Appendix

The Appendix is organized as follows. In Appendix A, we provide the gradients for the spectral entropy (5) and spectral hypentropy (4) mirror maps, and discuss the per-iteration computational cost of the corresponding mirror descent algorithms. In Appendix B, we provide proofs for the claims made in the main paper. In Appendix C, we present additional experiments addressing the problem of matrix completion.

A Mirror maps and gradients

We first consider rectangular matrix sensing with the spectral hypentropy mirror map

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

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 [7] 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 [7], see also [4]. 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 $\mathcal{O}(n^2 n')$ operations.

The singular value decomposition can be avoided if $n = n'$ and the sensing matrices $\mathbf{A}_i$'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 $y_i = \langle \mathbf{A}_i, \mathbf{X}^\star\rangle = \langle \frac{1}{2}(\mathbf{A}_i + \mathbf{A}_i^\top), \mathbf{X}^\star\rangle$ for all $i = 1, \ldots, m$. 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 (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.

B Proofs

In this section, we provide proofs for the claims made in the main paper.

B.1 Proof of Theorem 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 (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 (1) and the mirror descent update (2):

$$D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}_{t+1}) - D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}_t) = -\eta \langle \nabla f(\mathbf{X}_t), \mathbf{X}_t - \mathbf{X}' \rangle + D_{\Phi_\beta}(\mathbf{X}_t, \mathbf{X}_{t+1}), \tag{16}$$

where $\mathbf{X}'$ is any reference point. Letting $\mathbf{X}'$ be any global minimizer of f, the first term in (16) can be written as

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

where we used the assumption that there exists a matrix achieving zero training error, i.e. $\langle \mathbf{A}_i, \mathbf{X}' \rangle = y_i$ for all $i = 1, \ldots, m$. 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 $\mathcal{B}(\tau) = \{ \mathbf{X} \in \mathbb{R}^{n \times n'} : \|\mathbf{X}\|_* \leq \tau \}$, see Theorem 14 in [4]. Writing $\tau_t = \max\{ \|\mathbf{X}_t\|_*, \|\mathbf{X}_{t+1}\|_* \}$, we can bound the second term in (16) by

$$\begin{aligned}
D_{\Phi_\beta}(\mathbf{X}_t, \mathbf{X}_{t+1}) &= \Phi_\beta(\mathbf{X}_t) - \Phi_\beta(\mathbf{X}_{t+1}) - \langle \nabla \Phi_\beta(\mathbf{X}_{t+1}), \mathbf{X}_t - \mathbf{X}_{t+1} \rangle \\
&\leq \langle \nabla \Phi_\beta(\mathbf{X}_t) - \nabla \Phi_\beta(\mathbf{X}_{t+1}), \mathbf{X}_t - \mathbf{X}_{t+1} \rangle - \frac{1}{4(\tau_t + \beta n)} \|\mathbf{X}_t - \mathbf{X}_{t+1}\|_*^2 \\
&= \langle \eta \nabla f(\mathbf{X}_t), \mathbf{X}_t - \mathbf{X}_{t+1} \rangle - \frac{1}{4(\tau_t + \beta n)} \|\mathbf{X}_t - \mathbf{X}_{t+1}\|_*^2 \\
&\leq \eta \|\nabla f(\mathbf{X}_t)\|_2 \|\mathbf{X}_t - \mathbf{X}_{t+1}\|_* - \frac{1}{4(\tau_t + \beta n)} \|\mathbf{X}_t - \mathbf{X}_{t+1}\|_*^2 \\
&\leq \eta^2 (\tau_t + \beta n) \|\nabla f(\mathbf{X}_t)\|_2^2,
\end{aligned} \tag{17}$$

where we used strong convexity of Φ_β in the second line, the mirror descent update (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

$$\begin{aligned}
\|\nabla f(\mathbf{X}_t)\|_2^2 &= \left\| \frac{1}{m} \sum_{i=1}^{m} \big(\langle \mathbf{A}_i, \mathbf{X}_t \rangle - y_i \big) \mathbf{A}_i \right\|_2^2 \\
&\leq \left(\frac{1}{m} \sum_{i=1}^{m} \big| \langle \mathbf{A}_i, \mathbf{X}_t \rangle - y_i \big| \|\mathbf{A}_i\|_2 \right)^2 \\
&\leq \frac{1}{m} \sum_{i=1}^{m} \|\mathbf{A}_i\|_2^2 \cdot 2f(\mathbf{X}_t),
\end{aligned}$$

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 (17) to obtain

$$\|\mathbf{X}_t - \mathbf{X}_{t+1}\|_* \leq 4(\tau_t + \beta n) \eta \|\nabla f(\mathbf{X}_t)\|_2 \leq \frac{1}{2}(\tau_t + \beta n), \tag{18}$$

provided the step size η satisfies

$$\eta \leq \frac{1}{8\sqrt{2}} \left(\frac{1}{m} \sum_{i=1}^{m} \|\mathbf{A}_i\|_2^2 \cdot f(\mathbf{X}_t) \right)^{-1/2}. \tag{19}$$

We will show below that the upper bound in (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 (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

$$\eta \leq \frac{1}{4}\left(\frac{1}{m}\sum_{i=1}^m \|\mathbf{A}_i\|_2^2 \cdot \Big(\|\mathbf{X}_t\|_* + \beta n\Big)\right)^{-1}. \tag{20}$$

With this, the identity in (16) becomes

$$D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}_{t+1}) - D_{\Phi_\beta}(\mathbf{X}', \mathbf{X}_t) = -2\eta f(\mathbf{X}_t) + D_{\Phi_\beta}(\mathbf{X}_t, \mathbf{X}_{t+1}) \leq -\eta f(\mathbf{X}_t) \tag{21}$$

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 (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

$$\begin{aligned} D_{\Phi_\beta}(\mathbf{X}_\infty, \mathbf{X}_0) &= \sum_{i=1}^n \sigma_i \operatorname{arcsinh}\Big(\frac{\sigma_i}{\beta}\Big) - \sqrt{\sigma_i^2 + \beta^2} - n\beta \\ &= \sum_{i=1}^n \sigma_i \log\frac{1}{\beta} + \sigma_i \log\Big(\sigma_i + \sqrt{\sigma_i^2 + \beta^2}\Big) - \sqrt{\sigma_i^2 + \beta^2} - n\beta \end{aligned}$$

is minimized among all global minimizers of the empirical risk f, which is the quantity in (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 (19) and (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 $\mathbf{A}_i$'s and the observations y_i's.

Part 2: Proving the bound (7).
In order to show the bound (7), we first show that $f(\mathbf{X}_t)$ decreases monotonously. To this end, we verifiy that f is $\frac{1}{m}\sum_{i=1}^m \|\mathbf{A}_i\|_2^2$-smooth with respect to the nuclear norm. Indeed, ∇f is Lipschitz continuous with Lipschitz constant $\frac{1}{m}\sum_{i=1}^m \|\mathbf{A}_i\|_2^2$,

$$\begin{aligned} \|\nabla f(\mathbf{X}) - \nabla f(\mathbf{Y})\|_2 &= \left\|\frac{1}{m}\sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{X} - \mathbf{Y}\rangle \mathbf{A}_i\right\|_2 \\ &\leq \left\|\frac{1}{m}\sum_{i=1}^m \|\mathbf{A}_i\|_2 \|\mathbf{X} - \mathbf{Y}\|_* \mathbf{A}_i\right\|_2 \\ &\leq \frac{1}{m}\sum_{i=1}^m \|\mathbf{A}_i\|_2^2 \cdot \|\mathbf{X} - \mathbf{Y}\|_*, \end{aligned}$$

where we used the duality of the nuclear and spectral norms in the second line. Hence, we can bound

$$f(\mathbf{X}_{t+1}) \leq f(\mathbf{X}_t) + \langle \nabla f(\mathbf{X}_t), \mathbf{X}_{t+1} - \mathbf{X}_t\rangle + \frac{1}{2m}\sum_{i=1}^m \|\mathbf{A}_i\|_2^2 \cdot \|\mathbf{X}_{t+1} - \mathbf{X}_t\|_*^2.$$

If we can bound

$$\langle \nabla f(\mathbf{X}_t), \mathbf{X}_{t+1} - \mathbf{X}_t \rangle \le -\frac{1}{2m} \sum_{i=1}^{m} \|\mathbf{A}_i\|_2^2 \cdot \|\mathbf{X}_{t+1} - \mathbf{X}_t\|_*^2,$$

then this would show that $f(\mathbf{X}_t)$ is monotonously decreasing. Recall the proximal formulation of mirror descent (see e.g. [1]),

$$\mathbf{X}_{t+1} = \arg \min_{\mathbf{X} \in \mathbb{R}^{n \times n'}} \left\{ \langle \nabla f(\mathbf{X}_t), \mathbf{X} - \mathbf{X}_t \rangle + \frac{1}{\eta} D_{\Phi_\beta}(\mathbf{X}, \mathbf{X}_t) \right\}.$$

Since the quantity being minimized is zero for $\mathbf{X} = \mathbf{X}_t$, we obtain the upper bound

$$\begin{aligned} \langle \nabla f(\mathbf{X}_t), \mathbf{X}_{t+1} - \mathbf{X}_t \rangle &\le -\frac{1}{\eta} D_{\Phi_\beta}(\mathbf{X}_{t+1}, \mathbf{X}_t) \\ &= \frac{1}{\eta} \Big((\Phi_\beta(\mathbf{X}_t) - \Phi_\beta(\mathbf{X}_{t+1})) + \langle \nabla \Phi_\beta(\mathbf{X}_t), \mathbf{X}_{t+1} - \mathbf{X}_t \rangle \Big) \\ &\le -\frac{1}{4\eta(\tau_t + \beta n)} \|\mathbf{X}_{t+1} - \mathbf{X}_t\|_*^2 \\ &\le -\frac{1}{2m} \sum_{i=1}^{m} \|\mathbf{A}_i\|_2^2 \cdot \|\mathbf{X}_{t+1} - \mathbf{X}_t\|_*^2, \end{aligned}$$

where we used strong convexity of Φ_β for the second inequality, and the last inequality holds if the step size η satisfies inequality (20). This completes the proof that $f(\mathbf{X}_{t+1}) \le f(\mathbf{X}_t)$.

To show the bound (7), assume that it were violated for some $t > 0$. Since $f(\mathbf{X}_t)$ is non-increasing, this means that

$$f(\mathbf{X}_s) \ge f(\mathbf{X}_t) > \frac{D_{\Phi_\beta}(\mathbf{X}_\infty, \mathbf{X}_0)}{\eta t}$$

for all $s \le t$. The bound in (21) controls by how much the Bregman divergence must decrease in each iteration. Summing over the expression in (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 (7) must be satisfied for all $t > 0$. ☐

B.2 Proof of Theorem 2

The proof of Theorem 2 follows the same steps as the proof of Theorem 1 and uses the fact that the spectral entropy (5) is $(2\tau)^{-1}$ strongly convex with respect to the nuclear norm on the nuclear norm ball $\mathcal{B}_+(\tau) = \{\mathbf{X} \in \mathbb{S}_+^n : \|\mathbf{X}\|_* \le \tau\}$, for which we include a proof for completeness' sake.

Lemma 7 (Strong convexity of the spectral entropy)**.** *The spectral entropy* (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 7 closely follows the proof of strong convexity of the spectral hypentropy mirror map provided in [4]. We first introduce some notation. We denote by $\lambda(\mathbf{X})$ the vector of eigenvalues of a symmetric matrix $\mathbf{X} \in \mathbb{S}^n$. 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. [4],

$$\mathbf{X} = \mathbf{U} \operatorname{diag}[\lambda(\mathbf{X})] \mathbf{U}^\top \qquad \Rightarrow \qquad f(\mathbf{X}) = \mathbf{U} \operatorname{diag}[f(\lambda(\mathbf{X}))] \mathbf{U}^\top,$$

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 [6] 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 8 (Theorem 28 [6])**.** *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 [5] allows us to reduce the smoothness of matrix functions to the smoothness of functions taking vectors as argument.

Lemma 9 (Proposition 3.1 [5])**.** *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 $F : \mathbb{S}^n \to \mathbb{R}$ defined by $F(\mathbf{X}) = \operatorname{tr}(f(\mathbf{X}))$ is twice continuously differentiable and satisfies, for every $\mathbf{H} \in \mathbb{S}^n$,

$$D^2F(\mathbf{X})[\mathbf{H},\mathbf{H}] \leq c\operatorname{tr}(\mathbf{H}f''(\mathbf{X})\mathbf{H}).$$

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 9, we can bound, for any $\mathbf{X} = \nabla\Phi(\mathbf{Y})$ with $\mathbf{Y} \in \mathcal{B}_+(\tau)$,

$$\begin{aligned}
\sup_{\mathbf{H}\in\mathbb{S}^n:\|\mathbf{H}\|_2\leq 1} D^2\Phi^*(\mathbf{X})[\mathbf{H},\mathbf{H}] &\leq \sup_{\mathbf{H}\in\mathbb{S}^n:\|\mathbf{H}\|_2\leq 1} 2\operatorname{tr}(\mathbf{H}(\phi^*)''(\mathbf{X})\mathbf{H}) \\
&= \sup_{\mathbf{H}\in\mathbb{S}^n:\|\mathbf{H}\|_2\leq 1} 2\operatorname{tr}(\mathbf{H}^2(\phi^*)''(\mathbf{X})) \\
&\leq \sup_{\mathbf{H}\in\mathbb{S}^n:\|\mathbf{H}\|_2\leq 1} 2\langle\sigma^2(\mathbf{H}), \sigma((\phi^*)''(\mathbf{X}))\rangle,
\end{aligned}$$

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

$$\sup_{\mathbf{H}\in\mathbb{S}^n:\|\mathbf{H}\|_2\leq 1} D^2\Phi^*(\mathbf{X})[\mathbf{H},\mathbf{H}] \leq 2\sum_{i=1}^{n} 1\cdot\sigma_i(\mathbf{Y}) = 2\|\mathbf{Y}\|_* \leq 2\tau,$$

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 2 follows the exact same steps as the proof of Theorem 1, it is omitted to avoid repetition.

B.3 Proof of Theorem 3

Proof of Theorem 3. We begin by considering the rectangular case and prove the bound (11) in Theorem 3 for the spectral hypentropy mirror map (4). The proof of Theorem 3 is an adaption of and builds upon the proofs of Theorem 3.3 in [9] and Theorem 4 in [3]. First, we need to bound the nuclear norm of the matrix $\mathbf{X}_\infty$. Then, we follow [3, 9] 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 1, $\mathbf{X}_\infty$ minimizes the quantity in (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

$$\|\mathbf{X}_\infty\|_* \leq (1 + \Delta_\beta) \left(\|\mathbf{X}^\star\|_* + \frac{n\beta}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1} \right), \tag{22}$$

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 [3, 9]. Writing $\mathbf{R} = \mathbf{X}_\infty - \mathbf{X}^\star$, we can apply Lemma 3.4 from [9] 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 [9]. Together with (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

$$\begin{aligned}
\sum_{j \geq 2} \|\mathbf{R}_j\|_F &\leq \frac{1}{\sqrt{3r}} \sum_{j \geq 1} \|\mathbf{R}_j\|_* \\
&= \frac{1}{\sqrt{3r}} \|\mathbf{R}_c\|_* \\
&\leq \frac{\sqrt{2r}}{\sqrt{3r}} \|\mathbf{R}_0\|_F + \frac{\Delta_\beta \|\mathbf{X}^\star\|_* + (1+\Delta_\beta)\frac{n\beta}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}}{\sqrt{3r}},
\end{aligned} \tag{23}$$

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

$$\begin{aligned}
\left(\frac{1}{m}\sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{R} \rangle^2\right)^{1/2} &\geq \left(\frac{1}{m}\sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{R}_0 + \mathbf{R}_1 \rangle^2\right)^{1/2} - \sum_{j \geq 2}\left(\frac{1}{m}\sum_{i=1}^m \langle \mathbf{A}_i, \mathbf{R}_j \rangle^2\right)^{1/2} \\
&\geq (1-\delta)\|\mathbf{R}_0 + \mathbf{R}_1\|_F - \sum_{j \geq 2}(1+\delta)\|\mathbf{R}_j\|_F.
\end{aligned} \tag{24}$$

Since $\mathbf{R}_0$ is orthogonal to $\mathbf{R}_1$ (see Lemma 3.4 in [9]), 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 $\langle \mathbf{A}_i, \mathbf{R} \rangle = 0$ for all $i = 1, \ldots, m$. Hence, we can use (23) and rearrange (24) for $\|\mathbf{R}_0 + \mathbf{R}_1\|_F$ to obtain

$$\|\mathbf{R}_0 + \mathbf{R}_1\|_F \leq \left(1 - \sqrt{\frac{2}{3}} - \delta\left(1 + \sqrt{\frac{2}{3}}\right)\right)^{-1} (1+\delta) \frac{\Delta_\beta \|\mathbf{X}^\star\|_* + (1+\Delta_\beta)\frac{n\beta}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}}{\sqrt{3r}}.$$

Finally, this yields

$$\begin{aligned}
\|\mathbf{R}\|_F &\leq \|\mathbf{R}_0 + \mathbf{R}_1\|_F + \sum_{j \geq 2} \|\mathbf{R}_j\|_F \\
&\leq 2\left(1 - \sqrt{\frac{2}{3}} - \delta\left(1 + \sqrt{\frac{2}{3}}\right)\right)^{-1} \frac{\Delta_\beta \|\mathbf{X}^\star\|_* + (1+\Delta_\beta)\frac{n\beta}{\log \frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}}{\sqrt{3r}},
\end{aligned}$$

which completes the proof of the bound (11) in Theorem 3. The bound (12) can be shown following the same steps, and we omit the details to avoid repetition. ☐

B.4 Proof of Theorem 4

Proof of Theorem 4. As in Theorem 3, we first consider the rectangular case and show the bound (13) in Theorem 4. The proof of Theorem 4 combines the ideas from and closely follows the proofs of Theorem 2 in [8] and Theorem 7 in [2]. It was shown in Proposition 3 in [8] 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 [8].

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 $\Omega = \{(a_i, b_i)\}_{i=1}^m$ be a collection of indices sampled uniformly at random with replacement (possibly containing repetitions), and define the operator

$$\mathcal{R}_\Omega(\mathbf{X}) = \sum_{i=1}^m \langle \mathbf{e}_{a_i}\mathbf{e}_{b_i}^\top, \mathbf{X} \rangle \mathbf{e}_{a_i}\mathbf{e}_{b_i}^\top.$$

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 [8] that, with high probability,

$$\frac{nn'}{m}\left\|\mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_T - \frac{m}{nn'}\mathcal{P}_T\right\| \le \frac{1}{2}, \qquad \|\mathcal{R}_\Omega\| \le \frac{8}{3}\sqrt{c}\log n', \tag{25}$$

and that there exists a matrix $\mathbf{Y}$ in the range of $\mathcal{R}_\Omega$ satisfying

$$\|\mathcal{P}_T(\mathbf{Y}) - \mathbf{U}\mathbf{V}^\top\|_F \le \sqrt{\frac{r}{2n'}}, \qquad \|\mathcal{P}_{T^\perp}(\mathbf{Y})\|_F \le \frac{1}{2}, \tag{26}$$

see Section 4 in [8] 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

$$0 = \|\mathcal{R}_\Omega(\mathbf{R})\|_F \ge \|\mathcal{R}_\Omega\mathcal{P}_T(\mathbf{R})\|_F - \|\mathcal{R}_\Omega\mathcal{P}_{T^\perp}(\mathbf{R})\|_F,$$

where we used the reverse triangle inequality. Further, the first bound in (25) implies

$$\|\mathcal{R}_\Omega\mathcal{P}_T(\mathbf{R})\|_F^2 = \langle \mathbf{R}, \mathcal{P}_T\mathcal{R}_\Omega^2\mathcal{P}_T(\mathbf{R})\rangle \ge \langle \mathbf{R}, \mathcal{P}_T\mathcal{R}_\Omega\mathcal{P}_T(\mathbf{R})\rangle \ge \frac{m}{2nn'}\|\mathcal{P}_T(\mathbf{R})\|_F^2,$$

and, using the second bound in (25), we can bound $\|\mathcal{R}_\Omega\mathcal{P}_{T^\perp}(\mathbf{R})\|_F \le \frac{8}{3}\sqrt{c}\log(n')\|\mathcal{P}_{T^\perp}(\mathbf{R})\|_F$. Together, this implies

$$\|\mathcal{P}_{T^\perp}(\mathbf{R})\|_F \ge \sqrt{\frac{9m}{128cnn'\log^2 n'}}\|\mathcal{P}_T(\mathbf{R})\|_F \ge \sqrt{\frac{4.5r}{n'}}\|\mathcal{P}_T(\mathbf{R})\|_F. \tag{27}$$

Recalling the variational characterization of the nuclear norm $\|\mathbf{A}\|_* = \sup_{\mathbf{B}:\|\mathbf{B}\|\le 1}\langle \mathbf{A}, \mathbf{B}\rangle$, 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 (26). Then, we can bound

$$\begin{aligned}
\|\mathbf{X}^\star + \mathbf{R}\|_* &\ge \langle \mathbf{U}\mathbf{V}^\top + \mathbf{U}_\perp\mathbf{V}_\perp^\top, \mathbf{X}^\star + \mathbf{R}\rangle \\
&= \|\mathbf{X}^\star\|_* + \langle \mathbf{U}\mathbf{V}^\top + \mathbf{U}_\perp\mathbf{V}_\perp^\top, \mathbf{R}\rangle \\
&= \|\mathbf{X}^\star\|_* + \langle \mathbf{U}\mathbf{V}^\top + \mathbf{U}_\perp\mathbf{V}_\perp^\top - (\mathcal{P}_T(\mathbf{Y}) + \mathcal{P}_{T^\perp}(\mathbf{Y})), \mathcal{P}_T(\mathbf{R}) + \mathcal{P}_{T^\perp}(\mathbf{R})\rangle \\
&= \|\mathbf{X}^\star\|_* + \langle \mathbf{U}\mathbf{V}^\top - \mathcal{P}_T(\mathbf{Y}), \mathcal{P}_T(\mathbf{R})\rangle + \langle \mathbf{U}_\perp\mathbf{V}_\perp^\top - \mathcal{P}_{T^\perp}(\mathbf{Y}), \mathcal{P}_{T^\perp}(\mathbf{R})\rangle \\
&\ge \|\mathbf{X}^\star\|_* - \sqrt{\frac{r}{2n'}}\|\mathcal{P}_T(\mathbf{R})\|_F + \frac{1}{2}\|\mathcal{P}_{T^\perp}(\mathbf{R})\|_* \\
&\ge \|\mathbf{X}^\star\|_* + \frac{1}{6}\|\mathcal{P}_{T^\perp}(\mathbf{R})\|_*,
\end{aligned}$$

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 (26) and the definition of $\mathbf{U}_\perp, \mathbf{V}_\perp$, and the last line follows from the bound (27) and the fact that the Frobenius norm is bounded from above by the nuclear norm.

As in the proof of Theorem 3, we can bound the nuclear norm

$$\|\mathbf{X}_\infty\|_* \leq (1+\Delta_\beta)\bigg(\|\mathbf{X}^\star\|_* + \frac{n\beta}{\log\frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}\bigg),$$

where $\Delta_\beta = (\frac{\log(\|\mathbf{X}^\star\|_*/\beta)-1}{\log(1.05n)} - 1)^{-1}$. Hence, we can bound

$$\|\mathcal{P}_{T^\perp}(\mathbf{R})\|_F \leq \|\mathcal{P}_{T^\perp}(\mathbf{R})\|_* \leq 6\bigg(\Delta_\beta\|\mathbf{X}^\star\|_* + (1+\Delta_\beta)\frac{n\beta}{\log\frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}\bigg).$$

Using (27), we can also bound

$$\begin{aligned}\|\mathcal{P}_T(\mathbf{R})\|_F &\leq \sqrt{\frac{128cnn'\log^2 n'}{9m}}\|\mathcal{P}_{T^\perp}(\mathbf{R})\|_F \\ &\leq 6\bigg(\Delta_\beta\|\mathbf{X}^\star\|_* + (1+\Delta_\beta)\frac{n\beta}{\log\frac{\|\mathbf{X}^\star\|_*}{\beta} - 1}\bigg)\sqrt{\frac{128cnn'\log^2 n'}{9m}}.\end{aligned}$$

Putting everything together, we have

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

which completes the proof of the bound (13) in Theorem 4. The bound (14) can be shown following the same steps, and we omit the details to avoid repetition. ☐

B.5 Proof of Proposition 5

Proof. We begin by showing the first part of Proposition 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 A, the mirror descent update (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 $\mathbf{A}_i$'s, and hence also with the gradient

$$\nabla f(\mathbf{X}_0) = \frac{1}{m}\sum_{i=1}^m(\langle\mathbf{A}_i, \mathbf{X}_0\rangle - y_i)\mathbf{A}_i,$$

which is a linear combination of the $\mathbf{A}_i$'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 (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 (15) with initialization $\mathbf{U}_0 = \mathbf{X}_0$ and $\mathbf{V}_0 = \mathbf{0}$.

Proof of part 2.
We begin by studying the mirror descent update (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 A, the mirror descent update (2) becomes

$$\begin{aligned}\mathbf{X}_{t+1} &= \nabla\Phi_\beta^{-1}\big(\nabla\Phi_\beta(\mathbf{X}_t) - \eta\nabla f(\mathbf{X}_t)\big) \\ &= \frac{\beta}{2}\Bigg[\exp\Bigg(\log\Bigg(\frac{\mathbf{X}_t}{\beta} + \sqrt{\frac{\mathbf{X}_t^2}{\beta^2} + \mathbf{I}}\Bigg) - \eta\nabla f(\mathbf{X}_t)\Bigg) \\ &\qquad - \exp\Bigg(-\log\Bigg(\frac{\mathbf{X}_t}{\beta} + \sqrt{\frac{\mathbf{X}_t^2}{\beta^2} + \mathbf{I}}\Bigg) + \eta\nabla f(\mathbf{X}_t)\Bigg)\Bigg].\end{aligned}$$

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 $\mathbf{A}_i$'s, the mirror descent update can be written as

$$\begin{aligned} \mathbf{X}_{t+1} &= \frac{\beta}{2}\left[\left(\frac{\mathbf{X}^2}{\beta} + \sqrt{\frac{\mathbf{X}^2}{\beta^2} + \mathbf{I}}\right)\exp\bigl(-\eta\nabla f(\mathbf{X}_t)\bigr) - \left(-\frac{\mathbf{X}^2}{\beta} + \sqrt{\frac{\mathbf{X}^2}{\beta^2} + \mathbf{I}}\right)\exp\bigl(\eta\nabla f(\mathbf{X}_t)\bigr)\right] \\ &= \frac{1}{2}\Biggl[\exp\Bigl(-\eta\nabla f(\mathbf{X}_t)\Bigr)\frac{\mathbf{X}_t + \sqrt{\mathbf{X}_t^2 + \beta^2\mathbf{I}}}{2} + \frac{\mathbf{X}_t + \sqrt{\mathbf{X}_t^2 + \beta^2\mathbf{I}}}{2}\exp\Bigl(-\eta\nabla f(\mathbf{X}_t)\Bigr) \\ &\qquad + \exp\Bigl(\eta\nabla f(\mathbf{X}_t)\Bigr)\frac{-\mathbf{X}_t + \sqrt{\mathbf{X}_t^2 + \beta^2\mathbf{I}}}{2} + \frac{-\mathbf{X}_t + \sqrt{\mathbf{X}_t^2 + \beta^2\mathbf{I}}}{2}\exp\Bigl(\eta\nabla f(\mathbf{X}_t)\Bigr)\Biggr]. \end{aligned}$$

Since $\mathbf{X}_0 = \mathbf{0}$ is symmetric and commutes with all $\mathbf{A}_i$'s, this identity inductively shows that $\mathbf{X}_t$ is symmetric and commutes with all $\mathbf{A}_i$'s for all $t \geq 0$.

Next, consider the exponentiated gradient algorithm (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 $\mathbf{A}_i$'s, the update (15) implies that $\mathbf{U}_t$ and $\mathbf{V}_t$ commute with all $\mathbf{A}_i$'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 (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 (2) is equivalent to the exponentiated gradient algorithm (15) when the sensing matrices $\mathbf{A}_i$'s are symmetric and commute. $\square$

B.6 Further claims

In this section, we elaborate on and justify further claims made in the main paper.

First, we demonstrate that minimizing the quantity

$$\sum_{i=1}^{n} \sigma_i \log\frac{1}{\beta} + \sigma_i \log\Bigl(\sigma_i + \sqrt{\sigma_i^2 + \beta^2}\Bigr) - \sqrt{\sigma_i^2 + \beta^2} \tag{28}$$

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 [11], 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 (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 (28) as

$$\beta\sum_{i=1}^{n} z_i \log\Bigl(z_i + \sqrt{z_i^2 + 1}\Bigr) - \sqrt{z_i^2 + 1} = \beta\sum_{i=1}^{n} -1 + \frac{z_i^2}{2} + \mathcal{O}\bigl(z_i^2\bigr), \tag{29}$$

where we applied a Taylor expansion around $z_i = 0$. Hence, minimizing the quantity in (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 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 (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 (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 (15) and gradient descent with full-rank factorized parametrization coincide.

C Additional experiments for matrix completion

In this section, we present additional numerical simulations which consider matrix completion, i.e. the sensing matrices $\mathbf{A}_i$'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 6, with the difference that we choose step sizes $\mu = 2000$ and $\mu = 500$ for mirror descent and gradient descent, respectively, due to the lower spectral norm of the sensing matrices $\mathbf{A}_i$'s in matrix completion compared to matrix sensing with random Gaussian sensing matrices. As the experiments for Figure 1, the experiments for Figure 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.

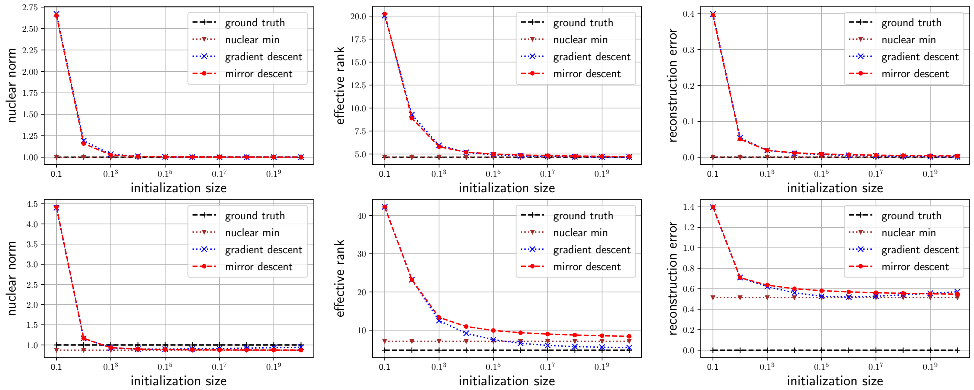

Figure 2: Nuclear norm, effective rank [10] and reconstruction error in matrix completion against initialization size α for $n = 50$ and $r = 5$. Top row: $m = 3nr$. Bottom row: $m = nr$.

We consider the nuclear norm $\|\mathbf{X}\|_*$, the effective rank defined in [10] 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 2 shows that the results in matrix completion qualitatively match the results in Figure 1 for matrix sensing with random Gaussian sensing matrices. In particular, with $m = 3nr$ observed entries (Figure 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 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}.

References

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

[2] E. J. Candès and Y. Plan. Matrix completion with noise. *Proceedings of the IEEE*, 98(6):925–936, 2010.

[3] M. Fazel, E. J. Candès, B. Recht, and P. Parrilo. Compressed sensing and robust recovery of low rank matrices. In *2008 42nd Asilomar Conference on Signals, Systems and Computers*, pages 1043–1047, 2008.

[4] 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.

[5] A. B. Juditsky and A. S. Nemirovski. Large deviations of vector-valued martingales in 2-smooth normed spaces. *arXiv preprint arXiv:0809.0813*, 2008.

[6] 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.

[7] A. S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(82):173–183, 1995.

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

[9] 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.

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

[11] 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.