Implicit Regularization in Matrix Sensing via Mirror Descent

Fan Wu
Department of Statistics
University of Oxford
fan.wu@stats.ox.ac.uk

Patrick Rebeschini
Department of Statistics
University of Oxford
patrick.rebeschini@stats.ox.ac.uk

Abstract

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.

1 Introduction

Matrix sensing represents a paradigm in modern statistics [8, 26, 27], with applications ranging from image compression [2] to collaborative filtering [18] and dimensionality reduction [34], 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 $y_i = \langle \mathbf{A}_i, \mathbf{X}^\star \rangle$, $i = 1, \dots, m$, where the sensing matrices $\mathbf{A}_i \in \mathbb{R}^{n\times n'}$ 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. [5, 15, 17, 23, 26, 27, 30]. 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. [7, 8, 16, 21, 22, 29, 31, 39].

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 [14] 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 $\mathbf{A}_i$'s commute [14], or when the sensing matrices satisfy a restricted isometry property [19]. Gradient descent with low-rank initialization within a specific "capture neighborhood" has been studied in [9], 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

35th Conference on Neural Information Processing Systems (NeurIPS 2021).

the lenses of deep matrix factorization in [3, 12, 20, 25], 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 [11].

Most theoretical results on implicit regularization in matrix sensing consider the continuous-time dynamics (i.e. gradient flow), e.g. [3, 9, 12, 14, 20, 25], or assume an infinitesimally small initialization, e.g. [14, 20]. Notable exceptions are [11], which makes a commutativity assumption on the data matrices in discrete time, and [19], 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 [27]. In the context of linear neural networks, the dependence of the implicit bias of gradient descent on the initialization has been studied in [4, 24, 35, 38]. For instance, linear diagonal networks with shared weights were considered in [35], 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 [38] 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.

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 $\mathbf{A}_i$'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 [10] 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 $\mathbf{A}_i$'s satisfy the restricted isometry property with restricted isometry constant smaller than some absolute constant [27], 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 [26]. 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 [26, 27]. A comparable result for gradient descent with full-rank factorized parametrization has been established in [19] 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 [13]. As a by-product, our proof of Theorem 1 yields an alternative proof of Theorem 1 in [13]. 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. [36, 37, 40], 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 $\mathbf{A}_i$'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 [1, 10, 33, 36]. 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 [14] 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 $\mathbf{A}_i$'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 [28] at the expense of a higher nuclear norm when fewer measurements are available, which supports the empirical observations in [3, 20].

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 $\mathbb{S}^n, \mathbb{S}^n_+ \subseteq \mathbb{R}^{n\times n}$ 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

$$D_\Phi(\mathbf{X}, \mathbf{Y}) = \Phi(\mathbf{X}) - \Phi(\mathbf{Y}) - \langle \nabla\Phi(\mathbf{Y}), \mathbf{X} - \mathbf{Y}\rangle. \tag{1}$$

Then, the mirror descent algorithm to minimize a function $f : \mathcal{D} \to \mathbb{R}$ is defined by the update

$$\nabla\Phi(\mathbf{X}_{t+1}) = \nabla\Phi(\mathbf{X}_t) - \eta_t \nabla f(\mathbf{X}_t), \tag{2}$$

where $\eta_t > 0$ is a sequence of step sizes. We approach the problem of matrix sensing, where we are given measurements $\{\mathbf{A}_i, y_i\}_{i=1}^m$, by minimizing the unregularized empirical risk

$$f(\mathbf{X}) = \frac{1}{2m}\sum_{i=1}^m \big(\langle \mathbf{A}_i, \mathbf{X}\rangle - y_i\big)^2 \tag{3}$$

using mirror descent equipped with the spectral hypentropy mirror map, which is defined as [10]

$$\Phi_\beta(\mathbf{X}) = \sum_{i=1}^n \sigma_i \operatorname{arcsinh}\left(\frac{\sigma_i}{\beta}\right) - \sqrt{\sigma_i^2 + \beta^2}, \tag{4}$$

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 A.

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 $\mathbb{S}^n$ in above definitions and consider the spectral entropy (using the convention $0\log 0 = 0$)

$$\Phi(\mathbf{X}) = \operatorname{tr}(\mathbf{X}\log\mathbf{X} - \mathbf{X}), \tag{5}$$

which is well-defined on the set of positive semidefinite matrices $\mathbb{S}^n_+$.

3 Algorithmic regularization of mirror descent

In this section, we show that mirror descent, equipped with the spectral hypentropy mirror map (4) in the rectangular case and the spectral entropy mirror map (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 $\mathbf{X} \in \mathbb{S}_+^n$.

The following Theorem applies to matrix sensing in the case of general rectangular matrices.

Theorem 1 (Rectangular case)**.** *Consider the mirror descent algorithm* (2) *with mirror map* (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 $\mathbf{A}_i$'s, observations y_i'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*

$$\sum_{i=1}^n \sigma_i \log \frac{1}{\beta} + \sigma_i \log\Big(\sigma_i + \sqrt{\sigma_i^2 + \beta^2}\Big) - \sqrt{\sigma_i^2 + \beta^2}, \tag{6}$$

where $\{\sigma_i\}_{i=1}^n$ denote the singular values of $\mathbf{X}_\infty$. We have, for any $t \geq 0$,

$$f(\mathbf{X}_t) \leq \frac{D_{\Phi_\beta}(\mathbf{X}_\infty, \mathbf{X}_0)}{\eta t}. \tag{7}$$

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:

$$D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}_{t+1}) - D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}_t) = -\eta\langle \nabla f(\mathbf{X}_t), \mathbf{X}_t - \mathbf{X}'\rangle + D_{\Phi_\beta}(\mathbf{X}_t, \mathbf{X}_{t+1}), \tag{8}$$

which follows from the definition of the Bregman divergence (1) and the mirror descent update (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 (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 (6) among all global minimizers of f.

The bound (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 1 can be found in Appendix B.1. □

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 (5) when optimizing over positive semidefinite matrices.

Theorem 2 (Positive semidefinite case)**.** *Consider the mirror descent algorithm* (2) *with mirror map* (5) *and initialization $\mathbf{X}_0 = \alpha\mathbf{I}$ for some $\alpha > 0$. Suppose that the step sizes satisfy $\eta_t \equiv \eta \leq c$, where $c > 0$ is a constant depending on the sensing matrices $\mathbf{A}_i$'s and observations y_i'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

$$\sum_{i=1}^{n} \left(\log \frac{1}{\alpha} - 1 \right) \lambda_i + \lambda_i \log \lambda_i, \tag{9}$$

where $\{\lambda_i\}_{i=1}^n$ denote the eigenvalues of $\mathbf{X}_\infty$. We have, for any $t \geq 0$,

$$f(\mathbf{X}_t) \leq \frac{D_\Phi(\mathbf{X}_\infty, \mathbf{X}_0)}{\eta t}. \tag{10}$$

Theorem 2 can be proved the same way as Theorem 1, see Appendix B.2. Neither Theorem requires any assumptions on the sensing matrices $\mathbf{A}_i$'s beyond the existence of a matrix $\mathbf{X}'$ achieving zero training error $f(\mathbf{X}') = 0$, which, for instance, is satisfied when $\{\mathbf{A}_i\}_{i=1}^m$ are linearly independent and $m \leq nn'$ in the rectangular case or $m \leq n(n+1)/2$ 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 [13]. Once we have established convergence of mirror descent towards a global minimizer of f, we could have alternatively invoked Theorem 1 of [13] 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 (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 [13], our analysis presents an alternative proof technique for Theorem 1 in [13] 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 (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 B.6 for further details. This can be seen as a matrix analogue of the result in [35], 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. [33, 36]. In the positive semidefinite case, minimizing the quantity in (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 (9) is equivalent to maximizing the nuclear norm.

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 $y_i = \langle \mathbf{A}_i, \mathbf{X}^\star \rangle$, $i = 1, \ldots, m$. Theorem 1 and Theorem 2 imply that mirror descent, equipped with the spectral hypentropy mirror map (4) or spectral entropy mirror map (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.

4.1 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 [27], and has also been used in [19] to show that gradient descent with full-rank factorized parametrization recovers $\mathbf{X}^\star$.

Definition 1 (Restricted isometry property [27]). *A set of matrices $\mathbf{A}_1, \dots, \mathbf{A}_m \in \mathbb{R}^{n \times n'}$ satisfies (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*

$$(1-\delta)\|\mathbf{X}\|_F \leq \left(\frac{1}{m}\sum_{i=1}^{m}\langle \mathbf{A}_i, \mathbf{X}\rangle^2\right)^{1/2} \leq (1+\delta)\|\mathbf{X}\|_F.$$

With this definition, we have the following recovery guarantee for mirror descent.

Theorem 3. *Assume that the set of measurement matrices $\{\mathbf{A}_i\}_{i=1}^m$ satisfies $(5r, \delta)$-restricted isometry property with $\delta \leq 0.1$. Then, the mirror descent algorithm described in Theorem 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*

$$\|\mathbf{X}_\infty - \mathbf{X}^\star\|_F \leq \frac{\Delta_\beta \|\mathbf{X}^\star\|_* + (1+\Delta_\beta)\frac{n\beta}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}}{C_\delta \sqrt{3r}}, \tag{11}$$

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 2 with $\alpha < \frac{\|\mathbf{X}^\star\|_*}{en}$ converges to a positive semidefinite matrix $\mathbf{X}_\infty$ that satisfies

$$\|\mathbf{X}_\infty - \mathbf{X}^\star\|_F \leq \frac{\Delta_\alpha \|\mathbf{X}^\star\|_*}{C_\delta \sqrt{3r}}, \tag{12}$$

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 3 is Theorem 1.1 in [19], 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 3 differs significantly from the proof of Theorem 1.1 in [19]. The proof of Theorem 1.1 in [19] involves an analysis of the trajectory of gradient descent, relating it to the population dynamics defined by the population risk $\mathbb{E}_{(\mathbf{A}_i)_{kl}\sim\mathcal{N}(0,1)}[f(\mathbf{U}_t\mathbf{U}_t^\top)] = \|\mathbf{U}_t\mathbf{U}_t^\top - \mathbf{X}^\star\|_F^2$, and identifying adaptive rank-r subspaces to which the iterates $\mathbf{U}_t$ are approximately confined. On the other hand, Theorem 3 follows from Theorem 1 and Theorem 2, which rely on a simple analysis of the evolution of the Bregman divergence, and employs arguments similar to the ones used in [27] 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 [19] 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. [16, 21, 29, 31, 39].

Sample complexity and condition number. The restricted isometry property assumption in Theorem 1.1 in [19] 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 $\mathbf{A}_i$'s belong to a general class of random matrices [27]. It is conjectured in [19] 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 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 [27] 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 [27].

Convergence speed and dependence on intialization size. The analysis along the trajectory of gradient descent in [19] 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 (7) and (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. [14, 20, 35].

We remark that, beyond the aforementioned result on implicit regularization in matrix sensing [19], 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 [32] and sparse phase retrieval [36, 37]. 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.

4.2 Matrix completion

In matrix completion, a subset of the entries of the matrix $\mathbf{X}^\star$ is observed, and the corresponding sensing matrices $\mathbf{A}_i$'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 [26] and gradient descent [22], for instance.

Definition 2 (Coherence [6])**.** *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 ($\{\mathbf{e}_i\}_{i=1}^n$ 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 [26].

Theorem 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 1 with $\beta < \frac{\|\mathbf{X}^\star\|_*}{1.05en}$ converges to a matrix $\mathbf{X}_\infty = \lim_{t \to \infty} \mathbf{X}_t$ that satisfies

$$\|\mathbf{X}_\infty - \mathbf{X}^\star\|_F \leq 6\left(\Delta_\beta \|\mathbf{X}^\star\|_* + (1+\Delta_\beta)\frac{n\beta}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}\right)\left(1 + \left(\frac{128cnn' \log^2 n'}{9m}\right)^{\frac{1}{2}}\right), \quad (13)$$

with probability at least $1 - 6\log(n')(n+n')^{2-2c} - (n')^{2-2\sqrt{c}}$, 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 2 with $\alpha < \frac{\|\mathbf{X}^\star\|_*}{en}$ converges to a positive semidefinite matrix $\mathbf{X}_\infty$ that satisfies

$$\|\mathbf{X}_\infty - \mathbf{X}^\star\|_F \leq 6\Delta_\alpha \|\mathbf{X}^\star\|_* \left(1 + \left(\frac{128cn^2 \log^2 n}{9m}\right)^{\frac{1}{2}}\right), \quad (14)$$

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 [22] 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 $r^3 n \log^3 n$.

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 [10, 33, 36],

and a general framework connecting mirror descent and reparametrized gradient descent was studied in [1]. 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 $\mathbf{A}_i \in \mathbb{R}^{n\times n}$, $i = 1, \dots, m$. Consider the following exponentiated gradient algorithm given by

$$\mathbf{X}_t = \mathbf{U}_t - \mathbf{V}_t$$
$$\mathbf{U}_{t+1} = \frac{\mathbf{U}_t e^{-\eta\nabla f(\mathbf{X}_t)} + e^{-\eta\nabla f(\mathbf{X}_t)}\mathbf{U}_t}{2}, \qquad \mathbf{V}_{t+1} = \frac{\mathbf{V}_t e^{\eta\nabla f(\mathbf{X}_t)} + e^{\eta\nabla f(\mathbf{X}_t)}\mathbf{V}_t}{2}. \tag{15}$$

When considering the full-rank parametrization $\mathbf{X} = \mathbf{U}\mathbf{U}^\top - \mathbf{V}\mathbf{V}^\top$, where $\mathbf{U}, \mathbf{V} \in \mathbb{R}^{n\times n}$, gradient descent on the variables $(\mathbf{U}, \mathbf{V})$ is a first-order approximation to the exponentiated gradient algorithm defined in (15), with the step size rescaled by a constant factor and the approximation being exact in the limit $\eta \to 0$, see Appendix B.6 for details. The gradient descent algorithm considered in [3, 14, 19, 20] can be obtained by initializing $\mathbf{V}_0 = \mathbf{0}$.

Proposition 5. *Assume that the sensing matrices $\mathbf{A}_i$'s are symmetric and commute. Then:*

1. *Mirror descent equipped with the spectral entropy mirror map* (5) *and any positive definite initialization $\mathbf{X}_0$ which commutes with all $\mathbf{A}_i$'s (e.g. $\mathbf{X}_0 = \alpha\mathbf{I}$ for some $\alpha > 0$) is equivalent to the exponentiated gradient algorithm defined in* (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) *with parameter $\beta > 0$ and initialization $\mathbf{X}_0 = \mathbf{0}$ is equivalent to the exponentiated gradient algorithm defined in* (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 [14]. Let $\widetilde{f}(\mathbf{U}) = f(\mathbf{U}\mathbf{U}^\top)$.

Corollary 6. *Assume that the sensing matrices $\mathbf{A}_i$'s are symmetric and commute, and that there exists a $\mathbf{X}' \in \mathbb{S}^n_+$ 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 [14] 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 [14], 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 2, which then carries over to gradient flow on the non-convex objective $\widetilde{f}$ via Proposition 5 (when the $\mathbf{A}_i$'s are symmetric and commute).

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 (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 [28] $\mathrm{effrank}(\mathbf{X}) = \exp(-\sum_{i=1}^n p_i \log p_i)$, 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 C 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 $\mathbf{A}_i = \frac{1}{2}(\mathbf{B}_i + \mathbf{B}_i^\top)$, where the entries of $\mathbf{B}_i$ 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, $\mathrm{argmin}\{\|\mathbf{X}\|_* : \mathbf{X} \succeq \mathbf{0}, f(\mathbf{X}) = 0\}$, using the

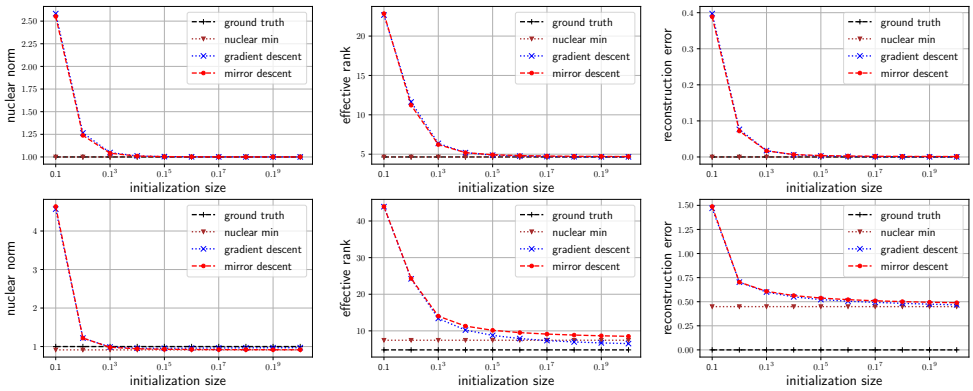

Figure 1: Nuclear norm, effective rank [28] and reconstruction error against initialization size α for $n = 50$ and $r = 5$. Top row: $m = 3nr$ measurements. Bottom row: $m = nr$ measurements.

`cvxopt` package. The experiments for Figure 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.

With $m = 3nr$ measurements (Figure 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 [28] and reconstruction error, even though the sensing matrices do not commute. When we have $m = nr$ measurements (Figure 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 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 [3, 20], which suggest that a notion of rank-minimization is involved in the implicit bias of gradient descent.

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 $\mathbf{X}^\star$ [26, 27].

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 A. 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 $\nabla f(\mathbf{X}) = \frac{1}{m}\sum_{i=1}^m(\langle \mathbf{A}_i, \mathbf{X}\rangle - y_i)\mathbf{A}_i$ requires $\mathcal{O}(mnn')$ operations, and $m > n$ is typically required to recover $\mathbf{X}^\star$. However, the inner product $\langle \mathbf{A}_i, \mathbf{X}\rangle$ 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. [8, 22].

Previous results on implicit regularization-based algorithms in matrix sensing [19], sparse linear regression [32] and sparse phase retrieval [36, 37] establish convergence speed guarantees and a poly-

nomial 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.

Acknowledgments and Disclosure of Funding

Fan Wu is supported by the EPSRC and MRC through the OxWaSP CDT programme (EP/L016710/1). Patrick Rebeschini was supported in part by the Alan Turing Institute under the EPSRC grant EP/N510129/1.

References

[1] E. Amid and M. K. Warmuth. Reparameterizing mirror descent as gradient descent. In *Advances in Neural Information Processing Systems*, pages 8430–8439, 2020.

[2] H. Andrews and C. Patterson. Singular value decomposition (SVD) image coding. *IEEE Transactions on Communications*, 24(4):425–432, 1976.

[3] S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, pages 7411–7422, 2019.

[4] S. Azulay, E. Moroshko, M. S. Nacson, B. Woodworth, N. Srebro, A. Globerson, and D. Soudry. On the implicit bias of initialization shape: Beyond infinitesmial mirror descent. *arXiv preprint arXiv:2102.09769*, 2021.

[5] 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.

[6] E. J. Candès and B. Recht. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

[7] 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.

[8] 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.

[9] A. Eftekhari and K. Zygalakis. Implicit regularization in matrix sensing: Initialization rank matters. *arXiv preprint arXiv:2008.12091*, 2021.

[10] U. Ghai, E. Hazan, and Y. Singer. Exponentiated gradient meets gradient descent. In *International Conference on Algorithmic Learning Theory*, volume 117, pages 386–407, 2020.

[11] G. Gidel, F. Bach, and S. Lacoste-Julien. Implicit regularization of discrete gradient dynamics in neural networks. In *Advances in Neural Information Processing Systems*, pages 3196–3206, 2019.

[12] 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.

[13] S. Gunasekar, J. Lee, D. Soudry, and N. Srebro. Characterizing implicit bias in terms of optimization geometry. In *Proceedings of the 35th International Conference on Machine Learning*, pages 1832–1841, 2018.

[14] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, pages 6151–6159, 2017.

[15] P. Jain, R. Meka, and I. Dhillon. Guaranteed rank minimization via singular value projection. In *Advances in Neural Information Processing Systems*, pages 937—-945, 2010.

[16] 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*, pages 4520–4528, 2016.

[17] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[18] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[19] 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*, pages 2–47, 2018.

[20] 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.

[21] 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.

[22] C. Ma, K. Wang, Y. Chi, and Y. Chen. Implicit regularization in nonconvex statistical estimation: Gradient descent converges linearly for phase retrieval and matrix completion. In *International Conference on Machine Learning*, pages 3345–3354, 2018.

[23] 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.

[24] 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*, pages 22182–22193, 2020.

[25] N. Razin and N. Cohen. Implicit regularization in deep learning may not be explainable by norms. In *Advances in Neural Information Processing Systems*, pages 21174–21187, 2020.

[26] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12:3413–3430, 2011.

[27] 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.

[28] O. Roy and M. Vetterli. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference*, pages 606–610, 2007.

[29] R. Sun and Z.-Q. Luo. Guaranteed matrix completion via non-convex factorization. *IEEE Transactions on Information Theory*, 65(2):742–769, 2016.

[30] 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.

[31] S. Tu, R. Boczar, M. Simchowitz, M. Soltanolkotabi, and B. Recht. Low-rank solutions of linear matrix equations via Procrustes flow. In *Proceedings of The 33rd International Conference on Machine Learning*, pages 964–973, 2016.

[32] T. Vaškevičius, V. Kanade, and P. Rebeschini. Implicit regularization for optimal sparse recovery. In *Advances in Neural Information Processing Systems*, pages 2968–2979, 2019.

[33] T. Vaškevičius, V. Kanade, and P. Rebeschini. The statistical complexity of early-stopped mirror descent. In *Advances in Neural Information Processing Systems*, volume 33, pages 253–264, 2020.

[34] 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.

[35] 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*, pages 3635–3673, 2020.

[36] F. Wu and P. Rebeschini. A continuous-time mirror descent approach to sparse phase retrieval. In *Advances in Neural Information Processing Systems*, pages 20192–20203, 2020.

[37] 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.

[38] 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.

[39] Q. Zheng and J. Lafferty. Convergence analysis for rectangular matrix completion using Burer-Monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

[40] 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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] See Sections 4.1 and 7.
   (c) Did you discuss any potential negative societal impacts of your work? [No] Our work is of theoretical nature and does not present any foreseeable societal consequences.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] See Sections 3, 4 and 5.
   (b) Did you include complete proofs of all theoretical results? [Yes] See Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] See supplemental material or `https://github.com/fawuuu/irmsmd`.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See Section 6.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] We consider the results for a single run to be more meaningful compared to the average obtained from multiple runs with error bars, since the estimates of mirror descent and gradient descent being close to each other for any individual run is a stronger statement than their means and standard deviations being close. This is analogous to $X, Y \sim \mathcal{N}(0, 1)$, which have the same mean and standard deviation, but each relization of X could differ significantly from Y.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See Section 6 and Appendix.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]

(e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]

5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]