d}{dt}\|\mathbf{X}(T_2)\|_2^2 > 0$, which contradicts the definition of T_2. Therefore, we must have $T_2 \geq T$.

The upper bound in (3.25) is an immediate consequence of (3.26): by the Cauchy-Schwarz inequality we have $\mathbf{X}(t)^\top \mathbf{x}^\star \leq \|\mathbf{X}(t)\|_2$, so, for $3\|\mathbf{X}(t)\|_2^2 - 1 > 0$, we obtain

$$\frac{2\|\mathbf{X}(t)\|_2}{3\|\mathbf{X}(t)\|_2^2 - 1} \geq \frac{2(\mathbf{X}(t)^\top \mathbf{x}^\star)}{3\|\mathbf{X}(t)\|_2^2 - 1} \geq \frac{1}{\sqrt{3}} \quad \Rightarrow \quad \|\mathbf{X}(t)\|_2 \leq \frac{1+\sqrt{2}}{\sqrt{3}} < \sqrt{2}$$

by solving the quadratic form.

Step 3: (3.26) continues to hold as long as (3.24) and (3.25) hold

The proof of (3.26) follows the same recipe as the two previous steps, although the calculations are more complicated. When $3\|\mathbf{X}(t)\|_2^2 - 1 \leq 0$, there is nothing to show. Otherwise, we can consider the ratio

$$R(t) = \frac{2(\mathbf{X}(t)^\top \mathbf{x}^\star)}{3\|\mathbf{X}(t)\|_2^2 - 1}$$

and show that it is bounded from below by $\frac{1}{\sqrt{3}}$ for all $t \leq T$.

Let $T_3 = \inf\{t \geq 0 : R(t) < \frac{1}{\sqrt{3}}\}$ and assume $T_3 < T$ as before. For notational simplicity, we will omit the argument T_3 in $X_i(T_3)$ in what follows. We can compute

$$\frac{d}{dt}R(T_3) = \sum_{i=1}^{n} -\sqrt{X_i^2 + \beta^2}\, \nabla F(\mathbf{X})_i \frac{2x_i^\star(3\|\mathbf{X}\|_2^2 - 1) - 2(\mathbf{X}^\top \mathbf{x}^\star) \cdot 6X_i}{(3\|\mathbf{X}\|_2^2 - 1)^2}. \tag{3.61}$$

By continuity, we have $R(T_3) = \frac{1}{\sqrt{3}}$. Then, for $X_i, x_i^\star > 0$, we have

$$\begin{aligned} & 2x_i^\star(3\|\mathbf{X}\|_2^2 - 1) - 2(\mathbf{X}^\top \mathbf{x}^\star) \cdot 6X_i > 0 \\ \Leftrightarrow \quad & X_i < \frac{1}{\sqrt{3}} x_i^\star, \end{aligned} \tag{3.62}$$

and the analogous result for $X_i, x_i^* < 0$. The idea to showing $\frac{d}{dt}R(t_3) > 0$ is to show that coordinates with small magnitude $|X_i| < \frac{1}{\sqrt{3}}|x_i^\star|$ are increasing in magnitude, and conversely coordinates with large magnitude $|X_i| > \frac{1}{\sqrt{3}}|x_i^\star|$ are decreasing in magnitude. To this end, we split the coordinates $i \in [n]$ into five subsets:

$$\begin{aligned} \mathcal{S}^c &= \{i \in [n] : x_i^\star = 0\}, \\ \mathcal{S}_1 &= \left\{ i \in \mathcal{S} : |X_i| < \left(\frac{1}{\sqrt{3}} - 0.1\right)|x_i^\star| \right\}, \\ \mathcal{S}_2 &= \left\{ i \in \mathcal{S} : \left(\frac{1}{\sqrt{3}} - 0.1\right)|x_i^\star| \leq |X_i| < \left(\frac{1}{\sqrt{3}} + 0.1\right)|x_i^\star| \right\}, \\ \mathcal{S}_3 &= \left\{ i \in \mathcal{S} : \left(\frac{1}{\sqrt{3}} + 0.1\right)|x_i^\star| \leq |X_i| < \frac{2}{\sqrt{3}}|x_i^\star| \right\}, \\ \mathcal{S}_4 &= \left\{ i \in \mathcal{S} : |X_i| \geq \frac{2}{\sqrt{3}}|x_i^\star| \right\}. \end{aligned}$$

We bound the sum (3.61) on each of these five sets.

- For $\mathcal{S}^c$, we have

$$\left| \sum_{i \in \mathcal{S}^c} \sqrt{X_i^2 + \beta^2}\, \nabla F(\mathbf{X})_i \cdot 2\Big(\mathbf{X}^\top \mathbf{x}^\star\Big) \cdot 6X_i \right| \leq \frac{2c_\star^2 \delta^2}{k} \|\mathbf{X}_{\mathcal{S}}\|_1 \Big(\mathbf{X}^\top \mathbf{x}^\star\Big),$$

  where we used the fact that $\sqrt{X_i^2 + \beta^2} \leq \sqrt{2}\delta$, $\|\mathbf{X}_{\mathcal{S}^c}\|_1^2 \leq \delta^2$ and that we can bound $|\nabla F(\mathbf{X})_i| \leq 0.1\|\mathbf{X}_{\mathcal{S}}\|_1 \frac{c_\star^2}{k}$ using Lemma 3.6 as in (3.60).

- For an $i \in \mathcal{S}_1$ with $x_i^\star > 0$, we have

$$\nabla f(\mathbf{X})_i = \Big(3\|\mathbf{X}\|_2^2 - 1\Big)X_i - 2\Big(\mathbf{X}^\top \mathbf{x}^\star\Big)x_i^\star \leq -0.2 \cdot \sqrt{3}\|\mathbf{X}_{\mathcal{S}}\|_1 \frac{c_\star^2}{k},$$

where we used that $3\|\mathbf{X}\|_2^2 - 1 = \sqrt{3} \cdot 2(\mathbf{X}^\top \mathbf{x}^\star)$, $X_i < (\frac{1}{\sqrt{3}} - 0.1)x_i^\star$ and (3.58). Together with the bound (3.60), this shows that $\nabla F(\mathbf{X})_i < 0$. Similarly, we can show $\nabla F(\mathbf{X})_i > 0$ for $i \in \mathcal{S}_1$ with $x_i^\star < 0$. Hence, recalling (3.62),

$$\sum_{i \in \mathcal{S}_1} -\sqrt{X_i^2 + \beta^2}\, \nabla F(\mathbf{X})_i \Big(2x_i^\star \Big(3\|\mathbf{X}\|_2^2 - 1\Big) - 2\Big(\mathbf{X}^\top \mathbf{x}^\star\Big) \cdot 6X_i\Big) \geq 0,$$

as each summand is non-neagtive.

- For $\mathcal{S}_3$, we can use the same argument as for $\mathcal{S}_1$ to show that

$$\sum_{i \in \mathcal{S}_3} -\sqrt{X_i^2 + \beta^2}\, \nabla F(\mathbf{X})_i \Big(2x_i^\star \Big(3\|\mathbf{X}\|_2^2 - 1\Big) - 2\Big(\mathbf{X}^\top \mathbf{x}^\star\Big) \cdot 6X_i\Big) \geq 0.$$

- For $\mathcal{S}_2$, we will show that the sum

$$\sum_{i \in \mathcal{S}_2} -\sqrt{X_i^2 + \beta^2}\, \nabla F(\mathbf{X})_i \Big(2x_i^\star \Big(3\|\mathbf{X}\|_2^2 - 1\Big) - 2\Big(\mathbf{X}^\top \mathbf{x}^\star\Big) \cdot 6X_i\Big)$$

is bounded from below. Let $i \in \mathcal{S}_2$ with $x_i^\star > 0$. If $X_i < \frac{1}{\sqrt{3}} x_i^\star$ and $\nabla F(\mathbf{X})_i < 0$, or if $X_i > \frac{1}{\sqrt{3}} x_i^\star$ and $\nabla F(\mathbf{X})_i > 0$, then the summand is non-negative, i.e. bounded from below by zero. For $i \in \mathcal{S}_2$ with $X_i < \frac{1}{\sqrt{3}} x_i^\star$ and $\nabla F(\mathbf{X})_i > 0$, we can bound

$$\begin{aligned}
&\sqrt{X_i^2 + \beta^2} \leq (1 + \beta)|X_i|, \\
&\nabla f(\mathbf{X})_i < 0, \\
&\nabla F(\mathbf{X})_i \leq \nabla f(\mathbf{X})_i + \Big|\nabla f(\mathbf{X})_i - \nabla F(\mathbf{X})_i\Big| \leq 0.1 \|\mathbf{X}_{\mathcal{S}}\|_1 \frac{c_\star^2}{k},
\end{aligned}$$

where we used the bound (3.60). Recalling the definition of $\mathcal{S}_2$, we have

$$2x_i^\star \Big(3\|\mathbf{X}\|_2^2 - 1\Big) - 2\Big(\mathbf{X}^\top \mathbf{x}^\star\Big) \cdot 6X_i = 12\Big(\mathbf{X}^\top \mathbf{x}^\star\Big)\left(\frac{1}{\sqrt{3}} x_i^\star - X_i\right) \leq 3\Big(\mathbf{X}^\top \mathbf{x}^\star\Big) X_i.$$

Putting this together, we can bound

$$\begin{aligned}
&-\sqrt{X_i^2 + \beta^2}\, \nabla F(\mathbf{x})_i \Big(2x_i^\star \Big(3\|\mathbf{X}\|_2^2 - 1\Big) - 2\Big(\mathbf{X}^\top \mathbf{x}^\star\Big) \cdot 6X_i\Big) \\
\geq &-\frac{0.3(1 + \beta)c_\star^2}{k} \|\mathbf{X}_{\mathcal{S}}\|_1 \Big(\mathbf{X}^\top \mathbf{x}^\star\Big) X_i^2.
\end{aligned}$$

Together with the analogous bounds for the cases $X_i > \frac{1}{\sqrt{3}} x_i^\star$ and $\nabla F(\mathbf{X})_i < 0$, and for $i \in \mathcal{S}_2$ with $x_i^\star < 0$, this yields

$$\begin{aligned}
&\sum_{i \in \mathcal{S}_2} -\sqrt{X_i^2 + \beta^2}\, \nabla F(\mathbf{X})_i \Big(2x_i^\star \Big(3\|\mathbf{X}\|_2^2 - 1\Big) - 2\Big(\mathbf{X}^\top \mathbf{x}^\star\Big) \cdot 6X_i\Big) \\
\geq &-\frac{0.3(1 + \beta)c_\star^2}{k} \|\mathbf{X}_{\mathcal{S}}\|_1 \Big(\mathbf{X}^\top \mathbf{x}^\star\Big) \|\mathbf{X}_{\mathcal{S}_2}\|_2^2.
\end{aligned}$$

- Finally, for $i \in \mathcal{S}_4$ we have

$$\begin{aligned}
&\sqrt{X_i^2+\beta^2} \geq |X_i|,\\
&|\nabla f(\mathbf{X})_i| = 2\Big(\mathbf{X}^\top \mathbf{x}^\star\Big)\Big|\sqrt{3}X_i - x_i^\star\Big| \geq 2\Big(\mathbf{X}^\top \mathbf{x}^\star\Big)|x_i^\star| \geq 2\|\mathbf{X}_{\mathcal{S}}\|_1 \frac{c_\star^2}{k},\\
&|\nabla F(\mathbf{X})_i| \geq |\nabla f(\mathbf{X})_i| - |\nabla f(\mathbf{X})_i - \nabla F(\mathbf{X})_i| \geq 1.9\|\mathbf{X}_{\mathcal{S}}\|_1 \frac{c_\star^2}{k},
\end{aligned}$$

  where we used the bound (3.60). Recalling the definition of $\mathcal{S}_4$, we can bound

$$\Big|2x_i^\star\Big(3\|\mathbf{X}\|_2^2 - 1\Big) - 2\Big(\mathbf{X}^\top \mathbf{x}^\star\Big)\cdot 6X_i\Big| = 12\Big(\mathbf{X}^\top \mathbf{x}^\star\Big)\left|\frac{1}{\sqrt{3}}x_i^\star - X_i\right| \geq 6\Big(\mathbf{X}^\top \mathbf{x}^\star\Big)X_i.$$

  Putting everything together, we can bound

$$\begin{aligned}
&\sum_{i\in\mathcal{S}_2} -\sqrt{X_i^2+\beta^2}\,\nabla F(\mathbf{X})_i\Big(2x_i^\star\Big(3\|\mathbf{X}\|_2^2-1\Big) - 2\Big(\mathbf{X}^\top\mathbf{x}^\star\Big)\cdot 6X_i\Big)\\
\geq &\frac{11.4c_\star^2}{k}\|\mathbf{X}_{\mathcal{S}}\|_1\Big(\mathbf{X}^\top\mathbf{x}^\star\Big)\|\mathbf{X}_{\mathcal{S}_4}\|_2^2.
\end{aligned}$$

Putting these five sums together, we have shown that $\frac{d}{dt}R(T_3) > 0$ if we can show that

$$\frac{11.4c_\star^2}{k}\|\mathbf{X}_{\mathcal{S}}\|_1\Big(\mathbf{X}^\top\mathbf{x}^\star\Big)\|\mathbf{X}_{\mathcal{S}_4}\|_2^2 \geq \left(\frac{0.3(1+\beta)c_\star^2}{k}\|\mathbf{X}_{\mathcal{S}_2}\|_2^2 + \frac{2c_\star^2\delta^2}{k}\right)\|\mathbf{X}_{\mathcal{S}}\|_1\Big(\mathbf{X}^\top\mathbf{x}^\star\Big)$$

Since $\delta \leq c_1/n$ is sufficiently small compared to $\|\mathbf{X}_{\mathcal{S}_4}\|_2$, this reduces to showing

$$11\|\mathbf{X}_{\mathcal{S}_4}\|_2^2 \geq 0.3(1+\beta)\|\mathbf{X}_{\mathcal{S}_2}\|_2^2. \tag{3.63}$$

Now, we can rearrange the equality $R(t) = \frac{1}{\sqrt{3}}$ to obtain

$$\begin{aligned}
&3\Big(\|\mathbf{X}_{\mathcal{S}_1}\|_2^2 + \|\mathbf{X}_{\mathcal{S}_2}\|_2^2 + \|\mathbf{X}_{\mathcal{S}_3}\|_2^2 + \|\mathbf{X}_{\mathcal{S}_4}\|_2^2 + \|\mathbf{X}_{\mathcal{S}^c}\|_2^2\Big) - 1\\
=&\sqrt{3}\cdot 2\Big(\mathbf{X}_{\mathcal{S}_1}^\top\mathbf{x}_{\mathcal{S}_1}^\star + \mathbf{X}_{\mathcal{S}_2}^\top\mathbf{x}_{\mathcal{S}_2}^\star + \mathbf{X}_{\mathcal{S}_3}^\top\mathbf{x}_{\mathcal{S}_3}^\star + \mathbf{X}_{\mathcal{S}_4}^\top\mathbf{x}_{\mathcal{S}_4}^\star\Big).
\end{aligned}$$

By definition, we have $3X_i^2 < \sqrt{3}\cdot 2X_i x_i^\star$ for $i \in S_1 \cup S_2 \cup S_3$, which gives

$$3\|\mathbf{X}_{\mathcal{S}_j}\|_2^2 < \sqrt{3}\cdot 2\mathbf{X}_{\mathcal{S}_j}^\top\mathbf{x}_{\mathcal{S}_j}^\star, \quad \text{for } j = 1,2,3.$$

Further, we have $\|\mathbf{X}_{\mathcal{S}^c}\|_2^2 \leq \delta$, so

$$3\|\mathbf{X}_{\mathcal{S}_4}\|_2^2 - 2\sqrt{3}\cdot\mathbf{X}_{\mathcal{S}_4}^\top\mathbf{x}_{\mathcal{S}_4}^\star > 1 - 3\delta + \Big(2\sqrt{3}\cdot\mathbf{X}_{\mathcal{S}_2}^\top\mathbf{x}_{\mathcal{S}_2}^\star - 3\|\mathbf{X}_{\mathcal{S}_2}\|_2^2\Big).$$

By the definition of $\mathcal{S}_2$, we have for $i \in \mathcal{S}_2$,

$$2\sqrt{3}X_i x_i^\star - 3X_i^2 \geq \left(\frac{2\sqrt{3}}{1/\sqrt{3}+0.1} - 3\right)X_i^2 \geq 2.2X_i^2.$$

Since $\mathbf{X}_{\mathcal{S}_4}^\top \mathbf{x}^\star_{\mathcal{S}_4} \geq 0$ by (3.24), this gives

$$3\|\mathbf{X}_{\mathcal{S}_4}\|_2^2 > 1 - 3\delta + 2.2\|\mathbf{X}_{\mathcal{S}_2}\|_2^2,$$

which shows that (3.63) holds, thus completing the proof of (3.26). □

Proof of Lemma 3.8 in the discrete-time case. We will prove the inequalities (3.24), (3.25) and (3.26) via induction. Let I_0 be the index for which we have the non-zero initialization $X_{I_0}^0 > 0$. As in the continuous-time case, we can assume without loss of generality that $x^\star_{I_0} > 0$ and show (3.24) with $\xi = +1$. The inequalities (3.24), (3.25) and (3.26) hold at $t = 0$ as in the continuous-time case.

Step 1: (3.24) continues to hold as long as (3.25) holds

For $t \geq 0$, let $i \in \mathcal{S}$ with $x_i^\star, X_i^t \geq 0$. We will show that also $X_i^{t+1} \geq 0$. Negative coordinates can be treated the same way, and for $i \notin \mathcal{S}$ there is nothing to show. Rearranging (3.40) shows that $X_i^{t+1} \geq 0$ if and only if

$$\frac{\beta^2}{\Big(\sqrt{(X_i^t)^2 + \beta^2} + X_i^t\Big)^2} = \frac{\sqrt{(X_i^t)^2 + \beta^2} - X_i^t}{\sqrt{(X_i^t)^2 + \beta^2} + X_i^t} \leq \exp\Big(-2\eta \nabla F(\mathbf{X}^t)_i\Big). \tag{3.64}$$

In order to show (3.64), we need to bound the gradient $\nabla F(\mathbf{X}^t)_i$ from above. Since by the induction hypothesis we have $\|\mathbf{X}_{\mathcal{S}}^t\|_1 = \|\mathbf{X}^t\|_1 - \|\mathbf{X}_{\mathcal{S}^c}^t\|_1 \geq \frac{1}{2}$ and by assumption $\delta \leq c_1/n$, we can apply the simplified bound in Lemma 3.6 to bound, for $t \leq T$,

$$|\nabla F(\mathbf{X}^t)_i - \nabla f(\mathbf{X}^t)_i| \leq \frac{c_\star^2 \|\mathbf{X}_{\mathcal{S}}^t\|_1}{k}$$

with probability $1 - c_p n^{-10}$, provided that c_s is sufficiently large. Further, we can bound, since (3.24) holds at time t by induction hypothesis,

$$(\mathbf{X}^t)^\top \mathbf{x}^\star \geq \|\mathbf{X}_{\mathcal{S}}^t\|_1 x^\star_{min} \geq \|\mathbf{X}_{\mathcal{S}}^t\|_1 \frac{c_\star}{\sqrt{k}},$$

which leads to

$$\begin{aligned} \nabla F(\mathbf{X}^t)_i &\leq \nabla f(\mathbf{X}^t)_i + |\nabla F(\mathbf{X}^t)_i - \nabla f(\mathbf{X}^t)_i| \\ &\leq \Big(3\|\mathbf{X}^t\|_2^2 - 1\Big) X_i^t - \Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big) x_i^\star. \end{aligned} \tag{3.65}$$

If $X_i^t \leq \frac{c_\star}{2\sqrt{3k}}$, then $\nabla F(\mathbf{X}^t)_i \leq 0$ by (3.26). In this case, (3.64) holds since

$$\frac{\beta^2}{\Big(\sqrt{(X_i^t)^2 + \beta^2} + X_i^t\Big)^2} \leq 1 \leq \exp\Big(-2\eta \nabla F(\mathbf{X}^t)_i\Big).$$

If on the other hand $X_i^t > \frac{c_\star}{2\sqrt{3k}}$, we can bound $\nabla F(\mathbf{X}^t)_i \le 5X_i^t$ since $\|\mathbf{X}^t\|_2^2 \le 2$, and

$$\frac{\beta^2}{\Big(\sqrt{(X_i^t)^2+\beta^2}+X_i^t\Big)^2} \le \frac{\beta^2}{4(X_i^t)^2} \le \exp\Big(-10\eta X_i^t\Big)$$

holds for all $\eta \le \frac{1}{\sqrt{200}}$, where we used that $X_i^t \le \sqrt{2}$ by (3.25) and $\log \frac{c_\star^2}{3k\beta^2} \ge 1$ since by assumption $\beta \le c_1/n$. This completes the proof that (3.64) holds. Repeating the same argument for $i \in \mathcal{S}$ with $x_i^\star, X_i^t \le 0$ shows that (3.24) is satisfied at $t+1$.

Step 2: (3.25) continues to hold as long as (3.24) and (3.26) hold

We begin by inductively showing the lower bound in (3.25). For $t \ge 0$, we write

$$\|\mathbf{X}^{t+1}\|_2^2 - \|\mathbf{X}^t\|_2^2 = \sum_{i=1}^n \Big(X_i^{t+1}+X_i^t\Big)\Big(X_i^{t+1}-X_i^t\Big),$$

and consider the partial sums for $i \in \mathcal{S}$ and $i \notin \mathcal{S}$ separately. For $i \notin \mathcal{S}$, we have

$$\sum_{i\notin\mathcal{S}} \Big(X_i^{t+1}+X_i^t\Big)\Big(X_i^{t+1}-X_i^t\Big) \ge -\sum_{i\notin\mathcal{S}} (X_i^t)^2 \ge -n\delta^2,$$

where we used $\|\mathbf{X}^t_{\mathcal{S}^c}\|_2 \le \sqrt{n}\|\mathbf{X}^t_{\mathcal{S}^c}\|_1 \le \sqrt{n}\delta$.

To bound the partial sum containing $i \in \mathcal{S}$, we distinguish two cases: $\|\mathbf{X}^t\|_2^2 - \frac{1}{3} \ge 2n\delta^2$ and $\|\mathbf{X}^t\|_2^2 - \frac{1}{3} < 2n\delta^2$. In the first case, let $i \in \mathcal{S}$ and without loss of generality $x_i^\star > 0$. If $X_i^{t+1} \ge X_i^t$, then the summand $(X_i^{t+1}+X_i^t)(X_i^{t+1}-X_i^t)$ is non-negative. If on the other hand $X_i^{t+1} < X_i^t$, then $\nabla F(\mathbf{X}^t)_i > 0$, and we can bound, recalling (3.43),

$$X_i^{t+1} - X_i^t \ge -\frac{3}{2}\eta \nabla F(\mathbf{X}^t)_i \sqrt{(X_i^t)^2+\beta^2} \ge -\frac{3}{2}\eta\Big(3\|\mathbf{X}^t\|_2^2 - 1\Big)X_i^t\sqrt{(X_i^t)^2+\beta^2},$$

where we used (3.65) for the last inequality. An analogous upper bound can be derived for negative coordinates with $x_i^\star < 0$, so, writing $\mathcal{S}' = \{i \in \mathcal{S} : |X_i^{t+1}| < |X_i^t|\}$ for the coordinates which decrease in magnitude, we can bound

$$\begin{aligned}\sum_{i\in\mathcal{S}'} \Big(X_i^{t+1}+X_i^t\Big)\Big(X_i^{t+1}-X_i^t\Big) &\ge -3\eta \sum_{i\in\mathcal{S}'} X_i^t\Big(3\|\mathbf{X}^t\|_2^2-1\Big)X_i^t\sqrt{(X_i^t)^2+\beta^2} \\ &\ge -\bigg(9\eta\sum_{i\in\mathcal{S}'}(X_i^t)^2\Big(|X_i^t|+\beta\Big)\bigg)\Big(\|\mathbf{X}^t\|_2^2-\frac{1}{3}\Big),\end{aligned}$$

where for the last inequality we used that $\sqrt{x^2+\beta^2} \le |x|+\beta$ for $\beta \ge 0$. Hence,

$$\|\mathbf{X}^{t+1}\|_2^2 - \|\mathbf{X}^t\|_2^2 \ge \frac{1}{3} - \|\mathbf{X}^t\|_2^2$$

holds for $\eta \leq \frac{1}{54}$, where we used that $\sum_{i\in\mathcal{S}'}(X_i^t)^2(|X_i^t|+\beta) \leq 3$ since the upper bound in (3.25) holds at time t by induction hypothesis.

Next, we consider the case $\|\mathbf{X}^t\|_2^2 - \frac{1}{3} < 2n\delta^2$. In this case, it follows from (3.65) that every coordinate $i \in \mathcal{S}$ must be increasing in magnitude, i.e. $|X_i^{t+1}| > |X_i^t|$ for all $i \in \mathcal{S}$. Assuming without loss of generality that $x_i^\star > 0$, we can use (3.44) to bound by how much X_i^t must at least increase:

$$\begin{aligned}
X_i^{t+1} - X_i^t &\geq -\frac{1}{2}\eta\nabla F(\mathbf{X}^t)_i\sqrt{(X_i^t)^2+\beta^2} \\
&\geq -\frac{1}{2}\eta\Big(\big(3\|\mathbf{X}^t\|_2^2 - 1\big)X_i^t - \big((\mathbf{X}^t)^\top\mathbf{x}^\star\big)x_i^\star\Big)\sqrt{(X_i^t)^2+\beta^2} \\
&\geq -\frac{1}{2}\eta\left(6n\delta^2 X_i^t - \frac{c_\star^2}{2k}\right)\sqrt{(X_i^t)^2+\beta^2} \\
&\geq \frac{c_\star^2\eta}{5k}X_i^t,
\end{aligned}$$

where we used (3.28) and the assumptions $x_{min}^\star \geq \frac{c_\star}{\sqrt{k}}$ and $\delta \leq c_1/n$. Hence, we have

$$\begin{aligned}
\sum_{i\in\mathcal{S}}\big(X_i^{t+1}+X_i^t\big)\big(X_i^{t+1}-X_i^t\big) &\geq \sum_{i\in\mathcal{S}} 2X_i^t\frac{c_\star^2\eta}{5k}X_i^t \\
&\geq \eta\frac{2c_\star^2}{5k}\left(\frac{1}{3} - (3+9\sigma)\sqrt{\frac{\log n}{m}} - \delta^2\right),
\end{aligned}$$

where we used the induction hypothesis that the lower bound in (3.25) is satisfied at time t along with the bound $\|\mathbf{X}_{\mathcal{S}^c}^t\|_2^2 \leq \|\mathbf{X}_{\mathcal{S}^c}^t\|_1^2 \leq \delta^2$. Thus, we have $\|\mathbf{X}^{t+1}\|_2^2 - \|\mathbf{X}^t\|_2^2 > 0$ if $\delta \leq c_1/n$ for $c_1 > 0$ sufficiently small.

The upper bound in (3.25) can be shown the same way as in the continuous-time case.

Step 3: (3.26) continues to hold as long as (3.24) and 3.25 hold

As in the continuous-time case, we can restrict our attention to the case $3\|\mathbf{X}^t\|_2^2 - 1 > 0$ (since otherwise there is nothing to show), and consider the ratio

$$R_t = \frac{2((\mathbf{X}^t)^\top\mathbf{x}^\star)}{3\|\mathbf{X}^t\|_2^2 - 1}.$$

As in the continuous-time case, we can show $R_0 \geq 1/\sqrt{3}$.

To show $R_{t+1} \geq 1/\sqrt{3}$ for any $t \geq 0$, we consider the two cases $R_t \geq 1/\sqrt{3} + 0.05$ and $1/\sqrt{3} \leq R_t < 1/\sqrt{3} + 0.05$. We assume $3\|\mathbf{X}^t\|_2^2 - 1 > 0$ for convenience's sake, as the other case $3\|\mathbf{X}^t\|_2^2 - 1 < 0$ can be treated the same way, but requires distinguishing cases at various points in the following argument, which we omit to avoid repetition.

Step 3, Case 1: $R_t \geq 1/\sqrt{3} + 0.05$

We consider four types of coordinates, defined by

$$\begin{aligned}
\mathcal{S}^c &= \{i \in [n] : x_i^\star = 0\} && \text{(off-support)} \\
\mathcal{S}_1 &= \left\{ i \in \mathcal{S} : |X_i^t| \leq \frac{1}{3}|x_i^\star| \right\} && \text{(small coordinates)} \\
\mathcal{S}_2 &= \left\{ i \in \mathcal{S} : |X_i^t| > \frac{1}{3}|x_i^\star|,\ |X_i^{t+1}| > |X_i^t| \right\} && \text{(large increasing coordinates)} \\
\mathcal{S}_3 &= \left\{ i \in \mathcal{S} : |X_i^t| > \frac{1}{3}|x_i^\star|,\ |X_i^{t+1}| \leq |X_i^t| \right\} && \text{(large decreasing coordinates)}
\end{aligned}$$

We will show $R_{t+1} \geq 1/\sqrt{3}$ by considering the following sequence of vectors defined by

$$\mathbf{X}^{(1)} = \mathbf{X}^t, \quad X_i^{(2)} = \begin{cases} X_i^{(1)} & i \notin \mathcal{S}^c \\ X_i^{t+1} & i \in \mathcal{S}^c \end{cases}, \quad X_i^{(3)} = \begin{cases} X_i^{(2)} & i \notin \mathcal{S}_2 \\ X_i^{t+1} & i \in \mathcal{S}_2 \end{cases},$$
$$X_i^{(4)} = \begin{cases} X_i^{(3)} & i \notin \mathcal{S}_3 \\ X_i^{t+1} & i \in \mathcal{S}_3 \end{cases}, \quad \mathbf{X}^{(5)} = \mathbf{X}^{t+1},$$

and bound the ratio $\frac{2((\mathbf{X}^{(i)})^\top \mathbf{x}^\star)}{3\|\mathbf{X}^{(i)}\|_2^2 - 1}$ for each $i = 2, ..., 5$.

Exchanging off-support coordinates

By assumption, we have $R_t \geq 1/\sqrt{3} + 0.05$, which can also be written as

$$\left(\frac{1}{\sqrt{3}} + 0.05\right)^{-1} \cdot 2\Big((\mathbf{X}^{(1)})^\top \mathbf{x}^\star\Big) \geq 3\|\mathbf{X}^{(1)}\|_2^2 - 1.$$

As shown in Step 2, $\|\mathbf{X}^{(2)}\|_2^2 \leq \|\mathbf{X}^{(1)}\|_2^2 + n\delta^2$, so we have, since $((\mathbf{X}^{(1)})^\top \mathbf{x}^\star) \geq \frac{c_\star}{2\sqrt{k}}$,

$$\left(\frac{1}{\sqrt{3}} + 0.04\right)^{-1} \cdot 2\Big((\mathbf{X}^{(2)})^\top \mathbf{x}^\star\Big) \geq 3\|\mathbf{X}^{(2)}\|_2^2 - 1, \tag{3.66}$$

provided that $\delta \leq c_1/n$ is sufficiently small.

Exchanging large increasing coordinates

Let $i \in \mathcal{S}_2$ with $x_i^\star > 0$. Analogous bounds for the case $x_i^\star < 0$ can be derived the same way. As $X_i^{t+1} > X_i^t$, we must have $\nabla F(\mathbf{X}^t)_i < 0$. First, we bound by how much X_i^t can increase in one iteration. We can write

$$\nabla f(\mathbf{X}^t)_i = \frac{2((\mathbf{X}^t)^\top \mathbf{x}^\star)}{R_t} X_i^t - 2\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big) x_i^\star = -2\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big)\left(x_i^\star - \frac{1}{R_t} X_i^t\right).$$

Using the simplified bound of Lemma 3.6, we have with probability $1 - c_p n^{-10}$,

$$|\nabla F(\mathbf{X}^t)_i - \nabla f(\mathbf{X}^t)_i| \leq 0.1\|\mathbf{X}_{\mathcal{S}}^t\|_1 \frac{c_\star^2}{k} \leq 0.1\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big) x_i^\star,$$

where we used (3.58). This gives the lower bound

$$\nabla F(\mathbf{X}^t)_i \geq -2\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big)\left(1.05x_i^\star - \frac{1}{R_t}X_i^t\right) \geq -1.05 \cdot 2\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big)x_i^\star, \tag{3.67}$$

since $x_i^\star > 0$ and hence $X_i^t \geq 0$ by (3.24). With this, we can use (3.43) to bound

$$\begin{aligned} X_i^{t+1} - X_i^t &\leq -\frac{3}{2}\eta \nabla F(\mathbf{X}^t)_i \sqrt{(X_i^t)^2 + \beta^2} \\ &\leq 1.7\eta \cdot 2\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big)x_i^\star |X_i^t|, \end{aligned} \tag{3.68}$$

where we used that $\sqrt{(X_i^t)^2 + \beta^2} \leq 1.05|X_i^t|$ for $i \in \mathcal{S}_2$. Further, we can use the upper bound in (3.25) to bound $X_i^{t+1} + X_i^t \leq 3$. With this, we obtain

$$\begin{aligned} \frac{2((\mathbf{X}^{(3)})^\top \mathbf{x}^\star)}{3\|\mathbf{X}^{(3)}\|_2^2 - 1} &\geq \frac{2((\mathbf{X}^{(2)})^\top \mathbf{x}^\star)}{3\|\mathbf{X}^{(2)}\|_2^2 - 1 + \sum_{i \in \mathcal{S}_2} 3(X_i^{t+1} + X_i^t)(X_i^{t+1} - X_i^t)} \\ &\geq \frac{2((\mathbf{X}^{(2)})^\top \mathbf{x}^\star)}{3\|\mathbf{X}^{(2)}\|_2^2 - 1 + 3 \cdot 3 \cdot 1.7 \cdot 2\eta((\mathbf{X}^t)^\top \mathbf{x}^\star)^2}, \end{aligned}$$

which is bounded from below by $1/\sqrt{3}$, provided that

$$\eta \leq \frac{1}{193} \leq \frac{1}{15.3((\mathbf{X}^t)^\top \mathbf{x}^\star)}\left(\sqrt{3} - \frac{3\|\mathbf{X}^{(2)}\|_2^2 - 1}{2((\mathbf{X}^{(2)})^\top \mathbf{x}^\star)}\right),$$

where we used that $(\mathbf{X}^{(2)})^\top \mathbf{x}^\star = (\mathbf{X}^t)^\top \mathbf{x}^\star \leq \sqrt{2}$ by (3.25) and inequality (3.66).

Exchanging large decreasing coordinates

Let $i \in \mathcal{S}_3$ with $x_i^\star > 0$. Analogous bounds for the case $x_i^\star < 0$ can be derived the same way. We first bound by how much X_i^t can decrease in one iteration. Following the same steps as in the derivation of the lower bound (3.67), we obtain the analogous upper bound

$$\nabla F(\mathbf{X}^t)_i \leq 2\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big)\left(\frac{1}{R_t}X_i^t - 0.95x_i^\star\right). \tag{3.69}$$

Since we are considering a positive decreasing coordinate, we must have $\nabla F(\mathbf{X}^t)_i > 0$ and therefore $X_i^t \geq 0.95R_t x_i^\star$. As before, we can use (3.43) to bound

$$\begin{aligned} X_i^{t+1} - X_i^t &\geq -\frac{3}{2}\eta \nabla F(\mathbf{X}^t)_i \sqrt{(X_i^t)^2 + \beta^2} \\ &\geq -3.2\eta\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big)\frac{1}{R_t}X_i^t\Big(X_i^t - 0.95R_t x_i^\star\Big), \end{aligned} \tag{3.70}$$

where we used that $\sqrt{(X_i^t)^2 + \beta^2} \leq 1.05|X_i^t|$ for $i \in \mathcal{S}_3$. Further, we can write

$$X_i^{t+1} \geq \left(1 - 3.2\eta\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big)\frac{1}{R_t}X_i^t\right)X_i^t + 3.2\eta\Big((\mathbf{X}^t)^\top \mathbf{x}^\star\Big)\frac{1}{R_t}X_i^t \cdot 0.95R_t x_i^\star \geq 0.95R_t x_i^\star,$$

provided $\eta \leq \frac{1}{6.4\sqrt{3}} \leq \frac{R_t}{3.2((\mathbf{X}^t)^\top \mathbf{x}^\star)X_i^t}$, where we used (3.25) to bound $((\mathbf{X}^t)^\top \mathbf{x}^\star)X_i^t \leq 2$ and $X_i^t \geq 0.95 R_t x_i^\star$. We can now bound

$$\begin{aligned}
3\|\mathbf{X}^{(4)}\|_2^2 - 1 &= 3\|\mathbf{X}^{(3)}\|_2^2 - 1 + \sum_{i \in \mathcal{S}_3} 3(X_i^{t+1} + X_i^t)(X_i^{t+1} - X_i^t) \\
&\leq 3\|\mathbf{X}^{(3)}\|_2^2 - 1 + \sum_{i \in \mathcal{S}_3} 3 \cdot 2 \cdot 0.95 R_t x_i^\star (X_i^{t+1} - X_i^t) \\
&\leq \sqrt{3}\Big(2((\mathbf{X}^{(3)})^\top \mathbf{x}^\star) + \sum_{i \in \mathcal{S}_3} 2x_i^\star (X_i^{t+1} - X_i^t)\Big) \\
&= \sqrt{3} \cdot 2\Big((\mathbf{X}^{(4)})^\top \mathbf{x}^\star\Big),
\end{aligned}$$

where the second inequality holds because $3 \cdot 0.95 R_t \geq \sqrt{3}$ for $R_t \geq \frac{1}{\sqrt{3}} + 0.05$, and $x_i^\star(X_i^{t+1} - X_i^t) < 0$.

Exchanging small coordinates

Let $i \in \mathcal{S}_1$ with $x_i^\star > 0$. Analogous bounds for the case $x_i^\star < 0$ can be derived the same way. Then, we have $\nabla F(\mathbf{X}^t)_i < 0$ by (3.69) as $X_i^t \leq \frac{1}{3} x_i^\star$, that is all small coordinates must increase in magnitude. Further, we can use (3.43) to bound

$$X_i^{t+1} + X_i^t \leq |X_i^{t+1} - X_i^t| + 2X_i^t \leq -\frac{3}{2}\eta \nabla F(\mathbf{X}^t)_i \sqrt{(X_i^t)^2 + \beta^2} + \frac{2}{3} x_i^\star \leq \frac{2}{\sqrt{3}} x_i^\star,$$

if $\eta \leq 0.3$, where we used that $(\mathbf{X}^t)^\top \mathbf{x}^\star \leq \sqrt{2}$ by (3.25), $\sqrt{(X_i^t)^2 + \beta^2} \leq 0.35$ by the definition of $\mathcal{S}_1$, and the lower bound (3.67) for $\nabla F(\mathbf{X}^t)_i$. Hence, we can bound

$$\begin{aligned}
3\|\mathbf{X}^{(5)}\|_2^2 - 1 &= 3\|\mathbf{X}^{(4)}\|_2^2 - 1 + \sum_{i \in \mathcal{S}_1} 3(X_i^{t+1} + X_i^t)(X_i^{t+1} - X_i^t) \\
&\leq \sqrt{3}\Big(2((\mathbf{X}^{(4)})^\top \mathbf{x}^\star) + \sum_{i \in \mathcal{S}_1} 2x_i^\star (X_i^{t+1} - X_i^t)\Big) \\
&= \sqrt{3} \cdot 2\Big((\mathbf{X}^{(5)})^\top \mathbf{x}^\star\Big).
\end{aligned}$$

This completes the proof of $R_{t+1} \geq 1/\sqrt{3}$ if $R_t \geq 1/\sqrt{3} + 0.05$.

Step 3, Case 2: $1/\sqrt{3} \leq R_t < 1/\sqrt{3} + 0.05$

The proof that $R_{t+1} \geq 1/\sqrt{3}$ largely follows the same steps as the previous case, although it requires a different sequence of vectors interpolating between $\mathbf{X}^t$ and $\mathbf{X}^{t+1}$. Define $\widetilde{\mathbf{X}}^{t+1}$ by

$$\nabla \Phi(\widetilde{\mathbf{X}}^{t+1}) = \nabla \Phi(\mathbf{X}^t) - \eta \nabla f(\mathbf{X}^t),$$

that is the vector obtained by applying one step of mirror descent with the population gradient. Replacing X_i^{t+1} by $\widetilde{X}_i^{t+1}$ in the definitions of the sets $\mathcal{S}_i$, consider the sequence

$$\mathbf{X}^{(1)} = \mathbf{X}^t, \quad X_i^{(2)} = \begin{cases} X_i^{(1)} & i \notin \mathcal{S}_2 \\ \widetilde{X}_i^{t+1} & i \in \mathcal{S}_2 \end{cases}, \quad X_i^{(3)} = \begin{cases} X_i^{(2)} & i \notin \mathcal{S}_3 \\ \widetilde{X}_i^{t+1} & i \in \mathcal{S}_3 \end{cases},$$
$$X_i^{(4)} = \begin{cases} X_i^{(3)} & i \notin \mathcal{S}_2 \cup \mathcal{S}_3 \\ X_i^{t+1} & i \in \mathcal{S}_2 \cup \mathcal{S}_3 \end{cases}, \quad X_i^{(5)} = \begin{cases} X_i^{(4)} & i \notin \mathcal{S}^c \\ X_i^{t+1} & i \in \mathcal{S}^c \end{cases}, \quad \mathbf{X}^{(6)} = \mathbf{X}^{t+1}.$$

Exchanging large increasing coordinates with population gradient

Let $i \in \mathcal{S}_2$ with $x_i^\star > 0$. Analogous bounds for the case $x_i^\star < 0$ can be established the same way. Since we are considering an increasing coordinate, the population gradient

$$\nabla f(\mathbf{X}^t)_i = -\Big(3\|\mathbf{X}^t\|_2^2 - 1\Big)\Big(R_t x_i^\star - X_i^t\Big)$$

must be negative, so we must have $X_i^t \le R_t x_i^\star$. The bound (3.68) becomes

$$\widetilde{X}_i^{t+1} - X_i^t \le 1.6\eta\Big(3\|\mathbf{X}^t\|_2^2 - 1\Big)\Big(R_t x_i^\star - X_i^t\Big)X_i^t,$$

and we can bound

$$\widetilde{X}_i^{t+1} \le \Big(1 + 1.6\eta\Big(3\|\mathbf{X}^t\|_2^2 - 1\Big)\Big(R_t x_i^\star - X_i^t\Big)\Big)X_i^t \le R_t x_i^\star$$

if $\eta \le \frac{1}{8\sqrt{2}} \le \frac{1}{1.6(3\|\mathbf{X}^t\|_2^2 - 1)X_i^t}$, as the expression is increasing in η. We can bound

$$\begin{aligned} \frac{2((\mathbf{X}^{(2)})^\top \mathbf{x}^\star)}{3\|\mathbf{X}^{(2)}\|_2^2 - 1} &= \frac{2((\mathbf{X}^t)^\top \mathbf{x}^\star) + \sum_{i\in\mathcal{S}_2} 2x_i^\star(\widetilde{X}_i^{t+1} - X_i^t)}{3\|\mathbf{X}^t\|_2^2 - 1 + \sum_{i\in\mathcal{S}_2} 3(\widetilde{X}_i^{t+1} + X_i^t)(\widetilde{X}_i^{t+1} - X_i^t)} \\ &\ge \frac{2((\mathbf{X}^t)^\top \mathbf{x}^\star) + B}{3\|\mathbf{X}^t\|_2^2 - 1 + 3R_t B}, \end{aligned}$$

where $B = \sum_{i\in\mathcal{S}_2} 2x_i^\star(\widetilde{X}_i^{t+1} - X_i^t)$. This ratio is bounded from below by $1/\sqrt{3}$ if

$$B \le \frac{2((\mathbf{X}^t)^\top \mathbf{x}^\star) - (3\|\mathbf{X}^t\|_2^2 - 1)/\sqrt{3}}{\sqrt{3}R_t - 1} = \frac{3\|\mathbf{X}^t\|_2^2 - 1}{\sqrt{3}},$$

which is satisfied as, using $|X_i^t| \le R_t|x_i^\star|$, we can bound

$$\begin{aligned} B &= \sum_{i\in\mathcal{S}_2} 2x_i^\star(\widetilde{X}_i^{t+1} - X_i^t) \\ &\le \sum_{i\in\mathcal{S}_2} 2x_i^\star \cdot 1.6\eta\Big(3\|\mathbf{X}^t\|_2^2 - 1\Big)\Big(R_t x_i^\star - X_i^t\Big)|X_i^t| \\ &\le \Big(3\|\mathbf{X}^t\|_2^2 - 1\Big) \cdot 3.2\eta R_t^2 \sum_{i\in\mathcal{S}_2} |x_i^\star|^3 \\ &\le \frac{3\|\mathbf{X}^t\|_2^2 - 1}{\sqrt{3}} \end{aligned}$$

for $\eta \le 0.45 \le \frac{1}{3.2\sqrt{3}R_t^2}$, where we used that $\sum_{i\in\mathcal{S}_2}|x_i^\star|^3 \le 1$ and $R_t \le 1/\sqrt{3}+0.05$.

Exchanging large decreasing coordinates with population gradient

Let $i \in \mathcal{S}_3$ with $x_i^\star > 0$. Analogous bounds for the case $x_i^\star < 0$ can be established the same way. Following the same steps as before, we can bound by how much X_i^t can decrease after one step of mirror descent using the population gradient, and the bound (3.70) becomes

$$X_i^{(3)} - X_i^t \ge -3.2\eta\Big((\mathbf{X}^t)^\top\mathbf{x}^\star\Big)\left(\frac{1}{R_t}X_i^t - x_i^\star\right)X_i^t.$$

As in the previous step, rearranging this inequality shows that $X_i^{(3)} \ge R_t x_i^\star$ provided that $\eta \le 0.39 \le \frac{R_t}{3.2((\mathbf{X}^t)^\top\mathbf{x}^\star)X_i^t}$. We will show

$$3\|\mathbf{X}^{(3)}\|_2^2 - 1 \le \sqrt{3}\cdot 2\Big((\mathbf{X}^{(3)})^\top\mathbf{x}^\star\Big) - \Delta,$$

where $\Delta > 0$ is a buffer term due to very large coordinates decreasing. To this end, define $\mathcal{S}_3' = \{i \in \mathcal{S}_3 : |X_i^t| \ge \frac{2}{\sqrt{3}+0.15}|x_i^\star|\}$. Then, since $R_t \le 1/\sqrt{3}+0.05$, we have

$$\begin{aligned}
&\frac{1}{1/\sqrt{3}+0.05}\cdot 2\Big((\mathbf{X}^t)^\top\mathbf{x}^\star\Big) \le 3\|\mathbf{X}^t\|_2^2 - 1\\
\Rightarrow\ & 3\sum_{i\in\mathcal{S}_3'}(X_i^t)^2 - \frac{2}{1/\sqrt{3}+0.05}X_i^t x_i^\star \ge 1 + \sum_{i\notin\mathcal{S}_3'}\frac{2}{1/\sqrt{3}+0.05}X_i^t x_i^\star - 3(X_i^t)^2 \ge 1 - 3\delta^2,
\end{aligned}$$

since $\frac{2}{1/\sqrt{3}+0.05}X_i^t x_i^\star - 3(X_i^t)^2 \ge 0$ if $i \notin \mathcal{S}_3' \cup \mathcal{S}^c$, and $\|\mathbf{X}_{\mathcal{S}^c}^t\|_2^2 \le \delta^2$. This implies

$$\sum_{i\in\mathcal{S}_3'}(X_i^t)^2 \ge \frac{1}{3}.$$

Next, we bound by how much coordinates in $\mathcal{S}_3'$ must at least decrease. Let $i \in \mathcal{S}_3'$ with $x_i^\star > 0$. We have

$$\begin{aligned}
\nabla f(\mathbf{X}^t)_i &= 2\Big((\mathbf{X}^t)^\top\mathbf{x}^\star\Big)\left(\frac{1}{R_t}X_i^t - x_i^\star\right)\\
&\ge 2\Big((\mathbf{X}^t)^\top\mathbf{x}^\star\Big)\left(\frac{2}{(\sqrt{3}+0.15)(\frac{1}{\sqrt{3}}+0.05)} - 1\right)x_i^\star\\
&\ge 2\Big((\mathbf{X}^t)^\top\mathbf{x}^\star\Big)\frac{2c_\star}{3\sqrt{k}}.
\end{aligned}$$

With this, we can bound, using (3.44),

$$\widetilde{X}_i^{t+1} - X_i^t \le -\frac{1}{2}\eta\nabla f(\mathbf{X}^t)_i X_i^t \le -\eta\cdot 2\Big((\mathbf{X}^t)^\top\mathbf{x}^\star\Big)\frac{c_\star}{3\sqrt{k}}X_i^t.$$

Following the same steps as in Case 1, we can show that $\widetilde{X}_i^{t+1}, X_i^t \geq R_t x_i^\star \geq x_i^\star/\sqrt{3}$ provided that $\eta \leq \frac{1}{6.4\sqrt{3}}$, where we used that $R_t \geq 1/\sqrt{3}$. With this, we can bound

$$\begin{aligned}
3\|\mathbf{X}^{(3)}\|_2^2 - 1 &= 3\|\mathbf{X}^{(2)}\|_2^2 - 1 + \sum_{i\in\mathcal{S}_3\setminus\mathcal{S}_3'} 3(\widetilde{X}_i^{t+1} + X_i^t)(\widetilde{X}_i^{t+1} - X_i^t) \\
&\quad + \sum_{i\in\mathcal{S}_3'} 3(\widetilde{X}_i^{t+1} + X_i^t)(\widetilde{X}_i^{t+1} - X_i^t) \\
&\leq \sqrt{3}\Big(2\Big((\mathbf{X}^{(2)})^\top \mathbf{x}^\star\Big) + \sum_{i\in\mathcal{S}_3\setminus\mathcal{S}_3'} 2x_i^\star(\widetilde{X}_i^{t+1} - X_i^t)\Big) \\
&\quad + \sum_{i\in\mathcal{S}_3'} 3(R_t+1)X_i^t(\widetilde{X}_i^{t+1} - X_i^t) \\
&\leq \sqrt{3}\cdot 2\Big((\mathbf{X}^{(3)})^\top \mathbf{x}^\star\Big) + \sum_{i\in\mathcal{S}_3'}\Big(\sqrt{3} + 3 - \Big(\sqrt{3} + 0.15\Big)\Big)X_i^t\Big(\widetilde{X}_i^{t+1} - X_i^t\Big),
\end{aligned}$$

where in the last line we used that $2|x_i^\star| \leq (\sqrt{3}+0.15)|X_i^t|$ for $i \in \mathcal{S}_3'$. Finally, we have

$$\begin{aligned}
\sum_{i\in\mathcal{S}_3'}(3-0.15)X_i^t(\widetilde{X}_i^{t+1} - X_i^t) &\leq -\sum_{i\in\mathcal{S}_3'} 2.85X_i^t\eta\cdot 2\Big((\mathbf{X}^t)^\top\mathbf{x}^\star\Big)\frac{c_\star}{3\sqrt{k}}X_i^t \\
&\leq -2.85\eta\cdot 2\Big((\mathbf{X}^t)^\top\mathbf{x}^\star\Big)\frac{c_\star}{3\sqrt{k}}\cdot\frac{1}{3} \\
&=: -\Delta,
\end{aligned}$$

that is

$$3\|\mathbf{X}^{(3)}\|_2^2 - 1 \leq \sqrt{3}\cdot 2\Big((\mathbf{X}^{(3)})^\top\mathbf{x}^\star\Big) - \Delta. \tag{3.71}$$

Exchanging large coordinates with empirical gradient

Let $i \in \mathcal{S}_2 \cup \mathcal{S}_3$ with $x_i^\star > 0$. Analogous bounds for the case $x_i^\star < 0$ can be established the same way.

In the following, we write $a = \frac{1}{2}(X_i^t + \sqrt{(X_i^t)^2+\beta^2})$, $b = \frac{1}{2}(-X_i^t + \sqrt{(X_i^t)^2+\beta^2})$, $G = \nabla F(\mathbf{X}^t)_i$ and $g = \nabla f(\mathbf{X}^t)_i$ for notational brevity. We can bound the ratio

$$\begin{aligned}
\frac{X_i^{t+1}}{\widetilde{X}_i^{t+1}} &= \frac{ae^{-\eta G} - be^{\eta G}}{ae^{-\eta g} - be^{\eta g}} \\
&\leq \frac{ae^{-\eta G}}{ae^{-\eta g} - be^{\eta g}} \\
&= \frac{ae^{-\eta G}}{ae^{-\eta g}} + \frac{ae^{-\eta G}be^{\eta g}}{ae^{-\eta g}(ae^{-\eta g} - be^{\eta g})} \\
&\leq \exp\left(\eta\frac{0.01c_\star((\mathbf{X}^t)^\top\mathbf{x}^\star)}{\sqrt{k}}\right) + \frac{3\sqrt{k}}{c_\star}\beta \\
&\leq 1 + 0.02c_\star\eta\frac{((\mathbf{X}^t)^\top\mathbf{x}^\star)}{\sqrt{k}},
\end{aligned}$$

where for the penultimate line we used the inequality $-x + \sqrt{x^2 + \beta^2} \le \beta$ for $x > 0$, the fact that $|X_i^t| \ge |x_i^\star|/3 \ge c_\star/(3\sqrt{k})$ for $i \in \mathcal{S}_2 \cup \mathcal{S}_3$, and that, as in (3.41), we can use Lemma 3.6 to bound, with probability $1 - c_p n^{-10}$,

$$|\nabla F(\mathbf{X}^t)_i - \nabla f(\mathbf{X}^t)_i| \le \frac{0.01 c_\star ((\mathbf{X}^t)^\top \mathbf{x}^\star)}{\sqrt{k}}.$$

Similarly, we can also bound $X_i^{t+1}/\widetilde{X}_i^{t+1} \ge 1 - 0.02 c_\star \eta \frac{((\mathbf{X}^t)^\top \mathbf{x}^\star)}{\sqrt{k}}$. With this, we have

$$\|\mathbf{X}^{(4)}\|_2^2 \le \left(1 + 0.02 c_\star \eta \frac{((\mathbf{X}^t)^\top \mathbf{x}^\star)}{\sqrt{k}}\right)^2 \|\mathbf{X}^{(3)}\|_2^2 \le \left(1 + 2.1 \cdot 0.02 c_\star \eta \frac{((\mathbf{X}^t)^\top \mathbf{x}^\star)}{\sqrt{k}}\right) \|\mathbf{X}^{(3)}\|_2^2,$$

provided that $\eta \le 5/(\sqrt{2} c_\star)$, and similarly

$$\Big((\mathbf{X}^{(4)})^\top \mathbf{x}^\star\Big) \ge \left(1 - 0.02 c_\star \eta \frac{((\mathbf{X}^t)^\top \mathbf{x}^\star)}{\sqrt{k}}\right) \Big((\mathbf{X}^{(3)})^\top \mathbf{x}^\star\Big).$$

Recalling the definition of Δ and using (3.71), we can now bound

$$\begin{aligned}
3\|\mathbf{X}^{(4)}\|_2^2 - 1 &\le 3\|\mathbf{X}^{(3)}\|_2^2 - 1 + 3\|\mathbf{X}^{(3)}\|_2^2 \cdot 2.1 \cdot 0.02 c_\star \eta \frac{((\mathbf{X}^t)^\top \mathbf{x}^\star)}{\sqrt{k}} \\
&\le 3\|\mathbf{X}^{(3)}\|_2^2 - 1 + \frac{\Delta}{2} \\
&\le \sqrt{3} \cdot 2\Big((\mathbf{X}^{(3)})^\top \mathbf{x}^\star\Big) - \frac{\Delta}{2} \\
&\le \sqrt{3} \cdot 2\Big((\mathbf{X}^{(4)})^\top \mathbf{x}^\star\Big) + \sqrt{3} \cdot 2\Big((\mathbf{X}^{(3)})^\top \mathbf{x}^\star\Big) \cdot 0.02 c_\star \eta \frac{((\mathbf{X}^t)^\top \mathbf{x}^\star)}{\sqrt{k}} - \frac{\Delta}{2} \\
&\le \sqrt{3} \cdot 2\Big((\mathbf{X}^{(4)})^\top \mathbf{x}^\star\Big) - \frac{\Delta}{4},
\end{aligned}$$

where the last inequality holds because we can show $\|\mathbf{X}^{(3)}\|_2^2 \le 2$ the same way as in the proof of the upper bound of (3.25).

Exchanging off-support coordinates

We have $\|\mathbf{X}^{(5)}\|_2^2 - \|\mathbf{X}^{(4)}\| \le n\delta^2 \le \Delta/4$ for $\delta \le c_1/n$ small enough, so

$$3\|\mathbf{X}^{(5)}\|_2^2 - 1 \le \sqrt{3} \cdot 2\Big((\mathbf{X}^{(5)})^\top \mathbf{x}^\star\Big).$$

Exchanging small coordinates

Following the same steps for small coordinates in the previous case, we can show that also

$$3\|\mathbf{X}^{(6)}\|_2^2 - 1 \le \sqrt{3} \cdot 2\Big((\mathbf{X}^{(6)})^\top \mathbf{x}^\star\Big),$$

which completes the proof that $R_{t+1} \ge 1/\sqrt{3}$ if $1/\sqrt{3} \le R_t < 1/\sqrt{3} + 0.05$. □

3.8.3 Technical lemmas

In this section, we collect technical lemmas and concentration bounds used in the proofs of Theorems 3.3 and 3.4 and the supporting Lemmas 3.6–3.8.

Theorem 3.9 (Proposition 34 [115]). *Let $g : \mathbb{R}^n \to \mathbb{R}$ be a Lipschitz continuous function with Lipschitz constant λ, i.e. $|g(\mathbf{x}) - g(\mathbf{y})| \leq \lambda \|\mathbf{x} - \mathbf{y}\|_2$ for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Let $\mathbf{A} \in \mathbb{R}^n$ be a standard normal random vector. Then, for any $\epsilon > 0$, we have*

$$\mathbb{P}\left[|g(\mathbf{A}) - \mathbb{E}[g(\mathbf{A})]| \geq \epsilon\right] \leq 2\exp\left(-\frac{\epsilon^2}{2\lambda^2}\right)$$

Theorem 3.10 (Theorems 3.6, 3.7 [34]). *Let X_i be independent random variables satisfying $|X_i| \leq \lambda$ for all $i \in [n]$. Let $X = \sum_{i=1}^n X_i$ and $\|X\| = \sqrt{\sum_{i=1}^n \mathbb{E}[X_i^2]}$. Then, for any $\epsilon > 0$, we have*

$$\mathbb{P}\left[|X - \mathbb{E}[X]| \geq \epsilon\right] \leq \exp\left(-\frac{\epsilon^2}{2(\|X\|^2 + \lambda\epsilon/3)}\right)$$

We state the following Lemma from [20] without proof. While the first and last inequality were not shown in [20], it can be done the same way as in the proof of Lemma A.5 in [20]. Convexity follows from the convexity of the operator norm.

Lemma 3.11 (Lemma A.5 [20]). *Let $\{\mathbf{A}_j\}_{j=1}^m$ be a collection of i.i.d. $\mathcal{N}(0, \mathbf{I}_k)$ random vectors. For any $t > 0$, let $\mathcal{A} \subseteq \mathbb{R}^{m \times k}$ be the set consisting of all $\{\mathbf{a}_j\}_{j=1}^m \in \mathbb{R}^{m \times k}$ satisfying the following:*

$$\|\mathbf{a}\|_{2\to 2} \leq \sqrt{m} + \sqrt{k} + t, \tag{3.72}$$

$$\|\mathbf{a}\|_{2\to 4} \leq (3m)^{\frac{1}{4}} + \sqrt{k} + t, \tag{3.73}$$

$$\|\mathbf{a}\|_{2\to 6} \leq (15m)^{\frac{1}{6}} + \sqrt{k} + t, \tag{3.74}$$

$$\|\mathbf{a}\|_{2\to 8} \leq (105m)^{\frac{1}{8}} + \sqrt{k} + t, \tag{3.75}$$

where we write

$$\|\mathbf{a}\|_{2\to p} = \sup_{\|x\|_2 \leq 1} \|\mathbf{a}\mathbf{x}\|_p.$$

Then, we have $P[\{\mathbf{A}_j\}_{j=1}^m \in \mathcal{A}] \geq 1 - 4\exp(-t^2/2)$. Further, the set $\mathcal{A}$ is convex.

The following lemma is a slight modification of the previous result.

Lemma 3.12. *Let $\{\mathbf{A}_j\}_{j=1}^m$ be a collection of i.i.d. $\mathcal{N}(0, \mathbf{I}_k)$ random vectors. For any $t > 0$, let $\mathcal{A} \subseteq \mathbb{R}^{m\times k}$ be the set consisting of all $\{\mathbf{a}_j\}_{j=1}^m \in \mathbb{R}^{m\times k}$ satisfying the following: For all $\mathbf{x} \in \mathbb{R}^k$ with $\|\mathbf{x}\|_2 = 1$,*

$$\left(\sum_{j=1}^m (\mathbf{a}_j^\top \mathbf{x})^4\right)^{\frac{1}{4}} \leq (3m)^{\frac{1}{4}} + \sqrt{8\log(2k)}\|\mathbf{x}\|_1 + t, \tag{3.76}$$

$$\left(\sum_{j=1}^m (\mathbf{a}_j^\top \mathbf{x})^6\right)^{\frac{1}{6}} \leq (15m)^{\frac{1}{6}} + \sqrt{8\log(2k)}\|\mathbf{x}\|_1 + t, \tag{3.77}$$

$$\left(\sum_{j=1}^m (\mathbf{a}_j^\top \mathbf{x})^8\right)^{\frac{1}{8}} \leq (105m)^{\frac{1}{8}} + \sqrt{8\log(2k)}\|\mathbf{x}\|_1 + t. \tag{3.78}$$

Then, we have $\mathbb{P}[\{\mathbf{A}_j\}_{j=1}^m \in \mathcal{A}] \geq 1 - 3\lceil\log_2\sqrt{k}\rceil \exp(-t^2/2)$. Further, the set $\mathcal{A}$ is convex.

Lemma 3.13. *Let $\mathbf{x}^\star \in \mathbb{R}^n$ be a k-sparse vector with $\|\mathbf{x}^\star\|_2 = 1$. Let $\{\mathbf{A}_j\}_{j=1}^m$ be a collection of i.i.d. $\mathcal{N}(0, \mathbf{I}_n)$ random vectors and $\{\varepsilon_j\}_{j=1}^m$ a collection of independent centered sub-exponential random variables with maximum sub-exponential norm $\sigma = \max_j \|\varepsilon_j\|_{\psi_1}$. There exist universal constants $c, c_s, c_p > 0$ such that if $m \geq c_s \max\{k^2\log^2 n,\ \log^5 n\}$, then, with probability at least $1 - c_p n^{-13}$,*

$$\left|\frac{1}{m}\sum_{j=1}^m A_{ji}^2 A_{js}A_{jl} - \mathbb{E}\Big[A_{1i}^2 A_{1s}A_{1l}\Big]\right| \leq c\sqrt{\frac{\log n}{m}} \quad \forall i, l, s \in [n], \tag{3.79}$$

$$\left|\frac{1}{m}\sum_{j=1}^m A_{ji}^2 A_{jl}(\mathbf{A}_{j,-i}^\top \mathbf{x}_{-i}^\star) - \mathbb{E}\Big[A_{1i}^2 A_{1l}(\mathbf{A}_{1,-i}^\top \mathbf{x}_{-i}^\star)\Big]\right| \leq c\sqrt{\frac{\log n}{m}} \quad \forall i, l \in [n], \tag{3.80}$$

$$\left|\frac{1}{m}\sum_{j=1}^m A_{ji}A_{jl}(\mathbf{A}_{j,-i}^\top \mathbf{x}_{-i}^\star)^2 - \mathbb{E}\Big[A_{1i}A_{1l}(\mathbf{A}_{1,-i}^\top \mathbf{x}_{-i}^\star)^2\Big]\right| \leq c\sqrt{\frac{\log n}{m}} \quad \forall i, l \in [n], \tag{3.81}$$

$$\left|\frac{1}{m}\sum_{j=1}^m A_{ji}^8 - 105\right| \leq c\sqrt{\frac{\log^5 n}{m}} \quad \forall i \in [n], \tag{3.82}$$

$$\left|\frac{1}{m}\sum_{j=1}^m \varepsilon_j A_{ji}A_{jl}\right| \leq c\sigma\sqrt{\frac{\log n}{m}} \quad \forall i, l \in [n], \tag{3.83}$$

$$\left|\frac{1}{m}\sum_{j=1}^m \varepsilon_j A_{jl}(\mathbf{A}_j^\top \mathbf{x}^\star)\right| \leq c\sigma\sqrt{\frac{\log n}{m}} \quad \forall l \in [n]. \tag{3.84}$$

Lemma 3.14. *Let $\mathbf{x}^\star \in \mathbb{R}^n$ be a k-sparse vector with $\|\mathbf{x}^\star\|_2 = 1$, and let $\{\mathbf{A}_j\}_{j=1}^m$ be a collection of i.i.d. $\mathcal{N}(0, \mathbf{I}_n)$ random vectors. There exist universal constants $c, c_s, c_p > 0$*

such that the following holds. Let $\mathcal{A} \subseteq \mathbb{R}^{m\times n}$ *be the set consisting of all* $\{\mathbf{a}_j\}_{j=1}^m \in \mathbb{R}^{m\times n}$ *satisfying the following:*

$$\left|\frac{1}{m}\sum_{j=1}^m (\mathbf{a}_j^\top \mathbf{x}^\star)^4 - 3\right| \le c\sqrt{\frac{\log n}{m}}$$
$$\left|\frac{1}{m}\sum_{j=1}^m (\mathbf{a}_j^\top \mathbf{x}^\star)^6 - 15\right| \le c\sqrt{\frac{\log^3 n}{m}}$$
$$\left|\frac{1}{m}\sum_{j=1}^m (\mathbf{a}_j^\top \mathbf{x}^\star)^8 - 105\right| \le c\sqrt{\frac{\log^5 n}{m}}.$$

Then, if $m \ge c_s \log^5 n$, *we have* $\mathbb{P}[\{\mathbf{A}_j\}_{j=1}^m \in \mathcal{A}] \ge 1 - c_p n^{-12}$. *Further, the set* $\mathcal{A}$ *is convex.*

Lemma 3.15. *Let* $\{\mathbf{A}_j\}_{j=1}^m$ *be a collection of i.i.d.* $\mathcal{N}(0, \mathbf{I}_k)$ *random vectors. There exist universal constants* $c, c_s, c_p > 0$ *such that if* $m \ge c_s \max\{k^2 \log^2 n,\ \log^5 n\}$, *where* $n \ge k$ *is any natural number, then the set* $\mathcal{A} \subseteq \mathbb{R}^{m\times k}$ *defined by*

$$\mathcal{A} = \left\{\{\mathbf{a}_j\}_{j=1}^m \in \mathbb{R}^{m\times k} : \frac{1}{m}\sum_{j=1}^m \left(\sum_{l\in\mathcal{L}} a_{jl}\right)^4 \le c|\mathcal{L}|^2 \text{ for all } \mathcal{L} \subseteq [k]\right\}$$

satisfies $\mathbb{P}[\{\mathbf{A}_j\}_{j=1}^m \in \mathcal{A}] \ge 1 - c_p n^{-11}$. *Further, the set* $\mathcal{A}$ *is convex.*

Lemma 3.16. *Let* $\mathbf{x}^\star \in \mathbb{R}^n$ *be a* k*-sparse vector, and let* $\mathcal{S} = \{i \in [n] : x_i^\star \ne 0\}$ *be its support. Let* $\{\mathbf{A}_j\}_{j=1}^m$ *be a collection of i.i.d.* $\mathcal{N}(0, \mathbf{I}_n)$ *random variables and* $\{\varepsilon_j\}_{j=1}^m$ *a collection of independent centered sub-exponential random variables with maximum sub-exponential norm* $\sigma = \max_j \|\varepsilon_j\|_{\psi_1}$. *There exist universal constants* $c_s, c_p > 0$ *such that, for any constant* $c_1 > 0$, *there is a* $c > 0$ *such that if* $m \ge c_s \max\{k^2 \log n,\ \log^5 n\}$, *then, with probability at least* $1 - c_p n^{-10}$, *the following holds: For any* $i \in [n]$, $\mathcal{K} \subseteq \mathcal{S}\backslash\{i\}$ *and* $\epsilon \ge c_1/m$, *let* $N_\epsilon^{\mathcal{K}}$ *be a smallest* ϵ*-net of* $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}_{\mathcal{K}^c} = \mathbf{0}, \|\mathbf{x}\|_2 = 1\}$. *Then, for any* $\mathcal{L} \subseteq \mathcal{K}$ *with* $|\mathcal{L}| \ge \frac{1}{2}|\mathcal{K}|$, *and* $\mathbf{x} \in N_\epsilon^{\mathcal{K}}$,

$$\left|\sum_{l\in\mathcal{L}} \frac{1}{m}\sum_{j=1}^m A_{ji}A_{jl}(\mathbf{A}_j^\top \mathbf{x})^2\right| \le c|\mathcal{L}|\frac{\log n}{\sqrt{m}},, \tag{3.85}$$
$$\left|\sum_{l\in\mathcal{L}} \frac{1}{m}\sum_{j=1}^m A_{ji}^2 A_{jl}(\mathbf{A}_{j,-l}^\top \mathbf{x}_{-l})\right| \le c|\mathcal{L}|\sqrt{\frac{\log n}{m}}, \tag{3.86}$$
$$\left|\sum_{l\in\mathcal{L}} \frac{1}{m}\sum_{j=1}^m \varepsilon_j A_{jl}(\mathbf{A}_{j,-l}^\top \mathbf{x}_{-l})\right| \le c\sigma|\mathcal{L}|\sqrt{\frac{\log n}{m}}, \tag{3.87}$$
$$\left|\sum_{l\in\mathcal{L}} \frac{1}{m}\sum_{j=1}^m A_{ji}A_{jl}(\mathbf{A}_j^\top \mathbf{x})(\mathbf{A}_{j,-i}^\top \mathbf{x}_{-i}^\star)\right| \le c|\mathcal{L}|\frac{\log n}{\sqrt{m}}. \tag{3.88}$$

Lemma 3.17. *Let $\mathbf{x}^\star \in \mathbb{R}^n$ be a k-sparse vector, and let $\mathcal{S} = \{i \in [n] : x_i^\star \neq 0\}$ be its support. Let $\{\mathbf{A}_j\}_{j=1}^m$ be a collection of i.i.d. $\mathcal{N}(0, \mathbf{I}_n)$ random vectors and $\{\varepsilon_j\}_{j=1}^m$ a collection of independent centered sub-exponential random variables with maximum sub-exponential norm $\sigma = \max_j \|\varepsilon_j\|_{\psi_1}$. Let $\mathcal{X} = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}_{\mathcal{S}^c} = \mathbf{0}, \|\mathbf{x}\|_2 \leq 1\}$. There exist universal constants $c, c_s, c_p > 0$ such that if $m \geq c_s \max\{k^2 \log^2 n, \log^5 n\}$, then the following holds with probability at least $1 - c_p n^{10}$: For all $i \in [n]$ and $\mathbf{x} \in \mathcal{X}$,*

$$\left|\frac{1}{m}\sum_{j=1}^m A_{ji}(\mathbf{A}_{j,-i}^\top \mathbf{x}_{-i})^3\right| \leq c\|\mathbf{x}\|_1 \frac{\log n}{\sqrt{m}}, \tag{3.89}$$

$$\left|\sum_{l\neq i} x_l \sum_{r\neq l,i} x_r \frac{1}{m}\sum_{j=1}^m A_{ji}^2 A_{jl} A_{jr}\right| \leq c\|\mathbf{x}\|_1 \sqrt{\frac{\log n}{m}}, \tag{3.90}$$

$$\left|\sum_{l\neq i} x_l \sum_{r\neq l,i} x_r \frac{1}{m}\sum_{j=1}^m \varepsilon_j A_{jl} A_{jr}\right| \leq c\sigma\|\mathbf{x}\|_1 \sqrt{\frac{\log n}{m}}, \tag{3.91}$$

$$\left|\frac{1}{m}\sum_{j=1}^m A_{ji}(\mathbf{A}_{j,-i}^\top \mathbf{x}_{-i})^2(\mathbf{A}_{j,-i}^\top \mathbf{x}_{-i}^\star)\right| \leq c\|\mathbf{x}\|_1 \frac{\log n}{\sqrt{m}}, . \tag{3.92}$$

In the following, we prove Lemmas 3.12–3.17.

Proof of Lemma 3.12. The proof follows that of Lemma A.5 of [20] closely. Since all three inequalities can be shown the same way, we only present the proof for (3.78). For any fixed $\lambda \geq 1$, define

$$\|\mathbf{a}\|_{2\to 8,\lambda} = \max\{\|\mathbf{a}\mathbf{x}\|_8 : \|\mathbf{x}\|_2 = 1, \|\mathbf{x}\|_1 \leq \lambda\}.$$

We will show that $\|\mathbf{A}\|_{2\to 8,\lambda} \leq (105m)^{\frac{1}{8}} + \sqrt{2\log(2k)}\lambda + t$ holds with probability at least $1 - 2\exp(-t^2/2)$. The result then follows by taking the union bound over $\lambda = 2^1, 2^2, \ldots, 2^{\lceil \log_2 \sqrt{k}\rceil}$ for each of the three inequalities (3.76)–(3.78), since we have $1 \leq \|\mathbf{x}\|_1 \leq \sqrt{k}$ and the union bound is chosen such that there always exists a λ with $\|\mathbf{x}\|_1 \leq \lambda \leq 2\|\mathbf{x}\|_1$.

Define $X_{\mathbf{u},\mathbf{v}} = \langle \mathbf{A}\mathbf{u}, \mathbf{v}\rangle$ on the set

$$T_\lambda = \{(\mathbf{u}, \mathbf{v}) : \mathbf{u} \in \mathbb{R}^k, \|\mathbf{u}\|_2 = 1, \|\mathbf{u}\|_1 \leq \lambda, \mathbf{v} \in \mathbb{R}^m, \|\mathbf{v}\|_{8/7} \leq 1\}.$$

Then, by Hölder's inequality, we have $\|\mathbf{A}\|_{2\to 8,\lambda} = \max_{(\mathbf{u},\mathbf{v})\in T_\lambda} X_{\mathbf{u},\mathbf{v}}$.

Define $Y_{\mathbf{u},\mathbf{v}} = \langle \mathbf{G}, \mathbf{u}\rangle + \langle \mathbf{H}, \mathbf{v}\rangle$, where $\mathbf{G} \in \mathbb{R}^k$ and $\mathbf{H} \in \mathbb{R}^m$ are independent standard normal random vectors.

We have, for any $(\mathbf{u}, \mathbf{v}), (\mathbf{u}', \mathbf{v}') \in T_\lambda$,

$$\mathbb{E}[|X_{\mathbf{u},\mathbf{v}} - X_{\mathbf{u}',\mathbf{v}'}|^2] = \|\mathbf{v}\|_2^2 + \|\mathbf{v}'\|_2^2 - 2\langle \mathbf{u}, \mathbf{u}' \rangle \langle \mathbf{v}, \mathbf{v}' \rangle,$$

and

$$\mathbb{E}[|Y_{\mathbf{u},\mathbf{v}} - Y_{\mathbf{u}',\mathbf{v}'}|^2] = 2 + \|\mathbf{v}\|_2^2 + \|\mathbf{v}'\|_2^2 - 2\langle \mathbf{u}, \mathbf{u}' \rangle - 2\langle \mathbf{v}, \mathbf{v}' \rangle,$$

where we used $\|\mathbf{u}\|_2 = \|\mathbf{u}'\|_2 = 1$. Therefore, we have

$$\mathbb{E}[|Y_{\mathbf{u},\mathbf{v}} - Y_{\mathbf{u}',\mathbf{v}'}|^2] - \mathbb{E}[|X_{\mathbf{u},\mathbf{v}} - X_{\mathbf{u}',\mathbf{v}'}|^2] = 2(1 - \langle \mathbf{u}, \mathbf{u}' \rangle)(1 - \langle \mathbf{v}, \mathbf{v}' \rangle) \geq 0,$$

where we used $\|\mathbf{v}\|_2 \leq \|\mathbf{v}\|_{8/7} \leq 1$ and $\|\mathbf{v}'\|_2 \leq \|\mathbf{v}'\|_{8/7} \leq 1$. By Proposition 33 of [115], this implies

$$\mathbb{E}\left[\max_{(\mathbf{u},\mathbf{v}) \in T_\lambda} X_{\mathbf{u},\mathbf{v}}\right] \leq \mathbb{E}\left[\max_{(\mathbf{u},\mathbf{v}) \in T_\lambda} Y_{\mathbf{u},\mathbf{v}}\right],$$

and hence

$$\begin{aligned}
\mathbb{E}[\|\mathbf{A}\|_{2\to 8,\lambda}] &\leq \mathbb{E}\left[\max_{(\mathbf{u},\mathbf{v}) \in T_\lambda} Y_{\mathbf{u},\mathbf{v}}\right] \\
&\leq \mathbb{E}\left[\max_{(\mathbf{u},\mathbf{v}) \in T_\lambda} \|\mathbf{G}\|_\infty \|\mathbf{u}\|_1 + \|\mathbf{H}\|_8 \|\mathbf{v}\|_{8/7}\right] \\
&\leq \sqrt{2\log(2k)}\lambda + (105m)^{\frac{1}{8}},
\end{aligned}$$

where we used Hölder's inequality in the second line, and the bound

$$\exp(t\mathbb{E}[\|\mathbf{G}\|_\infty]) \leq \mathbb{E}[\exp(t\|\mathbf{G}\|_\infty)] \leq \sum_{i=1}^{k} \mathbb{E}[\exp(t|G_i|)] \leq 2k\exp(t^2/2)$$

with $t = \sqrt{2\log(2k)}$ for the last inequality.

Finally, $\|\cdot\|_{2\to 8,\lambda}$ is a 1-Lipschitz function: let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{m\times k}$ and, without loss of generality, $\|\mathbf{a}\|_{2\to 8,\lambda} \geq \|\mathbf{b}\|_{2\to 8,\lambda}$. Then,

$$\begin{aligned}
\|\mathbf{a}\|_{2\to 8,\lambda} - \|\mathbf{b}\|_{2\to 8,\lambda} &= \max_{\|\mathbf{x}\|_2=1, \|\mathbf{x}\|_1 \leq \lambda} \|\mathbf{a}\mathbf{x}\|_8 - \max_{\|\mathbf{y}\|_2=1, \|\mathbf{y}\|_1 \leq \lambda} \|\mathbf{b}\mathbf{y}\|_8 \\
&\leq \max_{\|\mathbf{x}\|_2=1, \|\mathbf{x}\|_1 \leq \lambda} \|\mathbf{a}\mathbf{x}\|_8 - \|\mathbf{b}\mathbf{x}\|_8 \\
&\leq \max_{\|\mathbf{x}\|_2=1, \|\mathbf{x}\|_1 \leq \lambda} \|(\mathbf{a}-\mathbf{b})\mathbf{x}\|_8 \\
&\leq \max_{\|\mathbf{x}\|_2=1, \|\mathbf{x}\|_1 \leq \lambda} \|(\mathbf{a}-\mathbf{b})\mathbf{x}\|_2 \\
&\leq \|\mathbf{a}-\mathbf{b}\|_F,
\end{aligned}$$

where we write $\|\cdot\|_F$ for the Frobenius norm and used the fact that it is an upper bound to the ℓ_2-operator norm. Hence, an application of Theorem 3.9 yields

$$\mathbb{P}\Big[\|\mathbf{A}\|_{2\to 8,\lambda} < \sqrt{2\log(2k)}\lambda + (105m)^{\frac{1}{8}} + t\Big] \geq 1 - \exp(-t^2/2).$$

Taking the union bound over $\lambda = 2^1, \dots, 2^{\lceil \log_2 \sqrt{k}\rceil}$ completes the proof that (3.78) holds with probability $1 - \lceil \log_2 \sqrt{k}\rceil \exp(-t^2/2)$, since the union bound is chosen such that for any $\mathbf{x} \in \mathbb{R}^k$ with $\|\mathbf{x}\|_2 = 1$ there is a λ satisfying $\|\mathbf{x}\|_1 \leq \lambda \leq 2\|\mathbf{x}\|_1$. The inequalities (3.76) and (3.77) can be treated the same way, and another union bound gives the desired result.

Finally, convexity of $\mathcal{A}$ follows from an application of the Minkowski inequality. Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^{m\times k}$ satisfy (3.78), and $\alpha \in (0,1)$. Then, for any $\mathbf{x} \in \mathbb{R}$ with $\|\mathbf{x}\|_2 = 1$,

$$\begin{aligned}\left(\sum_{j=1}^m \left(\alpha \mathbf{a}^\top \mathbf{x} + (1-\alpha)\mathbf{b}^\top \mathbf{x}\right)^8\right)^{\frac{1}{8}} &\leq \alpha \left(\sum_{j=1}^m (\mathbf{a}^\top \mathbf{x})^8\right)^{\frac{1}{8}} + (1-\alpha)\left(\sum_{j=1}^m (\mathbf{b}^\top \mathbf{x})^8\right)^{\frac{1}{8}} \\ &\leq (105m)^{\frac{1}{8}} + \sqrt{8\log(2k)}\|\mathbf{x}\|_1 + t.\end{aligned}$$

Since (3.76) and (3.77) can be treated the same way, this implies $\alpha\mathbf{a} + (1-\alpha)\mathbf{b} \in \mathcal{A}$. □

Proof of Lemma 3.13. The proof of Lemma 3.13 relies on Theorem 3.10 and the following truncation, which allows us to consider bounded random variables. We begin by showing the inequality (3.79).

Let $i, l, s \in [n]$. Writing $B_j = \{\max\{|A_{ji}|, |A_{jl}|, |A_{js}|\} \leq \sqrt{64\log n}\}$, we split the term $A_{ji}^2 A_{jl} A_{js} = A_{ji}^2 A_{jl} A_{js}\mathbf{1}(B_j) + A_{ji}^2 A_{jl} A_{js}(1 - \mathbf{1}(B_j)) = Y_j + Z_j$, where by $\mathbf{1}(\cdot)$ we denote the indicator function. We will show that each of the two terms concentrates around its respective mean.

Since $\frac{1}{m}|Y_j| \leq \frac{1}{m} 64^2 \log^2 n$ is bounded, we can apply Theorem 3.10. We can bound

$$\sqrt{\sum_{j=1}^m \frac{1}{m^2}\mathbb{E}[Y_j^2]} = \sqrt{\sum_{j=1}^m \frac{1}{m^2}\mathbb{E}\Big[A_{ji}^4 A_{jl}^2 A_{js}^2 \mathbf{1}(B_j)\Big]} \leq \sqrt{\frac{105}{m}}.$$

Then, since $m \geq c_s \log^5 n$, we have, for $c > 0$ sufficiently large,

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{j=1}^m Y_j - \mathbb{E}[Y_j]\right| > \frac{c}{2}\sqrt{\frac{\log n}{m}}\right] \leq 2\exp\left(-\frac{c^2 \log n/(4m)}{2(\frac{105}{m} + 64^2 c\frac{\log^{5/2} n}{6m^{3/2}})}\right) \leq 2n^{-16}.$$

For the second term Z_j, we can use the Cauchy-Schwarz inequality to bound

$$\begin{aligned}\operatorname{Var}\left(\frac{1}{m}\sum_{j=1}^{m} Z_j\right) &\leq \frac{1}{m}\mathbb{E}\Big[A_{1i}^4 A_{1l}^2 A_{1s}^2 \mathbf{1}(B_j^c)\Big] \\ &\leq \frac{1}{m}\sqrt{\mathbb{E}\Big[A_{1i}^8 A_{1l}^4 A_{1s}^4\Big]\mathbb{P}\Big[\max\{|A_{1i}|,|A_{1l}|,|A_{1s}|\} > \sqrt{64\log n}\Big]} \\ &\leq \frac{c'}{mn^{16}},\end{aligned}$$

for an absolute constant $c' > 0$, and hence, by the Chebyshev inequality,

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{j=1}^{m} Z_j - \mathbb{E}[Z_j]\right| > \frac{c}{2}\sqrt{\frac{\log n}{m}}\right] \leq \frac{\frac{c'}{mn^{16}}}{\frac{c^2\log n}{4m}} = \frac{4c'}{c^2\log n}n^{-16} \leq n^{-16},$$

for $c \geq \sqrt{\frac{4c'}{\log n}}$.

This completes the proof that

$$\mathbb{P}\left[\left|\frac{1}{m}\sum_{j=1}^{m} A_{ji}^2 A_{jl}A_{js} - \mathbb{E}\Big[A_{1i}^2 A_{1l}A_{1s}\Big]\right| > c\sqrt{\frac{\log n}{m}}\right] \leq 3n^{-16}.$$

Taking the union bound over all $i, l, s \in [n]$ shows that (3.79) holds with probability at least $1 - 3n^{-13}$.

The inequalities (3.80)–(3.84) can be shown following the same steps. To show (3.82), we need to control higher order terms, which is the reason for the additional logarithmic factor. The bounds (3.83) and (3.84) can be shown the same way, as, for sub-exponential random variables, we have the standard tail bound $\mathbb{P}[|\varepsilon| > t] \leq \exp(1 - \tilde{c}t/\sigma)$ for all $t \geq 0$, where $\tilde{c} > 0$ is an absolute constant, and the bound on the second moment $\mathbb{E}[\varepsilon^2] \leq 4\sigma^2$. As the proof of each bound follows exactly the same steps, we omit the details to avoid repetition. □

Proof of Lemma 3.14. Using the fact that $\mathbf{A}_j^\top \mathbf{x}^\star \sim \mathcal{N}(0, \|\mathbf{x}^\star\|_2^2)$ are independent random variables, this lemma can be shown following the same steps as in the proof of Lemma 3.13. The convexity of $\mathcal{A}$ follows, as in Lemma 3.12, from the Minkowski inequality. □

Proof of Lemma 3.15. We prove Lemma 3.15 via induction over the size $|\mathcal{L}|$. For any $1 \leq \lambda \leq k$, define

$$\mathcal{A}_\lambda = \left\{\{\mathbf{a}_j\}_{j=1}^m \in \mathbb{R}^{m\times k} : \frac{1}{m}\sum_{j=1}^{m}\left(\sum_{l\in\mathcal{L}} a_{jl}\right)^4 \leq c|\mathcal{L}|^2 \text{ for all } \mathcal{L} \subseteq [k] \text{ with } |\mathcal{L}| = \lambda\right\}.$$

We will show that $\mathbb{P}[\{\mathbf{A}_j\}_{j=1}^m \in \mathcal{A}_\lambda] \geq 1 - c_1\lambda n^{-13}$ for some constant $c_1 > 0$, from which the result $\mathbb{P}[\{\mathbf{A}_j\}_{j=1}^m \in \mathcal{A}] \geq 1 - c_p n^{-11}$ follows by taking the union bound over all possible sizes λ. As in the proof of Lemma 3.12, convexity of $\mathcal{A}_\lambda$ follows from the Minkowski inequality. As the intersection of convex sets, $\mathcal{A}$ is also convex.

The base case $\lambda = 1$ follows from the bound (3.80) of Lemma 3.13. For the induction step, assume that we have already shown $\mathbb{P}[\{\mathbf{A}_j\}_{j=1}^m \in \mathcal{A}_\lambda] \geq 1 - c_1\lambda n^{-13}$ for some $1 \leq \lambda < k$. Let $r \in [k]$ and $\mathcal{L} \subseteq [k]\backslash\{r\}$ be a subset with $|\mathcal{L}| = \lambda$. We have

$$\begin{aligned}\frac{1}{m}\sum_{j=1}^m \left(A_{jr} + \sum_{l\in\mathcal{L}} A_{jl}\right)^4 &= \frac{1}{m}\sum_{j=1}^m A_{jr}^4 + \frac{4}{m}\sum_{j=1}^m \left(\sum_{l\in\mathcal{L}} A_{jl}\right) A_{jr}^3 + \frac{6}{m}\sum_{j=1}^m \left(\sum_{l\in\mathcal{L}} A_{jl}\right)^2 A_{jr}^2 \\ &\quad + \frac{4}{m}\sum_{j=1}^m \left(\sum_{l\in\mathcal{L}} A_{jl}\right)^3 A_{jr} + \frac{1}{m}\sum_{j=1}^m \left(\sum_{l\in\mathcal{L}} A_{jl}\right)^4 \end{aligned} \tag{3.93}$$

We need to show that, with probability $1 - c_1(\lambda+1)n^{-13}$, the sum in (3.93) is bounded by $c(|\mathcal{L}|+1)^2$ for all $r \in [k]$ and $\mathcal{L} \subseteq [k]\backslash\{r\}$ with $|\mathcal{L}| = \lambda$.

By the induction hypothesis, we can bound the last term with probability $1 - c_1\lambda n^{-13}$,

$$\frac{1}{m}\sum_{j=1}^m \left(\sum_{l\in\mathcal{L}} A_{jl}\right)^4 \leq c|\mathcal{L}|^2.$$

The first three terms in (3.93) can be bounded using Lemma 3.13. The following holds with probability at least $1 - c_2 n^{-13}$, where c_2 is the universal constant from Lemma 3.13. Using the assumption $m \geq c_s k^2 \log^2 n$, we can bound the first term by

$$\frac{1}{m}\sum_{j=1}^m A_{jr}^4 \leq 4,$$

the second term by

$$\frac{4}{m}\sum_{j=1}^m \left(\sum_{l\in\mathcal{L}} A_{jl}\right) A_{jr}^3 = 4\sum_{l\in\mathcal{L}} \frac{1}{m}\sum_{j=1}^m A_{jl}A_{jr}^3 \leq 4\left(4 + \frac{|\mathcal{L}|-1}{k}\right) \leq 20,$$

and the third term in (3.93) can be bounded, using the Cauchy-Schwarz inequality, by

$$\begin{aligned}\frac{6}{m}\sum_{j=1}^m \left(\sum_{l\in\mathcal{L}} A_{jl}\right)^2 A_{jr}^2 &\leq 6\sqrt{\frac{1}{m}\sum_{j=1}^m \left(\sum_{l\in\mathcal{L}} A_{jl}\right)^4}\sqrt{\frac{1}{m}\sum_{j=1}^m A_{jr}^4} \\ &\leq 6\sqrt{c|\mathcal{L}|^2}\sqrt{4} \\ &\leq 0.4c|\mathcal{L}|,\end{aligned}$$

for $c \geq 900$, where we used the induction hypothesis in the second line.

To bound the fourth term in (3.93), fix any $r \in [k]$ and subset $\mathcal{L} \subseteq [k]\backslash\{r\}$ with cardinality $|\mathcal{L}| = \lambda$, and consider the function

$$h_{r,\mathcal{L}}(\mathbf{a}) = \frac{4}{m}\sum_{j=1}^{m} a_{jr}\left(\sum_{l\in\mathcal{L}} a_{jl}\right)^3.$$

Then, we need to bound the probability of the event

$$B_r = \{|h_{r,\mathcal{L}}(\mathbf{A})| \leq 1.6c|\mathcal{L}| \text{ for all } \mathcal{L} \subseteq [k]\backslash\{r\} \text{ with } |\mathcal{L}| = \lambda\}$$

for all $r \in [k]$. To this end, we condition on $\mathbf{A}_{\cdot r} = \mathbf{a}_{\cdot r}$ using the formula

$$\mathbb{P}[B] = \int \mathbb{P}[B|\mathbf{A}_{\cdot r} = \mathbf{a}_{\cdot r}]\mu(\mathbf{a}_{\cdot r})d\mathbf{a}_{\cdot r},$$

which holds for any event B, where we write μ for the standard normal density.

To show that the conditional probability $\mathbb{P}[B_r|\mathbf{A}_{\cdot r} = \mathbf{a}_{\cdot r}]$ is close to one, we begin by showing that, for any fixed $\mathcal{L}$, $h_{r,\mathcal{L}}(\mathbf{A})|_{\mathbf{A}_{\cdot r}=\mathbf{a}_{\cdot r}}$ (making explicit the fact that we are conditioning on $\mathbf{A}_{\cdot r} = \mathbf{a}_{\cdot r}$) concentrates around its expectation $\mathbb{E}[h_{r,\mathcal{L}}(\mathbf{A})|\mathbf{A}_{\cdot r} = \mathbf{a}_{\cdot r}] = 0$. For brevity's sake, we omit explicitly writing the condition $\mathbf{A}_{\cdot r} = \mathbf{a}_{\cdot r}$ in what follows.

Step 1: Bound the fourth term in (3.93) conditioned on r, $\mathbf{A}_{\cdot r}$ and $\mathcal{L}$

Fix an $r \in [k]$, a vector $\mathbf{a}_{\cdot r}$ satisfying $\max_j |a_{jr}| \leq 6\sqrt{\log n}$ and a subset $\mathcal{L} \subseteq [k]\backslash\{r\}$ with cardinality $|\mathcal{L}| = \lambda$. The idea is to apply concentration of Lipschitz functions of Gaussian random variables to show that the fourth term in (3.93), $h_{r,\mathcal{L}}(\mathbf{A})$, is close to its expectation $\mathbb{E}[h_{r,\mathcal{L}}(\mathbf{A})] = 0$. However, as $h_{r,\mathcal{L}}$ is not globally Lipschitz continuous, we cannot directly apply Theorem 3.9. Similar to the proof of Theorem 3 in [43], we will first restrict $h_{r,\mathcal{L}}$ to a high probability event on which $h_{r,\mathcal{L}}$ is Lipschitz continuous. Then, we extend this restricted function to a function $\tilde{h}_{r,\mathcal{L}}$ on the entire space in a way such that $\tilde{h}_{r,\mathcal{L}}$ is globally Lipschitz continuous, and apply Theorem 3.9 to $\tilde{h}_{r,\mathcal{L}}$. This also provides a high probability bound for $h_{r,\mathcal{L}}(\mathbf{A})$, since, by construction, $\tilde{h}_{r,\mathcal{L}}(\mathbf{A}) = h_{r,\mathcal{L}}(\mathbf{A})$ with high probability.

Step 1, part (a): Bound the Lipschitz constant of $h_{r,\mathcal{L}}$ restricted to $\mathcal{A}_\lambda$

Restricted to $\mathcal{A}_\lambda$ and conditioned on $\mathbf{A}_{\cdot r} = \mathbf{a}_{\cdot r}$, we can bound the Lipschitz constant of $h_{r,\mathcal{L}}$ by the norm of its gradient. The norm of the gradient is an upper bound for

the Lipschitz constant by the mean-value theorem, since $\mathcal{A}_\lambda$ is a convex set. For any $l \in \mathcal{L}$, we have

$$\frac{\partial}{\partial a_{jl}} h_{r,\mathcal{L}}(\mathbf{a}) = \frac{12}{m} a_{jr} \left(\sum_{l' \in \mathcal{L}} a_{jl'} \right)^2.$$

Hence, we can bound

$$\begin{aligned} \|\nabla h_{r,\mathcal{L}}(\mathbf{a})\|_2^2 &= \sum_{l \in \mathcal{L}} \sum_{j=1}^m \left(\frac{\partial}{\partial a_{jl}} h_{r,\mathcal{L}}(\mathbf{a}) \right)^2 \\ &\leq \frac{144 |\mathcal{L}| \max_j a_{jr}^2}{m} \frac{1}{m} \sum_{j=1}^m \left(\sum_{l \in \mathcal{L}} a_{jl} \right)^4 \\ &\leq \frac{5184 |\mathcal{L}| \log n}{m} c |\mathcal{L}|^2 \end{aligned}$$

on $\mathcal{A}_\lambda$, where we used the assumption $\max_j a_{jr}^2 \leq 36 \log n$ and the induction hypothesis.

Step 1, part (b)**: Construct a globally Lipschitz continuous extension of** $h_{r,\mathcal{L}}$ Consider the following Lipschitz extension of the function $h_{r,\mathcal{L}}$:

$$\tilde{h}_{r,\mathcal{L}}(\mathbf{a}) = \inf_{\mathbf{a}' \in \mathcal{A}_\lambda} \Big((h_{r,\mathcal{L}}(\mathbf{a}') + \mathrm{Lip}(h_{r,\mathcal{L}}) \|\mathbf{a} - \mathbf{a}'\|_2 \Big),$$

where we write $\mathrm{Lip}(h_{r,\mathcal{L}}) = \sqrt{\frac{5184 c |\mathcal{L}|^3 \log n}{m}}$. By definition, we have $\tilde{h}_{r,\mathcal{L}} = h_{r,\mathcal{L}}$ on $\mathcal{A}_\lambda$, and it follows from an application of the triangle inequality that $\tilde{h}_{r,\mathcal{L}}$ is globally Lipschitz continuous with Lipschitz constant $\mathrm{Lip}(h_{r,\mathcal{L}})$ (see e.g. Theorem 7.2 of [83]).

We will show that $\tilde{h}_{r,\mathcal{L}}$ concentrates around its mean, which can potentially differ from the mean of $h_{r,\mathcal{L}}$. Since $h_{r,\mathcal{L}}$ and $\tilde{h}_{r,\mathcal{L}}$ differ only on $\mathcal{A}_\lambda^c$, which has probability less than $c_1 \lambda n^{-13}$, we can bound, using the Cauchy-Schwarz inequality,

$$\mathbb{E}[|h_{r,\mathcal{L}}(\mathbf{A})| \mathbf{1}_{\mathcal{A}_\lambda^c}(\mathbf{A})] \leq \sqrt{\mathbb{E}[h_{r,\mathcal{L}}(\mathbf{A})^2]} \sqrt{\mathbb{E}[\mathbf{1}_{\mathcal{A}_\lambda^c}(\mathbf{A})]} \leq \frac{c'}{n^5},$$

where we used that $\mathbb{E}[h_{r,\mathcal{L}}(\mathbf{A})^2] = 15|\mathcal{L}|^3/m$. Using $\tilde{h}_{r,\mathcal{L}}(\mathbf{a}) \leq \mathrm{Lip}(h_{r,\mathcal{L}}) \|\mathbf{a}\|_2$, we have

$$\mathbb{E}[|\tilde{h}_{r,\mathcal{L}}(\mathbf{A})| \mathbf{1}_{\mathcal{A}_\lambda^c}(\mathbf{A})] \leq \mathrm{Lip}(h_{r,\mathcal{L}}) \sqrt{\mathbb{E}[\|\mathbf{A}\|_2^2]} \sqrt{\mathbb{E}[\mathbf{1}_{\mathcal{A}_\lambda^c}(\mathbf{A})]} \leq \frac{c''}{n^4}.$$

All in all, this shows that

$$\Big| \mathbb{E}[h_{r,\mathcal{L}}(\mathbf{A})] - \mathbb{E}[\tilde{h}_{r,\mathcal{L}}(\mathbf{A})] \Big| \leq \frac{c_3}{n^4}$$

for a constant $c_3 > 0$. Hence, we can apply Theorem 3.9 to obtain

$$\begin{aligned}\mathbb{P}\Big[|\tilde{h}_{r,\mathcal{L}}(\mathbf{A})| > 1.6c|\mathcal{L}| \;\Big|\; \mathbf{A}_{\cdot r} = \mathbf{a}_{\cdot r}\Big] &\leq 2\exp\left(-\frac{(1.6c|\mathcal{L}| - c_3/n^4)^2}{2\frac{5184c|\mathcal{L}|^3\log n}{m}}\right) \\ &\leq 2\exp\left(-c_4\frac{m}{|\mathcal{L}|\log n}\right),\end{aligned}$$

for a constant $c_4 \leq \frac{c}{4050} - \frac{c_3}{3240n^4}$.

Step 2: Unravel the conditions: take union bounds and integrate over $\mathbf{a}_{\cdot\mathbf{r}}$

Let

$$\widetilde{B}_r = \Big\{|\tilde{h}_{r,\mathcal{L}}(\mathbf{A})| \leq 1.6c|\mathcal{L}| \text{ for all } \mathcal{L} \subseteq [k]\backslash\{r\} \text{ with } |\mathcal{L}| = \lambda\Big\}.$$

Since we assume $m \geq c_s k^2 \log^2 n$, we can take the union bound over all possible subsets $\mathcal{L} \subseteq [k]\backslash\{r\}$ with $|\mathcal{L}| = \lambda$ to obtain, using the upper bound $\binom{k}{\lambda} \leq (\frac{ek}{\lambda})^\lambda$,

$$\begin{aligned}\mathbb{P}\Big[\widetilde{B}_r \;\Big|\; \mathbf{A}_{\cdot r} = \mathbf{a}_{\cdot r}\Big] &\geq 1 - 2\exp\left(-c_4\frac{m}{|\mathcal{L}|\log n} + |\mathcal{L}|\log\frac{ek}{|\mathcal{L}|}\right) \\ &\geq 1 - 2\exp(-(c_s c_4 - 1)k\log n).\end{aligned}$$

Next, we integrate over all $\mathbf{a}_{\cdot r}$ satisfying $\max_j |a_{jr}| \leq 6\sqrt{\log n}$:

$$\begin{aligned}\mathbb{P}\Big[\widetilde{B}_r\Big] &\geq \int_{\{\max_j |a_{jr}| \leq 6\sqrt{\log n}\}} \mathbb{P}\Big[B_r \;\Big|\; \mathbf{A}_{\cdot r} = \mathbf{a}_{\cdot r}\Big]\mu(\mathbf{a}_{\cdot r})d\mathbf{a}_{\cdot r} \\ &\geq \Big(1 - 2\exp(-(c_s c_4 - 1)k\log n)\Big)\mathbb{P}\Big[\max_j |a_{jr}| \leq 6\sqrt{\log n}\Big] \\ &\geq 1 - (m+2)n^{-18},\end{aligned}$$

where we write μ for the standard normal density and the last line follows from standard Gaussian tail bounds and a union bound, provided that $(c_s c_4 - 1)k \geq 18$.

By construction, we have $h_{r,\mathcal{L}}(\mathbf{a}) = \tilde{h}_{r,\mathcal{L}}(\mathbf{a})$ for all $r \in [k]$ and subsets $\mathcal{L} \subseteq [k]\backslash\{r\}$ with cardinality $|\mathcal{L}| = \lambda$ on the set $\mathcal{A}_\lambda \cap \{\max_{j,r} |a_{jr}| \leq 6\sqrt{\log n}\}$. Hence, a union bound gives a lower bound on the probability that the fourth term in (3.93) is bounded by $1.6c|\mathcal{L}|$ for all $r \in [k]$ and $\mathcal{L} \subseteq [k]\backslash\{r\}$ with $|\mathcal{L}| = \lambda$:

$$\mathbb{P}\left[\bigcap_{r=1}^k B_r\right] \geq \mathbb{P}\left[\bigcap_{r=1}^k \widetilde{B}_r \cap \mathcal{A}_\lambda \cap \left\{\max_{j,r} |a_{jr}| \leq 6\sqrt{\log n}\right\}\right] \geq 1 - (m+2)n^{-17} - c_1\lambda n^{-13}.$$

Putting everything together, this completes the induction step (3.93),

$$\frac{1}{m}\sum_{j=1}^m \left(\sum_{l\in\mathcal{L}} A_{jl} + A_{jr}\right)^4 \leq c(|\mathcal{L}| + 1)^2 \quad \text{for all } r \in [k], \mathcal{L} \subseteq [k]\backslash\{r\} \text{ with } |\mathcal{L}| = \lambda,$$

holds with probability at least $1 - c_1(\lambda+1)n^{-13}$, provided that $c_1 \geq c_2 + (m+2)n^{-4}$, which is satisfied for a universal constant c_1 if $m \lesssim n^4$. The case $m \gtrsim n^4$ is simpler and can be shown following the same steps, writing probabilities in terms of m instead of n. □

Proof of Lemma 3.16. We will show that each of the inequalities (3.85)–(3.88) is satisfied with probability at least $1 - \frac{c_p}{4}n^{-10}$. The lemma then follows by taking a union bound. Throughout this proof, we will assume $m \leq n^{3/2}$. The other case $m > n^{3/2}$ is simpler and can be shown following the same steps, writing probabilities in terms of m instead of n.

Proof that (3.85) is satisfied with high probability

In order to show that (3.85) holds with high probability, we proceed as in the induction step of the proof of Lemma 3.15: for any index $i \in [n]$, any subset $\mathcal{K} \subseteq \mathcal{S}\backslash\{i\}$, any subset $\mathcal{L} \subseteq \mathcal{K}$ with $|\mathcal{L}| \geq \frac{1}{2}|\mathcal{K}|$, and any vector $\mathbf{x} \in N_\epsilon^{\mathcal{K}}$, we consider the function

$$h_{i,\mathcal{K},\mathcal{L},\mathbf{x}}(\mathbf{a}) = \sum_{l \in \mathcal{L}} \frac{1}{m} \sum_{j=1}^{m} a_{ji} a_{jl} (\mathbf{a}_j^\top \mathbf{x})^2.$$

Then, we define B_i as the following event:

$$B_i = \Big\{ |h_{i,\mathcal{K},\mathcal{L},\mathbf{x}}(\mathbf{A})| \leq c|\mathcal{L}| \frac{\log n}{\sqrt{m}} \quad \text{for all subsets } \mathcal{K} \subseteq \mathcal{S}\backslash\{i\}, \mathcal{L} \subseteq \mathcal{K} \text{ with } |\mathcal{L}| \geq \frac{1}{2}|\mathcal{K}| \text{ and } \mathbf{x} \in N_\epsilon^{\mathcal{K}} \Big\}.$$

To bound the probability of B_i, we condition on $\mathbf{A}_{\cdot i} = \mathbf{a}_{\cdot i}$ using the formula

$$\mathbb{P}[B] = \int \mathbb{P}[B | \mathbf{A}_{\cdot i} = \mathbf{a}_{\cdot i}] \mu(\mathbf{a}_{\cdot i}) d\mathbf{a}_{\cdot i},$$

which holds for any event B, where we write μ for the standard normal density.

To show that the conditional probability $\mathbb{P}[B_i | \mathbf{A}_{\cdot i} = \mathbf{a}_{\cdot i}]$ is close to one, we begin by showing that, for any fixed $\mathcal{K}$, $\mathcal{L}$ and $\mathbf{x}$ as described above, $h_{i,\mathcal{K},\mathcal{L},\mathbf{x},}(\mathbf{A})|_{\mathbf{A}_{\cdot i} = \mathbf{a}_{\cdot i}}$ (making explicit the fact that we are conditioning on $\mathbf{A}_{\cdot i} = \mathbf{a}_{\cdot i}$) concentrates around its expectation $\mathbb{E}[h_{i,\mathcal{K},\mathcal{L},\mathbf{x}}(\mathbf{A}) | \mathbf{A}_{\cdot i} = \mathbf{a}_{\cdot i}] = 0$. To simplify notation, we will omit the subscripts and write $h(\mathbf{A})$ in what follows.

Step 1: Bound $h(\mathbf{A})$ conditioned on $i, \mathcal{K}, \mathcal{L}, \mathbf{x}$ and $\mathbf{A}_{\cdot i}$

Let $\mathcal{A}_1$ be as in Lemma 3.15, and $\mathcal{A}_2$ as in Lemma 3.11 with $t = 5\sqrt{\log n}$ (more precisely,

we have $\mathcal{A}_1, \mathcal{A}_2 \subset \mathbb{R}^{m\times n}$, and we require the projections onto $\mathbb{R}^{m\times\mathcal{S}}$ to be as in the respective lemmas). Then, by these two Lemmas, we have $\mathbb{P}[\mathcal{A}_1 \cap \mathcal{A}_2] \geq 1 - c_2 n^{-11}$ for a universal constant $c_2 > 0$. As the intersection of convex sets, $\mathcal{A}_1 \cap \mathcal{A}_2$ is also convex.

Fix an $i \in [n]$, a subset $\mathcal{K} \subseteq \mathcal{S}\backslash\{i\}$, a subset $\mathcal{L} \subseteq \mathcal{K}$ with $|\mathcal{L}| \geq \frac{1}{2}|\mathcal{K}|$ and a vector $\mathbf{x} \in N_\epsilon^{\mathcal{K}}$. We condition on $\mathbf{A}_{\cdot i} = \mathbf{a}_{\cdot i}$ for a vector $\mathbf{a}_{\cdot i} \in \mathbb{R}^n$ satisfying $\max_j |a_{ji}| \leq 5\sqrt{\log n}$. As in the proof of Lemma 3.15, the idea is use Theorem 3.9 to show that $h(\mathbf{A})$ is close to its expectation. However, $h(\mathbf{A})$ is not globally Lipschitz continuous. We will show that the function h restricted to $\mathcal{A}_1 \cap \mathcal{A}_2$ is Lipschitz continuous. Then, we extend this restricted function to a globally Lipschitz continuous function $\tilde{h}$ on the entire space, and apply Theorem 3.9 to $\tilde{h}$. By construction, we have $h(\mathbf{A}) = \tilde{h}(\mathbf{A})$ with high probability, which therefore also yields a high probability bound for $h(\mathbf{A})$.

Step 1, part (a): Bound the Lipschitz constant of h restricted to $\mathcal{A}_1 \cap \mathcal{A}_2$
Restricted to $\mathcal{A}_1 \cap \mathcal{A}_2$, we can bound the Lipschitz constant of h by the norm of its gradient. The norm of the gradient is an upper bound for the Lipschitz constant by the mean-value theorem, since $\mathcal{A}_1 \cap \mathcal{A}_2$ is a convex set. Note that h only depends on a_{jr} for $r \in \mathcal{K}$, since we have $x_r = 0$ for $r \notin \mathcal{K}$. We can compute

$$\frac{\partial}{\partial a_{jr}} h(\mathbf{a}) = \begin{cases} \frac{2}{m} a_{ji} x_r (\mathbf{a}_j^\top \mathbf{x}) \sum_{l\in\mathcal{L}} a_{jl} & r \in \mathcal{K}\backslash\mathcal{L} \\ \frac{2}{m} a_{ji} x_r (\mathbf{a}_j^\top \mathbf{x}) \sum_{l\in\mathcal{L}} a_{jl} + \frac{1}{m} a_{ji} (\mathbf{a}_j^\top \mathbf{x})^2 & r \in \mathcal{L} \end{cases}$$

Using the inequality $(a+b)^2 \leq 2(a^2+b^2)$, we can bound

$$\begin{aligned} \|\nabla h(\mathbf{a})\|_2^2 = \sum_{r\in\mathcal{K}} \sum_{j=1}^m \left(\frac{\partial}{\partial a_{jr}} h(\mathbf{a})\right)^2 \leq & 8 \sum_{r\in\mathcal{K}} \sum_{j=1}^m \frac{1}{m^2} x_r^2 a_{ji}^2 \left(\sum_{l\in\mathcal{L}} a_{jl}\right)^2 (\mathbf{a}_j^\top \mathbf{x})^2 \\ & + 2 \sum_{r\in\mathcal{L}} \sum_{j=1}^m \frac{1}{m^2} a_{ji}^2 (\mathbf{a}_j^\top \mathbf{x})^4. \end{aligned} \tag{3.94}$$

We bound the two sums separately. For the first sum, we have

$$\begin{aligned} \sum_{r\in\mathcal{K}} \sum_{j=1}^m \frac{1}{m^2} x_r^2 a_{ji}^2 \left(\sum_{l\in\mathcal{L}} a_{jl}\right)^2 (\mathbf{a}_j^\top \mathbf{x})^2 &= \frac{1}{m} \sum_{r\in\mathcal{K}} x_r^2 \sum_{j=1}^m \frac{1}{m} a_{ji}^2 \left(\sum_{l\in\mathcal{L}} a_{jl}\right)^2 (\mathbf{a}_j^\top \mathbf{x})^2 \\ &\leq \left(\max_j a_{ji}^2\right) \frac{1}{m} \sum_{r\in\mathcal{K}} x_r^2 \sum_{j=1}^m \frac{1}{m} \left(\sum_{l\in\mathcal{L}} a_{jl}\right)^2 (\mathbf{a}_j^\top \mathbf{x})^2 \\ &\leq \frac{25 \log n}{m} \sum_{r\in\mathcal{K}} x_r^2 \sqrt{\frac{1}{m} \sum_{j=1}^m \left(\sum_{l\in\mathcal{L}} a_{jl}\right)^4} \sqrt{\frac{1}{m} \sum_{j=1}^m (\mathbf{a}_j^\top \mathbf{x})^4}, \end{aligned}$$

where we used Hölder's inequality in the second and the Cauchy-Schwarz inequality in the last line.

Remark 3.1. *The application of Hölder's inequality in the second line is the reason for the term* $\log^2 n$ *in the sample complexity. We expect that also* $\frac{1}{m}\sum_{j=1}^m a_{ji}^4(\mathbf{a}_j^T\mathbf{x})^4 = \mathcal{O}(1)$*, however Lemma 3.11 does not apply because of the extra factor* a_{ji}^4*, and the truncation method of Lemma 3.13 does not yield a probability small enough to allow for taking the union bound over all* $\mathbf{x} \in N_{\epsilon'}^{\mathcal{K}}$*.*

Since $\mathbf{a} \in \mathcal{A}_1 \cap \mathcal{A}_2$, we can apply Lemma 3.15 to bound

$$\sqrt{\frac{1}{m}\sum_{j=1}^m \left(\sum_{l\in\mathcal{L}} a_{jl}\right)^4} \leq c_3|\mathcal{L}|,$$

where c_3 is the universal constant from Lemma 3.15, and Lemma 3.11 yields

$$\sqrt{\frac{1}{m}\sum_{j=1}^m (\mathbf{a}_j^\top\mathbf{x})^4} \leq \sqrt{\frac{1}{m}\|\mathbf{a}_{\cdot\mathcal{K}}\|_{2\to 4}^4} \leq 11,$$

as $\|\mathbf{x}\|_2 = 1$. Putting this together, we can bound the first sum in (3.94),

$$\sum_{r\in\mathcal{K}}\sum_{j=1}^m \frac{1}{m^2}x_r^2 a_{ji}^2\left(\sum_{l\in\mathcal{L}} a_{jl}\right)^2 (\mathbf{a}_j^\top\mathbf{x})^2 \leq \frac{5\log n}{m}\sum_{r\in\mathcal{K}} 11c_3|\mathcal{L}|x_r^2 \leq 55c_3\frac{|\mathcal{L}|\log n}{m},$$

where we again used $\|\mathbf{x}\|_2 = 1$. For the second sum in (3.94), we can write

$$\sum_{r\in\mathcal{L}}\sum_{j=1}^m \frac{1}{m^2}a_{ji}^2(\mathbf{a}_j^\top\mathbf{x})^4 \leq \frac{5\log n}{m}|\mathcal{L}|\frac{1}{m}\sum_{j=1}^m(\mathbf{a}_j^\top\mathbf{x})^4 \leq 605\frac{|\mathcal{L}|\log n}{m},$$

where we used Lemma 3.11 to bound $\frac{1}{m}\sum_{j=1}^m(\mathbf{a}_j^\top\mathbf{x})^4 \leq 121$. All in all, we have shown

$$\|\nabla h(\mathbf{a})\|_2 = \left(\sum_{r\in\mathcal{K}}\sum_{j=1}^m\left(\frac{\partial}{\partial a_{jr}}h(\mathbf{a})\right)^2\right)^{\frac{1}{2}} \leq \sqrt{c_4\frac{|\mathcal{L}|\log n}{m}}$$

for $\mathbf{a} \in \mathcal{A}_1 \cap \mathcal{A}_2$ and $c_4 = 440c_3 + 1210$.

Step 1, part (b)**: Construct a globally Lipschitz continuous extension of** h

Consider the following Lipschitz extension of the function h:

$$\tilde{h}(\mathbf{a}) = \inf_{\mathbf{a}'\in\mathcal{A}_1\cap\mathcal{A}_2}\Big(h(\mathbf{a}') + \mathrm{Lip}(h)\|\mathbf{a}-\mathbf{a}'\|_2\Big),$$

where we write $\mathrm{Lip}(h) = \sqrt{c_4 \frac{|\mathcal{L}| \log n}{m}}$. By definition, we have $\tilde{h} = h$ on $\mathcal{A}_1 \cap \mathcal{A}_2$, and it follows from an application of the triangle inequality that $\tilde{h}$ is globally Lipschitz continuous with Lipschitz constant $\mathrm{Lip}(h)$ (see e.g. Theorem 7.2 of [83]).

We will show that $\tilde{h}$ concentrates around its mean, which can potentially differ from the mean of h. Since h and $\tilde{h}$ differ only on $(\mathcal{A}_1 \cap \mathcal{A}_2)^c$, which has probability at most $c_2 n^{-11}$, we can bound, using the Cauchy-Schwarz inequality,

$$\mathbb{E}[|h(\mathbf{A})| \mathbf{1}_{(\mathcal{A}_1 \cap \mathcal{A}_2)^c}(\mathbf{A})] \leq \sqrt{\mathbb{E}[h(\mathbf{A})^2]} \sqrt{\mathbb{E}[\mathbf{1}_{(\mathcal{A}_1 \cap \mathcal{A}_2)^c}(\mathbf{A})]} \leq \frac{c'}{n^5},$$

where we used that $\mathbb{E}[h(\mathbf{A})^2] \leq \sqrt{315} k^2/m$. Using $\tilde{h}(\mathbf{a}) \leq \mathrm{Lip}(h) \|\mathbf{a}\|_2$, we can bound

$$\mathbb{E}[|\tilde{h}(\mathbf{A})| \mathbf{1}_{(\mathcal{A}_1 \cap \mathcal{A}_2)^c}(\mathbf{A})] \leq \mathrm{Lip}(h) \sqrt{\mathbb{E}[\|\mathbf{A}\|_2^2]} \sqrt{\mathbb{E}[\mathbf{1}_{(\mathcal{A}_1 \cap \mathcal{A}_2)^c}(\mathbf{A})]} \leq \frac{c''}{n^3}.$$

All in all, this shows that

$$\Big| \mathbb{E}[h(\mathbf{A})] - \mathbb{E}[\tilde{h}(\mathbf{A})] \Big| \leq \frac{c_5}{n^3}.$$

for a constant $c_5 > 0$. Hence, we can apply Theorem 3.9 to obtain

$$\begin{aligned} \mathbb{P}\left[|\tilde{h}(\mathbf{A})| > c \frac{|\mathcal{L}| \log n}{\sqrt{m}} \;\middle|\; \mathbf{A}_{\cdot i} = \mathbf{a}_{\cdot i} \right] &\leq 2 \exp\left(- \frac{(c \frac{|\mathcal{L}|}{\sqrt{m}} \log n - c_5/n^3)^2}{2 c_4 \frac{|\mathcal{L}| \log n}{m}} \right) \\ &\leq 2 \exp(-c_6 |\mathcal{L}| \log n), \end{aligned}$$

for a universal constant $c_6 \leq \frac{c}{2c_4} - \frac{c c_5}{c_4} \frac{\sqrt{m}}{n^3}$, where we used the assumption $m \leq n^{3/2}$.

Step 2: Unravel the conditions: take union bounds and integrate over $\mathbf{a}_{\cdot i}$

Let

$$\begin{aligned} \widetilde{B}_{i,\lambda} = \Big\{ |\tilde{h}_{i,\mathcal{K},\mathcal{L},\mathbf{x}}(\mathbf{A})| \leq c \frac{|\mathcal{L}| \log n}{\sqrt{m}} \quad &\text{for all subsets } \mathcal{K} \subseteq \mathcal{S} \backslash \{i\} \text{ with } |\mathcal{K}| = \lambda, \\ &\mathcal{L} \subseteq \mathcal{K} \text{ with } |\mathcal{L}| \geq \frac{1}{2} |\mathcal{K}| \text{ and } \mathbf{x} \in N_\epsilon^{\mathcal{K}} \Big\}, \end{aligned}$$

and let $\widetilde{B}_i$ be the same event without the restriction $|\mathcal{K}| = \lambda$. The cardinality of the ϵ-net can be bounded by $|N_\epsilon^{\mathcal{K}}| \leq (3/\epsilon)^\lambda \leq (3m/c_1)^\lambda$. Taking union bounds over all $\mathbf{x} \in N_\epsilon^{\mathcal{K}}$, $\mathcal{L} \subseteq \mathcal{K}$ with $|\mathcal{L}| \geq \frac{1}{2}|\mathcal{K}|$ and $\mathcal{K} \subseteq \mathcal{S} \backslash \{i\}$ with fixed cardinality $|\mathcal{K}| = \lambda$, we obtain (using the upper bound $\binom{k}{\lambda} \leq (\frac{ek}{\lambda})^\lambda$)

$$\begin{aligned} \mathbb{P}\big[\widetilde{B}_{i,\lambda} \,\big|\big|\, \mathbf{A}_{\cdot i} = \mathbf{a}_{\cdot i} \big] &\geq 1 - 2 \exp\left(-\frac{c_6}{2} \lambda \log n + \lambda \log \frac{3m}{c_1} + \lambda \log 2 + \lambda \log \frac{ek}{\lambda} \right) \\ &\geq 1 - 2 \exp(-c_7 \lambda \log n) \end{aligned}$$

for a constant $c_7 > 0$ if $c_6 \geq 5 + 2\frac{\log(6e/c_1)}{\log n}$, where we used the assumption $m \leq n^{3/2}$. Taking the union bound over all possible choices for the cardinalities $\lambda = 1, ..., k$ gives

$$\mathbb{P}\Big[\widetilde{B}_i \mid \mathbf{A}_{\cdot i} = \mathbf{a}_{\cdot i}\Big] \geq 1 - \frac{2}{1 - 2\exp(-c_7 \log n)} \exp(-c_7 \log n) \geq 1 - 2.1 \exp(-c_7 \log n),$$

provided $n^{-c_7} \leq 0.1/4.2$. Next, we integrate over all $\mathbf{a}_{\cdot i}$ satisfying $\max_j |a_{ji}| \leq 5\sqrt{\log n}$:

$$\begin{aligned}
\mathbb{P}\Big[\widetilde{B}_i\Big] &\geq \int_{\{\max_j |a_{ji}| \leq 5\sqrt{\log n}\}} \mathbb{P}\Big[\widetilde{B}_i \;\Big|\; \mathbf{A}_{\cdot i} = \mathbf{a}_{\cdot i}\Big] \mu(\mathbf{a}_{\cdot i}) d\mathbf{a}_{\cdot i} \\
&\geq \Big(1 - 2.1 e^{-c_7 \log n}\Big) \mathbb{P}\Big[\max_j |a_{ji}| \leq 5\sqrt{\log n}\Big] \\
&\geq 1 - 4n^{-11},
\end{aligned}$$

where we write μ for the Gaussian density. The last line follows from standard Gaussian tail bounds and a union bound using the assumption $m \leq n^{3/2}$, provided that $c_7 \geq 11$.

On $\mathcal{A}_1 \cap \mathcal{A}_2 \cap \{\max_j |a_{ji}| \leq 5\sqrt{\log n}\}$, which is independent of $\mathcal{K}$, $\mathcal{L}$ and $\mathbf{x}$, we have, by construction, $h_{i,\mathcal{K},\mathcal{L},\mathbf{x}} = \tilde{h}_{i,\mathcal{K},\mathcal{L},\mathbf{x}}$ for all subsets $\mathcal{K} \subseteq \mathcal{S}\backslash\{i\}$, $\mathcal{L} \subseteq \mathcal{K}$ with $|\mathcal{L}| \geq \frac{1}{2}|\mathcal{K}|$ and $\mathbf{x} \in N_\epsilon^{\mathcal{K}}$. In this case, the bound on $\tilde{h}_{\mathcal{K},\mathcal{L},\mathbf{x}}(\mathbf{A})$ also applies to $h_{\mathcal{K},\mathcal{L},\mathbf{x}}(\mathbf{A})$, so that

$$\begin{aligned}
\mathbb{P}[B_i] &\geq \mathbb{P}\Big[\widetilde{B}_i \cap \mathcal{A}_1 \cap \mathcal{A}_2 \cap \Big\{\max_j |a_{ji}| \leq 5\sqrt{\log n}\Big\}\Big] \\
&\geq 1 - 4n^{-11} - c_2 n^{-11} - n^{-11} \\
&\geq 1 - \frac{c_p}{4} n^{-11},
\end{aligned}$$

for $c_p = 4c_2 + 20$. Finally, taking the union bound over all $i \in [n]$ establishes the bound (3.85) with probability at least $1 - \frac{c_p}{4} n^{-10}$.

Proof that (3.86)–(3.88) are satisfied with high probability

Inequalities (3.86) and (3.88) can be shown following the exact same steps as above for (3.85). The former is slightly simpler because of the lower order of the expression we need to control, which is also the reason for the term $\sqrt{\log n}$ in the bound (3.86) instead of $\log n$. The proof of inequality (3.87) is identical to the one of (3.86), as we can condition on the random variables $\{\varepsilon_j\}_{j=1}^m$ (the same way we conditioned on $\mathbf{A}_{\cdot i} = \mathbf{a}_{\cdot i}$ in Step 1 above) and use the standard tail bound $\mathbb{P}[|\varepsilon_j| > t] \leq \exp(1 - \tilde{c}t/\sigma^2)$ for sub-exponential random variables, where $\tilde{c} > 0$ is an absolute constant. We omit the details of the proof to avoid repetition. □

Proof of Lemma 3.17. We will show that each of the inequalities (3.89)–(3.91) is satisfied with probability at least $1 - \frac{c_p}{4} n^{-10}$. The lemma then follows by taking a union bound.

We begin by showing the bound (3.89). To show that (3.89) holds with high probability, we will show that it is satisfied for all $\{\mathbf{a}_j\}_{j=1}^m \in \mathcal{A} \subseteq \mathbb{R}^{m \times n}$, where $\mathcal{A}$ is defined as the set where the bounds from Lemmas 3.13 and 3.16 are satisfied, and such that the projection of $\mathcal{A}$ onto $\mathbb{R}^{m \times \mathcal{S}}$ satisfies Lemma 3.11 with $t = 5\sqrt{\log n}$. By the aforementioned lemmas, we have $\mathbb{P}[\{\mathbf{A}_j\}_{j=1}^m \in \mathcal{A}] \geq 1 - \frac{c_p}{4} n^{-10}$ for a universal constant $c_p > 0$.

Proof that (3.89) is satisfied with high probability

Let $\{\mathbf{a}_j\}_{j=1}^m \in \mathcal{A}$ and fix any $i \in [n]$. We denote the term that we need to bound in (3.89) by

$$g(\mathbf{x}) = \frac{1}{m} \sum_{j=1}^m a_{ji} (\mathbf{a}_{j,-i}^\top \mathbf{x}_{-i})^3,$$

and we consider the constrained optimization problem

$$\max_{\mathbf{x}} g(\mathbf{x}) \quad \text{s.t.} \quad \mathbf{x} \in \mathcal{X}_b, \tag{3.95}$$

where we write $\mathcal{X}_b = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}_{\mathcal{S}^c} = \mathbf{0}, \|\mathbf{x}\|_2^2 \leq 1, \|\mathbf{x}\|_1 \leq b\}$ for some $0 \leq b \leq \sqrt{k}$. We will show that any optimizer $\mathbf{x}' \in \mathcal{X}_b$ satisfies $g(\mathbf{x}') \leq cb\sqrt{\frac{\log n}{m}}$. By symmetry, we can also obtain the lower bound $g(\mathbf{x}') \geq -cb\sqrt{\frac{\log n}{m}}$ by considering the corresponding minimization problem. Since the above holds for any $0 \leq b \leq \sqrt{k}$, this would then complete the proof that inequality (3.89) is satisfied for all $\{\mathbf{a}_j\}_{j=1}^m \in \mathcal{A}$ and $i \in [n]$.

Since we are maximizing a continuous function over a compact set, a global maximum is attained at some feasible point $\mathbf{x}' \in \mathcal{X}_b$. Using the KKT conditions at $\mathbf{x}'$, we will bound the Lagrange multiplier $\mu_2^\star$ corresponding to the constraint $\|\mathbf{x}\|_1 \leq b$. This controls how the maximum attainable value $g(\mathbf{x}')$ can increase if we relax the constraint $\|\mathbf{x}\|_1 \leq b$ (for details see e.g. Section 5.6 of [16]), and integrating over b gives the desired result.

Step 1: Establish KKT conditions

Let $\mathbf{x}' \in \mathcal{X}_b$ be an optimal point of the constrained optimization problem (3.95). In order to establish the KKT conditions at $\mathbf{x}'$, we verify that the linear independence

constraint qualification (LICP) (see e.g. [93]) holds in this problem: the gradients of all active inequality constraints are linearly independent at any point $\mathbf{x}$. We restrict our attention to $b \neq \sqrt{s}$ for $s = 1, ..., k$, since $b = \sqrt{s}$ for some $s = 1, \ldots, k$ is a set of measure zero and can be ignored when integrating. If only one inequality constraint is binding, there is nothing to show. Otherwise, the gradients of the constraints, $\frac{\partial}{\partial \mathbf{x}}\|\mathbf{x}\|_2^2 = 2\mathbf{x}$ and $\frac{\partial}{\partial \mathbf{x}}\|\mathbf{x}\|_1 = \text{sign}(\mathbf{x})$ can only be linearly dependent if $x_i \in \{-\tilde{c}, 0, \tilde{c}\}$ for all coordinates $i \in [n]$, where $\tilde{c} > 0$ is some constant. But then $\|\mathbf{x}\|_2^2 = s\tilde{c}^2$ and $\|\mathbf{x}\|_1 = s\tilde{c}$, where s is the number of non-zero coordinates of $\mathbf{x}$. Since $b \neq \sqrt{s}$, the two constraints cannot be simultaneously binding.

From LICP it follows that the following KKT conditions are satisfied at $\mathbf{x}'$: writing

$$\mathcal{L}(\mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = g(\mathbf{x}) - \mu_1\Big(\|\mathbf{x}\|_2^2 - 1\Big) - \mu_2\Big(\|\mathbf{x}\|_1 - b\Big) + \sum_{r \in \mathcal{S}^c} \lambda_r x_r$$

for the Lagrangian, there exist Lagrange multipliers $\boldsymbol{\lambda}^*$ and $\boldsymbol{\mu}^\star$ satisfying

$$\nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}', \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star) = \mathbf{0} \qquad \text{stationarity} \tag{3.96}$$

$$\mathbf{x}'_{\mathcal{S}^c} = \mathbf{0}, \quad \|\mathbf{x}'\|_2^2 \leq 1, \quad \|\mathbf{x}'\|_1 \leq b \qquad \text{primal feasibility} \tag{3.97}$$

$$\mu_1^\star \geq 0, \quad \mu_2^\star \geq 0 \qquad \text{dual feasibility} \tag{3.98}$$

$$\mu_1^\star\Big(\|\mathbf{x}'\|_2^2 - 1\Big) = 0, \quad \mu_2^\star\Big(\|\mathbf{x}'\|_1 - b\Big) = 0 \qquad \text{complementary slackness} \tag{3.99}$$

Step 2: Bounding the Lagrange multiplier $\mu_2^\star$

Using the stationarity condition (3.96), we can compute, for any $l \in \mathcal{S}\backslash\{i\}$,

$$\begin{aligned} \frac{\partial}{\partial x_l}\mathcal{L}(\mathbf{x}', \boldsymbol{\lambda}^\star, \boldsymbol{\mu}^\star) &= \frac{3}{m}\sum_{j=1}^m a_{ji}a_{jl}(\mathbf{a}_{j,-i}^\top \mathbf{x}'_{-i})^2 - 2\mu_1^\star x_l' - \mu_2^\star \,\text{sign}(x_l') = 0 \\ \Rightarrow \qquad h_l(\mathbf{x}') &:= \frac{1}{m}\sum_{j=1}^m a_{ji}a_{jl}(\mathbf{a}_{j,-i}^\top \mathbf{x}'_{-i})^2 = \frac{2}{3}\mu_1^\star x_l' + \frac{1}{3}\mu_2^\star \,\text{sign}(x_l'), \end{aligned}$$

This identity implies that $h_l(\mathbf{x}')$ has the same sign as x_l' because of dual feasibility (3.98), and that $x_l' = 0$ if and only if $h_l(\mathbf{x}') = 0$ (or if $l \notin \mathcal{S}$). Rearranging for $\mu_2^\star$, we obtain, for any l with $x_l' \neq 0$,

$$\mu_2^\star = 3|h_l(\mathbf{x}')| - 2\mu_1^\star|x_l'| \leq 3|h_l(\mathbf{x}')|, \tag{3.100}$$

where we again used that $\mu_1^\star \geq 0$. Next, we show that

$$\min_{l: x_l' \neq 0} |h_l(\mathbf{x}')| \leq \frac{c \log n}{3\sqrt{m}}. \tag{3.101}$$

To this end, let $\mathcal{K} = \{r \in \mathcal{S} : x'_r \neq 0\}$ and let $\mathcal{L} = \{l \in \mathcal{K} : x'_l > 0\}$. Note that we must have $x'_i = 0$, as x_i does not contribute to the value of $g(\mathbf{x})$. We can assume without loss of generality that $|\mathcal{L}| \geq \frac{1}{2}|\mathcal{K}|$, as we can otherwise consider the set defined by $x'_l < 0$. Further, assume for notational simplicity that $\|\mathbf{x}'\|_2 = 1$; otherwise, we can replace $\mathbf{x}'$ by $\mathbf{x}'/\|\mathbf{x}'\|_2$, which makes every $|h_l(\mathbf{x}')|$ larger, since we have $\|\mathbf{x}'\|_2 \leq 1$ by primal feasibility (3.97).

Let $N_\epsilon^{\mathcal{K}}$ be an ϵ-net of the set $\{\mathbf{x} \in \mathbb{R}^n : \mathbf{x}_{\mathcal{K}^c} = \mathbf{0}, \|\mathbf{x}\|_2 = 1\}$, where $\epsilon = c \log n/(465\sqrt{m})$. Let $\mathbf{x} \in N_\epsilon^{\mathcal{K}}$ with $\|\mathbf{x} - \mathbf{x}'\|_2 \leq \epsilon$. Then, recalling $m \geq c_s \max\{k^2 \log^2 n,\ \log^5 n\}$, we have by inequality (3.85) of Lemma 3.16,

$$\sum_{l \in \mathcal{L}} h_l(\mathbf{x}) \leq c_1 |\mathcal{L}| \frac{\log n}{\sqrt{m}} \leq \frac{c|\mathcal{L}| \log n}{6\sqrt{m}},$$

provided that $c \geq 6c_1$, where c_1 is the universal constant from Lemma 3.16. Since $h_l(\mathbf{x}) > 0$ for all $l \in \mathcal{L}$, there must be an index $l \in \mathcal{L}$ with $h_l(\mathbf{x}) \leq c \log n/(6\sqrt{m})$ by the pigeonhole principle. By the Cauchy-Schwarz inequality, we can bound

$$\begin{aligned}
|h_l(\mathbf{x}) - h_l(\mathbf{x}')| &= \left| \frac{1}{m} \sum_{j=1}^m a_{ji} a_{jl} \Big(\mathbf{a}_{j,-i}^\top (\mathbf{x}_{-i} + \mathbf{x}'_{-i}) \Big) \Big(\mathbf{a}_{j,-i}^\top (\mathbf{x}_{-i} - \mathbf{x}'_{-i}) \Big) \right| \\
&\leq \left(\frac{1}{m} \sum_{j=1}^m a_{ji}^2 a_{jl}^2 \Big(\mathbf{a}_{j,-i}^\top (\mathbf{x}_{-i} + \mathbf{x}'_{-i}) \Big)^2 \right)^{\frac{1}{2}} \left(\frac{1}{m} \sum_{j=1}^m \Big(\mathbf{a}_{j,-i}^\top (\mathbf{x}_{-i} - \mathbf{x}'_{-i}) \Big)^2 \right)^{\frac{1}{2}} \\
&\leq \left(\frac{1}{m} \sum_{j=1}^m a_{ji}^8 \right)^{\frac{1}{8}} \left(\frac{1}{m} \sum_{j=1}^m a_{jl}^8 \right)^{\frac{1}{8}} \left(\frac{1}{m} \sum_{j=1}^m \Big(\mathbf{a}_{j,-i}^\top (\mathbf{x}_{-i} + \mathbf{x}'_{-i}) \Big)^4 \right)^{\frac{1}{4}} \\
&\quad \cdot \left(\frac{1}{m} \sum_{j=1}^m \Big(\mathbf{a}_{j,-i}^\top (\mathbf{x}_{-i} - \mathbf{x}'_{-i}) \Big)^2 \right)^{\frac{1}{2}} \\
&\leq 106^{\frac{1}{4}} \frac{(3m)^{\frac{1}{4}} + \sqrt{k} + 5\sqrt{\log n}}{m^{\frac{1}{4}}} \|\mathbf{x} + \mathbf{x}'\|_2 \frac{\sqrt{m} + \sqrt{k} + 5\sqrt{\log n}}{\sqrt{m}} \|\mathbf{x} - \mathbf{x}'\|_2 \\
&\leq \frac{c \log n}{6\sqrt{m}},
\end{aligned}$$

where we used Lemmas 3.11 and 3.13, and $\|\mathbf{x} - \mathbf{x}'\| \leq c \log n/(465\sqrt{m})$. Since $h_l(\mathbf{x}') \geq 0$ is non-negative, this completes the proof of (3.101). Combining (3.100) with (3.101), we obtain

$$\mu_2^\star \leq c \frac{\log n}{\sqrt{m}}.$$

Step 3: Showing the inequality (3.89)

In order to show the inequality (3.89), define the value function

$$v(b) = \max_{\mathbf{x}} \Big\{ g(\mathbf{x}) : \mathbf{x}_{\mathcal{S}^c} = \mathbf{0}, \|\mathbf{x}\|_2^2 \leq 1, \|\mathbf{x}\|_1 \leq b \Big\}.$$

We use the fact that $\mu_2^\star$ is the shadow price of the constraint $\|\mathbf{x}\|_1 \leq b$ (see e.g. Section 5.6 of [16]):

$$\frac{\partial}{\partial b} v(b) = \mu_2^\star(b).$$

By definition, we have $v(0) = 0$. For any $0 \leq b_0 \leq \sqrt{k}$, the fundamental theorem of calculus yields

$$v(b_0) = \int_0^{b_0} \frac{\partial}{\partial b} v(b) db = \int_0^{b_0} \mu_2^\star(b) db \leq \int_0^{b_0} c \frac{\log n}{\sqrt{m}} db < c b_0 \frac{\log n}{\sqrt{m}},$$

By definition, we have $g(\mathbf{x}) \leq v(b_0)$ for all $\mathbf{x} \in \mathcal{X}$ with $\|\mathbf{x}\|_1 \leq b_0$. Hence,

$$g(\mathbf{x}) \leq c b \frac{\log n}{\sqrt{m}}$$

must hold for any $\mathbf{x}$ with $\|\mathbf{x}\|_2^2 \leq 1$ and $\|\mathbf{x}\|_1 \leq b$, which completes the proof of (3.89).

Proof that (3.90)–(3.92) are satisfied with high probability

The proofs of (3.90), (3.91) and (3.92) follow the same steps as the proof of (3.89), using the bounds (3.86), (3.87) and (3.88) of Lemma 3.16, respectively, to bound the shadow price $\mu_2^\star$. We omit the details of the proof to avoid repetition. □

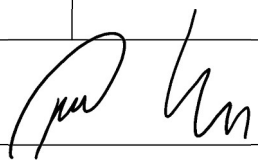
Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

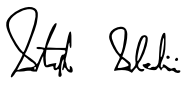
Title of Paper	Nearly minimax-optimal rates for noisy sparse phase retrieval via early-stopped mirror descent
Publication Status	□Submitted for Publication
Publication Details	"Nearly minimax-optimal rates for noisy sparse phase retrieval via early-stopped mirror descent", Fan Wu, Patrick Rebeschini. In preprint, 2021.

Student Confirmation

Student Name:	Fan Wu		
Contribution to the Paper	Derived main results and technical details of the manuscript and coded the numerical experiments. Wrote first draft of the manuscript, and wrote later versions of the paper alongside Patrick Rebeschini.		
Signature		Date	29/12/2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Patrick Rebeschini, Associate Professor		
Supervisor comments			
Signature		Date	30/12/2021

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 4

Implicit regularization in matrix sensing via mirror descent

We study discrete-time mirror descent applied to the unregularized empirical risk in matrix sensing. In both the general case of rectangular matrices and the particular case of positive semidefinite matrices, a simple potential-based analysis in terms of the Bregman divergence allows us to establish convergence of mirror descent—with different choices of the mirror maps—to a matrix that, among all global minimizers of the empirical risk, minimizes a quantity explicitly related to the nuclear norm, the Frobenius norm, and the von Neumann entropy. In both cases, this characterization implies that mirror descent, a first-order algorithm minimizing the unregularized empirical risk, recovers low-rank matrices under the same set of assumptions that are sufficient to guarantee recovery for nuclear-norm minimization. When the sensing matrices are symmetric and commute, we show that gradient descent with full-rank factorized parametrization is a first-order approximation to mirror descent, in which case we obtain an explicit characterization of the implicit bias of gradient flow as a by-product.

4.1 Introduction

Matrix sensing represents a paradigm in modern statistics [33, 97, 98], with applications ranging from image compression [5] to collaborative filtering [69] and dimensionality reduction [125], for instance. The goal is to recover a rank-r matrix $\mathbf{X}^\star \in \mathbb{R}^{n\times n'}$ from a set of linear measurements $y_j = \langle \mathbf{A}_j, \mathbf{X}^\star \rangle$, $j = 1, \dots, m$, where the sensing matrices

$\mathbf{A}_j \in \mathbb{R}^{n \times n'}$ are observed. This formulation includes the problem of matrix completion, where a subset of the entries of the matrix $\mathbf{X}^\star$ is observed.

Most of the literature on matrix sensing is based on some form of explicit regularization or rank constraint to encourage or enforce low-rankness of the estimated matrix. A popular approach is based on minimizing the nuclear norm or on using explicit regularization techniques based on the nuclear norm, e.g. [19, 61, 66, 81, 97, 98, 109]. Another popular approach is based on non-convex optimization and the low-rank factorization $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$ with matrices $\mathbf{U} \in \mathbb{R}^{n \times r}$, $\mathbf{V} \in \mathbb{R}^{n' \times r}$, where the factorization itself enforces the low-rankness of $\mathbf{X}$, e.g. [32, 33, 62, 79, 80, 107, 112, 141].

The literature on *implicit* regularization for matrix sensing is more recent and less well developed. When $\mathbf{X}^\star$ is assumed to be a positive semidefinite matrix, it was first empirically observed in [54] that mininizing the unregularized empirical risk using vanilla gradient descent with parametrization $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$ and random full-rank initialization close to zero yields a low nuclear norm solution, even when $\mathbf{U} \in \mathbb{R}^{n \times n}$, i.e. when no constraint on the rank of $\mathbf{X}$ is enforced. It was later proved that, with this parametrization, gradient descent minimizes the nuclear norm under the assumption that the sensing matrices $\mathbf{A}_j$'s commute [54], or when the sensing matrices satisfy a restricted isometry property [74]. Gradient descent with low-rank initialization within a specific "capture neighborhood" has been studied in [39], which ensures that the iterates of the algorithm stay low-rank. When $\mathbf{X}^\star$ is a general rectangular matrix, implicit regularization in matrix sensing has been studied through the lenses of deep matrix factorization in [6, 50, 75, 96], and empirical and theoretical evidence is provided which suggests that a notion of rank-minimization is involved in the implicit bias of gradient descent. However, these works do not establish an explicit characterization of the limiting point of the optimization algorithm, e.g. in the form of a quantity that is minimized among all minimizers of the empirical risk. In the special case of full-observation matrix sensing, gradient flow has been shown to learn solutions with gradually increasing rank [49].

Most theoretical results on implicit regularization in matrix sensing consider the continuous-time dynamics (i.e. gradient flow), e.g. [6, 39, 50, 54, 75, 96], or assume an infinitesimally small initialization, e.g. [54, 75]. Notable exceptions are [49], which

makes a commutativity assumption on the data matrices in discrete time, and [74], which assumes a restricted isometry property that leads to a suboptimal sample complexity when the sensing matrices belong to a general class of random matrices [98]. In the context of linear neural networks, the dependence of the implicit bias of gradient descent on the initialization has been studied in [11, 87, 126, 136]. For instance, linear diagonal networks with shared weights were considered in [126], where it was shown that gradient descent minimizes a quantity which corresponds to the ℓ_1-norm in the limit $\alpha \to 0$ and to the (weighted) ℓ_2-norm in the limit $\alpha \to \infty$, where α denotes the initialization size. This result was later generalized in [136] using a tensor formulation, which allows for architectures including linear diagonal networks and linear full-length convolutional networks. However, these results focus on vector-based notions of norm-like functions that do not capture matrix-based quantities typically of interest in matrix sensing.

4.1.1 Our contributions

We study the implicit bias of discrete-time mirror descent in matrix sensing in both the general case of rectangular matrices and the particular case of positive semidefinite matrices. Under the only assumption on the sensing matrices $\mathbf{A}_j$'s that there exists a matrix achieving zero training error, we characterize the limiting point as the matrix that minimizes a quantity which interpolates between the nuclear norm and Frobenius norm, parametrized by the mirror map parameter, in the rectangular case; and which is a linear combination of the nuclear norm and the negative von Neumann entropy, parametrized by the initialization size, in the positive semidefinite case. Compared to results on implicit regularization for gradient descent, our framework for mirror descent is simple, and the same analysis yields results for both the case of general rectangular and positive semidefinite matrices.

In the general case of rectangular matrices, we show that mirror descent, initialized at zero and equipped with the spectral hypentropy mirror map [48] parametrized by $\beta > 0$, among all global minimizers of the empirical risk, converges to a matrix that minimizes a quantity interpolating between the nuclear norm in the limit $\beta \to 0$ and the Frobenius norm in the limit $\beta \to \infty$. As a consequence, our result implies that, for $\beta \to 0$, mirror descent can recover a rank-r matrix $\mathbf{X}^\star$ under the same set of assumptions that is

sufficient for nuclear norm minimization to be successful, namely when the sensing matrices $\mathbf{A}_j$'s satisfy the restricted isometry property with restricted isometry constant smaller than some absolute constant [98], or if $\mathbf{X}^\star$ satisfies an incoherence condition and $r(n + n')$ (modulo constants and logarithmic term) random entries of $\mathbf{X}^\star$ are observed [97]. To the best of our knowledge, this is the first recovery guarantee for an implicit regularization-based algorithm that does not explicitly enforce low-rankness in general rectangular matrix sensing.

In the particular case of positive semidefinite matrices, we can alternatively consider the spectral entropy mirror map and show that mirror descent, initialized at $\alpha\mathbf{I}$ for any $\alpha > 0$, converges to a positive semidefinite matrix that minimizes a linear combination of the nuclear norm and the negative von Neumann entropy, where the relative weights are controlled by the initialization size α. While the limit $\alpha \to 0$ corresponds to minimizing the nuclear norm as in the case of rectangular matrices, the limit $\alpha \to \infty$ corresponds to *maximizing* the nuclear norm. This also translates into guaranteed recovery of a low-rank matrix $\mathbf{X}^\star$ for $\alpha \to 0$ under the same assumptions that are sufficient for nuclear norm minimization [97, 98]. A comparable result for gradient descent with full-rank factorized parametrization has been established in [74] under a stronger assumption on the restricted isometry constant, which is assumed to be smaller than a quantity depending on both the rank and the condition number of the matrix $\mathbf{X}^\star$ and translates into a sub-optimal sample complexity.

We establish our results using a potential-based analysis for mirror descent in terms of the Bregman divergence, which provides an alternative proof technique to characterize the limiting point of mirror descent compared to the analysis based on KKT optimality conditions used in [53]. As a by-product, our proof of Theorem 4.1 yields an alternative proof of Theorem 1 in [53]. The advantage of our approach is that convergence of mirror descent does not need to be assumed a-priori and can instead be established using convexity of the empirical risk. The analysis in terms of the Bregman divergence is not limited to convex settings and has been applied to non-convex problems, e.g. [128, 130, 142], and hence can be of more general interest to investigate the phenomenon of implicit bias.

In the case of square matrices, we show, assuming that the sensing matrices $\mathbf{A}_j$'s are symmetric and commute, that gradient descent with full-rank parametrization $\mathbf{X} = \mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top$, $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n\times n}$, is a first-order approximation to mirror descent equipped with the spectral hypentropy, where the initialization size corresponds to the mirror map parameter β. A similar connection between mirror descent and reparametrized gradient descent has been established in the vector-case [3, 48, 114, 128]. Similarly, we show that gradient descent with full-rank parametrization $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$, $\mathbf{U} \in \mathbb{R}^{n\times n}$, is a first-order approximation to mirror descent equipped with the spectral entropy when the sensing matrices are symmetric and commute, thus recovering a generalization of Theorem 1 in [54] which holds for any positive initialization size $\alpha > 0$ (rather than in the limit $\alpha \to 0$).

We present numerical simulations which suggest that, in some regimes, our results on the dependence on the initialization size α for mirror descent might be indicative for the behavior of gradient descent, even when the sensing matrices $\mathbf{A}_j$'s do not commute. More precisely, the final estimates of gradient descent and mirror descent closely track each other when the number of measurements is sufficiently large for nuclear norm minimization to recover a planted low-rank matrix, while gradient descent seems to put more emphasis on lowering the effective rank [99] at the expense of a higher nuclear norm when fewer measurements are available, which supports the empirical observations in [6, 75].

4.2 Background

We begin by introducing some notation used throughout this paper. We use boldface uppercase letters to denote matrices and boldface lowercase letters to denote vectors. We write $\|\cdot\|_*$, $\|\cdot\|_F$ and $\|\cdot\|_2$ for the nuclear, Frobenius and spectral norm, respectively, and denote by $\langle \mathbf{X}, \mathbf{Y}\rangle = \operatorname{tr}(\mathbf{X}^\top\mathbf{Y})$ the standard Frobenius inner product. We write $\mathcal{S}^n, \mathcal{S}^n_+ \subseteq \mathbb{R}^{n\times n}$ for the set of symmetric and positive semidefinite matrices, respectively. Without loss of generality, we will always assume $n \leq n'$.

We first give a brief overview of unconstrained mirror descent (with matrix arguments).

Let $\mathcal{D} \subseteq \mathbb{R}^{n\times n'}$ be an open convex set. We say that $\Phi : \mathcal{D} \to \mathbb{R}$ is a *mirror map*

if it is strictly convex, differentiable, and its gradient takes all possible values, i.e. $\{\nabla\Phi(\mathbf{X}) : \mathbf{X} \in \mathcal{D}\} = \mathbb{R}^{n\times n'}$.

Given a mirror map Φ, the associated *Bregman divergence* is defined as

$$D_\Phi(\mathbf{X}, \mathbf{Y}) = \Phi(\mathbf{X}) - \Phi(\mathbf{Y}) - \langle \nabla\Phi(\mathbf{Y}), \mathbf{X} - \mathbf{Y}\rangle. \tag{4.1}$$

Then, the mirror descent algorithm to minimize a function $f : \mathcal{D} \to \mathbb{R}$ is defined by the update

$$\nabla\Phi(\mathbf{X}^{t+1}) = \nabla\Phi(\mathbf{X}^t) - \eta_t \nabla f(\mathbf{X}^t), \tag{4.2}$$

where $\eta_t > 0$ is a sequence of step sizes. We approach the problem of matrix sensing, where we are given measurements $\{\mathbf{A}_j, y_j\}_{j=1}^m$, by minimizing the unregularized empirical risk

$$f(\mathbf{X}) = \frac{1}{2m}\sum_{j=1}^m \Big(\langle \mathbf{A}_j, \mathbf{X}\rangle - y_j\Big)^2 \tag{4.3}$$

using mirror descent equipped with the spectral hypentropy mirror map, which is defined as [48]

$$\Phi_\beta(\mathbf{X}) = \sum_{i=1}^n \sigma_i \operatorname{arcsinh}\left(\frac{\sigma_i}{\beta}\right) - \sqrt{\sigma_i^2 + \beta^2}, \tag{4.4}$$

for some $\beta > 0$, where $\{\sigma_i\}_{i=1}^n$ denote the singular values of $\mathbf{X}$. We provide expressions for the gradient $\nabla\Phi_\beta$ and discuss the per-iteration computational cost of the corresponding mirror descent algorithm in Appendix 4.8.1.

If the optimization is restricted to positive semidefinite matrices $\mathbf{X} \in \mathbb{R}^{n\times n}$, we can replace $\mathbb{R}^{n\times n'}$ with $\mathcal{S}^n$ in above definitions and consider the spectral entropy (using the convention $0 \log 0 = 0$)

$$\Phi(\mathbf{X}) = \operatorname{tr}(\mathbf{X}\log\mathbf{X} - \mathbf{X}), \tag{4.5}$$

which is well-defined on the set of positive semidefinite matrices $\mathcal{S}_+^n$.

4.3 Algorithmic regularization of mirror descent

In this section, we show that mirror descent, equipped with the spectral hypentropy mirror map (4.4) in the rectangular case and the spectral entropy mirror map (4.5) in the positive semidefinite case, converges to a global minimizer of the empirical risk f that minimizes a quantity which interpolates between the nuclear norm and the

Frobenius norm in the rectangular case, and is a linear combination of the nuclear norm and the negative von Neumann entropy $\operatorname{tr}(\mathbf{X}\log\mathbf{X}) = \sum_{i=1}^n \lambda_i \log \lambda_i$ in the positive semidefinite case, where $\{\lambda_i\}_{i=1}^n$ denote the eigenvalues of the matrix $\mathbf{X} \in \mathcal{S}_+^n$.

The following Theorem applies to matrix sensing in the case of general rectangular matrices.

Theorem 4.1 (Rectangular case)**.** *Consider the mirror descent algorithm* (4.2) *with mirror map* (4.4) *with parameter* $\beta > 0$ *and initialization* $\mathbf{X}^0 = \mathbf{0}$*. Suppose that the step sizes satisfy* $\eta_t \equiv \eta \leq c$*, where* $c > 0$ *is a constant depending on the sensing matrices* $\mathbf{A}_j$*'s, observations* y_j*'s and parameter* β*, and that there exists a matrix* $\mathbf{X}'$ *satisfying* $f(\mathbf{X}') = 0$*. Then, mirror descent converges to a matrix* $\mathbf{X}^\infty = \lim_{t\to\infty} \mathbf{X}^t$ *which, among all global minimizers of* f*, minimizes*

$$\sum_{i=1}^n \sigma_i \log \frac{1}{\beta} + \sigma_i \log\Big(\sigma_i + \sqrt{\sigma_i^2 + \beta^2}\Big) - \sqrt{\sigma_i^2 + \beta^2}, \tag{4.6}$$

where $\{\sigma_i\}_{i=1}^n$ *denote the singular values of* $\mathbf{X}^\infty$*. We have, for any* $t \geq 0$*,*

$$f(\mathbf{X}^t) \leq \frac{D_{\Phi_\beta}(\mathbf{X}^\infty, \mathbf{X}^0)}{\eta t}. \tag{4.7}$$

Proof sketch. The following identity is key to our proof and characterizes how the Bregman divergence between any reference point $\mathbf{X}'$ and the mirror descent iterates $\mathbf{X}^t$ evolves:

$$D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^{t+1}) - D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^t) = -\eta\langle \nabla f(\mathbf{X}^t), \mathbf{X}^t - \mathbf{X}'\rangle + D_{\Phi_\beta}(\mathbf{X}^t, \mathbf{X}^{t+1}), \tag{4.8}$$

which follows from the definition of the Bregman divergence (4.1) and the mirror descent update (4.2). Letting $\mathbf{X}'$ be any global minimizer of f, we can compute $\langle \nabla f(\mathbf{X}^t), \mathbf{X}^t - \mathbf{X}'\rangle = 2f(\mathbf{X}^t)$, where we used the assumption that there exists a matrix $\mathbf{X}'$ achieving zero training error $f(\mathbf{X}') = 0$. Using the strong convexity of the spectral hypentropy mirror map, we can bound $D_\Phi(\mathbf{X}^t, \mathbf{X}^{t+1})$ to show

$$D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^{t+1}) - D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^t) \leq -\eta f(\mathbf{X}^t),$$

for any global minimizer $\mathbf{X}'$ of f. Since the Bregman divergence $D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^t)$ is bounded from below by zero, this means that the empirical risk $f(\mathbf{X}^t)$ must converge to zero, which in turn implies that $\mathbf{X}^t$ converges to a global minimizer of f.

To see *which* global minimizer mirror descent converges to, observe that the difference in (4.8) does not depend on the reference point $\mathbf{X}'$, as long as $\mathbf{X}'$ is a global minimizer of f. This means that the Bregman divergence $D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^t)$ is decreased by the same amount for *all* global minimizers $\mathbf{X}'$, which then implies that $\mathbf{X}^t$ must converge to the global minimizer which is closest to $\mathbf{X}^0$ in terms of the Bregman divergence. From this observation it follows that $\mathbf{X}^\infty = \lim_{t\to\infty} \mathbf{X}^t$ minimizes the quantity in (4.6) among all global minimizers of f.

The bound (4.7) on the empirical risk of the last iterate $f(\mathbf{X}^t)$ can be shown using the smoothness of the empirical risk f and the strong convexity of the spectral hypentropy mirror map. A detailed proof of Theorem 4.1 can be found in Appendix 4.8.2. □

We remark that the step size η can be chosen independently from the parameter β if β is chosen from some interval bounded away from infinity, e.g. $\beta \in (0, 1)$.

An analogous result holds for mirror descent equipped with the spectral entropy mirror map (4.5) when optimizing over positive semidefinite matrices.

Theorem 4.2 (Positive semidefinite case)**.** *Consider the mirror descent algorithm* (4.2) *with mirror map* (4.5) *and initialization* $\mathbf{X}^0 = \alpha\mathbf{I}$ *for some* $\alpha > 0$*. Suppose that the step sizes satisfy* $\eta_t \equiv \eta \le c$*, where* $c > 0$ *is a constant depending on the sensing matrices* $\mathbf{A}_j$*'s and observations* y_j*'s, and that there exists a positive semidefinite matrix* $\mathbf{X}'$ *satisfying* $f(\mathbf{X}') = 0$*. Then, mirror descent converges to a positive semidefinite matrix* $\mathbf{X}^\infty = \lim_{t\to\infty} \mathbf{X}^t$ *which, among all global minimizers of* f*, minimizes*

$$\sum_{i=1}^{n} \left(\log \frac{1}{\alpha} - 1 \right) \lambda_i + \lambda_i \log \lambda_i, \tag{4.9}$$

where $\{\lambda_i\}_{i=1}^n$ *denote the eigenvalues of* $\mathbf{X}^\infty$*. We have, for any* $t \ge 0$*,*

$$f(\mathbf{X}^t) \le \frac{D_\Phi(\mathbf{X}^\infty, \mathbf{X}^0)}{\eta t}. \tag{4.10}$$

Theorem 4.2 can be proved the same way as Theorem 4.1, see Appendix 4.8.3. Neither Theorem requires any assumptions on the sensing matrices $\mathbf{A}_j$'s beyond the existence of a matrix $\mathbf{X}'$ achieving zero training error $f(\mathbf{X}') = 0$, which, for instance, is satisfied when $\{\mathbf{A}_j\}_{j=1}^m$ are linearly independent and $m \le nn'$ in the rectangular case or $m \le n(n+1)/2$ in the positive semidefinite case.

On the implicit bias of mirror descent. The implicit bias of mirror descent in linear models has previously been studied using KKT optimality conditions in [53]. Once we have established convergence of mirror descent towards a global minimizer of f, we could have alternatively invoked Theorem 1 of [53] to characterize the limiting point of mirror descent. Instead, our analysis of the Bregman divergence $D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^t)$ reveals that each iteration of the mirror descent algorithm (4.2) decreases $D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^t)$ by the same amount for all global minmimizers $\mathbf{X}'$ of f, provided that

$$\langle \nabla f(\mathbf{X}^t), \mathbf{X}^t - \mathbf{X}' \rangle = \langle \nabla f(\mathbf{X}^t), \mathbf{X}^t - \mathbf{X}'' \rangle$$

for any two global minimizers $\mathbf{X}'$ and $\mathbf{X}''$ of f. From this observation it immediately follows that mirror descent converges to the global minimizer of f that minimizes the quantity $D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^0)$ among all global minimizers of f. Since this equality is satisfied under the assumptions of Theorem 1 in [53], our analysis presents an alternative proof technique for Theorem 1 in [53] that does not use KKT optimality conditions and hence does not need to assume convergence of mirror descent.

On the mirror map parameter and initialization size. In the rectangular case, minimizing the quantity in (4.6) is equivalent to minimizing the nuclear norm in the limit $\beta \to 0$, and equivalent to minimizing the Frobenius norm in the limit $\beta \to \infty$, see Appendix 4.8.7 for further details. This can be seen as a matrix analogue of the result in [126], which shows for linear diagonal networks with shared weights that gradient descent minimizes a quantity interpolating between the ℓ_1 and (weighted) ℓ_2 norms. For a network with two layers, this architecture corresponds to the parametrization $\mathbf{x} = \mathbf{u} \odot \mathbf{u} - \mathbf{v} \odot \mathbf{v}$, where $\odot$ denotes the elementwise Hadamard product and with which gradient descent has been shown to be a first-order approximation to mirror descent equipped with the hypentropy mirror map in the vector-case, see e.g. [114, 128]. In the positive semidefinite case, minimizing the quantity in (4.9) also corresponds to minimizing the nuclear norm in the limit $\alpha \to 0$, while in the limit $\alpha \to \infty$ the coefficient $\log(1/\alpha) - 1$ tends to minus infinity, and minimizing the quantity in (4.9) is equivalent to maximizing the nuclear norm.

4.4 The estimation problem

In this section, we consider the estimation problem of reconstructing a rank-r matrix $\mathbf{X}^\star$ from a set of linear measurements $y_j = \langle \mathbf{A}_j, \mathbf{X}^\star \rangle$, $j = 1, \dots, m$. Theorem 4.1 and Theorem 4.2 imply that mirror descent, equipped with the spectral hypentropy mirror map (4.4) or spectral entropy mirror map (4.5), approximately minimizes the nuclear norm for a small mirror map parameter β or a small initialization size α. Hence, mirror descent, a first-order algorithm minimizing the unregularized empirical risk, can recover a low-rank matrix $\mathbf{X}^\star$ under the same set of assumptions which is sufficient for nuclear norm minimization to be successful.

4.4.1 Matrix sensing with restricted isometry property

The following restricted isometry property has been shown to be sufficient to guarantee recovery of low-rank matrices using nuclear norm minimization [98], and has also been used in [74] to show that gradient descent with full-rank factorized parametrization recovers $\mathbf{X}^\star$.

Definition 4.1 (Restricted isometry property [98])**.** *The matrices $\mathbf{A}_1, \dots, \mathbf{A}_m \in \mathbb{R}^{n \times n'}$ satisfy (r, δ)-restricted isometry property (RIP) if for any matrix $\mathbf{X} \in \mathbb{R}^{n \times n'}$ with rank at most r, we have*

$$(1 - \delta)\|\mathbf{X}\|_F \le \left(\frac{1}{m} \sum_{j=1}^{m} \langle \mathbf{A}_j, \mathbf{X} \rangle^2 \right)^{1/2} \le (1 + \delta)\|\mathbf{X}\|_F.$$

With this definition, we have the following recovery guarantee for mirror descent.

Theorem 4.3. *Assume that the set of measurement matrices $\{\mathbf{A}_j\}_{j=1}^m$ satisfies $(5r, \delta)$-restricted isometry property with $\delta \le 0.1$. Then, the mirror descent algorithm described in Theorem 4.1 with parameter $\beta < \frac{\|\mathbf{X}^\star\|_*}{1.05en}$ converges to a matrix $\mathbf{X}^\infty = \lim_{t \to \infty} \mathbf{X}^t$ that satisfies*

$$\|\mathbf{X}^\infty - \mathbf{X}^\star\|_F \le \frac{\Delta_\beta \|\mathbf{X}^\star\|_* + (1 + \Delta_\beta) \frac{n\beta}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}}{C_\delta \sqrt{3r}}, \tag{4.11}$$

where $C_\delta = \frac{1}{2}(1 - \sqrt{\frac{2}{3}} - \delta(1 + \sqrt{\frac{2}{3}}))$ and $\Delta_\beta = (\frac{\log(\|\mathbf{X}^\star\|_*/\beta) - 1}{\log(1.05n)} - 1)^{-1}$.

If, additionally, $n = n'$ and $\mathbf{X}^\star$ is positive semidefinite, then the mirror descent algorithm described in Theorem 4.2 with $\alpha < \frac{\|\mathbf{X}^\star\|_*}{en}$ converges to a positive semidefinite

matrix $\mathbf{X}^\infty$ *that satisfies*

$$\|\mathbf{X}^\infty - \mathbf{X}^\star\|_F \leq \frac{\Delta_\alpha \|\mathbf{X}^\star\|_*}{C_\delta \sqrt{3r}}, \tag{4.12}$$

where $\Delta_\alpha = (\frac{\log(\|\mathbf{X}^\star\|_*/\alpha)-1}{\log n} - 1)^{-1}$.

Among the existing results on implicit regularization in matrix sensing, the result that is perhaps most closely related to Theorem 4.3 is Theorem 1.1 in [74], which considers the positive semidefinite case of matrix sensing with full-rank parametrization $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$, $\mathbf{U} \in \mathbb{R}^{n\times n}$, and shows that gradient descent recovers $\mathbf{X}^\star$ under the assumption of restricted isometry with constant $\delta \leq c/(\kappa^3\sqrt{r}\log^2 n)$, where κ is the condition number of $\mathbf{X}^\star$ and $c > 0$ is an absolute constant.

The analysis that leads to Theorem 4.3 differs significantly from the proof of Theorem 1.1 in [74]. The proof of Theorem 1.1 in [74] involves an analysis of the trajectory of gradient descent, relating it to the population dynamics defined by the population risk $\mathbb{E}_{(\mathbf{A}_j)_{kl}\sim\mathcal{N}(0,1)}[f(\mathbf{U}^t(\mathbf{U}^t)^\top)] = \|\mathbf{U}^t(\mathbf{U}^t)^\top - \mathbf{X}^\star\|_F^2$, and identifying adaptive rank-r subspaces to which the iterates $\mathbf{U}^t$ are approximately confined. On the other hand, Theorem 4.3 follows from Theorem 4.1 and Theorem 4.2, which rely on a simple analysis of the evolution of the Bregman divergence, and employs arguments similar to the ones used in [98] to prove that RIP is sufficient to guarantee that nuclear norm minimization recovers $\mathbf{X}^\star$. Further, our analysis applies both to the case of rectangular and positive semidefinite matrices, while the results in [74] only hold for the full-rank parametrization $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$, $\mathbf{U} \in \mathbb{R}^{n\times n}$, for which an extension to general rectangular matrices does not seem to be trivial, see e.g. [62, 79, 107, 112, 141].

Sample complexity and condition number. The restricted isometry property assumption in Theorem 1.1 in [74] requires a restricted isometry constant $\delta \leq c/(\kappa^3\sqrt{r}\log^2 n)$ for an absolute constant $c > 0$, which leads to a sample complexity of $\mathcal{O}(\kappa^6 nr^2 \log^5 n)$ if the sensing matrices $\mathbf{A}_j$'s belong to a general class of random matrices [98]. It is conjectured in [74] that the dependence of δ on the rank r and condition number κ is not tight, and that δ only needs to be smaller than some absolute constant. On the other hand, Theorem 4.3 only requires $\delta \leq 0.1$, which is the same assumption that is sufficient to guarantee that nuclear norm minimization recovers low-rank matrices [98] and leads to a sample complexity of $\mathcal{O}((n+n')r\log(nn'))$. To

the best of our knowledge, our result for mirror descent is the first recovery guarantee for an implicit regularization-based algorithm for (rectangular) matrix sensing that only requires the same assumptions as nuclear norm minimization [98].

Convergence speed and dependence on intialization size. The analysis along the trajectory of gradient descent in [74] allows to establish convergence speed guarantees and a polynomial dependence of the estimation error $\|\mathbf{U}^t(\mathbf{U}^t)^\top - \mathbf{X}^\star\|_F$ on the initialization size α. On the other hand, we have no convergence speed guarantees for mirror descent (beyond the bounds on the empirical risk (4.7) and (4.10)), and the final estimation error $\|\mathbf{X}^\infty - \mathbf{X}^\star\|_F$ depends logarithmically on the parameters α and β, which can lead to an unpractically small value for α and β being required to reach some desired accuracy ε. Nonetheless, our result guarantees exact recovery of $\mathbf{X}^\star$ in the limit $\alpha, \beta \to 0$, which is a setting often considered in the literature, e.g. [54, 75, 126].

We remark that, beyond the aforementioned result on implicit regularization in matrix sensing [74], similar recovery results that include convergence speed guarantees and a polynomial dependence on the initialization size have also been established for implicit regularization-based algorithms in the context of sparse linear regression [113] and sparse phase retrieval [128, 130]. However, these results all require a sample complexity that scales quadratically in the respective notions of sparsity, and we leave it to future work to investigate whether it is possible to establish convergence guarantees that include a convergence speed analysis and depend polynomially on α and β when a sample complexity that scales only linearly in the rank r of $\mathbf{X}^\star$ is assumed.

4.4.2 Matrix completion

In matrix completion, a subset of the entries of the matrix $\mathbf{X}^\star$ is observed, and the corresponding sensing matrices $\mathbf{A}_j$'s do not satisfy the restricted isometry property. Instead, an incoherence condition together with a sufficient number of randomly observed entries have been used to guarantee recovery for nuclear norm minimization [97] and gradient descent [80], for instance.

Definition 4.2 (Coherence [24])**.** *Let $U \subseteq \mathbb{R}^n$ be a linear subspace of dimension r and $\mathbf{P}_U$ the orthogonal projection onto U. The coherence of U is defined as (here $\{\mathbf{e}_i\}_{i=1}^n$*

denotes the canonical basis)

$$\mu(U) = \frac{n}{r} \max_{1 \leq i \leq n} \|\mathbf{P}_U \mathbf{e}_i\|_2^2.$$

We have the following recovery guarantee for mirror descent under the same assumptions as in [97].

Theorem 4.4. *Let $\mathbf{X}^\star \in \mathbb{R}^{n \times n'}$ be a rank-r matrix with (compact) singular value decomposition $\mathbf{X}^\star = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n \times r}$, $\mathbf{V} \in \mathbb{R}^{n' \times r}$ and $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$. Assume that:*

A1 *The row and column spaces of $\mathbf{X}^\star$ have coherences bounded by $\mu_0 > 0$.*

A2 *The matrix $\mathbf{U}\mathbf{V}^\top$ has maximum entry bounded by $\mu_1 \sqrt{r/nn'}$ in absolute value for some $\mu_1 > 0$.*

Suppose that we observe m entries of $\mathbf{X}^\star$ with locations sampled uniformly at random. Then, if $m \geq 32c \max\{\mu_0^2, \mu_1\} r(n+n') \log^2(2n')$ for some $c > 1$, the mirror descent algorithm described in Theorem 4.1 with parameter $\beta < \frac{\|\mathbf{X}^\star\|_*}{1.05en}$ converges to a matrix $\mathbf{X}^\infty = \lim_{t \to \infty} \mathbf{X}^t$ that satisfies

$$\|\mathbf{X}^\infty - \mathbf{X}^\star\|_F \leq 6 \left(\Delta_\beta \|\mathbf{X}^\star\|_* + (1 + \Delta_\beta) \frac{n\beta}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1} \right) \left(1 + \left(\frac{128cnn' \log^2 n'}{9m} \right)^{\frac{1}{2}} \right), \tag{4.13}$$

with probability $1 - 6\log(n')(n+n')^{2-2c} - (n')^{2-2\sqrt{c}}$, where $\Delta_\beta = (\frac{\log(\|\mathbf{X}^\star\|_*/\beta) - 1}{\log(1.05n)} - 1)^{-1}$.

If, additionally, $n = n'$ and $\mathbf{X}^\star$ is positive semidefinite, then the mirror descent algorithm described in Theorem 4.2 with $\alpha < \frac{\|\mathbf{X}^\star\|_*}{en}$ converges to a positive semidefinite matrix $\mathbf{X}^\infty$ that satisfies

$$\|\mathbf{X}^\infty - \mathbf{X}^\star\|_F \leq 6\Delta_\alpha \|\mathbf{X}^\star\|_* \left(1 + \left(\frac{128cn^2 \log^2 n}{9m} \right)^{\frac{1}{2}} \right), \tag{4.14}$$

where $\Delta_\alpha = (\frac{\log(\|\mathbf{X}^\star\|_*/\alpha) - 1}{\log n} - 1)^{-1}$.

To the best of our knowledge, there are no recovery guarantees in matrix completion for an unregularized empirical risk minimization-based algorithm that does not explicitly enforce the low-rank constraint. For instance, when $\mathbf{X}^\star \in \mathbb{R}^{n \times n}$ is positive semidefinite, Theorem 2 in [80] establishes a recovery guarantee for gradient descent with low-rank parametrization $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$, where $\mathbf{U} \in \mathbb{R}^{n \times r}$, applied to the unregularized empirical risk with a sample requirement of order $r^3 n \log^3 n$.

4.5 Connection with gradient descent

In the vector-case, it has been established that gradient descent with parametrization $\mathbf{x} = \mathbf{u} \odot \mathbf{u} - \mathbf{v} \odot \mathbf{v}$ is a first-order approximation to mirror descent equipped with the hypentropy mirror map [48, 114, 128], and a general framework connecting mirror descent and reparametrized gradient descent has been studied in [3]. A natural question is whether such a connection also extends to the matrix-case.

In the following, we consider matrix sensing with square symmetric sensing matrices $\mathbf{A}_j \in \mathbb{R}^{n \times n}$, $j = 1, \dots, m$. Consider the following exponentiated gradient algorithm given by

$$\begin{gathered} \mathbf{X}^t = \mathbf{U}^t - \mathbf{V}^t \\ \mathbf{U}^{t+1} = \frac{\mathbf{U}^t e^{-\eta \nabla f(\mathbf{X}^t)} + e^{-\eta \nabla f(\mathbf{X}^t)} \mathbf{U}^t}{2}, \qquad \mathbf{V}^{t+1} = \frac{\mathbf{V}^t e^{\eta \nabla f(\mathbf{X}^t)} + e^{\eta \nabla f(\mathbf{X}^t)} \mathbf{V}^t}{2}. \end{gathered} \tag{4.15}$$

When considering the full-rank parametrization $\mathbf{X} = \mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top$, $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n \times n}$, gradient descent on the variables $(\mathbf{U}, \mathbf{V})$ is a first-order approximation to the exponentiated gradient algorithm defined in (4.15), with the step size rescaled by a constant factor and the approximation being exact in the limit $\eta \to 0$, see Appendix 4.8.7 for details. The gradient descent algorithm considered in [6, 54, 74, 75] can be obtained by initializing $\mathbf{V}^0 = \mathbf{0}$.

Proposition 4.5. *Assume that the sensing matrices $\mathbf{A}_j$'s are symmetric and commute. Then:*

1. *Mirror descent equipped with the spectral entropy mirror map* (4.5) *and any positive definite initialization $\mathbf{X}^0$ which commutes with all $\mathbf{A}_j$'s (e.g. $\mathbf{X}^0 = \alpha \mathbf{I}$ for some $\alpha > 0$) is equivalent to the exponentiated gradient algorithm defined in* (4.15) *with initialization $\mathbf{U}^0 = \mathbf{X}^0$ and $\mathbf{V}^0 = \mathbf{0}$.*
2. *Mirror descent equipped with the spectral hypentropy mirror map* (4.4) *with parameter $\beta > 0$ and initialization $\mathbf{X}^0 = \mathbf{0}$ is equivalent to the exponentiated gradient algorithm defined in* (4.15) *with initialization $\mathbf{U}^0 = \mathbf{V}^0 = \frac{1}{2}\beta \mathbf{I}$.*

As a Corollary, we obtain a generalization of Theorem 1 in [54]. Let $\widetilde{f}(\mathbf{U}) = f(\mathbf{U}\mathbf{U}^\top)$.

Corollary 4.6. *Assume that the sensing matrices $\mathbf{A}_j$'s are symmetric and commute, and that there exists a $\mathbf{X}' \in \mathcal{S}_+^n$ satisfying $f(\mathbf{X}') = 0$. Then, the gradient flow defined*

by $\frac{d\mathbf{U}^t}{dt} = -\nabla \widetilde{f}(\mathbf{U}^t)$ *and any initialization satisfying* $\mathbf{U}^0(\mathbf{U}^0)^\top = \alpha \mathbf{I}$ *converges to a matrix* $\mathbf{U}^\infty$ *minimizing*

$$\sum_{i=1}^{n} \left(\log \frac{1}{\alpha} - 1 \right) \lambda_i + \lambda_i \log \lambda_i$$

among all global minimizers of $\widetilde{f}$*, where* $\{\lambda_i\}_{i=1}^n$ *denote the eigenvalues of the matrix* $\mathbf{U}^\infty(\mathbf{U}^\infty)^\top$*.*

This result generalizes Theorem 1 in [54] in two ways: first, we obtain a precise characterization of the quantity that is minimized for any initialization size $\alpha > 0$, which indeed coincides with the nuclear norm in the limit $\alpha \to 0$. Second, convergence to a global minimizer is assumed in Theorem 1 in [54], which is non-trivial a-priori, since the optimization problem in $\mathbf{U}^t$ is non-convex. On the other hand, we show convergence of mirror descent in Theorem 4.2, which then carries over to gradient flow on the non-convex objective $\widetilde{f}$ via Proposition 4.5, assuming that the $\mathbf{A}_j$'s are symmetric and commute.

4.6 Numerical simulations

In this section, we present numerical simulations examining the dependence of the final estimates of mirror descent equipped with the spectral entropy (4.5) and of gradient descent with full-rank parametrization $\mathbf{X} = \mathbf{U}\mathbf{U}^\top$, $\mathbf{U} \in \mathbb{R}^{n \times n}$, on the initialization size for random Gaussian sensing matrices $\mathbf{A}_1, \dots, \mathbf{A}_m$. We evaluate the nuclear norm $\|\mathbf{X}\|_*$, the reconstruction error $\|\mathbf{X} - \mathbf{X}^\star\|_F$, and the effective rank [99] $\mathrm{effrank}(\mathbf{X}) = \exp(-\sum_{i=1}^n p_i \log p_i)$, where $p_i = \sigma_i / \|\mathbf{X}\|_*$, $i = 1, \dots, n$, denote the normalized singular values of $\mathbf{X}$. Numerical simulations for matrix completion are provided in Appendix 4.8.8 and yield similar results as for random Gaussian sensing matrices.

Our experimental setup is as follows. We generate a rank-r positive semidefinite matrix by sampling a random matrix $\mathbf{U}^\star \in \mathbb{R}^{n \times r}$ with i.i.d. $\mathcal{N}(0,1)$ entries, setting $\mathbf{X}^\star = \mathbf{U}^\star(\mathbf{U}^\star)^\top$ and normalizing $\|\mathbf{X}^\star\|_* = 1$. We generate m symmetric sensing matrices $\mathbf{A}_j = \frac{1}{2}(\mathbf{B}_j + \mathbf{B}_j^\top)$, where the entries of $\mathbf{B}_j$ are i.i.d. $\mathcal{N}(0,1)$. We run mirror descent and gradient descent with initialization $\mathbf{X}^0 = \alpha \mathbf{I}$ and constant step sizes $\mu = 1$ and $\mu = 0.25$, respectively, for $T = 5000$ iterations, and vary the initialization size α from 10^{-1} to 10^{-10}. For reference, we also include the ground truth $\mathbf{X}^\star$ and the estimate

obtained from minimizing the nuclear norm, $\operatorname{argmin}\{\|\mathbf{X}\|_* : \mathbf{X} \succeq \mathbf{0}, f(\mathbf{X}) = 0\}$, using the `cvxopt` package. The experiments for Figure 4.1 were implemented in Python 3.9 and took around 10 minutes on a machine with 1.1-GHz Intel Core i5 CPU and 8 GB of RAM.

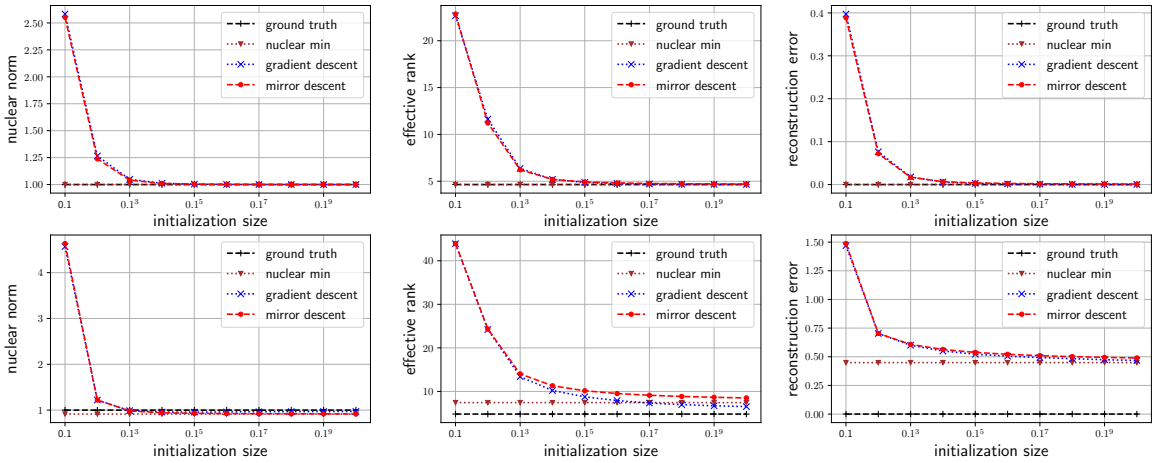


Figure 4.1: Nuclear norm, effective rank [99] and reconstruction error against initialization size α for $n = 50$ and $r = 5$. Top row: $m = 3nr$ measurements. Bottom row: $m = nr$ measurements.

With $m = 3nr$ measurements (Figure 4.1, top row), nuclear norm minimization coincides with $\mathbf{X}^\star$. In this case, our simulations show that the estimate of gradient descent closely tracks the estimate of mirror descent for all initialization sizes in terms of nuclear norm, effective rank [99] and reconstruction error, even though the sensing matrices do not commute. When we have $m = nr$ measurements (Figure 4.1, bottom row), nuclear norm minimization does not recover the planted matrix $\mathbf{X}^\star$, and our simulations show that gradient descent puts more emphasis on lowering the effective rank at the expense of a higher nuclear norm for initialization sizes smaller than 10^{-3}. Since Theorem 4.2 guarantees that mirror descent minimizes the nuclear norm in the limit $\alpha \to 0$, regardless of the number of measurements, this is in line with the observations in [6, 75], which suggest that a notion of rank-minimization is involved in the implicit bias of gradient descent.

4.7 Conclusion

In this paper, we analyzed discrete-time mirror descent for matrix sensing, equipped with the spectral hypentropy mirror map in the case of general rectangular matrices

and equipped with the spectral entropy mirror map in the particular case of positive semidefinite matrices. We showed that mirror descent minimizes a quantity that interpolates between the nuclear norm and Frobenius norm in the rectangular case, and is a linear combination of the nuclear norm and the negative von Neumann entropy in the positive semidefinite case. We used this result to show that mirror descent, a first-order algorithm minimizing the unregularized empirical risk that does not explicitly enforce low-rankness of its iterates, can recover a low-rank matrix $\mathbf{X}^\star$ under the same set of assumptions which is sufficient for nuclear norm minimization to recover $\mathbf{X}^\star$ [97, 98].

A downside of mirror descent compared to gradient descent with full-rank factorized parametrization, which is an alternative implicit regularization-based algorithm for matrix sensing, is its computational cost: the spectral hypentropy requires a singular value decomposition in each iteration, while the spectral entropy requires computing a matrix exponential in each iteration, see Appendix 4.8.1. For general sensing matrices, the computational cost of mirror descent is of the same order as that of gradient descent, since a singular value decomposition takes $\mathcal{O}(n^2n')$ operations and matrix exponentials can be computed in $\mathcal{O}(n^3)$ operations, while evaluating the gradient $\nabla f(\mathbf{X}) = \frac{1}{m}\sum_{j=1}^m(\langle \mathbf{A}_j, \mathbf{X}\rangle - y_j)\mathbf{A}_j$ requires $\mathcal{O}(mnn')$ operations, and $m > n$ is typically required to recover $\mathbf{X}^\star$. However, the inner product $\langle \mathbf{A}_j, \mathbf{X}\rangle$ can be computed in $\mathcal{O}(1)$ operations in matrix completion, in which case the computational cost of gradient descent with full-rank factorized parametrization is dominated by the multiplication of two $n \times n$ matrices. When $r \ll \min\{n, n'\}$, neither implicit regularization-based approach achieves the computational efficiency of gradient descent with low-rank factorized parametrization $\mathbf{X} = \mathbf{U}\mathbf{V}^\top$, where $\mathbf{U} \in \mathbb{R}^{n\times r}$, $\mathbf{V} \in \mathbb{R}^{n'\times r}$, see e.g. [33, 80].

Previous results on implicit regularization-based algorithms in matrix sensing [74], sparse linear regression [113] and sparse phase retrieval [128, 130] establish convergence speed guarantees and a polynomial dependence on the initialization size α, while a sample complexity that scales quadratically in the respective notions of sparsity is required. On the other hand, our results for mirror descent in matrix sensing require a sample complexity that scales linearly in the rank r of $\mathbf{X}^\star$, but do not establish any convergence speed guarantees, and the bound on the estimation error depends

logarithmically on α and β. We leave bridging this gap, i.e. establishing convergence speed guarantees with a polynomial dependence on α and β while only assuming a linear sample complexity, to future work.

4.8 Chapter appendix

The Appendix is organized as follows. In Appendix 4.8.1, we provide the gradients for the spectral entropy (4.5) and spectral hypentropy (4.4) mirror maps, and discuss the per-iteration computational cost of the corresponding mirror descent algorithms. In Appendices 4.8.2-4.8.7, we provide proofs for the claims made in the main paper. In Appendix 4.8.8, we present additional experiments addressing the problem of matrix completion.

4.8.1 Mirror maps and gradients

We first consider rectangular matrix sensing with the spectral hypentropy mirror map

$$\Phi_\beta(\mathbf{X}) = \sum_{i=1}^{n} \sigma_i \operatorname{arcsinh}\left(\frac{\sigma_i}{\beta}\right) - \sqrt{\sigma_i^2 + \beta^2},$$

where $\{\sigma_i\}_{i=1}^n$ denote the singular values of $\mathbf{X}$. Since $\Phi_\beta : \mathbb{R}^{n\times n'} \to \mathbb{R}$ is a function operating on the singular values of a matrix, we can use Theorem 3.1 in [72] to compute its gradient. Let $\mathbf{X} = \mathbf{U}\operatorname{diag}(\sigma_1, \ldots, \sigma_n)\mathbf{V}^\top$ be the singular value decomposition of the matrix $\mathbf{X}$, where $\operatorname{diag}(\sigma_1, \ldots, \sigma_n)$ denotes the diagonal matrix with diagonal elements $\sigma_1, \ldots, \sigma_n$. Then, we have

$$\nabla\Phi_\beta(\mathbf{X}) = \mathbf{U}\operatorname{diag}\Big(\operatorname{arcsinh}\Big(\frac{\sigma_1}{\beta}\Big), \ldots, \operatorname{arcsinh}\Big(\frac{\sigma_n}{\beta}\Big)\Big)\mathbf{V}^\top$$

by Theorem 3.1 in [72], see also [48]. This means that each step of mirror descent requires a singular value decomposition to compute

$$\mathbf{X}^{t+1} = \nabla\Phi_\beta^{-1}\Big(\nabla\Phi_\beta(\mathbf{X}^t) - \eta\nabla f(\mathbf{X}^t)\Big),$$

which takes $\mathcal{O}(n^2n')$ operations.

The singular value decomposition can be avoided if $n = n'$ and the sensing matrices $\mathbf{A}_j$'s are symmetric, which we can assume without loss of generality if $\mathbf{X}^\star \in \mathbb{R}^{n\times n}$ is symmetric, since then $y_j = \langle \mathbf{A}_j, \mathbf{X}^\star\rangle = \langle \frac{1}{2}(\mathbf{A}_j + \mathbf{A}_j^\top), \mathbf{X}^\star\rangle$ for all $j = 1, \ldots, m$. In

that case, the mirror descent iterates $\mathbf{X}^t$ stay symmetric for all $t \geq 0$, provided the initialization $\mathbf{X}^0$ is symmetric. Using the identity $\operatorname{arcsinh}(x) = \log(x + \sqrt{x^2+1})$, we can write

$$\Phi_\beta(\mathbf{X}^t) = \operatorname{tr}\left(\mathbf{X}^t \log\left(\frac{\mathbf{X}^t}{\beta} + \sqrt{\frac{(\mathbf{X}^t)^2}{\beta^2} + \mathbf{I}}\right) - \sqrt{(\mathbf{X}^t)^2 + \beta^2 \mathbf{I}}\right),$$

since all matrices in above expression are symmetric and simultaneously diagonalizable. In this case, the gradient of the spectral hypentropy can be written as

$$\nabla\Phi_\beta(\mathbf{X}^t) = \log\left(\frac{\mathbf{X}^t}{\beta} + \sqrt{\frac{(\mathbf{X}^t)^2}{\beta^2} + \mathbf{I}}\right),$$

and its inverse is given by

$$\nabla\Phi_\beta^{-1}(\mathbf{X}^t) = \beta\frac{e^{\mathbf{X}^t} - e^{-\mathbf{X}^t}}{2}.$$

Hence, for the mirror descent algorithm (4.2) we need to compute two matrix exponentials in each iteration. While computing matrix exponentials require $\mathcal{O}(n^3)$ operations, which is of the same order as a singular value decomposition, matrix exponentials are typically cheaper to compute in practice.

In the positive semidefinite case, the spectral entropy mirror map is given by

$$\Phi(\mathbf{X}) = \operatorname{tr}(\mathbf{X}\log\mathbf{X} - \mathbf{X}),$$

which has gradient given by

$$\nabla\Phi(\mathbf{X}) = \log\mathbf{X},$$

with inverse

$$\nabla\Phi^{-1}(\mathbf{X}) = \exp(\mathbf{X}).$$

Hence, mirror descent equipped with the spectral entropy mirror map requires computing a matrix exponential in each iteration, which requires $\mathcal{O}(n^3)$ operations.

4.8.2 Proof of Theorem 4.1

Proof. We begin by showing convergence of mirror descent to a global minimizer of the empirical risk f. The characterization of the limiting point follows immediately from the proof of convergence. Then, we show the bound (4.7) by showing that the empirical risk $f(\mathbf{X}^t)$ is monotonously decreasing.

Part 1: Convergence of mirror descent.

The following identity characterizes the evolution of the Bregman divergence and follows from its definition (4.1) and the mirror descent update (4.2):

$$D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^{t+1}) - D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^t) = -\eta\langle \nabla f(\mathbf{X}^t), \mathbf{X}^t - \mathbf{X}'\rangle + D_{\Phi_\beta}(\mathbf{X}^t, \mathbf{X}^{t+1}), \qquad (4.16)$$

where $\mathbf{X}'$ is any reference point. Letting $\mathbf{X}'$ be any global minimizer of f, the first term in (4.16) can be written as

$$\langle \nabla f(\mathbf{X}^t), \mathbf{X}^t - \mathbf{X}'\rangle = \frac{1}{m}\sum_{j=1}^m \big(\langle \mathbf{A}_j, \mathbf{X}^t\rangle - y_j\big)\langle \mathbf{A}_j, \mathbf{X}^t - \mathbf{X}'\rangle = 2f(\mathbf{X}^t),$$

where we used the assumption that there exists a matrix achieving zero training error, i.e. $\langle \mathbf{A}_j, \mathbf{X}'\rangle = y_j$ for all $j = 1, \dots, m$. The spectral hypentropy mirror map is $(2(\tau + \beta n))^{-1}$-strongly convex with respect to the nuclear norm $\|\cdot\|_*$ on the nuclear norm ball $\mathcal{B}(\tau) = \{\mathbf{X} \in \mathbb{R}^{n\times n'} : \|\mathbf{X}\|_* \le \tau\}$, see Theorem 14 in [48]. Writing $\tau_t = \max\{\|\mathbf{X}^t\|_*, \|\mathbf{X}^{t+1}\|_*\}$, we can bound the second term in (4.16) by

$$\begin{aligned}
D_{\Phi_\beta}(\mathbf{X}^t, \mathbf{X}^{t+1}) &= \Phi_\beta(\mathbf{X}^t) - \Phi_\beta(\mathbf{X}^{t+1}) - \langle \nabla\Phi_\beta(\mathbf{X}^{t+1}), \mathbf{X}^t - \mathbf{X}^{t+1}\rangle \\
&\le \langle \nabla\Phi_\beta(\mathbf{X}^t) - \nabla\Phi_\beta(\mathbf{X}^{t+1}), \mathbf{X}^t - \mathbf{X}^{t+1}\rangle - \frac{1}{4(\tau_t + \beta n)}\|\mathbf{X}^t - \mathbf{X}^{t+1}\|_*^2 \\
&= \langle \eta\nabla f(\mathbf{X}^t), \mathbf{X}^t - \mathbf{X}^{t+1}\rangle - \frac{1}{4(\tau_t + \beta n)}\|\mathbf{X}^t - \mathbf{X}^{t+1}\|_*^2 \\
&\le \eta\|\nabla f(\mathbf{X}^t)\|_2\|\mathbf{X}^t - \mathbf{X}^{t+1}\|_* - \frac{1}{4(\tau_t + \beta n)}\|\mathbf{X}^t - \mathbf{X}^{t+1}\|_*^2 \\
&\le \eta^2(\tau_t + \beta n)\|\nabla f(\mathbf{X}^t)\|_2^2, \qquad (4.17)
\end{aligned}$$

where we used strong convexity of Φ_β in the second line, the mirror descent update (4.2) in the third line, the fact that the spectral norm $\|\cdot\|_2$ is the dual norm to the nuclear norm $\|\cdot\|_*$ in the fourth line, and we optimized a quadratic function in $\|\mathbf{X}^t - \mathbf{X}^{t+1}\|_*$ to obtain the last inequality.

The spectral norm of the gradient ∇f can be bounded in terms of the empirical risk f: we have

$$\begin{aligned}
\|\nabla f(\mathbf{X}^t)\|_2^2 &= \bigg\|\frac{1}{m}\sum_{j=1}^m \big(\langle \mathbf{A}_j, \mathbf{X}^t\rangle - y_j\big)\mathbf{A}_j\bigg\|_2^2 \\
&\le \left(\frac{1}{m}\sum_{j=1}^m \big|\langle \mathbf{A}_j, \mathbf{X}^t\rangle - y_j\big|\|\mathbf{A}_j\|_2\right)^2 \\
&\le \frac{1}{m}\sum_{j=1}^m \|\mathbf{A}_j\|_2^2 \cdot 2f(\mathbf{X}^t),
\end{aligned}$$

where we used the triangle inequality in the second and the Cauchy-Schwarz inequality in the last line. Using the non-negativity of the Bregman divergence, we can rearrange the penultimate inequality in (4.17) to obtain

$$\|\mathbf{X}^t - \mathbf{X}^{t+1}\|_* \leq 4(\tau_t + \beta n)\eta\|\nabla f(\mathbf{X}^t)\|_2 \leq \frac{1}{2}(\tau_t + \beta n), \tag{4.18}$$

provided the step size η satisfies

$$\eta \leq \frac{1}{8\sqrt{2}}\left(\frac{1}{m}\sum_{j=1}^{m}\|\mathbf{A}_j\|_2^2 \cdot f(\mathbf{X}^t)\right)^{-1/2}. \tag{4.19}$$

We will show below that the upper bound in (4.19) is uniformly bounded from below by a constant $c > 0$, i.e. we can indeed choose a constant step size $\eta_t \equiv \eta \leq c$. If $\|\mathbf{X}^{t+1}\|_* > \|\mathbf{X}^t\|_*$, then the reverse triangle inequality yields

$$\|\mathbf{X}^{t+1}\|_* - \|\mathbf{X}^t\|_* \leq \|\mathbf{X}^t - \mathbf{X}^{t+1}\|_* \leq \frac{1}{2}\Big(\|\mathbf{X}^{t+1}\|_* + \beta n\Big),$$

which can be rearranged to $\|\mathbf{X}^{t+1}\|_* \leq 2\|\mathbf{X}^t\|_* + \beta n$, so that also $\tau_t \leq 2\|\mathbf{X}^t\|_* + \beta n$. Hence, the second term in (4.16) can be bounded by

$$D_{\Phi_\beta}(\mathbf{X}^t, \mathbf{X}^{t+1}) \leq \eta^2(\tau_t + \beta n)\|\nabla f(\mathbf{X}^t)\|_2^2 \leq \eta f(\mathbf{X}^t),$$

provided that the step size η also satisfies

$$\eta \leq \frac{1}{4}\left(\frac{1}{m}\sum_{j=1}^{m}\|\mathbf{A}_j\|_2^2 \cdot \Big(\|\mathbf{X}^t\|_* + \beta n\Big)\right)^{-1}. \tag{4.20}$$

With this, the identity in (4.16) becomes

$$D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^{t+1}) - D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^t) = -2\eta f(\mathbf{X}^t) + D_{\Phi_\beta}(\mathbf{X}^t, \mathbf{X}^{t+1}) \leq -\eta f(\mathbf{X}^t) \tag{4.21}$$

for any global minimizer $\mathbf{X}'$ of f. Since the Bregman divergence $D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^t)$ is bounded from below by zero, this means that the empirical risk $f(\mathbf{X}^t)$ must converge to zero, which in turn implies that $\mathbf{X}^t$ converges to a global minimizer of f.

To see *which* global minimizer mirror descent converges to, observe that the difference in (4.8) does not depend on the reference point $\mathbf{X}'$, as long as $\mathbf{X}'$ is a global minimizer of f. This means that the Bregman divergence $D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}^t)$ is decreased by the same amount for *all* global minimizers $\mathbf{X}'$, which then implies that $\mathbf{X}^t$ must converge to the

global minimizer which is closest to $\mathbf{X}^0$ in terms of the Bregman divergence. Hence, writing $\{\sigma_i\}_{i=1}^n$ for the singular values of $\mathbf{X}^\infty = \lim_{t\to\infty} \mathbf{X}^t$ and using the identity $\operatorname{arcsinh}(x) = \log(x + \sqrt{x^2+1})$, the quantity

$$\begin{aligned} D_{\Phi_\beta}(\mathbf{X}^\infty, \mathbf{X}^0) &= \sum_{i=1}^n \sigma_i \operatorname{arcsinh}\left(\frac{\sigma_i}{\beta}\right) - \sqrt{\sigma_i^2+\beta^2} - n\beta \\ &= \sum_{i=1}^n \sigma_i \log\frac{1}{\beta} + \sigma_i \log\left(\sigma_i + \sqrt{\sigma_i^2+\beta^2}\right) - \sqrt{\sigma_i^2+\beta^2} - n\beta \end{aligned}$$

is minimized among all global minimizers of the empirical risk f, which is the quantity in (4.6) modulo the constant $n\beta$.

Finally, since we show convergence of $\mathbf{X}^t$, this means that the nuclear norm $\|\mathbf{X}^t\|_*$ and the empirical risk $f(\mathbf{X}^t)$ stay bounded for all $t \geq 0$. This implies that, in order to satisfy inequalities (4.19) and (4.20), we can indeed choose a constant step size $\eta_t \equiv \eta \leq c$, where the constant $c > 0$ depends on the spectral norm of the sensing matrices $\mathbf{A}_j$'s and the observations y_j's.

Part 2: Proving the bound (4.7)**.**
In order to show the bound (4.7), we first show that $f(\mathbf{X}^t)$ decreases monotonously. To this end, we verifiy that f is $\frac{1}{m}\sum_{j=1}^m \|\mathbf{A}_j\|_2^2$-smooth with respect to the nuclear norm. Indeed, ∇f is Lipschitz continuous with Lipschitz constant $\frac{1}{m}\sum_{j=1}^m \|\mathbf{A}_j\|_2^2$,

$$\begin{aligned} \|\nabla f(\mathbf{X}) - \nabla f(\mathbf{Y})\|_2 &= \left\| \frac{1}{m}\sum_{j=1}^m \langle \mathbf{A}_j, \mathbf{X}-\mathbf{Y}\rangle \mathbf{A}_j \right\|_2 \\ &\leq \left\| \frac{1}{m}\sum_{j=1}^m \|\mathbf{A}_j\|_2 \|\mathbf{X}-\mathbf{Y}\|_* \mathbf{A}_j \right\|_2 \\ &\leq \frac{1}{m}\sum_{j=1}^m \|\mathbf{A}_j\|_2^2 \cdot \|\mathbf{X}-\mathbf{Y}\|_*, \end{aligned}$$

where we used the duality of the nuclear and spectral norms in the second line. Hence, we can bound

$$f(\mathbf{X}^{t+1}) \leq f(\mathbf{X}^t) + \langle \nabla f(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{X}^t\rangle + \frac{1}{2m}\sum_{j=1}^m \|\mathbf{A}_j\|_2^2 \cdot \|\mathbf{X}^{t+1} - \mathbf{X}^t\|_*^2.$$

If we can bound

$$\langle \nabla f(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{X}^t\rangle \leq -\frac{1}{2m}\sum_{j=1}^m \|\mathbf{A}_j\|_2^2 \cdot \|\mathbf{X}^{t+1} - \mathbf{X}^t\|_*^2,$$

then this would show that $f(\mathbf{X}^t)$ is monotonously decreasing. Recall the proximal formulation of mirror descent (see e.g. [15]),

$$\mathbf{X}^{t+1} = \arg\min_{\mathbf{X}\in\mathbb{R}^{n\times n'}} \left\{ \langle \nabla f(\mathbf{X}^t), \mathbf{X} - \mathbf{X}^t \rangle + \frac{1}{\eta} D_{\Phi_\beta}(\mathbf{X}, \mathbf{X}^t) \right\}.$$

Since the quantity being minimized is zero for $\mathbf{X} = \mathbf{X}^t$, we obtain the upper bound

$$\begin{aligned}
\langle \nabla f(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{X}^t \rangle &\leq -\frac{1}{\eta} D_{\Phi_\beta}(\mathbf{X}^{t+1}, \mathbf{X}^t) \\
&= \frac{1}{\eta} \Big((\Phi_\beta(\mathbf{X}^t) - \Phi_\beta(\mathbf{X}^{t+1})) + \langle \nabla \Phi_\beta(\mathbf{X}^t), \mathbf{X}^{t+1} - \mathbf{X}^t \rangle \Big) \\
&\leq -\frac{1}{4\eta(\tau_t + \beta n)} \|\mathbf{X}^{t+1} - \mathbf{X}^t\|_*^2 \\
&\leq -\frac{1}{2m} \sum_{j=1}^{m} \|\mathbf{A}_j\|_2^2 \cdot \|\mathbf{X}^{t+1} - \mathbf{X}^t\|_*^2,
\end{aligned}$$

where we used strong convexity of Φ_β for the second inequality, and the last inequality holds if the step size η satisfies inequality (4.20). This completes the proof that $f(\mathbf{X}^{t+1}) \leq f(\mathbf{X}^t)$.

To show the bound (4.7), assume that it were violated for some $t > 0$. Since $f(\mathbf{X}^t)$ is non-increasing, this means that

$$f(\mathbf{X}^s) \geq f(\mathbf{X}^t) > \frac{D_{\Phi_\beta}(\mathbf{X}^\infty, \mathbf{X}^0)}{\eta t}$$

for all $s \leq t$. The bound in (4.21) controls by how much the Bregman divergence must decrease in each iteration. Summing over the expression in (4.21), we obtain

$$\begin{aligned}
D_{\Phi_\beta}(\mathbf{X}^\infty, \mathbf{X}^t) &= D_{\Phi_\beta}(\mathbf{X}^\infty, \mathbf{X}^0) + \sum_{s=0}^{t-1} D_{\Phi_\beta}(\mathbf{X}^\infty, \mathbf{X}^{s+1}) - D_{\Phi_\beta}(\mathbf{X}^\infty, \mathbf{X}^s) \\
&< D_{\Phi_\beta}(\mathbf{X}^\infty, \mathbf{X}^0) - \sum_{s=0}^{t-1} \eta \frac{D_{\Phi_\beta}(\mathbf{X}^\infty, \mathbf{X}^0)}{\eta t} \\
&= 0,
\end{aligned}$$

which contradicts the non-negativity of the Bregman divergence and therefore shows that the bound (4.7) must be satisfied for all $t > 0$. □

4.8.3 Proof of Theorem 4.2

The proof of Theorem 4.2 follows the same steps as the proof of Theorem 4.1 and uses the fact that the spectral entropy (4.5) is $(2\tau)^{-1}$ strongly convex with respect to the

nuclear norm on the nuclear norm ball $\mathcal{B}_+(\tau) = \{\mathbf{X} \in \mathcal{S}^n_+ : \|\mathbf{X}\|_* \leq \tau\}$, for which we include a proof for completeness' sake.

Lemma 4.7 (Strong convexity of the spectral entropy)**.** *The spectral entropy* (4.5) *is* $(2\tau)^{-1}$*-strongly convex with respect to the nuclear norm* $\|\cdot\|_*$ *on the nuclear norm ball* $\mathcal{B}_+(\tau)$*.*

Proof. The proof of Lemma 4.7 closely follows the proof of strong convexity of the spectral hypentropy mirror map provided in [48]. We first introduce some notation. We denote by $\lambda(\mathbf{X})$ the vector of eigenvalues of a symmetric matrix $\mathbf{X} \in \mathcal{S}^n$. For a function $f : \mathbb{R} \to \mathbb{R}$, we denote by $f(\mathbf{X})$ the standard lifting of scalar functions to symmetric matrices, see e.g. [48],

$$\mathbf{X} = \mathbf{U}\operatorname{diag}[\lambda(\mathbf{X})]\mathbf{U}^\top \qquad \Rightarrow \qquad f(\mathbf{X}) = \mathbf{U}\operatorname{diag}[f(\lambda(\mathbf{X}))]\mathbf{U}^\top,$$

where f is applied to the vector $\lambda(\mathbf{X})$ componentwise.

In order to show that the spectral entropy Φ is $(2\tau)^{-1}$-strongly convex with respect to the nuclear norm $\|\cdot\|_*$ on $\mathcal{B}_+(\tau)$, we use the duality of strong convexity and smoothness and show instead that the Fenchel conjugate Φ^* is 2τ-smooth with respect to the spectral norm $\|\cdot\|_2$ on the set $\nabla\Phi(\mathcal{B}_+(\tau))$.

The following Theorem from [64] relates the conjugate of rotationally invariant matrix functions, i.e. functions that can be written as $\Psi(\mathbf{X}) = (\psi \circ \lambda)(\mathbf{X})$, where $\psi : \mathbb{R}^n \to \mathbb{R}$, to the conjugate of the vector function ψ.

Theorem 4.8 (Theorem 28 [64])**.** *Let* $g : \mathbb{R}^n \to \mathbb{R}$ *be a symmetric function, i.e. invariant under permutations of its argument. Then,*

$$(g \circ \lambda)^* = g^* \circ \lambda.$$

With this, we can compute the Fenchel conjugate of Φ. We have

$$\Phi(\mathbf{X}) = \operatorname{tr}(\mathbf{X}\log\mathbf{X} - \mathbf{X}) = \sum_{i=1}^n \lambda_i(\mathbf{X})\log\lambda_i(\mathbf{X}) - \lambda_i(\mathbf{X}) =: \sum_{i=1}^n \phi(\lambda_i(\mathbf{X})),$$

so that

$$\Phi^*(\mathbf{X}) = \sum_{i=1}^n \phi^*(\lambda_i(\mathbf{X})) = \sum_{i=1}^n e^{\lambda_i(\mathbf{X})},$$

since the conjugate of the scalar function $\phi(x) = x \log x - x$ is given by $\phi^*(x) = e^x$.

The following Lemma from [63] allows us to reduce the smoothness of matrix functions to the smoothness of functions taking vectors as argument.

Lemma 4.9 (Proposition 3.1 [63])**.** *Let $f : \mathbb{R}_+ \to \mathbb{R}$ be a twice continuously differentiable function and $c > 0$ a constant such that, for all $b > a > 0$,*

$$\frac{f'(b) - f'(a)}{b - a} \leq c \frac{f''(a) + f''(b)}{2}.$$

Then, the function $F : \mathcal{S}^n \to \mathbb{R}$ defined by $F(\mathbf{X}) = \operatorname{tr}(f(\mathbf{X}))$ is twice continuously differentiable and satisfies, for every $\mathbf{H} \in \mathcal{S}^n$,

$$D^2 F(\mathbf{X})[\mathbf{H}, \mathbf{H}] \leq c \operatorname{tr}(\mathbf{H} f''(\mathbf{X}) \mathbf{H}).$$

We can now analyze the smoothness of Φ^*. By the mean-value theorem, we have for some $c \in [a, b]$,

$$\frac{(\phi^*)'(b) - (\phi^*)'(a)}{b - a} = (\phi^*)''(c) \leq (\phi^*)''(a) + (\phi^*)''(b).$$

Then, by Lemma 4.9, we can bound, for any $\mathbf{X} = \nabla \Phi(\mathbf{Y})$ with $\mathbf{Y} \in \mathcal{B}_+(\tau)$,

$$\begin{aligned}
\sup_{\mathbf{H} \in \mathcal{S}^n : \|\mathbf{H}\|_2 \leq 1} D^2 \Phi^*(\mathbf{X})[\mathbf{H}, \mathbf{H}] &\leq \sup_{\mathbf{H} \in \mathcal{S}^n : \|\mathbf{H}\|_2 \leq 1} 2 \operatorname{tr}(\mathbf{H} (\phi^*)''(\mathbf{X}) \mathbf{H}) \\
&= \sup_{\mathbf{H} \in \mathcal{S}^n : \|\mathbf{H}\|_2 \leq 1} 2 \operatorname{tr}(\mathbf{H}^2 (\phi^*)''(\mathbf{X})) \\
&\leq \sup_{\mathbf{H} \in \mathcal{S}^n : \|\mathbf{H}\|_2 \leq 1} 2 \langle \sigma^2(\mathbf{H}), \sigma((\phi^*)''(\mathbf{X})) \rangle,
\end{aligned}$$

where we write $\sigma(\mathbf{X})$ for the vector of singular values of a matrix $\mathbf{X}$. The equality follows from commutativity of the trace, and the last inequality follows from von Neumann's trace inequality $\operatorname{tr}(\mathbf{A}^\top \mathbf{B}) \leq \langle \sigma(\mathbf{A}), \sigma(\mathbf{B}) \rangle$. By definition, we have $\sigma_i^2(\mathbf{H}) \leq 1$ and $\sigma_i((\phi^*)''(\mathbf{X})) = \sigma_i(\mathbf{Y})$ for all $i = 1, \ldots, n$, so that we can bound

$$\sup_{\mathbf{H} \in \mathcal{S}^n : \|\mathbf{H}\|_2 \leq 1} D^2 \Phi^*(\mathbf{X})[\mathbf{H}, \mathbf{H}] \leq 2 \sum_{i=1}^n 1 \cdot \sigma_i(\mathbf{Y}) = 2 \|\mathbf{Y}\|_* \leq 2\tau,$$

which completes the proof that Φ^* is 2τ-smooth with respect to the spectral norm on $\nabla \Phi(\mathcal{B}_+(\tau))$. □

Since the rest of the proof of Theorem 4.2 follows the exact same steps as the proof of Theorem 4.1, it is omitted to avoid repetition.

4.8.4 Proof of Theorem 4.3

Proof of Theorem 4.3. We begin by considering the rectangular case and prove the bound (4.11) in Theorem 4.3 for the spectral hypentropy mirror map (4.4). The proof of Theorem 4.3 is an adaption of and builds upon the proofs of Theorem 3.3 in [98] and Theorem 4 in [41]. First, we need to bound the nuclear norm of the matrix $\mathbf{X}^\infty$. Then, we follow [41, 98] and use the RIP-assumption to bound the deviation $\|\mathbf{X}^\infty - \mathbf{X}^\star\|_F$.

Step 1: Bound the nuclear norm $\|\mathbf{X}^\infty\|_*$.

If $\|\mathbf{X}^\infty\|_* \leq \|\mathbf{X}^\star\|_*$, then we have a suitable upper bound for the nuclear norm $\|\mathbf{X}^\infty\|_*$. Hence, assume that $\|\mathbf{X}^\infty\|_* > \|\mathbf{X}^\star\|_*$. By Theorem 4.1, $\mathbf{X}^\infty$ minimizes the quantity in (4.6) among all global minimizers of the empirical risk f which, in particular, include $\mathbf{X}^\star$. Writing σ_i and μ_i for the singular values of $\mathbf{X}^\infty$ and $\mathbf{X}^\star$, respectively, we can bound

$$\begin{aligned}
&\sum_{i=1}^n \sigma_i \log \frac{\|\mathbf{X}^\star\|_*}{\beta} + \sigma_i \log \frac{\sigma_i + \sqrt{\sigma_i^2 + \beta^2}}{\|\mathbf{X}^\star\|_*} - \sqrt{\sigma_i^2 + \beta^2} \\
\leq &\sum_{i=1}^n \mu_i \log \frac{\|\mathbf{X}^\star\|_*}{\beta} + \mu_i \log \frac{\mu_i + \sqrt{\mu_i^2 + \beta^2}}{\|\mathbf{X}^\star\|_*} - \sqrt{\mu_i^2 + \beta^2}.
\end{aligned}$$

For any $x, \beta > 0$, we have

$$x \leq \sqrt{x^2 + \beta^2} \leq x + \beta.$$

Rearranging above inequality for the nuclear norm $\|\mathbf{X}^\infty\|_*$, we obtain the upper bound

$$\begin{aligned}
\|\mathbf{X}^\infty\|_* &\leq \|\mathbf{X}^\star\|_* + \frac{1}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1} \sum_{i=1}^n \mu_i \log \frac{\mu_i + \sqrt{\mu_i^2 + \beta^2}}{\|\mathbf{X}^\star\|_*} - \sigma_i \log \frac{\sigma_i + \sqrt{\sigma_i^2 + \beta^2}}{\|\mathbf{X}^\star\|_*} + \beta \\
&\leq \|\mathbf{X}^\star\|_* + \frac{1}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1} \left(\|\mathbf{X}^\infty\|_* \log 2.1 - \|\mathbf{X}^\infty\|_* \log \frac{2\|\mathbf{X}^\infty\|_*}{n\|\mathbf{X}^\star\|_*} + n\beta \right) \\
&\leq \|\mathbf{X}^\star\|_* + \frac{1}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1} \left(\|\mathbf{X}^\infty\|_* \log(1.05n) + n\beta \right),
\end{aligned}$$

where we used the assumptions $\beta \leq \frac{\|\mathbf{X}^\star\|_*}{1.05en}$ and $\|\mathbf{X}^\star\|_* < \|\mathbf{X}^\infty\|_*$, and for the second inequality we used the fact that the constrained optimization problem

$$\text{optimize } \sum_{i=1}^n x_i \log\left(x_i + \sqrt{x_i^2 + \beta^2}\right) \qquad \text{s.t. } \sum_{i=1}^n x_i = K, \quad x_i \geq 0 \text{ for all } i = 1, \ldots, n$$

attains a maximum when $x_i = K$ for exactly one $i \in \{1, \ldots, n\}$, and attains a minimum when all $x_i = K/n$ are equal. Hence, again using the assumption $\beta < \frac{\|\mathbf{X}^\star\|_*}{1.05en}$, we can

bound

$$\|\mathbf{X}^\infty\|_* \leq (1+\Delta_\beta)\left(\|\mathbf{X}^\star\|_* + \frac{n\beta}{\log\frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}\right), \tag{4.22}$$

where $\Delta_\beta = (\frac{\log(\|\mathbf{X}^\star\|_*/\beta)-1}{\log(1.05n)} - 1)^{-1} > 0$, since $\beta \leq \frac{\|\mathbf{X}^\star\|_*}{1.05en}$.

Step 2: Bound the reconstruction error $\|\mathbf{X}^\infty - \mathbf{X}^\star\|_F$.

With this, we can now proceed as in [41, 98]. Writing $\mathbf{R} = \mathbf{X}^\infty - \mathbf{X}^\star$, we can apply Lemma 3.4 from [98] to the matrices $\mathbf{X}^\star$ and $\mathbf{R}$ to decompose $\mathbf{R} = \mathbf{R}_0 + \mathbf{R}_c$, where $\operatorname{rank}(\mathbf{R}_0) \leq 2\operatorname{rank}(\mathbf{X}^\star)$, $\mathbf{X}^\star\mathbf{R}_c^\top = \mathbf{0}$ and $(\mathbf{X}^\star)^\top\mathbf{R}_c = \mathbf{0}$. We can bound

$$\|\mathbf{X}^\star + \mathbf{R}\|_* \geq \|\mathbf{X}^\star + \mathbf{R}_c\|_* - \|\mathbf{R}_0\|_* = \|\mathbf{X}^\star\|_* + \|\mathbf{R}_c\|_* - \|\mathbf{R}_0\|_*,$$

where the inequality follows from the triangle inequality, and the equality holds since $\mathbf{X}^\star\mathbf{R}_c^\top = \mathbf{0}$ and $(\mathbf{X}^\star)^\top\mathbf{R}_c = \mathbf{0}$ together imply that the nuclear norm decomposes, see e.g. Lemma 2.3 of [98]. Together with (4.22), this implies

$$\|\mathbf{R}_c\|_* \leq \|\mathbf{R}_0\|_* + \Delta_\beta\|\mathbf{X}^\star\|_* + (1+\Delta_\beta)\frac{n\beta}{\log\frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}.$$

Next, we partition $\mathbf{R}_c$ into a sum of matrices $\mathbf{R}_1, \mathbf{R}_2, \ldots, \mathbf{R}_{\lceil\frac{n}{3r}\rceil}$, with each being of rank at most $3r$. Letting $\mathbf{R}_c = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be the singular value decomposition of $\mathbf{R}_c$, where the diagonal elements of Σ are in non-increasing order $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n \geq 0$, define $\mathbf{R}_i = \mathbf{U}_{I_i}\mathbf{\Sigma}_{I_i}\mathbf{V}_{I_i}^\top$, where $I_i = \{3r(i-1)+1, \ldots, 3ri\}$. By construction, we have

$$\sigma_k \leq \frac{1}{3r}\sum_{j\in I_i}\sigma_j \qquad \text{for all } k \in I_{i+1},\ i \in \left\{1, \ldots, \left\lceil\frac{n}{3r}\right\rceil\right\},$$

which implies $\|\mathbf{R}_{i+1}\|_F^2 \leq \frac{1}{3r}\|\mathbf{R}_i\|_*^2$. With this, we can bound

$$\begin{aligned}
\sum_{i\geq 2}\|\mathbf{R}_i\|_F &\leq \frac{1}{\sqrt{3r}}\sum_{i\geq 1}\|\mathbf{R}_i\|_* \\
&= \frac{1}{\sqrt{3r}}\|\mathbf{R}_c\|_* \\
&\leq \frac{\sqrt{2r}}{\sqrt{3r}}\|\mathbf{R}_0\|_F + \frac{\Delta_\beta\|\mathbf{X}^\star\|_* + (1+\Delta_\beta)\frac{n\beta}{\log\frac{\|\mathbf{X}^\star\|_*}{\beta}-1}}{\sqrt{3r}},
\end{aligned} \tag{4.23}$$

where for the last inequality we used that $\operatorname{rank}(\mathbf{R}_0) \leq 2r$. Since $\mathbf{R}_0 + \mathbf{R}_1$ is at most of rank $5r$, we can use the triangle inequality and the restricted isometry property to bound

$$\begin{aligned}
\left(\frac{1}{m}\sum_{j=1}^m\langle\mathbf{A}_j, \mathbf{R}\rangle^2\right)^{1/2} &\geq \left(\frac{1}{m}\sum_{j=1}^m\langle\mathbf{A}_j, \mathbf{R}_0+\mathbf{R}_1\rangle^2\right)^{1/2} - \sum_{i\geq 2}\left(\frac{1}{m}\sum_{j=1}^m\langle\mathbf{A}_j, \mathbf{R}_i\rangle^2\right)^{1/2} \\
&\geq (1-\delta)\|\mathbf{R}_0+\mathbf{R}_1\|_F - \sum_{i\geq 2}(1+\delta)\|\mathbf{R}_i\|_F.
\end{aligned} \tag{4.24}$$

Since $\mathbf{R}_0$ is orthogonal to $\mathbf{R}_1$ (see Lemma 3.4 in [98]), we have $\|\mathbf{R}_0+\mathbf{R}_1\|_F \geq \|\mathbf{R}_0\|_F$. By definition, we have $f(\mathbf{X}^\star) = f(\mathbf{X}^\infty) = 0$, which implies $\langle \mathbf{A}_j, \mathbf{R}\rangle = 0$ for all $j = 1, \ldots, m$. Hence, we can use (4.23) and rearrange (4.24) for $\|\mathbf{R}_0 + \mathbf{R}_1\|_F$ to obtain

$$\|\mathbf{R}_0 + \mathbf{R}_1\|_F \leq \left(1 - \sqrt{\frac{2}{3}} - \delta\left(1 + \sqrt{\frac{2}{3}}\right)\right)^{-1}\left(1+\delta\right)\frac{\Delta_\beta\|\mathbf{X}^\star\|_* + (1+\Delta_\beta)\frac{n\beta}{\log\frac{\|\mathbf{X}^\star\|_*}{\beta}-1}}{\sqrt{3r}}.$$

Finally, this yields

$$\begin{aligned}\|\mathbf{R}\|_F &\leq \|\mathbf{R}_0 + \mathbf{R}_1\|_F + \sum_{i\geq 2}\|\mathbf{R}_i\|_F \\ &\leq 2\left(1 - \sqrt{\frac{2}{3}} - \delta\left(1 + \sqrt{\frac{2}{3}}\right)\right)^{-1}\frac{\Delta_\beta\|\mathbf{X}^\star\|_* + (1+\Delta_\beta)\frac{n\beta}{\log\frac{\|\mathbf{X}^\star\|_*}{\beta}-1}}{\sqrt{3r}},\end{aligned}$$

which completes the proof of the bound (4.11) in Theorem 4.3. The bound (4.12) can be shown following the same steps, and we omit the details to avoid repetition. □

4.8.5 Proof of Theorem 4.4

Proof of Theorem 4.4. As in Theorem 4.3, we first consider the rectangular case and show the bound (4.13) in Theorem 4.4. The proof of Theorem 4.4 combines the ideas from and closely follows the proofs of Theorem 2 in [97] and Theorem 7 in [23]. It was shown in Proposition 3 in [97] that it suffices to consider a setting where the entries are sampled independently and uniformly with replacement. We first introduce some notation necessary for the proof. A more detailed background on the following quantities can be found in [97].

We use calligraphic letters to denote linear operators on matrices, for instance, we denote the identity operator by $\mathcal{I}$. We define the spectral norm of an operator as $\|\mathcal{A}\| = \sup_{\mathbf{X}:\|\mathbf{X}\|_F\leq 1}\|\mathcal{A}(\mathbf{X})\|_F$. Let $\Omega = \{(a_j, b_j)\}_{j=1}^m$ be a collection of indices sampled uniformly at random with replacement (possibly containing repetitions), and define the operator

$$\mathcal{R}_\Omega(\mathbf{X}) = \sum_{j=1}^m \langle \mathbf{e}_{a_j}\mathbf{e}_{b_j}^\top, \mathbf{X}\rangle \mathbf{e}_{a_j}\mathbf{e}_{b_j}^\top.$$

Let $\mathbf{X}^\star = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be the singular value decomposition of $\mathbf{X}^\star$, and let $\mathbf{u}_k$ (resp. $\mathbf{v}_k$) be the k-th column of $\mathbf{U}$ (resp. $\mathbf{V}$), and define the subspaces $U = \operatorname{span}(\mathbf{u}_1, \ldots, \mathbf{u}_r)$ and $V = \operatorname{span}(\mathbf{v}_1, \ldots, \mathbf{v}_r)$. Let T be the linear space spanned by elements of the form $\mathbf{u}_k\mathbf{y}^\top$

and $\mathbf{x}\mathbf{v}_k^\top$, $k = 1, \ldots, r$, where $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^{n'}$ are arbitrary vectors, and let $T^\perp$ be its orthogonal complement. The orthogonal projection onto the subspace T is given by

$$\mathcal{P}_T(\mathbf{X}) = \mathbf{P}_U\mathbf{X} + \mathbf{X}\mathbf{P}_V - \mathbf{P}_U\mathbf{X}\mathbf{P}_V,$$

where $\mathbf{P}_U$ and $\mathbf{P}_V$ are the orthogonal projections onto U and V, respectively. Then, the orthogonal projection onto $T^\perp$ is given by

$$\mathcal{P}_{T^\perp}(\mathbf{X}) = (\mathcal{I} - \mathcal{P}_T)(\mathbf{X}).$$

It has been shown in [97] that, with high probability,

$$\frac{nn'}{m}\left\|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_T - \frac{m}{nn'}\mathcal{P}_T\right\| \leq \frac{1}{2}, \qquad \|\mathcal{R}_\Omega\| \leq \frac{8}{3}\sqrt{c}\log n', \tag{4.25}$$

and that there exists a matrix $\mathbf{Y}$ in the range of $\mathcal{R}_\Omega$ satisfying

$$\|\mathcal{P}_T(\mathbf{Y}) - \mathbf{U}\mathbf{V}^\top\|_F \leq \sqrt{\frac{r}{2n'}}, \qquad \|\mathcal{P}_{T^\perp}(\mathbf{Y})\|_F \leq \frac{1}{2}, \tag{4.26}$$

see Section 4 in [97] for a proof of these statements.

Let $\mathbf{R} = \mathbf{X}^\infty - \mathbf{X}^\star$. Since the subspaces T and $T^\perp$ are orthogonal by construction, we have

$$\|\mathbf{R}\|_F^2 = \|\mathcal{P}_T(\mathbf{R})\|_F^2 + \|\mathcal{P}_{T^\perp}(\mathbf{R})\|_F^2,$$

so the goal is to bound the two terms on the right hand side of above identity. Since both $\mathbf{X}^\star$ and $\mathbf{X}^\infty$ are global minimizers of the empirical risk f, we have

$$0 = \|\mathcal{R}_\Omega(\mathbf{R})\|_F \geq \|\mathcal{R}_\Omega\mathcal{P}_T(\mathbf{R})\|_F - \|\mathcal{R}_\Omega\mathcal{P}_{T^\perp}(\mathbf{R})\|_F,$$

where we used the reverse triangle inequality. Further, the first bound in (4.25) implies

$$\|\mathcal{R}_\Omega\mathcal{P}_T(\mathbf{R})\|_F^2 = \langle \mathbf{R}, \mathcal{P}_T\mathcal{R}_\Omega^2\mathcal{P}_T(\mathbf{R})\rangle \geq \langle \mathbf{R}, \mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_T(\mathbf{R})\rangle \geq \frac{m}{2nn'}\|\mathcal{P}_T(\mathbf{R})\|_F^2,$$

and we can bound $\|\mathcal{R}_\Omega\mathcal{P}_{T^\perp}(\mathbf{R})\|_F \leq \frac{8}{3}\sqrt{c}\log(n')\|\mathcal{P}_{T^\perp}(\mathbf{R})\|_F$ by using the second bound in (4.25),. Together, this implies

$$\|\mathcal{P}_{T^\perp}(\mathbf{R})\|_F \geq \sqrt{\frac{9m}{128cnn'\log^2 n'}}\|\mathcal{P}_T(\mathbf{R})\|_F \geq \sqrt{\frac{4.5r}{n'}}\|\mathcal{P}_T(\mathbf{R})\|_F. \tag{4.27}$$

Recalling the variational characterization of the nuclear norm $\|\mathbf{A}\|_* = \sup_{\mathbf{B}:\|\mathbf{B}\|\leq 1}\langle \mathbf{A}, \mathbf{B}\rangle$, we can choose matrices $\mathbf{U}_\perp$ and $\mathbf{V}_\perp$ such that $[\mathbf{U}, \mathbf{U}_\perp]$ and $[\mathbf{V}, \mathbf{V}_\perp]$ are orthogonal matrices and $\langle \mathbf{U}_\perp \mathbf{V}_\perp^\top, \mathcal{P}_{T^\perp}(\mathbf{R})\rangle = \|\mathcal{P}_{T^\perp}(\mathbf{R})\|_*$. Let $\mathbf{Y}$ be as in (4.26). Then,

$$\begin{aligned}
\|\mathbf{X}^\star + \mathbf{R}\|_* &\geq \langle \mathbf{U}\mathbf{V}^\top + \mathbf{U}_\perp \mathbf{V}_\perp^\top, \mathbf{X}^\star + \mathbf{R}\rangle \\
&= \|\mathbf{X}^\star\|_* + \langle \mathbf{U}\mathbf{V}^\top + \mathbf{U}_\perp \mathbf{V}_\perp^\top, \mathbf{R}\rangle \\
&= \|\mathbf{X}^\star\|_* + \langle \mathbf{U}\mathbf{V}^\top + \mathbf{U}_\perp \mathbf{V}_\perp^\top - (\mathcal{P}_T(\mathbf{Y}) + \mathcal{P}_{T^\perp}(\mathbf{Y})), \mathcal{P}_T(\mathbf{R}) + \mathcal{P}_{T^\perp}(\mathbf{R})\rangle \\
&= \|\mathbf{X}^\star\|_* + \langle \mathbf{U}\mathbf{V}^\top - \mathcal{P}_T(\mathbf{Y}), \mathcal{P}_T(\mathbf{R})\rangle + \langle \mathbf{U}_\perp \mathbf{V}_\perp^\top - \mathcal{P}_{T^\perp}(\mathbf{Y}), \mathcal{P}_{T^\perp}(\mathbf{R})\rangle \\
&\geq \|\mathbf{X}^\star\|_* - \sqrt{\frac{r}{2n'}}\|\mathcal{P}_T(\mathbf{R})\|_F + \frac{1}{2}\|\mathcal{P}_{T^\perp}(\mathbf{R})\|_* \\
&\geq \|\mathbf{X}^\star\|_* + \frac{1}{6}\|\mathcal{P}_{T^\perp}(\mathbf{R})\|_*,
\end{aligned}$$

where the first line follows from the varitional characterization of the nuclear norm, the third line from the fact that $\mathbf{Y}$ and $\mathbf{R}$ are orthogonal since $\mathbf{Y}$ is in the range and $\mathbf{R}$ in the kernel of $\mathcal{R}_\Omega$, the fourth line from the fact that T and $T^\perp$ are, by construction, orthogonal subspaces, the fifth line from the bound (4.26) and the definition of $\mathbf{U}_\perp, \mathbf{V}_\perp$, and the last line follows from the bound (4.27) and the fact that the Frobenius norm is bounded from above by the nuclear norm.

As in the proof of Theorem 4.3, we can bound the nuclear norm

$$\|\mathbf{X}^\infty\|_* \leq (1 + \Delta_\beta)\left(\|\mathbf{X}^\star\|_* + \frac{n\beta}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}\right),$$

where $\Delta_\beta = (\frac{\log(\|\mathbf{X}^\star\|_*/\beta)-1}{\log(1.05n)} - 1)^{-1}$. Hence, we can bound

$$\|\mathcal{P}_{T^\perp}(\mathbf{R})\|_F \leq \|\mathcal{P}_{T^\perp}(\mathbf{R})\|_* \leq 6\left(\Delta_\beta\|\mathbf{X}^\star\|_* + (1 + \Delta_\beta)\frac{n\beta}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}\right).$$

Using (4.27), we can also bound

$$\begin{aligned}
\|\mathcal{P}_T(\mathbf{R})\|_F &\leq \sqrt{\frac{128cnn'\log^2 n'}{9m}}\|\mathcal{P}_{T^\perp}(\mathbf{R})\|_F \\
&\leq 6\left(\Delta_\beta\|\mathbf{X}^\star\|_* + (1 + \Delta_\beta)\frac{n\beta}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}\right)\sqrt{\frac{128cnn'\log^2 n'}{9m}}.
\end{aligned}$$

Putting everything together, we have

$$\|\mathbf{R}\|_F \leq 6\left(\Delta_\beta\|\mathbf{X}^\star\|_* + (1 + \Delta_\beta)\frac{n\beta}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}\right)\left(1 + \left(\frac{128cnn'\log^2 n'}{9m}\right)^{\frac{1}{2}}\right),$$

which completes the proof of the bound (4.13) in Theorem 4.4. The bound (4.14) can be shown following the same steps, and we omit the details to avoid repetition. □

4.8.6 Proof of Proposition 4.5

Proof. We begin by showing the first part of Proposition 4.5.

Proof of part 1.

Recalling the expressions for the gradient of the spectral entropy mirror map $\nabla\Phi$ and its inverse $\nabla\Phi^{-1}$ provided in Appendix 4.8.1, the mirror descent update (4.2) becomes

$$\mathbf{X}^{t+1} = \exp\Big(\log \mathbf{X}^t - \eta \nabla f(\mathbf{X}^t)\Big).$$

By assumption, $\mathbf{X}^0$ commutes with all sensing matrices $\mathbf{A}_j$'s, and hence also with the gradient

$$\nabla f(\mathbf{X}^0) = \frac{1}{m}\sum_{j=1}^{m}(\langle \mathbf{A}_j, \mathbf{X}^0\rangle - y_j)\mathbf{A}_j,$$

which is a linear combination of the $\mathbf{A}_j$'s. Further, note that if two matrices $\mathbf{A}$ and $\mathbf{B}$ commute, then the matrices $\log \mathbf{A}$ and $\exp(\mathbf{A})$ also commute with $\mathbf{B}$. By induction, this implies that $\log \mathbf{X}^t$ and $\nabla f(\mathbf{X}^t)$ commute for all $t \geq 0$, and we therefore have

$$\exp\Big(\log \mathbf{X}^t - \eta \nabla f(\mathbf{X}^t)\Big) = \mathbf{X}^t \exp\Big(-\eta \nabla f(\mathbf{X}^t)\Big) = \exp\Big(-\eta \nabla f(\mathbf{X}^t)\Big)\mathbf{X}^t,$$

where we used the fact that $e^{\mathbf{A}+\mathbf{B}} = e^{\mathbf{A}}e^{\mathbf{B}}$ if the matrices $\mathbf{A}$ and $\mathbf{B}$ commute. Hence, the mirror descent update (4.2) is equivalent to

$$\mathbf{X}^{t+1} = \frac{1}{2}\Big(\mathbf{X}^t \exp\Big(-\eta \nabla f(\mathbf{X}^t)\Big) + \exp\Big(-\eta \nabla f(\mathbf{X}^t)\Big)\mathbf{X}^t\Big),$$

which is exactly the exponentiated gradient algorithm defined in (4.15) with initialization $\mathbf{U}^0 = \mathbf{X}^0$ and $\mathbf{V}^0 = \mathbf{0}$.

Proof of part 2.

We begin by considering the mirror descent update (4.2). Recalling the expressions for the gradient of the spectral hypentropy mirror map $\nabla\Phi_\beta$ and its inverse $\nabla\Phi_\beta^{-1}$ for symmetric matrices we derived in Appendix 4.8.1, the mirror descent update (4.2) becomes

$$\begin{aligned}
\mathbf{X}^{t+1} &= \nabla\Phi_\beta^{-1}\Big(\nabla\Phi_\beta(\mathbf{X}^t) - \eta\nabla f(\mathbf{X}^t)\Big) \\
&= \frac{\beta}{2}\Bigg[\exp\left(\log\left(\frac{\mathbf{X}^t}{\beta} + \sqrt{\frac{(\mathbf{X}^t)^2}{\beta^2} + \mathbf{I}}\right) - \eta\nabla f(\mathbf{X}^t)\right) \\
&\qquad - \exp\left(-\log\left(\frac{\mathbf{X}^t}{\beta} + \sqrt{\frac{(\mathbf{X}^t)^2}{\beta^2} + \mathbf{I}}\right) + \eta\nabla f(\mathbf{X}^t)\right)\Bigg].
\end{aligned}$$

Assuming that $\mathbf{X}^t$ is symmetric, we can write $\mathbf{X}^t = \mathbf{B}\mathbf{D}\mathbf{B}^\top$ by the spectral theorem, where $\mathbf{B}$ is an orthogonal matrix and $\mathbf{D}$ a diagonal matrix. Then, we have

$$\begin{aligned}
-\log\left(\frac{\mathbf{X}^t}{\beta} + \sqrt{\frac{(\mathbf{X}^t)^2}{\beta^2} + \mathbf{I}}\right) &= -\mathbf{B}\log\left(\frac{\mathbf{D}}{\beta} + \sqrt{\frac{\mathbf{D}^2}{\beta^2} + \mathbf{I}}\right)\mathbf{B}^\top \\
&= \mathbf{B}\log\left(-\frac{\mathbf{D}}{\beta} + \sqrt{\frac{\mathbf{D}^2}{\beta^2} + \mathbf{I}}\right)\mathbf{B}^\top \\
&= \log\left(-\frac{\mathbf{X}^t}{\beta} + \sqrt{\frac{(\mathbf{X}^t)^2}{\beta^2} + \mathbf{I}}\right),
\end{aligned}$$

since we have $(x + \sqrt{x^2+1})^{-1} = -x + \sqrt{x^2+1}$ for all $x \in \mathbb{R}$. Assuming that $\mathbf{X}^t$ (and hence also $\log(\mathbf{X}^t/\beta + \sqrt{(\mathbf{X}^t/\beta)^2 + \mathbf{I}})$) commutes with all $\mathbf{A}_j$'s, the mirror descent update can be written as

$$\begin{aligned}
\mathbf{X}^{t+1} &= \frac{\beta}{2}\left[\left(\frac{\mathbf{X}^2}{\beta} + \sqrt{\frac{\mathbf{X}^2}{\beta^2} + \mathbf{I}}\right)\exp\Big(-\eta\nabla f(\mathbf{X}^t)\Big) - \left(-\frac{\mathbf{X}^2}{\beta} + \sqrt{\frac{\mathbf{X}^2}{\beta^2} + \mathbf{I}}\right)\exp\Big(\eta\nabla f(\mathbf{X}^t)\Big)\right] \\
&= \frac{1}{2}\Bigg[\exp\Big(-\eta\nabla f(\mathbf{X}^t)\Big)\frac{\mathbf{X}^t + \sqrt{(\mathbf{X}^t)^2 + \beta^2\mathbf{I}}}{2} + \frac{\mathbf{X}^t + \sqrt{(\mathbf{X}^t)^2 + \beta^2\mathbf{I}}}{2}\exp\Big(-\eta\nabla f(\mathbf{X}^t)\Big) \\
&\quad + \exp\Big(\eta\nabla f(\mathbf{X}^t)\Big)\frac{-\mathbf{X}^t + \sqrt{(\mathbf{X}^t)^2 + \beta^2\mathbf{I}}}{2} + \frac{-\mathbf{X}^t + \sqrt{(\mathbf{X}^t)^2 + \beta^2\mathbf{I}}}{2}\exp\Big(\eta\nabla f(\mathbf{X}^t)\Big)\Bigg].
\end{aligned}$$

Since $\mathbf{X}^0 = \mathbf{0}$ is symmetric and commutes with all $\mathbf{A}_j$'s, this identity inductively shows that $\mathbf{X}^t$ is symmetric and commutes with all $\mathbf{A}_j$'s for all $t \geq 0$.

Next, consider the exponentiated gradient algorithm (4.15) with initialization $\mathbf{U}^0 = \mathbf{V}^0 = \frac{1}{2}\beta\mathbf{I}$. Since the initializations $\mathbf{U}^0$ and $\mathbf{V}^0$ both commute with all $\mathbf{A}_j$'s, the update (4.15) implies that $\mathbf{U}^t$ and $\mathbf{V}^t$ commute with all $\mathbf{A}_j$'s for all $t \geq 0$. Then, we have

$$\mathbf{U}^{t+1}\mathbf{V}^{t+1} = \frac{e^{-\eta\nabla f(\mathbf{X}^t)}\mathbf{U}^t + \mathbf{U}^t e^{-\eta\nabla f(\mathbf{X}^t)}}{2} \cdot \frac{e^{\eta\nabla f(\mathbf{X}^t)}\mathbf{V}^t + \mathbf{V}^t e^{\eta\nabla f(\mathbf{X}^t)}}{2} = \mathbf{U}^t\mathbf{V}^t,$$

that is the product $\mathbf{U}^t\mathbf{V}^t = \mathbf{U}^0\mathbf{V}^0 = \frac{1}{4}\beta^2\mathbf{I}$ stays constant for all $t \geq 0$. Since the matrix exponential of a symmetric matrix is always positive definite, the update (4.15) also implies that $\mathbf{U}^t$ and $\mathbf{V}^t$ stay positive definite for all $t \geq 0$, so that $\mathbf{U}^t$ and $\mathbf{V}^t$ are invertible. Together with the definition $\mathbf{X}^t = \mathbf{U}^t - \mathbf{V}^t$, we can solve for

$$\mathbf{U}^t = \frac{\mathbf{X}^t + \sqrt{(\mathbf{X}^t)^2 + \beta^2\mathbf{I}}}{2}, \qquad \mathbf{V}^t = \frac{-\mathbf{X}^t + \sqrt{(\mathbf{X}^t)^2 + \beta^2\mathbf{I}}}{2},$$

which completes the proof that mirror descent (4.2) is equivalent to the exponentiated gradient algorithm (4.15) when the sensing matrices $\mathbf{A}_j$'s are symmetric and commute.

□

4.8.7 Further claims

In this section, we elaborate on and justify further claims made in the main paper.

First, we demonstrate that minimizing the quantity

$$\sum_{i=1}^{n} \sigma_i \log \frac{1}{\beta} + \sigma_i \log\Big(\sigma_i + \sqrt{\sigma_i^2 + \beta^2}\Big) - \sqrt{\sigma_i^2 + \beta^2} \tag{4.28}$$

corresponds to minimizing the nuclear norm in the limit $\beta \to 0$ and to minimizing the Frobenius norm in the limit $\beta \to \infty$, see also [126], which showed the analogous result in the vector-case.

In the limit $\beta \to 0$, the term $\log \frac{1}{\beta}$ converges to infinity, hence minimizing the quantity in (4.28) corresponds to minimizing the nuclear norm $\sum_{i=1}^{n} \sigma_i$.

In the limit $\beta \to \infty$, we can substitute $z_i = \sigma_i/\beta$ and write the expression in (4.28) as

$$\beta \sum_{i=1}^{n} z_i \log\Big(z_i + \sqrt{z_i^2 + 1}\Big) - \sqrt{z_i^2 + 1} = \beta \sum_{i=1}^{n} -1 + \frac{z_i^2}{2} + \mathcal{O}\big(z_i^2\big), \tag{4.29}$$

where we applied a Taylor expansion around $z_i = 0$. Hence, minimizing the quantity in (4.28) corresponds to minimizing the Frobenius norm $(\sum_{i=1}^{n} \sigma_i^2)^{1/2}$ in the limit $\beta \to \infty$.

Next, we demonstrate that gradient descent with full-rank factorized parametrization $\mathbf{X} = \mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top$, where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n \times n}$, is a first order approximation to the exponentiated gradient algorithm defined in (4.15), with the step size rescaled by a factor 4 and the approximation being exact in the limit $\eta \to 0$, i.e. the continuous-time algorithms are equivalent.

First, using the first-order approximation $e^{\eta \mathbf{A}} = \mathbf{I} + \eta \mathbf{A} + \mathcal{O}(\eta^2)$, the exponentiated gradient algorithm becomes

$$\mathbf{X}^t = \mathbf{U}^t - \mathbf{V}^t$$
$$\mathbf{U}^{t+1} \approx \mathbf{U}^t - \eta \frac{\mathbf{U}^t \nabla f(\mathbf{X}^t) + \nabla f(\mathbf{X}^t)\mathbf{U}^t}{2}, \qquad \mathbf{V}^{t+1} \approx \mathbf{V}^t + \eta \frac{\mathbf{V}^t \nabla f(\mathbf{X}^t) + \nabla f(\mathbf{X}^t)\mathbf{V}^t}{2},$$

where we omitted higher order $\mathcal{O}(\eta^2)$ terms. On the other hand, the update for gradient descent is given by

$$\mathbf{X}^t = \mathbf{U}^t (\mathbf{U}^t)^\top - \mathbf{V}^t (\mathbf{V}^t)^\top$$
$$\mathbf{U}^{t+1} = \mathbf{U}^t - 2\eta \nabla f(\mathbf{X}^t), \qquad \mathbf{V}^{t+1} = \mathbf{V}^t + 2\eta \nabla f(\mathbf{X}^t).$$

With this, we can compute $\mathbf{U}^{t+1}(\mathbf{U}^{t+1})^\top = \mathbf{U}^t + 2\eta(\mathbf{U}^t\nabla f(\mathbf{X}^t) + \nabla f(\mathbf{X}^t)\mathbf{U}^t) + \mathcal{O}(\eta^2)$, so gradient descent with full-rank parametrization $\mathbf{X} = \mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top$, $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n\times n}$, is indeed a first-order approximation of the exponentiated gradient algorithm defined in (4.15), with the step size rescaled by a factor 4. Hence, in the limit $\eta \to 0$, the differentials $\frac{d\mathbf{X}^t}{dt} = \lim_{\eta\to 0} \frac{\mathbf{X}^{t+\eta}-\mathbf{X}^t}{\eta}$ of the exponentiated gradient algorithm (4.15) and gradient descent with full-rank factorized parametrization coincide.

4.8.8 Additional experiments for matrix completion

In this section, we present additional numerical simulations which consider matrix completion, i.e. the sensing matrices $\mathbf{A}_j$'s each have exactly one random entry set to one and all other entries set to zero. The remaining exeperimental setup is as described in Section 4.6, with the difference that we choose step sizes $\eta = 2000$ and $\eta = 500$ for mirror descent and gradient descent, respectively, due to the lower spectral norm of the sensing matrices $\mathbf{A}_j$'s in matrix completion compared to matrix sensing with random Gaussian sensing matrices. As the experiments for Figure 4.1, the experiments for Figure 4.2 were implemented in Python 3.9 and took around 10 minutes on a machine with 1.1-GHz Intel Core i5 CPU and 8 GB of RAM.

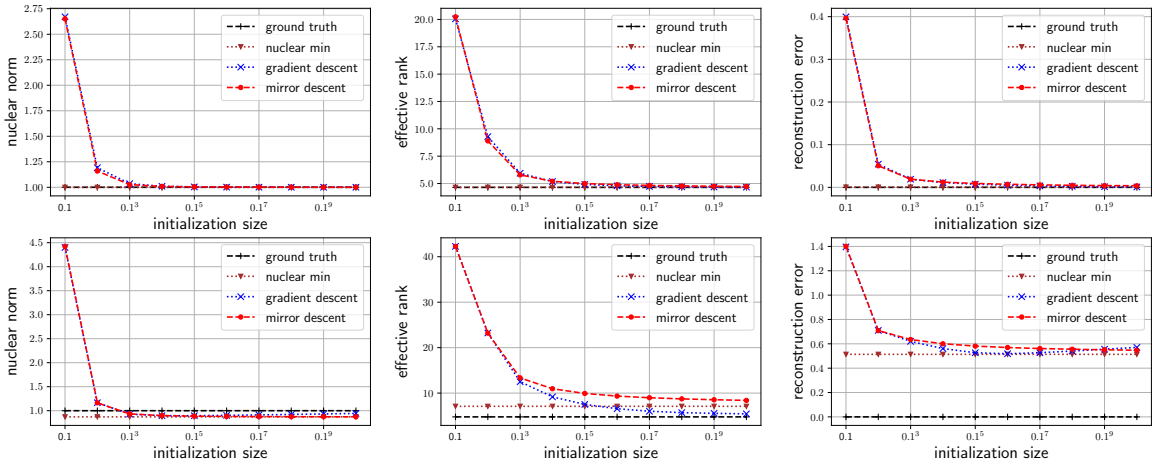


Figure 4.2: Nuclear norm, effective rank [99] and reconstruction error in matrix completion against initialization size α for $n = 50$ and $r = 5$. Top row: $m = 3nr$. Bottom row: $m = nr$.

We consider the nuclear norm $\|\mathbf{X}\|_*$, the effective rank defined in [99] and the reconstruction error $\|\mathbf{X} - \mathbf{X}^\star\|_F$ of the estimates from mirror descent, gradient descent and nuclear norm minimization, and compare these quantities to the ground truth $\mathbf{X}^\star$.

Figure 4.2 shows that the results in matrix completion qualitatively match the results in Figure 4.1 for matrix sensing with random Gaussian sensing matrices. In particular, with $m = 3nr$ observed entries (Figure 4.2, top row), nuclear norm minimization recovers the planted matrix $\mathbf{X}^\star$ and the estimates of mirror descent and gradient descent closely track each other in terms of the quantities considered. When only $m = nr$ entries are observed (Figure 4.2, bottom row), nuclear norm minimization does not recover the planted matrix $\mathbf{X}^\star$, and we observe that gradient descent puts more emphasis on lowering the effective rank at the expense of a (slightly) higher nuclear norm for initialization sizes smaller than 10^{-3}.

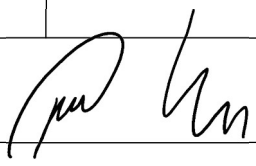
Statement of Authorship for joint/multi-authored papers for PGR thesis

To appear at the end of each thesis chapter submitted as an article/paper

The statement shall describe the candidate's and co-authors' independent research contributions in the thesis publications. For each publication there should exist a complete statement that is to be filled out and signed by the candidate and supervisor **(only required where there isn't already a statement of contribution within the paper itself).**

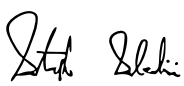
Title of Paper	Implicit regularization in matrix sensing via mirror descent
Publication Status	□Published
Publication Details	"Implicit regularization in matrix sensing via mirror descent", Fan Wu, Patrick Rebeschini. In Advances in Neural Information Processing Systems, 2021

Student Confirmation

Student Name:	Fan Wu		
Contribution to the Paper	Derived main results and technical details of the manuscript and coded the numerical experiments. Wrote first draft of the manuscript, and wrote later versions of the paper alongside Patrick Rebeschini. Wrote first draft of the response to the reviewers, and implemented feedback from the reviewers.		
Signature		Date	29/12/2021

Supervisor Confirmation

By signing the Statement of Authorship, you are certifying that the candidate made a substantial contribution to the publication, and that the description described above is accurate.

Supervisor name and title:	Patrick Rebeschini, Associate Professor		
Supervisor comments			
Signature		Date	30/12/2021

This completed form should be included in the thesis, at the end of the relevant chapter.

Chapter 5

Conclusion

In this thesis, we have studied mirror descent in two canonical high-dimensional statistical models of interest and showed that mirror descent, equipped with an appropriate mirror map and properly initialized, can provably recover an underlying low-dimensional object. In Chapters 2 and 3, we considered the problem of sparse phase retrieval and began by empirically studying Hadamard Wirtinger flow (HWF), a simple iterative algorithm that, unlike existing approaches to sparse phase retrieval, does not rely on explicit regularization in the form of ℓ_1-penalty terms, thresholding steps, sparsity-inducing priors or restrictions to low-dimensional subspaces to reconstruct a sparse signal vector. In particular, this means that, to run HWF, we do not require knowledge of the sparsity level k nor need to tune any parameters that control the sparsity level of the estimates. We demonstrated in numerical experiments that HWF can reconstruct sparse signals from fewer measurements than existing gradient-based methods, and a partial theoretical analysis of the first step of HWF revealed that one step of HWF can be used as a support recovery tool which, when combined with existing algorithms such as SPARTA, yields an end-to-end procedure for sparse phase retrieval with a potentially improved sample complexity. Concretely, the sample complexity depends on the maximum signal component and matches the best known sample requirements of $\mathcal{O}(k^2 \log n)$ for general k-sparse signals, while the sample complexity improves to $\mathcal{O}(k \log n)$ if the signal contains at least one large component.

Since HWF is a first-order approximation to mirror descent equipped with the hypentropy mirror map, we adopted the mirror descent framework in Chapter 3 and conducted

a full convergence analysis in the more general setting of noisy sparse phase retrieval. Specifically, assuming that the sample size is at least $\mathcal{O}(\max\{k^2 \log^2 n, \log^5 n\})$ and that the minimum signal component is of order $\|\mathbf{x}^\star\|_2/\sqrt{k}$, we established a quantitative control on a variational coherence property along the path of mirror descent and showed that the estimates of mirror descent remain sparse up to a prescribed stopping time. This allowed us to (i) show that early-stopped mirror descent achieves the nearly minimax-optimal rate of convergence of $\frac{\sigma}{\|\mathbf{x}^\star\|_2^2}\sqrt{\frac{k \log n}{m}}$, and to (ii) fully characterize the convergence behavior of mirror descent which can be divided into an initial warm-up stage followed by a stage where we have linear convergence towards the signal. Beyond the application of sparse phase retrieval, we hope that this type of analysis based on a quantitative version of variational coherence along the path traced by the algorithm might inspire similar approaches to investigate the convergence and implicit regularization properties of optimization algorithms in other non-convex settings.

Our analysis of HWF and mirror descent in sparse phase retrieval motivates a few possible directions for future research, which we summarize in the following.

Relaxing assumptions. Compared to an existing result that establishes a nearly minimax-optimal rate of converge for a tractable algorithm in noisy sparse phase retrieval [20], our result requires a sample complexity of $\mathcal{O}(\max\{k^2 \log^2 n, \log^5 n\})$ as well as an additional assumption on the minimum signal component. Provided that $k^2 \geq \log^3 n$, our sample requirement reads $\mathcal{O}(k^2 \log^2 n)$ and is worse by a factor $\log n$ compared to the result in [20]. We expect that both the additional logarithmic factor and the assumption on the minimum signal component are artifacts of our proof technique and not necessary for mirror descent in practice. In a nutshell, our analysis is based on proving that all support coordinates are being fitted before any of the off-support coordinates becomes large. To this end, we first analyzed the population-level dynamics, and then used concentration inequalities to bound the deviation of the finite-sample dynamics from the population-level dynamics. The additional logarithmic factor is due to the need to bound a fourth-order Gaussian polynomial over the unit sphere of k-sparse vectors, see Remark 3.1 in Section 3.8.3, while the assumption on the minimum signal component is needed since the rate at which each coordinate estimate grows depends on the corresponding signal component $x_i^\star$. Related to the sample

complexity, the empirical and partial theoretical results in Chapter 2 suggest that the sample complexity of mirror descent may depend on the maximum signal component. However, this dependency is not present in our analysis for mirror descent, since a quadratic sample complexity is needed to establish concentration of the finite-sample dynamics around their population-level version. Directly analyzing the finite-sample dynamics is challenging because the gradients of the empirical risk in sparse phase retrieval involve higher-order Gaussian polynomials with complicated dependencies both between different coordinates and between different iterates of the algorithm.

Computational complexity. Our proposed mirror descent approach to sparse phase retrieval consists of a simple initialization, followed by iterative gradient updates using the hypentropy mirror map. Given the simplicity of the algorithm, we expect it to be possible to combine mirror descent with additional algorithmic principles in order to improve its performance in practice. For instance, using a spectral initialization instead of our indicator vector initialization, we would obtain an initial estimate that correlates with the underlying signal $\mathbf{x}^\star$ (more precisely a correlation of $\Omega(1)$). We expect that this would allow us to skip the initial warm-up phase of mirror descent and thus reduce the iteration complexity by a factor $\sqrt{k}$. Another possibility would be to introduce thresholding steps, which can help keep the off-support coordinates of the estimates of mirror descent small and thus speed up the convergence of the algorithm, cf. stages (i) and (ii), part (a) in Section 3.8.1.

In Chapter 4, we studied mirror descent applied to the unregularized empirical risk in matrix sensing. A potential-based analysis in terms of the Bregman divergence yielded an explicit characterization of the limiting point of mirror descent in terms of the nuclear norm, the Frobenius norm, and the negative von Neumann entropy. While existing algorithms for matrix sensing often rely on some form of explicit regularization based on the nuclear norm or a low-rank factorized parametrization to incorporate the low-rankness structure, we can use aforementioned characterization to show that mirror descent, equipped with an appropriate mirror map, recovers low-rank matrices under the same set of assumptions which is sufficient for nuclear norm minimization to be exact [97, 98] and, in particular, relaxes existing requirements of an implicit regularization-based algorithm for matrix completion [74]. Further, we

show that gradient descent with full-rank factorized parametrization is a first-order approximation to mirror descent in the special case where the sensing matrices are symmetric and commute, and we present numerical experiments which suggest that mirror descent is well-approximated by gradient descent in some regimes, even when the sensing matrices do not commute.

Convergence speed analysis. Compared to our analysis of mirror descent in sparse phase retrieval, we achieve a sample requirement that scales linearly instead of quadratically in the respective notion of sparsity, i.e. the rank r in matrix sensing. The reason for this is because we do not need to analyze the population dynamics and show that the finite-sample dynamics concentrate around their population version, but can directly consider the finite-sample dynamics due to the convexity of the empirical risk in matrix sensing. The inner product corresponding to the variational coherence property is guaranteed to be non-negative for convex objectives, but not establishing a quantitative control on the inner product means that we do not obtain any convergence speed guarantees. In the case of positive semidefinite matrices, we expect it to be possible to adapt the analysis of [74] to mirror descent, since the matrix operations exp and log, which correspond to $\nabla\Phi^{-1}$ and $\nabla\Phi$ for the spectral entropy mirror map, leave the eigenspaces of symmetric matrices invariant, and we expect a similar adaptive decomposition of the space of positive semidefinite matrices $\mathbb{S}^n_+$ into a rank-r subspace, to which $\mathbf{X}^t$ is approximately confined, and a rank-$(n-r)$ subspace where $\mathbf{X}^t$ remains small, to be possible. However, such an analysis is likely to require a sample complexity that scales quadratically in r (and additionally depends on the condition number κ). A full convergence speed analysis with a sub-quadratic sample complexity is, if feasible, likely to require a novel analysis and is left to future work.

The noisy model. Noisy matrix sensing has been studied extensively for both convex and non-convex formulations that minimize the nuclear norm or utilize a low-rank factorized parametrization, see e.g. [23, 30, 41, 67, 80]. As in the case of sparse phase retrieval, we expect early stopping to become necessary for mirror descent if we allow the measurements to be corrupted by noise, since a global minimizer of the *unregularized* empirical risk is typically not a good estimate in the noisy model. However, our analysis for matrix sensing does not extend to the noisy case easily, as our recovery guarantees

only hold for the limiting point of mirror descent. We expect that an analysis as in [74] can be extended to the noisy case more easily, since the requirement on the RIP constant can be adapted to include the noise level in order to show that the iterations $\mathbf{X}^t$ stay sufficiently small within a rank-$(n-r)$ subspace, similarly to how we included the noise level in the sample requirement in Chapter 3 in order to show concentration of the finite-sample dynamics around their population version, which we used to show that off-support coordinates stay sufficiently small up to a prescribed stopping time.

Connection with gradient descent. Given the well-established connection between reparametrized gradient descent and mirror descent in the vector-case [3, 48, 114, 128], a natural question is whether a similar connection also extends to the matrix-case. While Proposition 4.5 in Chapter 4 answers this question affirmatively, it only holds under the assumption that the sensing matrices commute, which is a restrictive condition that is not satisfied in most applications of interest. However, our numerical experiments suggest that mirror descent is well approximated by gradient descent in some regimes where the sample size is sufficiently large, even when the sensing matrices do not commute. To gain some intuition, consider the positive semidefinite case in Proposition 4.5. For the proof, we require the matrices $\mathbf{X}^t$ and $\nabla f(\mathbf{X}^t)$ to commute, which is satisfied for the population-level dynamics as long as the initialization $\mathbf{X}_0$ commutes with $\mathbf{X}^\star$ (e.g. an initialization proportional to the identity matrix). To see this, we can compute $\mathbb{E}_{\mathbf{A}}[\nabla f(\mathbf{X})] = \mathbf{X} - \mathbf{X}^\star$, and since the operations exp and log leave the eigenspaces of symmetric matrices invariant, it follows by induction that $\mathbf{X}^t$ commutes with $\mathbf{X}^\star$ for all $t \geq 0$. This population-level analysis indicates that, indeed, mirror descent might be well-approximated by gradient descent when the number of samples is sufficiently large. However, controlling the deviation between mirror descent and gradient descent in the finite-sample case requires bounding terms involving the commutator $[\mathbf{X}^t, \nabla f(\mathbf{X}^t)]$, which is challenging due to the complicated depencies that accumulate across iterations and the higher-order terms involved.

Bibliography

[1] J. Abernethy, E. Hazan, and A. Rakhlin. Competing in the dark: An efficient algorithm for bandit linear optimization. In *Conference on Learning Theory*, pp. 263–274, 2008.

[2] A. Agarwal, S. Negahban, and M. J. Wainwright. Fast global convergence of gradient methods for high-dimensional statistical recovery. *Annals of Statistics*, 40(5):2452–2482, 2012.

[3] E. Amid and M. K. Warmuth. Reparameterizing mirror descent as gradient descent. In *Advances in Neural Information Processing Systems*, pp. 8430–8439, 2020.

[4] E. Amid and M. K. Warmuth. Winnowing with gradient descent. In *Conference on Learning Theory*, pp. 163–182, 2020.

[5] H. Andrews and C. Patterson. Singular value decomposition (SVD) image coding. *IEEE Transactions on Communications*, 24(4):425–432, 1976.

[6] S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 7411–7422, 2019.

[7] J.-Y. Audibert and S. Bubeck. Regret bounds and minimax policies under partial monitoring. *Journal of Machine Learning Research*, 11(94):2785–2836, 2010.

[8] J.-Y. Audibert, S. Bubeck, and G. Lugosi. Regret in online combinatorial optimization. *Mathematics of Operations Research*, 39(1):31–45, 2013.

[9] N. Azizan and B. Hassibi. Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. In *International Conference on Learning Representations*, 2019.

[10] N. Azizan, S. Lale, and B. Hassibi. Stochastic mirror descent on overparameterized nonlinear models. *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–11, 2021.

[11] S. Azulay, E. Moroshko, M. S. Nacson, B. E. Woodworth, N. Srebro, A. Globerson, and D. Soudry. On the implicit bias of initialization shape: Beyond infinitesimal mirror descent. In *International Conference on Machine Learning*, pp. 468–477, 2021.

[12] F. Bach. *Learning Theory from First Principles*. Draft available online under `https://www.di.ens.fr/~fbach/ltfp_book.pdf`, 2021.

[13] R. Balan, P. Casazza, and D. Edidin. On signal reconstruction without phase. *Computational Harmonic Analysis*, 20(3):345–356, 2006.

[14] M. Bayati and A. Montanari. The dynamics of message passing on dense graphs, with applications to compressed sensing. *IEEE Transactions on Information Theory*, 57(2):764–785, 2011.

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

[16] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004.

[17] S. Bubeck. Convex optimization: Algorithms and complexity. *Foundations and Trends in Machine Learning*, 8:231–358, 2015.

[18] O. Bunk, A. Diaz, F. Pfeiffer, C. David, B. Schmitt, D. K. Satapathy, and J. F. Veen. Diffractive imaging for periodic samples: Retrieving one-dimensional concenctration profiles across microfluidic channels. *Acta Crystallographica Section A: Foundations of Crystallography*, 63(4):306–314, 2007.

[19] J. F. Cai, E. J. Candès, and Z. Shen. A singular value thresholding algorithm for matrix completion. *SIAM Journal on Optimization*, 20(4):1956–1982, 2010.

[20] T. Cai, X. Li, and Z. Ma. Optimal rates of convergence for noisy sparse phase retrieval via thresholded Wirtinger flow. *Annals of Statistics*, 44(5):2221–2251, 2016.

[21] E. J. Candès and X. Li. Solving quadratic equations via PhaseLift when there are about as many equations as unknowns. *Foundations of Computational Mathematics*, 14(5):1017–1026, 2012.

[22] E. J. Candès, X. Li, and M. Soltanolkotabi. Phase retrieval via Wirtinger flow: Theory and algorithms. *IEEE Transactions on Information Theory*, 61(4):1985–2007, 2015.

[23] E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[24] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[25] E. J. Candès, T. Strohmer, and V. Voroninski. PhaseLift: Exact and stable signal recovery from magnitude measurements via convex programming. *Communications on Pure and Applied Mathematics*, 66(8):1241–1274, 2013.

[26] D. Chafaï, O. Guédon, G. Lecué, and A. Pajor. *Interactions between compressed sensing random matrices and high dimensional geometry*. Société Mathématique de France, Paris, 2012.

[27] G. Chen and M. Teboulle. Convergence analysis of a proximal-like minimization algorithm using Bregman functions. *SIAM Journal on Mathematical Analysis*, 3(3):538–543, 1993.

[28] Y. Chen and E. J. Candès. Solving random quadratic systems of equations is nearly as easy as solving linear systems. In *Advances in Neural Information Processing Systems*, pp. 739–747, 2015.

[29] Y. Chen, Y. Chi, J. Fan, and C. Ma. Gradient descent with random initialization: Fast global convergence for nonconvex phase retrieval. *Mathematical Programming*, 176(1–2):5–37, 2019.

[30] Y. Chen, Y. Chi, J. Fan, C. Ma, and Y. Yan. Noisy matrix completion: Understanding statistical guarantees for convex relaxation via nonconvex optimization. *SIAM Journal on Optimization*, 30(4):3098–3121, 2020.

[31] Y. Chen, Y. Chi, and A. J. Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7):4034–4059, 2015.

[32] Y. Chen and M. J. Wainwright. Fast low-rank estimation by projected gradient descent: General statistical and algorithmic guarantees. *arXiv preprint arXiv:1509.03025*, 2015.

[33] Y. Chi, Y. M. Lu, and Y. Chen. Non-convex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269, 2019.

[34] F. Chung and L. Lu. Concentration inequalities and martingale inequalities: A survey. *Internet Mathematics*, 3(1):79–127, 2006.

[35] G. M. Church. Genomic sequencing. *Proceedings of the National Academy of Sciences*, 81(7):1991–1995, 1984.

[36] J. C. Dainty and J. R. Fienup. Phase retrieval and image reconstruction for astronomy. In H. Stark, editor, *Image Recovery: Theory and Application*, chapter 7, pp. 231–275. Academic Press, New York, 1987.

[37] C. D. Dang and G. Lan. Stochastic block mirror descent methods for nonsmooth and stochastic optimization. *SIAM Journal on Optimization*, 25(2):856–881, 2015.

[38] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[39] A. Eftekhari and K. Zygalakis. Implicit regularization in matrix sensing: Initialization rank matters. *arXiv preprint arXiv:2008.12091*, 2021.

[40] Y. C. Eldar and S. Mendelson. Phase retrieval: Stability and recovery guarantees. *Applied and Computational Harmonic Analysis*, 36(3):473–494, 2014.

[41] M. Fazel, E. J. Candès, B. Recht, and P. Parrilo. Compressed sensing and robust recovery of low rank matrices. In *42nd Asilomar Conference on Signals, Systems and Computers*, pp. 1043–1047, 2008.

[42] J. R. Fienup. Phase retrieval algorithms: A comparison. *Applied Optics*, 21(15):2758–2769, 1982.

[43] D. J. Fresen. Variations and extensions of the Gaussian concentration inequality. *arXiv preprint arXiv:1812.10938*, 2018.

[44] Y. Freund and R. E. Shapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

[45] R. W. Gerchberg and W. O. Saxton. A practical algorithm for the determination of phase from image and diffraction. *Optik*, 35:237–246, 1972.

[46] S. Ghadimi and G. Lan. Stochastic first- and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

[47] S. Ghadimi, G. Lan, and H. Zhang. Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1–2):267–305, 2016.

[48] U. Ghai, E. Hazan, and Y. Singer. Exponentiated gradient meets gradient descent. In *International Conference on Algorithmic Learning Theory*, pp. 386–407, 2020.

[49] G. Gidel, F. Bach, and S. Lacoste-Julien. Implicit regularization of discrete gradient dynamics in neural networks. In *Advances in Neural Information Processing Systems*, pp. 3196–3206, 2019.

[50] D. Gissin, S. Shalev-Shwartz, and A. Daniely. The implicit bias of depth: How incremental learning drives generalization. In *International Conference on Learning Representations*, 2020.

[51] T. Goldstein and C. Studer. PhaseMax: Convex phase retrieval via basis pursuit. *IEEE Transactions on Information Theory*, 64(4):2675–2689, 2018.

[52] A. Graves, A. Mohamed, and G. Hinton. Speech recognition with deep recurrent neural networks. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 6645–6649, 2013.

[53] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *International Conference on Machine Learning*, pp. 1832–1841, 2018.

[54] S. Gunasekar, B. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pp. 6151–6159, 2017.

[55] P. Hand and V. Voroninski. Compressed sensing from phaseless Gaussian measurements via linear programming in the natural parameter spaces. *arXiv preprint arXiv:1611.05985*, 2016.

[56] P. D. Hoff. Lasso, fractional norm and structured sparse estimation using a Hadamard product parametrization. *Computational Statistics & Data Analysis*, 115:186–198, 2017.

[57] K. Jaganathan, Y. C. Eldar, and B. Hassibi. Phase retrieval: An overview of recent developments. In A. Stern, editor, *Optical Compressive Imaging*, chapter 13, pp. 263–296. Taylor Francis Group, Boca Raton, FL, 2016.

[58] K. Jaganathan, S. Oymak, and B. Hassibi. Sparse phase retrieval: Convex algorithms and limitations. In *Proceedings of IEEE International Symposium on Information Theory*, pp. 1022–1026, 2013.

[59] G. Jagatap and C. Hedge. Fast, sample-efficient algorithms for structured phase retrieval. In *Advances in Neural Information Processing Systems*, pp. 4917–4927, 2017.

[60] G. Jagatap and C. Hegde. Sample-efficient algorithms for recovering structured signals from magnitude-only measurements. *IEEE Transactions on Information Theory*, 65(7):4434–4456, 2019.

[61] P. Jain, R. Meka, and I. Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pp. 937–945, 2010.

[62] C. Jin, S. M. Kakade, and P. Netrapalli. Provable efficient online matrix completion via non-convex stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pp. 4520–4528, 2016.

[63] A. B. Juditsky and A. S. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

[64] S. M. Kakade, S. Shalev-Schwartz, and A. Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13(59):1865–1890, 2012.

[65] S. Kazemi, B. Yonel, and B. Yazici. Unrolled Wirtinger flow with deep priors for phaseless imaging. *arXiv preprint arXiv:2108.01735*, 2021.

[66] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[67] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from noisy entries. *Journal of Machine Learning Research*, 11:2057–2078, 2010.

[68] J. Kivinen and M. K. Warmuth. Exponentiated gradient versus gradient descent for linear predictors. *Information and Computation*, 132(1):1–63, 1997.

[69] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[70] W. Kotłowski and G. Neu. Bandit principal component analysis. In *Conference on Learning Theory*, pp. 1994–2024, 2019.

[71] G. Lecué and S. Mendelson. Minimax rate of convergence and the performance of empirical risk minimization in phase recovery. *Electronic Journal of Probability*, 20:1–29, 2015.

[72] A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(82):173–183, 1995.

[73] X. Li and V. Voroninski. Sparse signal recovery from quadratic measurements via convex programming. *SIAM Journal on Mathematical Analysis*, 45(5):3019–3033, 2013.

[74] Y. Li, T. Ma, and H. Zhang. Algorithmic regularization in over-parametrized matrix sensing and neural networks with quadratic activation. In *Conference on Learning Theory*, pp. 2–47, 2018.

[75] Z. Li, Y. Luo, and K. Lyu. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.

[76] N. Littlestone. Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine Learning*, 2(4):285–318, 1988.

[77] Z. Liu, S. Ghosh, and J. Scarlett. Towards sample-optimal compressive phase retrieval with sparse and generative priors. *arXiv preprint arXiv:2106.15358*, 2021.

[78] D. Lu and Q. Weng. A survey of image classification methods and techniques for improving classification performance. *International Journal of Remote Sensing*, 28(5):823–870, 2007.

[79] C. Ma, Y. Li, and Y. Chi. Beyond Procrustes: Balancing-free gradient descent for asymmetric low-rank matrix sensing. *arXiv preprint arXiv:2101.05113*, 2021.

[80] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval, matrix completion, and blind deconvolution. *Foundations of Computational Mathematics*, 20(3):451–632, 2020.

[81] S. Ma, D. Goldfarb, and L. Chen. Fixed point and Bregman iterative methods for matrix rank minimization. *Mathematical Programming*, 128(1–2):321–353, 2011.

[82] S. Mahadevan and B. Liu. Sparse Q-learning with mirror descent. In *Uncertainty in Artificial Intelligence*, pp. 564–573, 2012.

[83] P. Mattila. *Geometry of sets and measures in Euclidean spaces. Fractals and rectifiability.* Cambridge University Press, Cambridge, 1995.

[84] A. D. McRae, J. K. Romberg, and M. A. Davenport. Optimal convex lifted sparse phase retrieval and PCA with an atomic matrix norm regularizer. *arXiv preprint arXiv:2111.04652*, 2021.

[85] R. Millane. Phase retrieval in crystallography and optics. *Journal of the Optical Society of America A*, 7(3):394–411, 1990.

[86] W. Montgomery and S. Levine. Guided policy search via approximate mirror descent. In *Advances in Neural Information Processing Systems*, pp. 4015–4023, 2016.

[87] E. Moroshko, B. Woodworth, S. Gunasekar, J. D. Lee, N. Srebro, and D. Soudry. Implicit bias in deep linear classification: Initialization scale vs training accuracy. In *Advances in Neural Information Processing Systems*, pp. 22182–22193, 2020.

[88] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

[89] A. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization*. Wiley, New York, 1983.

[90] P. Netrapalli, P. Jain, and S. Sanghavi. Phase retrieval using alternating minimization. *IEEE Transactions on Signal Processing*, 63(18):4814–4826, 2015.

[91] H. Ohlsson, A. Y. Yang, R. Dong, and S. S. Sastry. CPRL–an extension of compressive sensing to the phase retrieval problem. In *Advances in Neural Information Processing Systems*, pp. 1367–1375, 2012.

[92] F. Orabona, K. Crammer, and N. Cesa-Bianchi. A generalized online mirror descent with applications to classification and regression. *Machine Learning*, 99:411–435, 2015.

[93] D. W. Peterson. A review of constraint qualifications in finite-dimensional spaces. *SIAM Review*, 15(3):639–654, 1973.

[94] T. Qiu and D. P. Palomar. Undersampled sparse phase retrieval via majorization–minimization. *IEEE Transactions on Signal Processing*, 65(22):5957–5969, 2017.

[95] G. Raskutti, M. J. Wainwright, and B. Yu. Early stopping and non-parametric regression: An optimal data-dependent stopping rule. *Journal of Machine Learning Research*, 15(1):335–366, 2014.

[96] N. Razin and N. Cohen. Implicit regularization in deep learning may not be explainable by norms. In *Advances in Neural Information Processing Systems*, pp. 21174–21187, 2020.

[97] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[98] B. Recht, M. Fazel, and P. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[99] O. Roy and M. Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference*, pp. 606–610, 2007.

[100] Y. Schechtman, A. Beck, and Y. C. Eldar. GESPAR: Efficient phase retrieval of sparse signals. *IEEE Transactions on Signal Processing*, 62(4):928–938, 2014.

[101] Y. Schechtman, Y. C. Eldar, O. Cohen, H. N. Chapman, J. Miao, and M. Segev. Phase retrieval with application to optical imaging: A contemporary overview. *IEEE Signal Processing Magazine*, 32(3):87–109, 2014.

[102] P. Schniter and S. Rangan. Compressive phase retrieval via generalized approximate message passing. *IEEE Transactions on Signal Processing*, 63(4):1043–1055, 2015.

[103] S. Shalev-Schwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2011.

[104] S. Shalev-Schwartz and A. Tewari. Stochastic methods for ℓ_1-regularized loss minimization. *Journal of Machine Learning Research*, 12:1865–1892, 2011.

[105] J. F. Sturm. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software*, 11-12:625–653, 1999.

[106] J. Sun, Q. Qu, and J. Wright. A geometric analysis of phase retrieval. *Foundations of Computational Mathematics*, 18(5):1131–1198, 2018.

[107] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 65(2):742–769, 2016.

[108] R. Tibshirani. Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[109] K.-C. Toh and S. Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of Optimization*, 6:615–640, 2010.

[110] M. Tomar, L. Shani, Y. Efroni, and M. Ghavamzadeh. Mirror descent policy optimization. *arXiv preprint arXiv:2005.09814*, 2020.

[111] R. S. Tsay. *Analysis of financial time series*. John Wiley & Sons, Hoboken, New Jersey, 2005.

[112] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via Procrustes flow. In *International Conference on Machine Learning*, pp. 964–973, 2016.

[113] T. Vaškevičius, V. Kanade, and P. Rebeschini. Implicit regularization for optimal sparse recovery. In *Advances in Neural Information Processing Systems*, pp. 2968–2979, 2019.

[114] T. Vaškevičius, V. Kanade, and P. Rebeschini. The statistical complexity of early-stopped mirror descent. In *Advances in Neural Information Processing Systems*, pp. 253–264, 2020.

[115] R. Vershynin. Introduction to the non-asymptotic analysis of random matrices. In Y. Eldar and G. Kutyniok, editors, *Compressed sensing, theory and applications*, chapter 5, pp. 210–268. Cambridge University Press, Cambridge, 2012.

[116] V. Voroninski and Z. Xu. A strong restricted isometry property, with an application to phaseless compressed sensing. *Applied and Computational Harmonic Analysis*, 40(2):386–395, 2016.

[117] M. J. Wainwright. *High-Dimensional Statistics: A non-asymptotic Viewpoint*, volume 48. Cambridge University Press, Cambridge, 2019.

[118] I. Waldspurger. Phase retrieval with random Gaussian sensing vectors by alternating projections. *IEEE Transactions on Information Theory*, 64(5):3301–3312, 2018.

[119] I. Waldspurger, A. d'Aspremont, and S. Mallat. Phase recovery, MaxCut and complex semidefinite programming. *Mathematical Programming*, 149(1-2):47–81, 2015.

[120] A. Walther. The question of phase retrieval in optics. *Optica Acta*, 10(1):41–49, 1963.

[121] G. Wang, G. B. Giannakis, and Y. C. Eldar. Solving systems of random quadratic equations via truncated amplitude flow. *IEEE Transactions on Information Theory*, 64(2):773–794, 2017.

[122] G. Wang, L. Zhang, G. B. Giannakis, M. Akçakaya, and J. Chen. Sparse phase retrieval via truncated amplitude flow. *IEEE Transactions on Signal Processing*, 66(2):479–491, 2018.

[123] G. A. Watson. Characterization of the subdifferential of some matrix norms. *Linear Algebra and Applications*, 170:1039–1053, 1992.

[124] Y. Wei, F. Yang, and M. J. Wainwright. Early stopping for kernel boosting algorithms: A general analysis with localized complexities. *IEEE Transactions on Information Theory*, 65(10):6685–6703, 2019.

[125] K. Q. Weinberger and L. K. Saul. Unsupervised learning of image manifolds by semidefinite programming. *International Journal of Computer Vision*, 70(1):77–90, 2006.

[126] B. Woodworth, S. Gunasekar, J. D. Lee, E. Moroshko, P. Savarese, I. Golan, D. Soudry, and N. Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673, 2020.

[127] J. Wright and Y. Ma. *High-Dimensional Data Analysis with Low-Dimensional Models: Principles, Computation, and Applications*. Cambridge University Press, Cambridge, 2022 (to appear).

[128] F. Wu and P. Rebeschini. A continuous-time mirror descent approach to sparse phase retrieval. In *Advances in Neural Information Processing Systems*, pp. 20192–20203, 2020.

[129] F. Wu and P. Rebeschini. A continuous-time mirror descent approach to sparse phase retrieval. *arXiv 2010.10168*, 2020.

[130] F. Wu and P. Rebeschini. Hadamard Wirtinger flow for sparse phase retrieval. In *International Conference on Artificial Intelligence and Statistics*, pp. 982–990, 2021.

[131] F. Wu and P. Rebeschini. Implicit regularization in matrix sensing via mirror descent. In *Advances in Neural Information Processing Systems*, pp. 20192–20203, 2021.

[132] F. Wu and P. Rebeschini. Nearly minimax-optimal rates for noisy sparse phase retrieval via early-stopped mirror descent. *arXiv preprint arXiv:2105.03678*, 2021.

[133] Z. Yang, C. Zhang, and L. Xie. Robust compressive phase retrieval via L1 minimization with application to image reconstruction. *arXiv preprint arXiv:1302.0081*, 2013.

[134] Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315, 2007.

[135] Z. Yuan, H. Wang, and Q. Wang. Phase retrieval via sparse Wirtinger flow. *Journal of Computational and Applied Mathematics*, 355:162–173, 2019.

[136] C. Yun, S. Krishnan, and H. Mobahi. A unifying view on implicit bias in training linear neural networks. In *International Conference on Learning Representations*, 2021.

[137] H. Zhang, Y. Zhou, Y. Liang, and Y. Chi. A nonconvex approach for phase retrieval: Reshaped Wirtinger flow and incremental algorithms. *Journal of Machine Learning Research*, 18(141):1–35, 2017.

[138] L. Zhang, G. Wang, G. B. Giannakis, and J. Chen. Compressive phase retrieval via reweighted amplitude flow. *IEEE Transactions on Signal Processing*, 66(19):5029–5040, 2018.

[139] T. Zhang and B. Yu. Boosting with early stopping: Convergence and consistency. *Annals of Statistics*, 33(4):1538–1579, 2005.

[140] P. Zhao, Y. Yang, and Q.-C. He. Implicit regularization via Hadamard product over-parametrization in high-dimensional linear regression. *arXiv preprint arXiv:1903.09367*, 2019.

[141] Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

[142] Z. Zhou, P. Mertikopoulos, N. Bambos, S. P. Boyd, and P. W. Glynn. On the convergence of mirror descent beyond stochastic convex programming. *SIAM Journal on Optimization*, 30(1):687–716, 2020.