Analysis of Singular Value Thresholding Algorithm for Matrix Completion*

Yunwen Lei[1] and Ding-Xuan Zhou[2]

[1]Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong
yunwen.lei@hotmail.com
[2]School of Data Science and Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong
mazhou@cityu.edu.hk

Abstract

This paper provides analysis for convergence of the singular value thresholding algorithm for solving matrix completion and affine rank minimization problems arising from compressive sensing, signal processing, machine learning, and related topics. A necessary and sufficient condition for the convergence of the algorithm with respect to the Bregman distance is given in terms of the step size sequence $\{\delta_k\}_{k\in\mathbb{N}}$ as $\sum_{k=1}^{\infty}\delta_k=\infty$. Concrete convergence rates in terms of Bregman distances and Frobenius norms of matrices are presented. Our novel analysis is carried out by giving an identity for the Bregman distance as the excess gradient descent objective function values and an error decomposition after viewing the algorithm as a mirror descent algorithm with a non-differentiable mirror map.

Keywords: matrix completion, singular value thresholding, mirror descent, Bregman distance
AMS Subject Classifications: 68Q32, 93E35

1 Introduction

Matrix completion and affine rank minimization are important research problems arising from numerous applications in various fields including compressive sensing, signal processing, machine learning, computer vision and control [6, 7, 18]. A simple and efficient first-order method for solving these problems is the singular value thresholding (SVT) algorithm introduced in [5].

Let $\mathcal{A}$ be a linear transformation mapping $n_1\times n_2$ matrices to $\mathbb{R}^m$ and $b\in\mathbb{R}^m$. SVT aims to find a low-rank solution to the linear system $\mathcal{A}(X)=b$ by iteratively producing a sequence of matrix pairs $\{(X^k,Y^k)\}_{k\in\mathbb{N}}$ as

$$\begin{cases} Y^{k+1}=Y^k+\delta_k\mathcal{A}^*(b-\mathcal{A}(X^k)),\\ X^{k+1}=\mathcal{D}_\tau(Y^{k+1}),\end{cases}\tag{1}$$

where $\mathcal{A}^*$ denotes the adjoint of $\mathcal{A}$, $X^1=Y^1$ is the zero matrix in $\mathbb{R}^{n_1\times n_2}$ and $\{\delta_k\}_{k\in\mathbb{N}}$ is a sequence of positive step sizes. Here $\mathcal{D}_\tau(Y^{k+1})$ is a soft-thresholding operator at level $\tau>0$ to be defined in (4) below, acting on the matrix Y^{k+1} to produce a low-rank approximation $X^{k+1}=\mathcal{D}_\tau(Y^{k+1})$. Due

*The work described in this paper is partially supported by the National Natural Science Foundation of China (Grant No. 61806091), and the NSFC/RGC Joint Research Scheme [RGC Project No. N_CityU120/14 and NSFC Project No. 11461161006].

to the ability of producing low-rank solutions with the soft-thresholding operator, SVT was shown to be extremely efficient at addressing problems with low-rank optimal solutions such as recommender systems [5]. It was shown in [5] that SVT is equivalent to the gradient descent algorithm applied to the dual problem of

$$\min_{X\in\mathbb{R}^{n_1\times n_2}} \left[\Psi(X) := \tau\|X\|_* + \frac{1}{2}\|X\|_F^2\right] \quad \text{subject to} \quad \mathcal{A}(X) = b, \tag{2}$$

where $\|X\|_* = \|\sigma(X)\|_1$ and $\|X\|_F = \|\sigma(X)\|_2$ are the nuclear norm and Frobenius norm of X, respectively. Here $\sigma(X)$ denotes the vector of all singular values of X in nonincreasing order and $\|x\|_p = [\sum_{i=1}^d |x_i|^p]^{\frac{1}{p}}$ denotes the ℓ_p-norm of $x = (x_i)_{i=1}^d \in \mathbb{R}^d$. Based on this interpretation, it was further shown that the sequence $\{X^k\}$ converges to the unique solution $X^\star$ of the optimization problem (2) with the error satisfying $\sum_{k=1}^\infty \|X^k - X^\star\|_F^2 < \infty$, provided that the linear system $\mathcal{A}(X) = b$ is consistent and that the step size sequence is bounded above and below from 0 satisfying $0 < \inf_k \delta_k \leq \sup_k \delta_k < \frac{2}{\|\mathcal{A}\|^2}$, where $\|\mathcal{A}\|$ is the operator norm of $\mathcal{A}$ defined by $\|\mathcal{A}\| = \sup\limits_{X\in\mathbb{R}^{n_1\times n_2}:\|X\|_F\leq 1} \|\mathcal{A}(X)\|_2$.

In this paper, we refine the existing convergence analysis of SVT in terms of both convergence conditions and convergence rates. We shall show that $\{X^k\}$ converges to the unique solution $X^\star$ of the optimization problem

$$\min_{X\in\mathbb{R}^{n_1\times n_2}} \Psi(X) \quad \text{subject to} \quad \mathcal{A}(X) = b_0, \tag{3}$$

with respect to the Bregman distance if and only if the step size sequence $\{\delta_k\}_{k\in\mathbb{N}}$ satisfies $\sum_{k=1}^\infty \delta_k = \infty$, under the mild assumption that the orthogonal projection b_0 of b onto the range of $\mathcal{A}$ is nonzero. This gives a precise characterization on the convergence of SVT, while only sufficient conditions for the convergence of SVT were considered in the literature. Then we shall establish a convergence rate $\|X^{T+1} - X^\star\|_F^2 = O(\frac{1}{\sum_{k=1}^T \delta_k})$, which gives the order $O(\frac{1}{T})$ in the general case $0 < \inf_k \delta_k \leq \sup_k \delta_k < \frac{2}{\|\mathcal{A}\|^2}$. This improves the previous convergence result $\sum_{k=1}^\infty \|X^k - X^*\|_F^2 < \infty$ under the same condition with no explicit convergence rates [5]. Our convergence rate discussion is based on a key identity on the Bregman distance between X^T and $X^\star$ and the excess objective function values of the dual problem of (3) in gradient descent at step T. Our discussion in getting the necessary condition $\sum_{k=1}^\infty \delta_k = \infty$ is based on a novel error decomposition for the excess Bregman distance after interpreting SVT as a specific mirror descent algorithm with a non-differentiable mirror map. Our basic idea with this error decomposition is to control the Bregman distance between X_k and X^* from below by making full use of the smoothness of the objective function. The new interpretation of SVT also opens the door of studying SVT in the mirror descent framework [2, 12]. Notice the above definition of b_0 also allows us to remove the assumption on the consistency of the linear system $\mathcal{A}(X) = b$ considered in the literature.

2 Main Results

Before stating our main results, we define the operator $\mathcal{D}_\tau$. Let $Y = U\Sigma V^*$ be a singular value decomposition of a matrix $Y \in \mathbb{R}^{n_1\times n_2}$ of rank r, where U and V are $n_1 \times r$ and $n_2 \times r$ matrices with orthonormal columns, respectively, and $\Sigma = \text{diag}(\{\sigma_1, \ldots, \sigma_r\})$ is the $r \times r$ diagonal matrix with the main diagonal entries $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ being the positive singular values of Y. The singular value shrinkage operator $\mathcal{D}_\tau$ at level τ is defined [5] by

$$\mathcal{D}_\tau(Y) = U\mathcal{D}_\tau(\Sigma)V^*, \tag{4}$$

where

$$D_\tau(\Sigma) = \text{diag}\big(\{(\sigma_1 - \tau)_+, \ldots, (\sigma_r - \tau)_+\}\big)$$

and $(t)_+ = \max(0, t)$.

Observe from the definition (3) of $X^\star$ that $X^\star = 0$ is equivalent to $b_0 = 0$. Since b_0 is the projection of b onto the range of $\mathcal{A}$, we know that $b - b_0$ is orthogonal to the range of $\mathcal{A}$ and thereby $\mathcal{A}^*(b - b_0) = 0$. So from the definition (1) of SVT, we see that in this special case, for any choice of the step size sequence, $X^k = 0$ and $Y^k = 0$ for all $k \in \mathbb{N}$, and the convergence holds obviously.

Our first main result provides a necessary and sufficient condition for the convergence of $\{X^k\}$ to $X^\star$ with respect to the Bregman distance when the trivial case $b_0 = 0$ is excluded. We denote $\langle X, Y\rangle = \sum_{i=1}^{n_1}\sum_{j=1}^{n_2} X_{ij}Y_{ij}$ the standard inner product between the matrices $X = (X_{ij}) \in \mathbb{R}^{n_1\times n_2}$ and $Y = (Y_{ij}) \in \mathbb{R}^{n_1\times n_2}$, and the subdifferential of a function $f : \mathbb{R}^{n_1\times n_2} \to \mathbb{R}$ at $X \in \mathbb{R}^{n_1\times n_2}$ as

$$\partial f(X) = \{Y \in \mathbb{R}^{n_1\times n_2} : f(\widetilde{X}) \geq f(X) + \langle \widetilde{X} - X, Y\rangle,\ \forall \widetilde{X} \in \mathbb{R}^{n_1\times n_2}\}.$$

If f is convex, the **Bregman distance** between X and $\widetilde{X}$ under f and $\widetilde{Y} \in \partial f(\widetilde{X})$ is defined as

$$D_f^{\widetilde{Y}}(X, \widetilde{X}) = f(X) - f(\widetilde{X}) - \langle X - \widetilde{X}, \widetilde{Y}\rangle.$$

If f is differentiable, then $\partial f(X)$ consists of $\nabla f(X)$, the gradient of f at X.

Now we can state our first main result as follows.

Theorem 1. *Let $\{(X^k, Y^k)\}_{k\in\mathbb{N}}$ be produced by* (1) *and $b_0 \neq 0$. Then the following statements hold.*

(a) If $\sup_k \delta_k < \frac{1}{2\|\mathcal{A}\|^2}$, then

$$\lim_{T\to\infty} D_\Psi^{Y^T}(X^\star, X^T) = 0 \textit{ if and only if } \sum_{k=1}^{\infty} \delta_k = \infty.$$

(b) If $\sup_k \delta_k < \frac{2}{\|\mathcal{A}\|^2}$, then

$$\left\|X^{T+1} - X^\star\right\|_F^2 \leq \widetilde{C}\Big[\sum_{k=1}^{T} \delta_k\Big]^{-1}, \quad \forall T \in \mathbb{N},$$

where $\widetilde{C}$ is a constant independent of T.

The necessity part of (a) of Theorem 1 will be proved by Proposition 5 in Section 3 while the sufficiency part of (a) and (b) follows from Proposition 9 in Section 4. We see from Theorem 1 that when $0 < \inf_k \delta_k \leq \sup_k \delta_k < \frac{2}{\|\mathcal{A}\|^2}$, there holds $\left\|X^{T+1} - X^\star\right\|_F^2 = O(1/T)$. Theorem 1 also applies to the linearized Bregman iteration for compressive sensing [4, 22].

Our second main result, to be proved in Section 3, is a monotonic property of the sequence $\{X^k\}$ in terms of the least squares error $F(X)$ used often in learning theory and defined for $X \in \mathbb{R}^{n_1\times n_2}$ by $F(X) = \frac{1}{2}\|\mathcal{A}(X) - b\|_2^2$.

Theorem 2. *Let $\{(X^k, Y^k)\}_{k\in\mathbb{N}}$ be produced by* (1) *with the step-size sequence $\{\delta_k\}_{k\in\mathbb{N}}$ satisfying $0 < \delta_k \leq \frac{1}{\|\mathcal{A}\|^2}$ for every $k \in \mathbb{N}$. Then the following statements hold.*

(a) $F(X^{k+1}) \leq F(X^k)$ for $k \in \mathbb{N}$.

(b) $X^\star$ is a minimizer of F over $\mathbb{R}^{n_1\times n_2}$.

(c) The following inequality holds for all $T \in \mathbb{N}$

$$F(X^{T+1}) - F(X^\star) = \frac{1}{2}\|\mathcal{A}(X^{T+1} - X^\star)\|_2^2 \leq \Psi(X^\star)\Big[\sum_{k=1}^{T} \delta_k\Big]^{-1}. \tag{5}$$

Some of our ideas in the above results can be used to analyze some other thresholding algorithms such as those derived from spectral algorithms [1, 8, 9]. It would be interesting to establish learning theory analysis [14, 15, 20, 21] for SVT algorithms in a noisy setting.

3 Necessity of Convergence

Our proof of the necessity part of (a) of Theorem 1 is based on interpreting SVT as a specific instantiation of mirror descent algorithms, a class of algorithms performing gradient descent in the dual space mapped from the primal space by the subgradient of the mirror map [2, 16]. This interpretation enables us to use arguments for mirror descent algorithms to analyze the convergence of SVT. However, standard analysis for mirror descent algorithms requires the mirror map to be differentiable, which is not the case for SVT having the **non-differentiable** mirror map Ψ. We use Bregman distances to overcome the difficulty. Our analysis can be extended to study SVT in the online setting [11, 13].

Our analysis needs some basic facts about convex functions. A function $f : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}$ is said to be σ-**strongly convex** with $\sigma > 0$ if $D_f^{\widetilde{Y}}(X, \widetilde{X}) \geq \frac{\sigma}{2}\|X - \widetilde{X}\|_F^2$ for any $X, \widetilde{X} \in \mathbb{R}^{n_1 \times n_2}$ and $\widetilde{Y} \in \partial f(\widetilde{X})$. It is said to be L-**strongly smooth** if it is differentiable and $D_f^{\nabla f(\widetilde{X})}(X, \widetilde{X}) \leq \frac{L}{2}\|X - \widetilde{X}\|_F^2$ for any $X, \widetilde{X}$. We denote $f^*(Y) = \sup_{X \in \mathbb{R}^{n_1 \times n_2}} \big[\langle X, Y\rangle - f(X)\big]$ the Fenchel (convex) conjugate of f.

Lemma 3. *For a convex function $f : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}$, the following statements hold.*

(a) $f^{**} = f$ and

$$\partial f^*(Y) = \{X : Y \in \partial f(X)\}, \quad \forall Y \in \mathbb{R}^{n_1 \times n_2}.$$

(b) For $\beta > 0$, the function f is β-strongly convex if and only if f^* is $\frac{1}{\beta}$-strongly smooth.

(c) If there exists a constant $L > 0$ such that

$$\|\nabla f(X) - \nabla f(\widetilde{X})\|_F \leq L\|X - \widetilde{X}\|_F \tag{6}$$

for all $X, \widetilde{X} \in \mathbb{R}^{n_1 \times n_2}$, then we have

$$f(X) \leq f(\widetilde{X}) + \langle X - \widetilde{X}, \nabla f(\widetilde{X})\rangle + \frac{L}{2}\|X - \widetilde{X}\|_F^2.$$

Part (a) of Lemma 3 on the duality between f and its Fenchel conjugate f^* can be found in [3]. Part (b) on the duality between strong convexity and strong smoothness can be found in [10]. Part (c) is a standard result in relating the Lipschitz continuity of ∇F to the strong smoothness of F (see, e.g., [17, 23]).

The idea of applying Bregman distances to SVT has been introduced in the literature. For example, it can be found in [5] that

$$D_\Psi^{\widetilde{Y}}(X, \widetilde{X}) \geq \frac{1}{2}\|X - \widetilde{X}\|_F^2 \tag{7}$$

for all $X, \widetilde{X} \in \mathbb{R}^{n_1 \times n_2}, \widetilde{Y} \in \partial\Psi(\widetilde{X})$.

We observe the relation $X^k = \nabla\Psi^*(Y^k)$ for SVT, which is a novelty of our analysis.

Lemma 4. *The sequence $\{(X^k, Y^k)\}_k$ produced by* (1) *satisfies $Y^k \in \partial\Psi(X^k)$ and $X^k = \nabla\Psi^*(Y^k)$, and Ψ^* is differentiable. Hence from $\nabla F(X) = \mathcal{A}^*(\mathcal{A}(X) - b)$, we have*

$$Y^{k+1} = Y^k - \delta_k \nabla F(X^k) = Y^k - \delta_k \mathcal{A}^*\big(\mathcal{A}(\nabla\Psi^*(Y^k)) - b\big). \tag{8}$$

Proof. The gradient of F reads directly as $\nabla F(X) = \mathcal{A}^*(\mathcal{A}(X) - b)$. It was shown in [5] that for each $\tau > 0$ and $Y \in \mathbb{R}^{n_1 \times n_2}$, the singular value shrinkage operator obeys $\mathcal{D}_\tau(Y) = \arg\min_X \frac{1}{2}\|X - Y\|_F^2 + \tau\|X\|_*$. It follows that the second equation in (1) for Y^k is equivalent to

$$X^k = \arg\min_{X \in \mathbb{R}^{n_1 \times n_2}} \frac{1}{2}\|X - Y^k\|_F^2 + \tau\|X\|_*.$$

Combining this with the optimality condition implies $0 \in X^k - Y^k + \tau\partial\|X^k\|_*$. That is, $Y^k \in \partial\Psi(X^k)$. By Part (a) of Lemma 3, this implies $X^k \in \partial\Psi^*(Y^k)$. But (7) shows that Ψ is 1-strongly convex, which implies that Ψ^* is differentiable according to Part (b) of Lemma 3. Therefore, $X^k = \nabla\Psi^*(Y^k)$. This proves the desired statement. □

Now we can carry out the novel analysis stated in the following proposition which proves the necessity part of Theorem 1.

Proposition 5. *Let $\{(X^k, Y^k)\}_{k\in\mathbb{N}}$ be produced by* (1). *If $b_0 \neq 0$ and for some $\kappa > 0$, the step-size sequence $\{\delta_k\}_{k\in\mathbb{N}}$ satisfies $0 < \delta_k \leq \frac{1}{(2+\kappa)\|\mathcal{A}\|^2}$ for every $k \in \mathbb{N}$, then $D_\Psi^{Y^T}(X^\star, X^T) > 0$ for $T \in \mathbb{N}$ and*

$$\sum_{k=1}^{T} \delta_k \geq \frac{\log \frac{\Psi(X^\star)}{D_\Psi^{Y^{T+1}}(X^\star, X^{T+1})}}{(2+\kappa)\|\mathcal{A}\|^2 \log \frac{2+\kappa}{\kappa}}. \tag{9}$$

In particular, $\lim_{T\to\infty} D_\Psi^{Y^T}(X^\star, X^T) = 0$ implies $\sum_{k=1}^\infty \delta_k = \infty$.

Proof. Let us first analyze how the Bregman distance is reduced in one step iteration of SVT.

Let $k \in \mathbb{N}$. By Lemma 4 and the definition of the Bregman distance, for $X \in \mathbb{R}^{n_1\times n_2}$, we have

$$\begin{aligned}\Delta_k(X) &:= D_\Psi^{Y^k}(X, X^k) - D_\Psi^{Y^{k+1}}(X, X^{k+1}) \\ &= \big[\Psi(X) - \Psi(X^k) - \langle X - X^k, Y^k\rangle\big] - \big[\Psi(X) - \Psi(X^{k+1}) - \langle X - X^{k+1}, Y^{k+1}\rangle\big] \\ &= \Psi(X^{k+1}) - \Psi(X^k) + \langle X - X^{k+1}, Y^{k+1}\rangle - \langle X - X^k, Y^k\rangle. \end{aligned} \tag{10}$$

Notice that

$$D_\Psi^{Y^{k+1}}(X^k, X^{k+1}) = \Psi(X^k) - \Psi(X^{k+1}) - \langle X^k - X^{k+1}, Y^{k+1}\rangle.$$

Hence, by Lemma 4

$$\begin{aligned}\Delta_k(X) &= \langle X - X^k, Y^{k+1} - Y^k\rangle - D_\Psi^{Y^{k+1}}(X^k, X^{k+1}) \\ &= -\delta_k\langle X - X^k, \nabla F(X^k)\rangle - D_\Psi^{Y^{k+1}}(X^k, X^{k+1}).\end{aligned}$$

Setting $X = X^\star$, we have

$$\Delta_k(X^\star) = -\delta_k\langle X^\star - X^k, \nabla F(X^k)\rangle - D_\Psi^{Y^{k+1}}(X^k, X^{k+1}). \tag{11}$$

To estimate the inner product in (11), we apply Part (c) of Lemma 3 to the function F whose gradient satisfies the Lipschitz condition as

$$\begin{aligned}\|\nabla F(X) - \nabla F(\widetilde{X})\|_F &= \|\mathcal{A}^*(\mathcal{A}(X) - b) - \mathcal{A}^*(\mathcal{A}(\widetilde{X}) - b)\|_F \\ &= \|\mathcal{A}^*(\mathcal{A}(X - \widetilde{X}))\|_F \leq \|\mathcal{A}\|^2\|X - \widetilde{X}\|_F.\end{aligned} \tag{12}$$

Setting $X = X^\star$, $\widetilde{X} = X^k$ yields

$$F(X^\star) - F(X^k) \leq \langle X^\star - X^k, \nabla F(X^k)\rangle + \frac{\|\mathcal{A}\|^2}{2}\|X^k - X^\star\|_F^2, \tag{13}$$

while the choice of $X = X^k$, $\widetilde{X} = X^\star$ gives

$$F(X^k) - F(X^\star) \leq \langle X^k - X^\star, \nabla F(X^\star)\rangle + \frac{\|\mathcal{A}\|^2}{2}\|X^k - X^\star\|_F^2.$$

Recall that $\mathcal{A}^*(b - b_0) = 0$. It follows that

$$\nabla F(X^\star) = \mathcal{A}^*(\mathcal{A}(X^\star) - b) = \mathcal{A}^*(b_0 - b) = 0, \tag{14}$$

and

$$F(X^k) - F(X^\star) \leq \frac{\|\mathcal{A}\|^2}{2}\|X^k - X^\star\|_F^2.$$

Combining this with (11) and (13) tells us that

$$\Delta_k(X^\star) \leq \delta_k\|\mathcal{A}\|^2\|X^k - X^\star\|_F^2 - D_\Psi^{Y^{k+1}}(X^k, X^{k+1}) \leq \delta_k\|\mathcal{A}\|^2\|X^k - X^\star\|_F^2.$$

But $\|X^k - X^\star\|_F^2 \leq 2D_\Psi^{Y^k}(X^\star, X^k)$ according to (7). Then we have

$$D_\Psi^{Y^{k+1}}(X^\star, X^{k+1}) \geq (1 - 2\delta_k \|\mathcal{A}\|^2) D_\Psi^{Y^k}(X^\star, X^k). \tag{15}$$

Now we need the restriction $0 < \delta_k \leq \frac{1}{(2+\kappa)\|\mathcal{A}\|^2}$ with $\kappa > 0$ on the step size sequence. Denote $a = \frac{2+\kappa}{2} \log \frac{2+\kappa}{\kappa}$ and apply the elementary inequality

$$1 - x \geq \exp(-ax), \quad \forall 0 < x \leq \frac{2}{2+\kappa}.$$

Then we see from (15) that

$$D_\Psi^{Y^{k+1}}(X^\star, X^{k+1}) \geq \exp\left\{-2a\delta_k \|\mathcal{A}\|^2\right\} D_\Psi^{Y^k}(X^\star, X^k).$$

Applying this inequality iteratively for $k = 1, \ldots, T$ yields

$$D_\Psi^{Y^{T+1}}(X^\star, X^{T+1}) \geq \prod_{k=1}^{T} \exp\left\{-2a\delta_k \|\mathcal{A}\|^2\right\} D_\Psi^{Y^1}(X^\star, X^1).$$

Since $Y^1 = X^1 = 0$, we have $D_\Psi^{Y^1}(X^\star, X^1) = \Psi(X^\star) > 0$ by our assumption of $b_0 \neq 0$. So $D_\Psi^{Y^{T+1}}(X^\star, X^{T+1}) > 0$ and

$$2a\|\mathcal{A}\|^2 \sum_{k=1}^{T} \delta_k \geq \log \Psi(X^\star) - \log D_\Psi^{Y^{T+1}}(X^\star, X^{T+1}).$$

This verifies the desired lower bound on $\sum_{k=1}^T \delta_k$. The proof is complete. □

We are in a position to prove our second main result.

Proof of Theorem 2. We follow (10), but decompose $\Delta_k(X)$ in a different way by means of $D_\Psi^{Y^k}(X^{k+1}, X^k) = \Psi(X^{k+1}) - \Psi(X^k) - \langle X^{k+1} - X^k, Y^k \rangle$ to get

$$\Delta_k(X) = \langle X - X^{k+1}, Y^{k+1} - Y^k \rangle + D_\Psi^{Y^k}(X^{k+1}, X^k).$$

By (8), $Y^{k+1} - Y^k = -\delta_k \nabla F(X^k)$. To be consistent with the gradient at X^k, we separate $X - X^{k+1}$ into $X - X^k + X^k - X^{k+1}$ and decompose $\Delta_k(X)$ as

$$\Delta_k(X) = -\delta_k \langle X - X^k, \nabla F(X^k) \rangle + \left\{\delta_k \langle X^{k+1} - X^k, \nabla F(X^k) \rangle + D_\Psi^{Y^k}(X^{k+1}, X^k)\right\}. \tag{16}$$

The inner product in the above last term can be estimated by applying Part (c) of Lemma 3 to the function F satisfying (12) as

$$\langle X^{k+1} - X^k, \nabla F(X^k) \rangle \geq F(X^{k+1}) - F(X^k) - \frac{\|\mathcal{A}\|^2}{2} \|X^k - X^{k+1}\|_F^2.$$

But

$$D_\Psi^{Y^k}(X^{k+1}, X^k) \geq \frac{1}{2} \|X^{k+1} - X^k\|_F^2$$

according to (7). Putting these lower bounds into the last term of (16) and applying the bound $\langle X - X^k, \nabla F(X^k) \rangle \leq F(X) - F(X^k)$ derived from the convexity of F, we find

$$\begin{aligned}\Delta_k(X) &\geq -\delta_k \left[F(X) - F(X^k)\right] \\ &\quad + \left\{\delta_k \left[F(X^{k+1}) - F(X^k)\right] - \frac{\delta_k \|\mathcal{A}\|^2}{2} \|X^k - X^{k+1}\|_F^2 + \frac{1}{2} \|X^{k+1} - X^k\|_F^2\right\} \\ &= \delta_k \left[F(X^{k+1}) - F(X)\right] + \frac{1 - \delta_k \|\mathcal{A}\|^2}{2} \|X^{k+1} - X^k\|_F^2.\end{aligned}$$

By the assumption on the step size, $\delta_k\|\mathcal{A}\|^2 \leq 1$. Therefore, the following inequality holds for all $X \in \mathbb{R}^{n_1\times n_2}$

$$\delta_k[F(X^{k+1}) - F(X)] \leq D_\Psi^{Y^k}(X, X^k) - D_\Psi^{Y^{k+1}}(X, X^{k+1}). \tag{17}$$

Then the property $F(X^{k+1}) \leq F(X^k)$ stated in Part (a) follows by setting $X = X^k$ in (17) because $D_\Psi^{Y^k}(X^k, X^k) = 0$ and $D_\Psi^{Y^{k+1}}(X^k, X^{k+1}) \geq 0$.

The statement in Part (b) follows immediately from (14). In fact, from the orthogonality of $b - b_0$ and the range of $\mathcal{A}$ and $\mathcal{A}(X^\star) = b_0$, we see the following well known relation in learning theory

$$\begin{aligned} F(X) &= \frac{1}{2}\|\mathcal{A}(X - X^\star) + \mathcal{A}(X^\star) - b\|_2^2 = \frac{1}{2}\|\mathcal{A}(X - X^\star)\|_2^2 + \frac{1}{2}\|\mathcal{A}(X^\star) - b\|_2^2 \\ &= \frac{1}{2}\|\mathcal{A}(X - X^\star)\|_2^2 + F(X^\star). \end{aligned} \tag{18}$$

To prove the statement in Part (c), we apply the monotonicity $F(X^{k+1}) \leq F(X^k)$ derived in Part (a) and find

$$F(X^{T+1}) - F(X) \leq \frac{\sum_{k=1}^T \delta_k[F(X^{k+1}) - F(X)]}{\sum_{\tilde{k}=1}^T \delta_{\tilde{k}}}.$$

Taking the summation of (17) from $k = 1$ to T gives

$$\begin{aligned} \sum_{k=1}^T \delta_k[F(X^{k+1}) - F(X)] &\leq \sum_{k=1}^T \left[D_\Psi^{Y^k}(X, X^k) - D_\Psi^{Y^{k+1}}(X, X^{k+1})\right] \\ &= D_\Psi^{Y^1}(X, X^1) - D_\Psi^{Y^{T+1}}(X, X^{T+1}). \end{aligned}$$

But $-D_\Psi^{Y^{T+1}}(X, X^{T+1}) \leq 0$ and $D_\Psi^{Y^1}(X, X^1) = \Psi(X)$ since $X^1 = Y^1 = 0$. Hence

$$F(X^{T+1}) - F(X) \leq \frac{\Psi(X)}{\sum_{k=1}^T \delta_k}, \quad \forall X \in \mathbb{R}^{n_1\times n_2}.$$

In particular, taking $X = X^\star$ and applying (18), we get (5). This completes the proof of Theorem 2. □

4 Sufficiency of Convergence

This section presents the proof for the sufficiency part of (a) and (b) of Theorem 1. Our analysis is based on the observation that SVT can be viewed as a gradient descent algorithm applied to the dual problem of (3), hence results on gradient descent algorithms can be applied. Here we apply the following standard estimates for the convergence of the gradient descent method applied to smooth optimization problems. The proof is given in the appendix for completeness.

Lemma 6. *Suppose $f : \mathbb{R}^m \to \mathbb{R}$ is convex and L-strongly smooth with $\lambda^\star$ being a minimizer. Let $\{\lambda^k\}_{k\in\mathbb{N}}$ be the following sequence produced by the gradient descent algorithm*

$$\lambda^1 = 0, \quad \lambda^{k+1} = \lambda^k - \delta_k \nabla f(\lambda^k), \qquad k \in \mathbb{N} \tag{19}$$

with a step size sequence $\{\delta_k > 0\}_{k\in\mathbb{N}}$. Then the following statements hold.

(a) If $\sup_k \delta_k < 2/L$, then there exists a constant $\widetilde{C}$

$$f(\lambda^{T+1}) - f(\lambda^\star) \leq \widetilde{C}\Big[\sum_{k=1}^T \delta_k\Big]^{-1}. \tag{20}$$

(b) If $\sup_k \delta_k \leq 1/L$, then (20) holds with $\widetilde{C} = \|\lambda^\star\|_2^2/2$.

The following lemma shows how SVT can be viewed as a gradient descent algorithm applied to the dual of (3). Part (a) establishes the dual problem of the optimization problem (3), and Part (b) shows that the sequence $\{Y^k\}$ coincides with $\{\mathcal{A}^*(\lambda^k)\}_{k\in\mathbb{N}}$ with $\{\lambda_k\}$ produced by applying the gradient descent algorithm (19) to the function G given in Part (a). This lemma was presented in [5] when $\mathcal{A}$ is an orthogonal projector and the system $\mathcal{A}(X) = b$ is consistent. It is extended here to the general linear transformation $\mathcal{A}$ allowing for inconsistent systems with b replaced by its orthogonal projection onto the range of $\mathcal{A}$.

Lemma 7. *(a) The Lagrangian dual problem of* (3) *is*

$$\min_{\lambda\in\mathbb{R}^m} G(\lambda), \quad \textit{where } G(\lambda) := \Psi^*(\mathcal{A}^*(\lambda)) - \langle \lambda, b_0\rangle. \tag{21}$$

(b) If $\{(X^k, Y^k)\}_{k\in\mathbb{N}}$ is produced by (1)*, and $\{\lambda^k\}_{k\in\mathbb{N}}$ is produced by applying the gradient descent algorithm* (19) *to the function G, then we have $Y^k = \mathcal{A}^*(\lambda^k)$ for $k \in \mathbb{N}$.*

Proof. The Lagrangian dual problem of (3) is

$$\begin{aligned}
&\max_{\lambda\in\mathbb{R}^m} \min_{X\in\mathbb{R}^{n_1\times n_2}} \big[\Psi(X) - \langle\lambda, \mathcal{A}(X)\rangle + \langle\lambda, b_0\rangle\big] \\
&= \max_{\lambda\in\mathbb{R}^m} \Big[- \max_{X\in\mathbb{R}^{n_1\times n_2}} \big[\langle X, \mathcal{A}^*(\lambda)\rangle - \Psi(X)\big] + \langle\lambda, b_0\rangle\Big] \\
&= \max_{\lambda\in\mathbb{R}^m} \Big[- \Psi^*(\mathcal{A}^*(\lambda)) + \langle\lambda, b_0\rangle\Big] = - \min_{\lambda\in\mathbb{R}^m} \big[\Psi^*(\mathcal{A}^*(\lambda)) - \langle\lambda, b_0\rangle\big] \\
&= - \min_{\lambda\in\mathbb{R}^m} G(\lambda),
\end{aligned}$$

where in the second identity we have used the definition of Fenchel conjugate. This proves (21).

When the gradient descent algorithm (19) is applied to the function G defined in (21), we see by the chain rule that the gradient equals

$$\nabla G(\lambda) = \nabla\big(\Psi^*(\mathcal{A}^*(\lambda)) - \langle\lambda, b_0\rangle\big) = \mathcal{A}\big((\nabla\Psi^*)(\mathcal{A}^*(\lambda))\big) - b_0. \tag{22}$$

So the sequence $\{\lambda^k\}_{k\in\mathbb{N}}$ produced by (19) translates to

$$\lambda^{k+1} = \lambda^k - \delta_k[\mathcal{A}((\nabla\Psi^*)(\mathcal{A}^*(\lambda^k))) - b_0]. \tag{23}$$

Applying the transformation $\mathcal{A}^*$ to both sides and noticing $\mathcal{A}^* b_0 = \mathcal{A}^* b$ yield the following identity for all $k \in \mathbb{N}$

$$\begin{aligned}
\mathcal{A}^*(\lambda^{k+1}) &= \mathcal{A}^*(\lambda^k) - \delta_k \mathcal{A}^*\big(\mathcal{A}((\nabla\Psi^*)(\mathcal{A}^*(\lambda^k))) - b_0\big) \\
&= \mathcal{A}^*(\lambda^k) - \delta_k \mathcal{A}^*\big(\mathcal{A}((\nabla\Psi^*)(\mathcal{A}^*(\lambda^k))) - b\big).
\end{aligned}$$

This iteration relation for the sequence $\{\mathcal{A}^*(\lambda^k)\}_{k\in\mathbb{N}}$ is exactly the same as (8) in Lemma 4 for the sequence $\{Y^k\}_{k\in\mathbb{N}}$. This together with the initial conditions $Y^1 = 0, \mathcal{A}^*(\lambda^1) = 0$ tells us that $Y^k = \mathcal{A}^*(\lambda^k)$ for $k \in \mathbb{N}$. The proof of the lemma is complete. □

Combining Lemma 6 and Lemma 7 enables us to bound the excess dual objective value $G(\lambda^{T+1}) - G(\lambda^\star)$ in terms of $\sum_{k=1}^T \delta_k$. What is left for estimating $D_\Psi^{Y^{T+1}}(X^\star, X^{T+1})$ to prove the sufficiency part of Theorem 1 is to find a relation between the excess dual objective value $G(\lambda^{T+1}) - G(\lambda^\star)$ and the Bregman distance $D_\Psi^{Y^{T+1}}(X^\star, X^{T+1})$. This is given in the following key identity which provides an elegant scheme to transfer decay rates of excess dual objective values to those for the Bregman distance of primal variables.

Lemma 8. *If $\{(X^k, Y^k)\}_{k\in\mathbb{N}}$ is produced by* (1)*, and $\{\lambda^k\}_{k\in\mathbb{N}}$ is produced by applying the gradient descent algorithm* (19) *to the function G, then there exists some $\lambda^\star \in \mathbb{R}^m$ such that $\mathcal{A}^*(\lambda^\star) \in \partial\Psi(X^\star)$ and*

$$D_\Psi^{Y^k}(X^\star, X^k) = G(\lambda^k) - G(\lambda^\star).$$

Proof. Since $X^\star$ is an optimal point of the problem (3) with only linear constraints, the existence of Lagrange multipliers (e.g., Corollary 28.2.2 in [19]) and the first-order optimality condition imply the existence of $\lambda^\star \in \mathbb{R}^m$ satisfying

$$Y^\star := \mathcal{A}^*(\lambda^\star) \in \partial\Psi(X^\star). \tag{24}$$

Together with Part (a) of Lemma 3 and Lemma 4, this implies that

$$X^\star = \nabla\Psi^*(\mathcal{A}^*(\lambda^\star)) = \nabla\Psi^*(Y^\star). \tag{25}$$

Since Ψ is convex, we know (see, e.g., Proposition 3.3.4 in [3]) that for any $X \in \mathbb{R}^{n_1\times n_2}$,

$$Y \in \partial\Psi(X) \Longrightarrow \Psi^*(Y) = \langle X, Y\rangle - \Psi(X).$$

Applying this implication to the pairs $(X^\star, Y^\star)$ in (24) and (X^k, Y^k) in Lemma 4 satisfying $Y^k \in \partial\Psi(X^k)$, we know that

$$\begin{aligned}
D_\Psi^{Y^k}(X^\star, X^k) &= \Psi(X^\star) - \Psi(X^k) - \langle X^\star - X^k, Y^k\rangle \\
&= \Psi(X^\star) - \langle X^\star, Y^\star\rangle + \langle X^\star, Y^\star\rangle - \Psi(X^k) - \langle X^\star - X^k, Y^k\rangle \\
&= -\Psi^*(Y^\star) + \langle X^\star, Y^\star - Y^k\rangle + \Psi^*(Y^k) \\
&= \Psi^*(Y^k) - \Psi^*(Y^\star) - \langle Y^k - Y^\star, \nabla\Psi^*(Y^\star)\rangle,
\end{aligned}$$

where we have used (25) in the last equality. But $Y^k = \mathcal{A}^*(\lambda^k)$ according to Part (b) of Lemma 7. Then we see from the definition of the function G that $D_\Psi^{Y^k}(X^\star, X^k)$ equals

$$\begin{aligned}
&\Psi^*(\mathcal{A}^*(\lambda^k)) - \Psi^*(\mathcal{A}^*(\lambda^\star)) - \langle \mathcal{A}^*(\lambda^k - \lambda^\star), \nabla\Psi^*(Y^\star)\rangle \\
&= \Psi^*(\mathcal{A}^*(\lambda^k)) - \Psi^*(\mathcal{A}^*(\lambda^\star)) - \langle \lambda^k - \lambda^*, b_0\rangle + \langle \lambda^k - \lambda^\star, b_0\rangle - \langle \mathcal{A}^*(\lambda^k - \lambda^\star), \nabla\Psi^*(Y^\star)\rangle \\
&= G(\lambda^k) - G(\lambda^\star) + \langle \lambda^k - \lambda^\star, b_0 - \mathcal{A}(\nabla\Psi^*(Y^\star))\rangle.
\end{aligned}$$

This together with the identities (25) and $\mathcal{A}(X^\star) = b_0$ implies

$$\begin{aligned}
D_\Psi^{Y^k}(X^\star, X^k) &= G(\lambda^k) - G(\lambda^\star) + \langle \lambda^k - \lambda^\star, b_0 - \mathcal{A}(X^\star)\rangle \\
&= G(\lambda^k) - G(\lambda^\star).
\end{aligned}$$

The proof of the lemma is complete. □

Now we can prove the sufficiency part of (a) and (b) of Theorem 1 by presenting the following more general estimate.

Proposition 9. *Let $\{(X^k, Y^k)\}_{k\in\mathbb{N}}$ be produced by* (1) *with a positive step-size sequence $\{\delta_k\}$ satisfying* $\sup_k \delta_k < \frac{2}{\|\mathcal{A}\|^2}$. *Then we have*

$$D_\Psi^{Y^{T+1}}(X^\star, X^{T+1}) \le \widetilde{C}\Big[\sum_{k=1}^{T} \delta_k\Big]^{-1}, \tag{26}$$

where $\widetilde{C}$ is a constant independent of T. Furthermore, if $\sup_k \delta_k \le \frac{1}{\|\mathcal{A}\|^2}$, *then* (26) *holds with $\widetilde{C} = \|\lambda^\star\|_2^2/2$, where $\lambda^\star$ is an element in $\mathbb{R}^m$ satisfying $\mathcal{A}^*(\lambda^\star) \in \partial\Psi(X^\star)$.*

Proof. Recall the expression (22) for the gradient of G. Take the vector $\lambda^\star$ given in Lemma 8. The identity (25) implies

$$\nabla G(\lambda^\star) = \mathcal{A}(\nabla\Psi^*(\mathcal{A}^*(\lambda^\star))) - b_0 = \mathcal{A}(X^\star) - b_0 = 0$$

and therefore $\lambda^\star$ minimizes G.

By (7), Ψ is 1-strongly convex. So its Fenchel conjugate Ψ^* is 1-strongly smooth according to Part (b) of Lemma 3. It follows that for $\lambda, \tilde{\lambda} \in \mathbb{R}^m$,

$$\begin{aligned} G(\lambda) - G(\tilde{\lambda}) &= \Psi^*(\mathcal{A}^*(\lambda)) - \langle \lambda, b_0 \rangle - \Psi^*(\mathcal{A}^*(\tilde{\lambda})) + \langle \tilde{\lambda}, b_0 \rangle \\ &\le \langle \nabla \Psi^*(\mathcal{A}^*(\tilde{\lambda})), \mathcal{A}^*(\lambda - \tilde{\lambda}) \rangle + \frac{1}{2} \|\mathcal{A}^*(\lambda - \tilde{\lambda})\|_F^2 - \langle \lambda - \tilde{\lambda}, b_0 \rangle \\ &= \langle \lambda - \tilde{\lambda}, \mathcal{A}((\nabla \Psi^*)(\mathcal{A}^*(\tilde{\lambda}))) - b_0 \rangle + \frac{1}{2} \|\mathcal{A}^*(\lambda - \tilde{\lambda})\|_F^2 \\ &\le \langle \lambda - \tilde{\lambda}, \nabla G(\tilde{\lambda}) \rangle + \frac{\|\mathcal{A}\|^2}{2} \|\lambda - \tilde{\lambda}\|_2^2, \end{aligned}$$

where in the last step we have used (22) and the definition of operator norm. It tells us that the function $G(\lambda)$ is $\|\mathcal{A}\|^2$-strongly smooth. So we apply Lemma 8 and Lemma 6 (a) and know that when $\{\delta_k\}_k$ satisfies $\sup_k \delta_k \le \frac{2}{\|\mathcal{A}\|^2}$, the following inequality holds with a constant $\widetilde{C}$ independent of T

$$D_{\Psi}^{Y^{T+1}}(X^\star, X^{T+1}) = G(\lambda^{T+1}) - G(\lambda^\star) \le \widetilde{C} \Big[\sum_{k=1}^{T} \delta_k \Big]^{-1}.$$

According to Lemma 8 and Lemma 6 (b), the constant $\widetilde{C}$ can be chosen to be $\|\lambda^\star\|_2^2/2$ if $\delta_k \le \frac{1}{\|\mathcal{A}\|^2}$. The proof is complete. □

Appendix. Proof of Lemma 6

We first prove part (a). Since $\sup_k \delta_k < 2/L$, there exists a $\gamma \in (0, 2)$ such that $\delta_k \le (2 - \gamma)/L$ for all $k \in \mathbb{N}$. According to the iteration (19), we know

$$\|\lambda^{k+1} - \lambda^\star\|_2^2 = \|\lambda^k - \lambda^\star\|_2^2 + \delta_k^2 \|\nabla f(\lambda^k)\|_2^2 - 2\delta_k \langle \lambda^k - \lambda^\star, \nabla f(\lambda^k) \rangle. \tag{27}$$

Since f is convex and L-strongly smooth, the co-coercivity of ∇f implies (see, e.g., Theorem 2.1.5 in [17]

$$\|\nabla f(\lambda^k) - \nabla f(\lambda^\star)\|_2^2 \le L \langle \lambda^k - \lambda^\star, \nabla f(\lambda^k) - \nabla f(\lambda^\star) \rangle.$$

Plugging this inequality back into (27) and using $\nabla f(\lambda^\star) = 0$, we derive

$$\begin{aligned} \|\lambda^{k+1} - \lambda^\star\|_2^2 &\le \|\lambda^k - \lambda^\star\|_2^2 + (L\delta_k - 2)\delta_k \langle \lambda^k - \lambda^\star, \nabla f(\lambda^k) \rangle \\ &\le \|\lambda^k - \lambda^\star\|_2^2 + (L\delta_k - 2)\delta_k \big(f(\lambda^k) - f(\lambda^\star) \big) \\ &\le \|\lambda^k - \lambda^\star\|_2^2 - \gamma \delta_k \big(f(\lambda^k) - f(\lambda^\star) \big), \end{aligned}$$

where we have used the Jensen inequality and $\delta_k < 2/L$ in the second inequality and $\delta_k \le (2 - \gamma)/L$ in the last inequality. It then follows that

$$\begin{aligned} \min_{1 \le k \le T} f(\lambda^k) - f(\lambda^\star) &\le \frac{\gamma \sum_{k=1}^T \delta_k \big(f(\lambda^k) - f(\lambda^\star) \big)}{\gamma \sum_{k=1}^T \delta_k} \\ &\le \frac{\sum_{k=1}^T \big[\|\lambda^k - \lambda^\star\|_2^2 - \|\lambda^{k+1} - \lambda^\star\|_2^2 \big]}{\gamma \sum_{k=1}^T \delta_k} \le \frac{\|\lambda^\star\|_2^2}{\gamma \sum_{k=1}^T \delta_k}. \end{aligned} \tag{28}$$

Furthermore, it follows from Lemma 3 (c) and the iteration (19) that

$$\begin{aligned} f(\lambda^{k+1}) &\le f(\lambda^k) + \langle \lambda^{k+1} - \lambda^k, \nabla f(\lambda^k) \rangle + \frac{L \|\lambda^{k+1} - \lambda^k\|_2^2}{2} \\ &= f(\lambda^k) - \delta_k \big(1 - 2^{-1} L \delta_k \big) \|\nabla f(\lambda^k)\|_2^2. \end{aligned} \tag{29}$$

The assumption $\delta_k < 2/L$ then implies $f(\lambda^{k+1}) \leq f(\lambda^k)$ for all $k \in \mathbb{N}$. This monotonicity together with (28) then shows (20) with $\widetilde{C} = \|\lambda^\star\|_2^2/\gamma$.

We now prove part (b) under the assumption $\sup_k \delta_k \leq 1/L$. An application of the Jensen inequality in (29) then implies

$$\begin{aligned}
f(\lambda^{k+1}) &\leq f(\lambda^\star) + \langle \lambda^k - \lambda^\star, \nabla f(\lambda^k)\rangle - \delta_k\big(1 - 2^{-1}L\delta_k\big)\|\nabla f(\lambda^k)\|_2^2 \\
&\leq f(\lambda^\star) + \langle \lambda^k - \lambda^\star, \nabla f(\lambda^k)\rangle - 2^{-1}\delta_k\|\nabla f(\lambda^k)\|_2^2 \\
&= f(\lambda^\star) + \frac{\|\lambda^k - \lambda^\star\|_2^2 - \|\lambda^{k+1} - \lambda^\star\|_2^2}{2\delta_k},
\end{aligned}$$

where we have used $\delta_k \leq 1/L$ in the second inequality and (27) in the last identity. It then follows that

$$\min_{1\leq k\leq T} f(\lambda^{k+1}) - f(\lambda^\star) \leq \frac{2\sum_{k=1}^T \delta_k\big[f(\lambda^{k+1}) - f(\lambda^\star)\big]}{2\sum_{k=1}^T \delta_k} \leq \frac{\|\lambda^\star\|_2^2}{2\sum_{k=1}^T \delta_k},$$

which together with the monotonicity of $f(\lambda^k)$ implies the stated inequality (20) with $\widetilde{C} = \|\lambda^\star\|_2^2/2$. The proof of Lemma 6 is complete.

References

[1] F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *Journal of Complexity*, 23(1):52–72, 2007.

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

[3] J. M. Borwein and A. S. Lewis. *Convex analysis and nonlinear optimization: theory and examples.* Springer Science & Business Media, 2010.

[4] J.-F. Cai, S. Osher, and Z. Shen. Convergence of the linearized bregman iteration for ℓ_1-norm minimization. *Mathematics of Computation*, 78(268):2127–2136, 2009.

[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] E. J. Candès and T. Tao. The power of convex relaxation: near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56(5):2053–2080, 2010.

[8] L. L. Gerfo, L. Rosasco, F. Odone, E. D. Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Computation*, 20(7):1873–1897, 2008.

[9] Z.-C. Guo, D.-H. Xiang, X. Guo, and D.-X. Zhou. Thresholded spectral algorithms for sparse approximations. *Analysis and Applications*, 15(03):433–455, 2017.

[10] S. M. Kakade, S. Shalev-Shwartz, and A. Tewari. Regularization techniques for learning with matrices. *Journal of Machine Learning Research*, 13:1865–1890, 2012.

[11] Y. Lei and D.-X. Zhou. Analysis of online composite mirror descent algorithm. *Neural Computation*, 29(3):825–860, 2017.

[12] Y. Lei and D.-X. Zhou. Convergence of online mirror descent. *Applied and Computational Harmonic Analysis*, 2018. doi: https://doi.org/10.1016/j.acha.2018.05.005.

[13] Y. Lei and D.-X. Zhou. Learning theory of randomized sparse Kaczmarz method. *SIAM Journal on Imaging Sciences*, 11(1):547–574, 2018.

[14] J. Lin and D.-X. Zhou. Learning theory of randomized Kaczmarz algorithm. *Journal of Machine Learning Research*, 16:3341–3365, 2015.

[15] H. Q. Minh. Infinite-dimensional log-determinant divergences between positive definite trace class operators. *Linear Algebra and its Applications*, 528:331–383, 2017.

[16] A.-S. Nemirovsky and D.-B. Yudin. *Problem complexity and method efficiency in optimization.* John Wiley & Sons, 1983.

[17] Y. Nesterov. *Introductory lectures on convex optimization: a basic course.* Springer Science & Business Media, 2013.

[18] B. Recht, M. Fazel, and P. A. Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM Review*, 52(3):471–501, 2010.

[19] R. T. Rockafellar. *Convex analysis.* Princeton University Press, 1960.

[20] I. Steinwart and A. Christmann. *Support vector machines.* Springer Science & Business Media, 2008.

[21] T. Strohmer and R. Vershynin. A randomized Kaczmarz algorithm with exponential convergence. *Journal of Fourier Analysis and Applications*, 15(2):262, 2009.

[22] W. Yin, S. Osher, D. Goldfarb, and J. Darbon. Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing. *SIAM Journal on Imaging Sciences*, 1(1):143–168, 2008.

[23] Y. Ying and D.-X. Zhou. Unregularized online learning algorithms with general loss functions. *Applied and Computational Harmonic Analysis*, 42(2):224—-244, 2017.