SIAM J. IMAGING SCIENCES
Vol. 11, No. 1, pp. 547–574

Learning Theory of Randomized Sparse Kaczmarz Method*

Yunwen Lei† and Ding-Xuan Zhou†

Abstract. In this paper we propose an online learning algorithm, a general randomized sparse Kaczmarz method, for generating sparse approximate solutions to linear systems and present learning theory analysis for its convergence. Under a mild assumption covering the case of noisy random measurements in the sampling process or nonlinear regression function, we show that the algorithm converges in expectation if and only if the step size sequence $\{\eta_t\}_{t\in\mathbb{N}}$ satisfies $\lim_{t\to\infty}\eta_t = 0$ and $\sum_{t=1}^{\infty}\eta_t = \infty$. Convergence rates are also obtained and linear convergence is shown to be impossible under the assumption of positive variance of the sampling process. A sufficient condition for almost sure convergence is derived with an additional restriction $\sum_{t=1}^{\infty}\eta_t^2 < \infty$. Our novel analysis is performed by interpreting the randomized sparse Kaczmarz method as a special online mirror descent algorithm with a *nondifferentiable* mirror map and using the Bregman distance. The sufficient and necessary conditions are derived by establishing a restricted variant of strong convexity for the involved generalization error and using the special structures of the soft-thresholding operator.

Key words. linearized Bregman iteration, randomized sparse Kaczmarz algorithm, online learning, learning theory, Bregman distance

AMS subject classifications. 68Q32, 93E35

DOI. 10.1137/17M1136225

1. Introduction.

In this big data era, scalable algorithms are desired for various learning tasks arising from practical applications. Online learning and stochastic gradient descent are useful tools to fulfill this scalability need. They can be used for handling big data or data arriving naturally in sequential ways and play an important role in deep learning. In this paper we study an online learning algorithm for generating sparse approximate solutions to linear systems, motivated by linearized Bregman iteration and sparse Kaczmarz algorithms.

Linearized Bregman iteration is a simple and fast algorithm to solve basis pursuit problems and has found wide applications in image processing, compressive sensing, and machine learning [12, 28, 29, 31]. To solve a linear system $Aw = y$ with $A \in \mathbb{R}^{n\times d}$ and $y \in \mathbb{R}^n$, the linearized Bregman iteration with a threshold parameter $\lambda \geq 0$ produces a sequence $\{(w_t, v_t)\}_{t\in\mathbb{N}}$ of vector pairs in $\mathbb{R}^d$ with initial vectors $w_1 = v_1 = 0 \in \mathbb{R}^d$ as

$$\begin{cases} v_{t+1} = v_t - \eta_t A^\top (Aw_t - y), \\ w_{t+1} = S_\lambda(v_{t+1}), \end{cases} \tag{1.1}$$

*Received by the editors June 26, 2017; accepted for publication (in revised form) January 2, 2018; published electronically February 20, 2018.

http://www.siam.org/journals/siims/11-1/M113622.html

Funding: The work of the authors was partially supported by the Research Grants Council of Hong Kong (Project CityU 11304114) and by National Natural Science Foundation of China under grants 11461161006 and 11471292. The work of the second author was partially sponsored by Shanghai Jiaotong University and the Ministry of Education.

†Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong (yunwelei@cityu.edu.hk, mazhou@cityu.edu.hk).

where $\{\eta_t\}_{t\in\mathbb{N}}$ is a positive step size sequence, $A^\top$ is the transpose of A and $S_\lambda : \mathbb{R}^d \to \mathbb{R}^d$ is the soft-thresholding operator defined componentwisely in terms of the soft-thresholding function $S_\lambda : \mathbb{R} \to \mathbb{R}$ given by $S_\lambda(v) := \operatorname{sgn}(v)\max(|v| - \lambda, 0)$. Here $\operatorname{sgn}(a)$ is the sign of $a \in \mathbb{R}$. If the linear system $Aw = y$ is consistent, $\lambda > 0$, and the step size sequence is constant bounded by the reciprocal of the largest eigenvalue of $AA^\top$, it was shown in [6, 7, 28] that the sequence $\{w_t\}_t$ produced by (1.1) converges to the unique solution of the optimization problem

$$\min_{w\in\mathbb{R}^d} \Psi(w) := \lambda\|w\|_1 + 2^{-1}\|w\|_2^2 \quad \text{subject to} \quad Aw = y, \tag{1.2}$$

where $\|w\|_p = [\sum_{i=1}^d |w(i)|^p]^{\frac{1}{p}}$ is the ℓ_p-norm of $w = (w(i))_{i=1}^d = [w(1), \dots, w(d)]^\top \in \mathbb{R}^d$ for $p \geq 1$.

Along a different direction, motivated by sparsity and scalable algorithms for handling big data, the classical Kaczmarz algorithm [16] for solving linear systems has recently attracted much attention in nonuniform sampling [26, 32] and in learning theory [19]. The most recent one is the sparse Kaczmarz algorithm proposed in [20, 21] by processing an example at each iteration as

$$\begin{cases} v_{t+1} = v_t - \delta_t a_{r(t)}, \\ w_{t+1} = S_\lambda(v_{t+1}). \end{cases} \tag{1.3}$$

Here $a_{r(t)}$ is the transpose of the $r(t)$th row of the matrix A, $\{r(t)\}_t$ is a sequence called the admissible control sequence satisfying some conditions such as the periodic condition $r(t) = t \bmod (n+1)$, and δ_t is a residual error term which is chosen in [20] to be

$$\delta_t = \arg\min_{\delta\in\mathbb{R}} \frac{1}{2}\left\|S_\lambda(v_t - \delta a_{r(t)})\right\|_2^2 + \delta y_{r(t)}$$

and in [21] to be similar to that in the randomized Kaczmarz algorithm [26] as

$$\delta_t = \frac{\langle w_t, a_{r(t)}\rangle - y_{r(t)}}{\|a_{r(t)}\|_2^2}, \tag{1.4}$$

where $\langle\cdot,\cdot\rangle$ is the dot product or Euclidean inner product in $\mathbb{R}^d$. Under the consistency assumption of the linear system $Aw = y$ and the above choices of δ_t, it was shown in [20, 21] that $\{w_t\}_{t\in\mathbb{N}}$ produced by (1.3) with $\lambda > 0$ converges to the solution of (1.2). A randomized sparse Kaczmarz method (RSK) was considered in [23] by taking iterations (1.3) with randomly chosen $r(t)$ from a distribution on $\{1, \dots, n\}$ and taking the residual error form (1.4), and linear convergence was established under the consistency assumption again.

In this paper we propose an online learning algorithm, a general RSK, based on motivations from the above two learning algorithms and stochastic gradient descent algorithms used recently in deep learning as scalable methods. This algorithm is able to perform learning tasks using sequentially arriving data or big data since each iteration involves only a single example. The relaxation of step sizes allows the algorithm to handle noisy data, to which the convergence analysis of the special RSK given in [23] does not apply. Let $\mathcal{X}$ (the input set) be a nonempty measurable subset of $\mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$ be the output set.

Definition 1. *Let $\{z_t := (x_t, y_t)\}_t \subset \mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ be a sequence of input-output pairs. The RSK produces a sequence of vector pairs $\{(w_t, v_t)\}_{t\in\mathbb{N}}$ defined iteratively with the initial pair $w_1 = v_1 = 0$ by*

$$\begin{cases} v_{t+1} = v_t - \eta_t(\langle w_t, x_t\rangle - y_t)x_t, \\ w_{t+1} = S_\lambda(v_{t+1}), \end{cases} \tag{1.5}$$

where $\{\eta_t\}$ is a sequence of positive relaxation parameters or step sizes.

This algorithm is an online version of the linearized Bregman iteration (1.1) modified with a step size sequence $\{\eta_t\}_t$. It is more general than the sparse Kaczmarz algorithm (1.3) considered in [21] which is a special case in the sense that the input set $\mathcal{X} = \{\frac{a_r}{\|a_r\|_2}\}_{r=1}^n$ takes the special choice consisting of the normalized row vectors of A and the step size sequence takes the special constant sequence $\eta_t \equiv 1$. If $\lambda = 0$, then our algorithm recovers the randomized Kaczmarz algorithm studied in [19].

We present learning theory analysis for the RSK (1.5) by assuming throughout the paper that $\{z_t = (x_t, y_t)\}_{t\in\mathbb{N}}$ are independently drawn from a Borel probability measure ρ on $\mathcal{Z}$. Different from the results in [6, 7, 20, 21, 28], our analysis does not require the consistency of the linear system $Aw = y$. So we do not need the sample value y to be exactly equal to the conditional mean of ρ at x, nor do we require the conditional mean (called the regression function of ρ below) to be a linear function. To handle this general case, a relaxation parameter sequence $\{\eta_t\}_t$ is necessary, as shown in our earlier work [19] for the special case with $\lambda = 0$ corresponding to the randomized Kaczmarz algorithm.

To state the limit of the vector sequence $\{w_t\}_{t\in\mathbb{N}}$ defined by (1.5), we denote by $Z = (X, Y)$ a random sample drawn from ρ and by $C_{\rho_X} = \mathbb{E}_Z[XX^\top]$ the covariance matrix of the marginal distribution ρ_X of ρ on $\mathcal{X}$, where $\mathbb{E}_Z$ is the expectation with respect to Z. Then the linear equation $C_{\rho_X} w = \mathbb{E}_Z[XY]$ is consistent and we denote its solution set as

$$W^* := \{w \in \mathbb{R}^d : C_{\rho_X} w = \mathbb{E}_Z[XY]\}. \tag{1.6}$$

Our target vector (the limit of $\{w_t\}$) is now defined by

$$w^* = \arg\min_{w\in W^*} \Psi(w). \tag{1.7}$$

Under a mild assumption ($w^* \neq 0$ and (2.1) below) on positive variance of the sampling process, we establish a sufficient and necessary condition as $\lim_{t\to\infty} \eta_t = 0$ and $\sum_{t=1}^\infty \eta_t = \infty$ for the convergence in expectation of the vector sequence $\{w_t\}_t$ produced by (1.5) to the target vector defined by (1.7). We also provide convergence rates of $\mathbb{E}[\|w_T - w^*\|_2^2]$ and show that these rates cannot be of order $O(T^{-q})$ for any $q > 2$. A sufficient condition with an additional constraint $\sum_{t=1}^\infty \eta_t^2 < \infty$ for the almost sure convergence is also provided. Our analysis is carried out based on error decompositions for the one-step progress of the RSK after interpreting it as a special online mirror descent algorithm [17, 22] with a *nondifferentiable* mirror map. Our main novelty in this paper is to bound a Bregman distance by means of the generalization error (to be defined below), which overcomes the difficulty caused by the nondifferentiability of the mirror map Ψ and making the method in [18] not applicable. The sufficient conditions are then established by showing that the generalization error

satisfies a restricted variant of strong convexity after observing that w^* defined in (1.7) is indeed the closest element to w_t in the solution set W^* in the sense of a Bregman distance. The necessary conditions are derived by observing that the convergence of $\{w_t\}_t$ implies that of $\{v_t\}_t$ restricted to the support of w^*, using special structures of the soft-thresholding operator.

2. Main results. We assume throughout the paper that $\mathcal{X} \subseteq \{x \in \mathbb{R}^d : \|x\|_2 \leq R\}$ for some $R > 0$ and that the target vector w^* defined by (1.7) is not the zero vector. Let $I = \{i \in \{1, \ldots, d\} : w^*(i) \neq 0\}$ be the support of w^* and we denote by $w(I) = (w(i))_{i \in I}$ the restriction of $w \in \mathbb{R}^d$ onto the index set I. For convenience we list in Table 1 the key notation used in this paper.

Our first main result, Theorem 1, gives a sufficient and necessary condition for the convergence in expectation of $\{\|w_t - w^*\|_2^2\}_t$ to 0. Its sufficiency part will be proved by Proposition 10 in section 4, for which the mild assumption (2.1) is not needed. The necessity part will be proved in section 5.

Theorem 1. *Let $\{(w_t, v_t)\}_{t \in \mathbb{N}}$ be the sequence generated by (1.5) and w^* defined by (1.7). Assume*

$$\inf_{w \in \mathbb{R}^d} \mathbb{E}_Z[\|(\langle w, X\rangle - Y)X(I)\|_2] > 0. \tag{2.1}$$

Then $\lim_{T \to \infty} \mathbb{E}_{z_1, \ldots, z_{T-1}}[\|w_T - w^*\|_2^2] = 0$ if and only if the step size sequence satisfies

$$\lim_{t \to \infty} \eta_t = 0 \quad \text{and} \quad \sum_{t=1}^{\infty} \eta_t = \infty. \tag{2.2}$$

In this case, we have

$$\sum_{T=1}^{\infty} \sqrt{\mathbb{E}_{z_1, \ldots, z_{T-1}}[\|w_T - w^*\|_2^2]} = \infty. \tag{2.3}$$

Our second main result, to be proved in section 6, gives a sufficient condition for the almost sure convergence of the RSK by imposing an additional constraint $\sum_{t=1}^{\infty} \eta_t^2 < \infty$.

Theorem 2. *Let $\{(w_t, v_t)\}_{t \in \mathbb{N}}$ be the sequence generated by (1.5) and w^* defined by (1.7). If the step size sequence satisfies*

$$\sum_{t=1}^{\infty} \eta_t = \infty \quad \text{and} \quad \sum_{t=1}^{\infty} \eta_t^2 < \infty, \tag{2.4}$$

then we have $\lim_{t \to \infty} \|w_t - w^*\|_2^2 = 0$ almost surely.

Condition (2.4) also appears in the literature to study the almost sure convergence of stochastic gradient descent algorithms [5]. It was commonly used in investigating online learning algorithms (e.g., [27, 30]). The second part of this condition implies $\lim_{t \to \infty} \eta_t = 0$.

Table 1

Notation and the number of the page where a notation appears first.

Notation	Meaning	Page
n	the number of examples	547
d	the dimension of the input space	547
A	a measurement-matrix in $\mathbb{R}^{n\times d}$	547
y	a vector of measurements in $\mathbb{R}^d$	547
λ	a threshold parameter	547
$\{\eta_t\}$	step size sequence	548
S_λ	the soft-thresholding function associated to the parameter λ	548
sgn	the sign operator	548
$\|\cdot\|_p$	ℓ_p-norm	548
$\Psi(w)$	$\lambda\|w\|_1 + 2^{-1}\|w\|_2^2$	548
$\langle\cdot,\cdot\rangle$	the Euclidean inner product in $\mathbb{R}^d$	548
$\mathcal{X},\mathcal{Y},\mathcal{Z}$	the input space, output space, and sample space	549
$z_t = (x_t, y_t)$	the tth example with $x_t \in \mathcal{X}$ and $y_t \in \mathcal{Y}$	549
$\{(w_t, v_t)\}$	the sequence of vector pairs in the algorithm	549
ρ	the probability measure defined over $\mathcal{Z}$	549
ρ_X	the marginal distribution of ρ on $\mathcal{X}$	549
C_{ρ_X}	the covariance matrix	549
$\mathbb{E}_Z$	the expectation with respect to Z	549
W^*	the set defined by (1.6)	549
w^*	the target vector defined by (1.7)	549
R	an upper bound on the radius of the input space	550
I	the support of w^*	550
$w(I)$	subvector of w with indices in I	550
$D_g^v(\tilde{w}, w)$	the Bregman distance between $\tilde{w}$ and w induced by g	552
$\partial g(w)$	the subdifferential of g at w	552
$L^2_{\rho_X}$	the L^2 space with respect to the marginal distribution ρ_X	553
f_ρ	the regression function	553
f_w	the linear function associated to the vector w	554
$\mathcal{H}$	the hypothesis space consisting of linear functions	554
$\mathcal{E}(w)$	the generalization error of f_w	554
$\sigma_{\min}(C_{\rho_X})$	the smallest positive eigenvalue of C_{ρ_X}	556
$\sigma_{\max}(C_{\rho_X})$	the largest positive eigenvalue of C_{ρ_X}	556
w_t^*	the orthogonal projection of w_t onto W^*	557
$\operatorname{dist}(S_1, S_2)$	the distance between the sets S_1 and S_2	557
κ	$\min_{i\in I}\|w^*(i)\|$	562
$\|I\|$	cardinality of I	562

So we see that (2.4) is stronger than (2.2). Condition (2.2) is found to be a necessary and sufficient condition for the convergence of the randomized Kaczmarz algorithm in [19]. Let us give some intuition on the necessity of this condition in Theorem 1. From the definition of the RSK (1.5), we see that the convergence of $\{\|w_t - w^*\|_2^2\}_t$ should imply that of $v_{t+1}(I) - v_t(I) = -\eta_t\big(\langle w_t, x_t\rangle - y_t\big)x_t(I)$ to the zero vector, from which together with (2.1) we expect $\lim_{t\to\infty}\eta_t = 0$. As we will show in Remark 1, the sequence $\{w_t\}$ satisfies $\|w_t\|_2^2 = O(\sum_{k=1}^t \eta_k)$. Hence the requirement $\sum_{k=1}^t \eta_k = \infty$ is needed for $\{w_t\}$ to explore vectors in the space $\mathbb{R}^d$.

Our third main result, to be proved in section 6, is on convergence rates of the RSK.

Theorem 3. *Let $\{(w_t, v_t)\}_{t\in\mathbb{N}}$ be the sequence generated by (1.5) and w^* defined by (1.7).*

(a) *If we take the step size sequence as $\eta_t = \eta_1 t^{-\theta}$ with $0 < \eta_1 \le (2R^2)^{-1}$ and $0 < \theta < 1$, then there exists a constant C_1 independent of T such that*

$$\mathbb{E}_{z_1,\dots,z_T}\left[\|w_{T+1} - w^*\|_2^2\right] \le C_1 T^{-\theta} \quad \forall T \in \mathbb{N}. \tag{2.5}$$

(b) *There exist constants $\tilde{a} > 0$ and $C_2 > 0$ independent of T such that with the step size sequence $\eta_t = \frac{2}{(t+1)\tilde{a}}$,*

$$\mathbb{E}_{z_1,\dots,z_T}\left[\|w_{T+1} - w^*\|_2^2\right] \le C_2 T^{-1} \quad \forall T \ge 4R^2\tilde{a}^{-1}. \tag{2.6}$$

Part (b) of Theorem 3 shows that $\mathbb{E}_{z_1,\dots,z_{T-1}}[\|w_T - w^*\|_2^2]$ may decay with the rate $O(T^{-1})$, though it cannot decay with a rate $O(T^{-q})$ for any $q > 2$ under the assumption (2.1) according to (2.3) in Theorem 1, in sharp contrast with the linear convergence [23] of the RSK in the consistency case with $\mathcal{X} = \{\frac{a_r}{\|a_r\|_2}\}_{r=1}^n$ and $y_{r(t)} = \langle w^*, a_{r(t)}\rangle$ in (1.4). This contrast is due to the positivity assumption (2.1) on the variance of the sampling process, which can be further seen in the one-step progress analysis below (see (4.13)). It would be interesting to investigate whether a rate $O(T^{-q})$ with $1 < q \le 2$ is possible for the RSK.

Our theoretical findings will be verified in the last section by some experimental results on tomography reconstruction in image processing.

Let us mention that our analysis and algorithm might be extended to the setting with functional data where the input set $\mathcal{X}$ becomes $L^2[0,1]$, the space of square integrable functions on $[0,1]$, each example x_t is a function in $L^2[0,1]$, and the inner product $\langle w_t, x_t\rangle$ is the one in $L^2[0,1]$. We leave detailed analysis to our further study.

The optimization problem (1.2) has been studied in a vast literature (e.g., [1, 2, 4, 7]). We focus on the general RSK (1.5) with motivations from computing sparse approximations, the recent work [20, 21, 23] in the special case with exact linear measurements, and designing scalable algorithms for handling big data or data arriving naturally in sequential ways. It would be interesting to extend our learning theory analysis to other methods as we have done for learning algorithms with kernels and general loss functions [11, 13, 15, 24].

3. Error decompositions and technical tools.

Our error decompositions for the RSK are based on two main tools, Bregman distance and excess generalization error.

3.1. Bregman distance associated with nondifferentiable mirror map.

The first main tool for our analysis of the RSK is the Bregman distance associated with the convex mirror map Ψ.

Definition 2. *For a convex function $g : \mathbb{R}^d \to \mathbb{R}$, we define its subdifferential at $w \in \mathbb{R}^d$ as*

$$\partial g(w) = \left\{v \in \mathbb{R}^d : g(\tilde{w}) - g(w) \ge \langle \tilde{w} - w, v\rangle \text{ for any } \tilde{w} \in \mathbb{R}^d\right\}.$$

For $v \in \partial g(w)$, we define the associated Bregman distance between $\tilde{w} \in \mathbb{R}^d$ and w by

$$D_g^v(\tilde{w}, w) = g(\tilde{w}) - g(w) - \langle \tilde{w} - w, v \rangle.$$

We used the Bregman distance in our earlier work [17, 18] on mirror descent algorithms where the mirror map is differentiable and $\partial g(w)$ consists of the gradient of g at w only. The mirror map $\Psi(w) = \lambda\|w\|_1 + 2^{-1}\|w\|_2^2$ in this paper is not differentiable, and our previous approach does not work. But this mirror map has its special explicit form involving only the 2-norm square and 1-norm and provides us some desired rich information: while its 2-norm square component yields its 1-strong convexity with respect to $\|\cdot\|_2$ in the sense that

$$D_\Psi^v(\tilde{w}, w) \geq \frac{1}{2}\|\tilde{w} - w\|_2^2 \qquad \forall \tilde{w}, w \in \mathbb{R}^d, v \in \partial\Psi(w), \tag{3.1}$$

with an additional 1-norm term the soft-thresholding operator S_λ is exactly the gradient of the Fenchel dual of Ψ as used in [20]. It follows that $v_{t+1} \in \partial\Psi(w_{t+1})$ which gives a useful choice for a subgradient to define the Bregman distance $D_\Psi^{v_t}(w, w_t)$. Moreover, we can see from (3.1) and (3.8) below the following upper and lower bounds for the Bregman distance between w^* and w_t defined by (1.5):

$$\frac{1}{2}\|w^* - w_t\|_2^2 \leq D_\Psi^{v_t}(w^*, w_t) \leq 2\lambda\sqrt{d}\|w^* - w_t\|_2 + \frac{1}{2}\|w^* - w_t\|_2^2.$$

For our analysis of estimating the expected Bregman distance $\mathbb{E}_{z_1,\ldots,z_{T-1}}[D_\Psi^{v_T}(w^*, w_T)]$, we shall refine these bounds by means of the second main tool, excess generalization error given in terms of the generalization error described in the next subsection.

3.2. Excess generalization error in the space of linear functions. In the standard least squares regression setting of learning theory [8], the generalization error of a measurable function $f : \mathcal{X} \to \mathcal{Y}$ is defined as $\mathcal{E}(f) = \frac{1}{2}\int_{\mathcal{Z}}(f(X) - Y)^2 d\rho$, which uses the least squares loss for taking the expected error to measure the generalization ability of the estimator f when predicting outcome values beyond the given sample points. In $L^2_{\rho_X}$, the L^2 space with respect to the marginal distribution ρ_X of ρ on $\mathcal{X}$, with the norm $\|f\|_{L^2_{\rho_X}} = \left(\int_{\mathcal{X}} |f(x)|^2 d\rho_X\right)^{1/2}$, the generalization error is minimized by the regression function $f_\rho : \mathcal{X} \to \mathcal{Y}$ defined by $f_\rho(x) = \int_{\mathcal{Y}} y d\rho(y|x)$, the mean of the conditional distribution $\rho(\cdot|x)$ at $x \in \mathcal{X}$. Moreover, the excess generalization error $\mathcal{E}(f) - \mathcal{E}(f_\rho)$ of an estimator f can be expressed in terms of the norm of the difference function $f - f_\rho$ as

$$2\big(\mathcal{E}(f) - \mathcal{E}(f_\rho)\big) = \int_{\mathcal{X}} (f(x) - f_\rho(x))^2 d\rho_X = \|f - f_\rho\|^2_{L^2_{\rho_X}} \qquad \forall f \in L^2_{\rho_X}.$$

A hypothesis space $\mathcal{H}$ of functions on $\mathcal{X}$ is often taken for implementing learning algorithms according to various learning tasks. A typical example is a reproducing kernel Hilbert space with a kernel chosen accordingly. For the purpose of analyzing the approximation of the target vector $w^* \in \mathbb{R}^d$ by the sequence $\{w_t\}_t$ of vectors in $\mathbb{R}^d$, we observe that the linear function f_w induced by a vector $w \in \mathbb{R}^d$ as $f_w(x) = \langle w, x \rangle$ has its $L^2_{\rho_X}$-norm square as

$$\|f_w\|^2_{L^2_{\rho_X}} = \int_{\mathcal{X}} |\langle w, x \rangle|^2 d\rho_X = \int_{\mathcal{X}} w^\top x x^\top w d\rho_X = w^\top C_{\rho_X} w.$$

This expression leads us to study the error $\|w_t - w^*\|_2$ from the approximation of the linear function f_{w^*} by f_{w_t} in the function space $L^2_{\rho_X}$. Take the hypothesis space consisting of homogeneous linear functions

$$\mathcal{H} = \{f_w : \mathcal{X} \to \mathcal{Y}\}_{w\in\mathbb{R}^d}, \text{where } f_w(x) = \langle w, x\rangle \text{ for } x \in \mathcal{X}.$$

For $w \in \mathbb{R}^d$, the doubled generalization error of the linear function $f_w \in \mathcal{H}$ equals

$$2\mathcal{E}(f_w) = \int_{\mathcal{Z}} (\langle w, x\rangle - y)^2 d\rho = \int_{\mathcal{Z}} w^\top x x^\top w - 2w^\top xy + y^2 d\rho = w^\top C_{\rho_X} w - 2w^\top \mathbb{E}[XY] + \mathbb{E}[Y^2].$$

Combining this with the constraint $C_{\rho_X} w = \mathbb{E}[XY]$ satisfied by any vector w in the set W^* defined by (1.6) tells us that f_w minimizes the generalization error over $\mathcal{H}$:

$$f_w = \arg\min_{f\in\mathcal{H}} \mathcal{E}(f) = \arg\min_{f\in\mathcal{H}} \|f - f_\rho\|^2_{L^2_{\rho_X}} \qquad \forall w \in W^*.$$

But the hypothesis space $\mathcal{H}$ is a finite-dimensional closed subspace of the Hilbert space $L^2_{\rho_X}$. So for any vector $w \in W^*$, the function f_w is a best approximation of f_ρ from $\mathcal{H}$. Moreover, for $w, \tilde{w} \in W^*$, the definition of W^* implies $C_{\rho_X}(w - \tilde{w}) = 0$ and thereby $f_w = f_{\tilde{w}}$:

$$\|f_w - f_{\tilde{w}}\|^2_{L^2_{\rho_X}} = \|(w - \tilde{w})^\top x\|^2_{L^2_{\rho_X}} = (w - \tilde{w})^\top C_{\rho_X}(w - \tilde{w}) = 0.$$

The vector w^* is a particular vector in the subspace $\mathcal{H}$ which minimizes the functional Ψ. The special form of the functional Ψ allows w^* to capture some a priori information on sparsity of the model. When $f_\rho \notin \mathcal{H}$, we have $f_{w^*} \neq f_\rho$ and say that the sampling process is nonlinear. When $\mathcal{E}(f_\rho) > 0$ no matter whether $f_\rho \in \mathcal{H}$ or not, we say that the sampling process is noisy. It is clear that the assumption (2.1) implies $\mathcal{E}(f_{w^*}) > 0$, which amounts to saying that the sampling process generated by ρ is either noisy or nonlinear [19].

Now we focus on the hypothesis space $\mathcal{H}$ and denote $\mathcal{E}(w) = \mathcal{E}(f_w)$ for $w \in \mathbb{R}^d$. Since a minimizer of the excess generalization error $\mathcal{E}(f_w) - \mathcal{E}(f_\rho)$ also minimizes $\mathcal{E}(f_w) - \mathcal{E}(f_{w^*}) = \mathcal{E}(w) - \mathcal{E}(w^*)$, we will estimate this last quantity in the following.

3.3. Error decompositions and novel estimates.

Our learning theory analysis of the RSK is based on the following two error decompositions for the one-step progress of the Bregman distance:

$$D_\Psi^{v_{t+1}}(w, w_{t+1}) - D_\Psi^{v_t}(w, w_t) = \langle w - w_{t+1}, v_t - v_{t+1}\rangle - D_\Psi^{v_t}(w_{t+1}, w_t), \tag{3.2}$$

$$D_\Psi^{v_{t+1}}(w, w_{t+1}) - D_\Psi^{v_t}(w, w_t) = \langle w - w_t, v_t - v_{t+1}\rangle + D_\Psi^{v_{t+1}}(w_t, w_{t+1}). \tag{3.3}$$

These identities can be found in the literature (e.g., [2]). They can be easily seen from the expression involving Bregman distances,

$$D_\Psi^{v_{t+1}}(w, w_{t+1}) - D_\Psi^{v_t}(w, w_t) = \Psi(w_t) - \Psi(w_{t+1}) - \langle w - w_{t+1}, v_{t+1}\rangle + \langle w - w_t, v_t\rangle, \tag{3.4}$$

by setting $w - w_t = w - w_{t+1} + w_{t+1} - w_t$ in the last term and $w - w_{t+1} = w - w_t + w_t - w_{t+1}$ in the middle term together with the definition of the Bregman distance. We shall use the

first error decomposition (3.2) for deriving sufficient conditions and the second decomposition (3.3) for deriving necessary conditions for the convergence of the RSK.

One technical novelty in our analysis of the RSK is to bound the Bregman distance $D_\Psi^{v_t}(w^*, w_t)$ by the quantity $\mathcal{E}(w_t) - \mathcal{E}(w^*)$, which will be proved in the next section.

Lemma 4. *Let $\{(w_t, v_t)\}_{t\in\mathbb{N}}$ be defined by (1.5) and w^* by (1.7). Then there exists a constant C_3 independent of t such that almost surely*

$$D_\Psi^{v_t}(w^*, w_t) \le C_3[\mathcal{E}(w_t) - \mathcal{E}(w^*)] \qquad \forall t \in \mathbb{N}. \tag{3.5}$$

The other technical novelty is to show in the following lemma, to be proved in the next section, that w^* is indeed the closest element to w_t in W^* almost surely, based on the key observation that v_t belongs to the range of C_{ρ_X} almost surely. Identity (3.7) establishes a generalized Pythagoras formula and estimate (3.8) controls the Bregman distance between w^* and w_t by the Euclidean distance between w and w_t for any $w \in W^*$.

Lemma 5. *Let $\{(w_t, v_t)\}_{t\in\mathbb{N}}$ be defined by (1.5) and w^* by (1.7). For $t \in \mathbb{N}$, w^* is the closest element to w_t in W^* with respect to the Bregman distance induced by Ψ almost surely meaning that*

$$w^* = \arg\min_{w\in W^*} D_\Psi^{v_t}(w, w_t). \tag{3.6}$$

Furthermore, there exists some $v^* \in \partial\Psi(w^*)$ such that there holds almost surely

$$D_\Psi^{v^*}(w, w^*) + D_\Psi^{v_t}(w^*, w_t) = D_\Psi^{v_t}(w, w_t), \qquad w \in W^*. \tag{3.7}$$

We also have almost surely

$$D_\Psi^{v_t}(w^*, w_t) \le 2\lambda\sqrt{d}\|w - w_t\|_2 + \frac{1}{2}\|w - w_t\|_2^2 \qquad \forall w \in W^*. \tag{3.8}$$

4. Sufficiency of convergence in expectation. This section presents the proof of the sufficiency of Theorem 1, after we prove Lemmas 5 and 4 stated in the previous section.

To prove Lemma 5, we need the following optimality condition for constrained optimization problems with linear constraints which follows from a direct application of Lagrange multipliers.

Lemma 6. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be a convex function, $A \in \mathbb{R}^{n\times d}$, and $y \in \mathbb{R}^n$. A point w^* is a solution to the minimization problem with linear constraints*

$$\min_{w\in\mathbb{R}^d} \{f(w) : Aw = y\} \tag{4.1}$$

if and only if there is some $\lambda^* \in \mathbb{R}^n$ such that

$$Aw^* = y, \quad A^\top\lambda^* \in \partial f(w^*).$$

Proof of Lemma 5. For any vector u in the kernel of C_{ρ_X}, we have

$$\mathbb{E}_{x_t}[(u^\top x_t)^2] = \mathbb{E}_{x_t}[u^\top x_t x_t^\top u] = u^\top C_{\rho_X} u = 0,$$

from which we know that $u^\top x_t = 0$ almost surely. Hence almost surely x_t is orthogonal to the kernel of C_{ρ_X} and therefore belongs to the range of C_{ρ_X}. It then follows by induction from $v_1 = 0$ and the definition of v_t in (1.5) that v_t also belongs to the range of C_{ρ_X} almost surely. So there exists some $\lambda_t \in \mathbb{R}^d$ such that $v_t = C_{\rho_X}\lambda_t$ almost surely. According to the definition of w^* in (1.7) and the optimality condition given in Lemma 6, we know the existence of some $\lambda^* \in \mathbb{R}^d$ such that $C_{\rho_X}\lambda^* \in \partial\Psi(w^*)$. According to the definition of the Bregman distance and $v_t \in \partial\Psi(w_t)$, we get almost surely

$$C_{\rho_X}(\lambda^* - \lambda_t) \in \partial\Psi(w^*) - v_t = \partial\left(\Psi(w) - \Psi(w_t) - \langle w - w_t, v_t\rangle\right)|_{w=w^*} = \partial D_\Psi^{v_t}(w, w_t)|_{w=w^*},$$

from which and Lemma 6 we verify (3.6).

Denote $v^* := C_{\rho_X}\lambda^* \in \partial\Psi(w^*)$. Let $w \in W^*$. Then $C_{\rho_X}(w - w^*) = 0$ and there holds almost surely

$$\langle w - w^*, v_t - v^*\rangle = \langle w - w^*, C_{\rho_X}(\lambda_t - \lambda^*)\rangle = \langle C_{\rho_X}(w - w^*), \lambda_t - \lambda^*\rangle = 0. \tag{4.2}$$

Combining this with the definition of the Bregman distance, we find almost surely

$$\begin{aligned} D_\Psi^{v^*}(w, w^*) + D_\Psi^{v_t}(w^*, w_t) - D_\Psi^{v_t}(w, w_t) &= \Psi(w) - \Psi(w^*) - \langle w - w^*, v^*\rangle \\ &\quad + \Psi(w^*) - \Psi(w_t) - \langle w^* - w_t, v_t\rangle - \Psi(w) + \Psi(w_t) + \langle w - w_t, v_t\rangle \\ &= \langle w - w^*, v_t - v^*\rangle = 0. \end{aligned}$$

This establishes identity (3.7).

To prove inequality (3.8), we apply (3.6) and know almost surely that for any $w \in W^*$,

$$\begin{aligned} D_\Psi^{v_t}(w^*, w_t) \le D_\Psi^{v_t}(w, w_t) &= \Psi(w) - \Psi(w_t) - \langle w - w_t, v_t\rangle \\ &= \lambda\|w\|_1 - \lambda\|w_t\|_1 + \frac{1}{2}\|w\|_2^2 - \frac{1}{2}\|w_t\|_2^2 - \langle w - w_t, v_t\rangle. \end{aligned}$$

Note that $\|w\|_2^2 = \|w - w_t + w_t\|_2^2 = \|w - w_t\|_2^2 + \|w_t\|_2^2 + 2\langle w - w_t, w_t\rangle$. Hence

$$D_\Psi^{v_t}(w, w_t) = \lambda\|w\|_1 - \lambda\|w_t\|_1 + \frac{1}{2}\|w - w_t\|_2^2 + \langle w - w_t, w_t - v_t\rangle. \tag{4.3}$$

By the definition (1.5), $w_t - v_t = S_\lambda(v_t) - v_t$. But the soft-thresholding function satisfies $\|S_\lambda(v) - v\|_\infty \le \lambda$. So $\langle w - w_t, w_t - v_t\rangle \le \lambda\|w - w_t\|_1$. Also, $\lambda\|w\|_1 - \lambda\|w_t\|_1 \le \lambda\|w - w_t\|_1$. Therefore,

$$D_\Psi^{v_t}(w^*, w_t) \le 2\lambda\|w - w_t\|_1 + \frac{1}{2}\|w - w_t\|_2^2 \le 2\lambda\sqrt{d}\|w - w_t\|_2 + \frac{1}{2}\|w - w_t\|_2^2 \tag{4.4}$$

almost surely. This verifies (3.8). The proof of Lemma 5 is complete. ∎

To prove Lemma 4, we need two auxiliary results. The first relates the quantity $\mathcal{E}(w_t) - \mathcal{E}(w^*)$ of w_t to its Euclidean distance to the set W^*. Let $\sigma_{\min}(C_{\rho_X})$ and $\sigma_{\max}(C_{\rho_X})$ be the smallest and the largest positive eigenvalue of C_{ρ_X}, respectively.

Lemma 7. *Let $\{(w_t, v_t)\}_t$ be defined by (1.5) and w^* by (1.7), and let w_t^* be the orthogonal projection of w_t onto W^* in $\mathbb{R}^d$ defined by*

$$w_t^* = \arg\min_{w \in W^*} \|w - w_t\|_2^2. \tag{4.5}$$

Then $\mathcal{E}(w_t^*) = \mathcal{E}(w^*)$ and

$$\frac{\sigma_{\min}(C_{\rho_X})}{2}\|w_t - w_t^*\|_2^2 \le \mathcal{E}(w_t) - \mathcal{E}(w^*) \le \frac{\sigma_{\max}(C_{\rho_X})}{2}\|w_t - w_t^*\|_2^2. \tag{4.6}$$

Proof. The identity $C_{\rho_X} w = \mathbb{E}_Z[XY]$ satisfied by both $w_t^* \in W^*$ and w^* implies $C_{\rho_X}(w_t^* - w^*) = 0$ and thereby

$$\begin{aligned} 2[\mathcal{E}(w_t^*) - \mathcal{E}(w^*)] &= \mathbb{E}_Z\left[\left(\langle w_t^* - w^*, X\rangle + \langle w^*, X\rangle - Y\right)^2\right] - \mathbb{E}_Z\left[(\langle w^*, X\rangle - Y)^2\right] \\ &= \mathbb{E}_Z\left[\langle w_t^* - w^*, X\rangle^2\right] + 2\mathbb{E}_Z\left[\langle w_t^* - w^*, X\rangle(\langle w^*, X\rangle - Y)\right] \\ &= \mathbb{E}_Z\left[(w_t^* - w^*)^\top X X^\top (w_t^* - w^*)\right] + 2\left\langle w_t - w^*, \mathbb{E}_X[XX^\top]w^* - \mathbb{E}_Z[XY]\right\rangle = 0, \end{aligned}$$

which verifies the first desired identity $\mathcal{E}(w_t^*) = \mathcal{E}(w^*)$.

In the same way, from the identity $C_{\rho_X} w_t^* = C_{\rho_X} w^* = \mathbb{E}[XY]$ again we find

$$\begin{aligned} 2[\mathcal{E}(w_t) - \mathcal{E}(w^*)] = 2[\mathcal{E}(w_t) - \mathcal{E}(w_t^*)] &= \mathbb{E}_Z\left[\left(\langle w_t - w_t^*, X\rangle + \langle w_t^*, X\rangle - Y\right)^2\right] - \mathbb{E}_Z\left[(\langle w_t^*, X\rangle - Y)^2\right] \\ &= \mathbb{E}_Z\left[\langle w_t - w_t^*, X\rangle^2\right] + 2\mathbb{E}_Z\left[\langle w_t - w_t^*, X\rangle(\langle w_t^*, X\rangle - Y)\right] \\ &= \mathbb{E}_Z[\langle w_t - w_t^*, X\rangle^2] = (w_t - w_t^*)^\top C_{\rho_X}(w_t - w_t^*). \end{aligned}$$

By Lemma 6 and the subdifferential formula $\partial\left(\|w - w_t\|_2^2\right)|_{w_t^*} = 2(w_t^* - w_t)$ we know that $w_t^* - w_t$ belongs to the range of C_{ρ_X} and is therefore orthogonal to the kernel of C_{ρ_X}, from which we see

$$\sigma_{\min}(C_{\rho_X})\|w_t - w_t^*\|_2^2 \le (w_t - w_t^*)^\top C_{\rho_X}(w_t - w_t^*) \le \sigma_{\max}(C_{\rho_X})\|w_t - w_t^*\|_2^2.$$

Then (4.6) follows. The proof is complete. ∎

The second auxiliary result is about distances between points and sets. The distance between two sets $S_1, S_2 \subset \mathbb{R}^d$ is defined as

$$\operatorname{dist}(S_1, S_2) = \inf_{u \in S_1, v \in S_2} \|u - v\|_2.$$

The distance between a point $w \in \mathbb{R}^d$ and a set $S \subset \mathbb{R}^d$ is $\operatorname{dist}(w, S) = \operatorname{dist}(\{w\}, S)$. Part (a) of the following lemma shows that the distance between a point and the intersection of two polyhedrons can be uniformly controlled by the maximum of its distance to these two polyhedrons. Here a set $S \in \mathbb{R}^d$ is said to be a polyhedron if it is the solution set of a linear inequality system, i.e., there exist a matrix $B \in \mathbb{R}^{m\times d}$ with $m \in \mathbb{N}$ and a vector $b \in \mathbb{R}^m$ such that $S = \{w \in \mathbb{R}^d : Bw \le b\}$. Part (b) shows that the distance between a compact set and a closed set is positive provided that they do not intersect. Part (a) can be found in Corollary 5.26 of [1], while part (b) is a standard result on metric spaces.

Lemma 8.

(a) *Let S_1, S_2 be two polyhedrons in $\mathbb{R}^d$. If $S_1 \cap S_2 \neq \emptyset$, then there exists some positive constant γ_{S_1,S_2} depending only on S_1, S_2 such that*

$$\operatorname{dist}(w, S_1 \cap S_2) \leq \gamma_{S_1,S_2} \max\big\{\operatorname{dist}(w, S_1), \operatorname{dist}(w, S_2)\big\} \qquad \forall w \in \mathbb{R}^d.$$

(b) *Let $H_1 \subset \mathbb{R}^d$ be compact and $H_2 \subset \mathbb{R}^d$ be closed. If $H_1 \cap H_2 = \emptyset$, then* $\operatorname{dist}(H_1, H_2) > 0$.

Based on the above two auxiliary results, we can now prove Lemma 4.

Proof of Lemma 4. Let $t \in \mathbb{N}$. We first consider the case $\|w_t\|_2 > 2\|w^*\|_2$. In this case, $\|w^* - w_t\|_2 \geq \|w^*\|_2$. Let w_t^* be defined by (4.5). Applying (3.8) with $w = w_t^*$ implies almost surely

$$\frac{1}{2}\|w^*\|_2^2 \leq \frac{1}{2}\|w^* - w_t\|_2^2 \leq D_\Psi^{v_t}(w^*, w_t) \leq 2\lambda\sqrt{d}\|w_t^* - w_t\|_2 + \frac{1}{2}\|w_t^* - w_t\|_2^2.$$

From this quadratic inequality concerning $\|w_t^* - w_t\|_2$ we derive almost surely

$$\|w_t^* - w_t\|_2 \geq -2\lambda\sqrt{d} + \sqrt{4\lambda^2 d + \|w^*\|_2^2} := c_1.$$

Plugging the above inequality back into (3.8) with $w = w_t^*$ gives almost surely

$$\begin{aligned} D_\Psi^{v_t}(w^*, w_t) &\leq 2\lambda\sqrt{d}\|w_t^* - w_t\|_2 + \frac{1}{2}\|w_t^* - w_t\|_2^2 \\ &\leq \left(2\lambda\sqrt{d}c_1^{-1} + \frac{1}{2}\right)\|w_t^* - w_t\|_2^2. \end{aligned}$$

Combining this with (4.6) yields almost surely

$$D_\Psi^{v_t}(w^*, w_t) \leq \left(4\lambda\sqrt{d}c_1^{-1} + 1\right)\sigma_{\min}^{-1}(C_{\rho_X})\big[\mathcal{E}(w_t) - \mathcal{E}(w^*)\big]. \tag{4.7}$$

Then we consider the case $\|w_t\|_2 \leq 2\|w^*\|_2$. For a vector $w \in \mathbb{R}^d$ with support I_w, we define a set (the half plane on its support)

$$P_w = \big\{\tilde{w} \in \mathbb{R}^d : \tilde{w}(i) = 0 \text{ for } i \notin I_w \text{ and } \tilde{w}(i)w(i) \geq 0 \text{ for } i \in I_w\big\}$$

and take

$$H_w = P_w \cap B_{2\|w^*\|_2},$$

where $B_r = \{w \in \mathbb{R}^d : \|w\|_2 \leq r\}$ denotes the ℓ_2-ball of radius $r \geq 0$. It is clear from $\|w_t\|_2 \leq 2\|w^*\|_2$ that $w_t \in H_{w_t}$.

We prove (3.5) according to whether H_{w_t} intersects W^*.

If $H_{w_t} \cap W^* = \emptyset$, then

$$\operatorname{dist}(H_{w_t}, W^*) \geq c_2 := \min_{w \in \mathbb{R}^d : H_w \cap W^* = \emptyset} \operatorname{dist}(H_w, W^*).$$

Since there are only 3^d different support sets corresponding to different signs of w, and $H_w \cap W^* = \emptyset$ implies $\operatorname{dist}(H_w, W^*) > 0$ by part (b) of Lemma 8, we know that c_2 is the minimum

of at most 3^d positive numbers and is therefore positive $c_2 > 0$. But $w_t \in H_{w_t}$ and $w_t^* \in W^*$. So $\|w_t - w_t^*\|_2 \geq \operatorname{dist}(H_{w_t}, W^*) \geq c_2$. Again, by (3.8), (4.6), and the identity $\mathcal{E}(w_t^*) = \mathcal{E}(w^*)$, we have almost surely

$$D_\Psi^{v_t}(w^*, w_t) \leq \left(2\lambda\sqrt{d}c_2^{-1} + \frac{1}{2}\right)\|w_t^* - w_t\|_2^2 \leq \left(4\lambda\sqrt{d}c_2^{-1} + 1\right)\sigma_{\min}^{-1}(C_{\rho_X})\left[\mathcal{E}(w_t) - \mathcal{E}(w^*)\right]. \tag{4.8}$$

If $H_{w_t} \cap W^* \neq \emptyset$, then it is clear that $P_{w_t} \cap W^* \neq \emptyset$. In this case, we define

$$\widetilde{w}_t = \arg\min_{w \in P_{w_t} \cap W^*} \|w - w_t\|_2^2.$$

Since $\widetilde{w}_t \in P_{w_t}$, we know that $\widetilde{w}_t(i) = w_t(i) = 0$ for $i \notin I_{w_t}$ and $\widetilde{w}_t(i)w_t(i) \geq 0$ for $i \in I_{w_t}$. This together with (4.3) for $\widetilde{w}_t \in W^*$ implies

$$D_\Psi^{v_t}(\widetilde{w}_t, w_t) = \lambda \sum_{i \in I_{w_t}} [|\widetilde{w}_t(i)| - |w_t(i)|] + \sum_{i \in I_{w_t}} [\widetilde{w}_t(i) - w_t(i)][w_t(i) - v_t(i)] + \frac{1}{2}\|\widetilde{w}_t - w_t\|_2^2.$$

But for $i \in I_{w_t}$, $w_t(i) \neq 0$, and $w_t(i) = S_\lambda(v_t(i))$. Hence $|v_t(i)| > \lambda$ and $v_t(i) = (|w_t(i)| + \lambda)\operatorname{sgn}(w_t(i))$. Also, $|\widetilde{w}_t(i)| - |w_t(i)| = \widetilde{w}_t(i)\operatorname{sgn}(\widetilde{w}_t(i)) - w_t(i)\operatorname{sgn}(w_t(i)) = (\widetilde{w}_t(i) - w_t(i))\operatorname{sgn}(w_t(i))$. Therefore,

$$D_\Psi^{v_t}(\widetilde{w}_t, w_t) = \lambda \sum_{i \in I_{w_t}} (\widetilde{w}_t(i) - w_t(i))\operatorname{sgn}(w_t(i)) - \lambda \sum_{i \in I_{w_t}} (\widetilde{w}_t(i) - w_t(i))\operatorname{sgn}(w_t(i)) + \frac{1}{2}\|\widetilde{w}_t - w_t\|_2^2,$$

which equals $\frac{1}{2}\|\widetilde{w}_t - w_t\|_2^2$. It is clear that P_w is a polyhedron for any $w \in \mathbb{R}^d$. The set W^* is also a polyhedron. According to part (a) of Lemma 8, for any P_w satisfying $P_w \cap W^* \neq \emptyset$, there exists a constant $\gamma_{P_w} > 0$ depending only on P_w and W^* such that

$$\operatorname{dist}(w_t, P_w \cap W^*) \leq \gamma_{P_w} \max\left\{\operatorname{dist}(w_t, P_w), \operatorname{dist}(w_t, W^*)\right\}. \tag{4.9}$$

Define $\gamma = \max_{w: P_w \cap W^* \neq \emptyset} \gamma_{P_w}$, which is finite since there are only 3^d different support sets P_w. According to (3.6), (4.9), and the proved identity $D_\Psi^{v_t}(\widetilde{w}_t, w_t) = \frac{1}{2}\|\widetilde{w}_t - w_t\|_2^2$, we get almost surely

$$\begin{aligned} D_\Psi^{v_t}(w^*, w_t) &\leq D_\Psi^{v_t}(\widetilde{w}_t, w_t) = \frac{1}{2}\|\widetilde{w}_t - w_t\|_2^2 = \frac{1}{2}\operatorname{dist}^2(w_t, P_{w_t} \cap W^*) \\ &\leq \frac{1}{2}\gamma^2 \max\left\{\operatorname{dist}^2(w_t, P_{w_t}), \operatorname{dist}^2(w_t, W^*)\right\} = \frac{1}{2}\gamma^2 \operatorname{dist}^2(w_t, W^*) \\ &\leq \sigma_{\min}^{-1}(C_{\rho_X})\gamma^2[\mathcal{E}(w_t) - \mathcal{E}(w^*)], \end{aligned}$$

where we have used the fact $w_t \in P_{w_t}$ in the last identity and (4.6) in the last inequality (note $\mathcal{E}(w^*) = \mathcal{E}(w)\ \forall\ w \in W^*$). Combining this with (4.7) and (4.8) verifies the desired bound by setting

$$C_3 = \sigma_{\min}^{-1}(C_{\rho_X}) \max\left\{4\lambda\sqrt{d}c_1^{-1} + 1, 4\lambda\sqrt{d}c_2^{-1} + 1, \gamma^2\right\}. \tag{4.10}$$

It is clear that C_3 is independent of t. The proof of Lemma 4 is complete. ∎

We shall also need the following elementary inequality which was essentially used in [19, 27, 30]. The proof is given in the appendix.

Lemma 9. *Let $\{\eta_t\}_{t\in\mathbb{N}}$ be a sequence of nonnegative numbers such that $\lim_{t\to\infty}\eta_t = 0$ and $\sum_{t=1}^{\infty}\eta_t = \infty$. Let $a > 0$ and $t_1 \in \mathbb{N}$ such that $\eta_t < a^{-1}$ for any $t \geq t_1$. Then we have $\lim_{T\to\infty}\sum_{t=t_1}^{T}\eta_t^2\prod_{k=t+1}^{T}(1-a\eta_k) = 0$.*

With the above lemmas, we are now in position to present the following convergence result, which is more general than the sufficiency part of Theorem 1 by means of (3.1) since condition (2.1) is not required.

Proposition 10. *Let $\{(w_t, v_t)\}_{t\in\mathbb{N}}$ be the sequence generated by (1.5) and w^* defined by (1.7). If the step size sequence satisfies (2.2), then $\lim_{T\to\infty}\mathbb{E}_{z_1,\ldots,z_{T-1}}[D_\Psi^{v_T}(w^*, w_T)] = 0$.*

Proof. Let $\gamma > 0$ be arbitrarily fixed. Since $\lim_{t\to\infty}\eta_t = 0$, we can find some $t_1 \in \mathbb{N}$ such that $\eta_t \leq \min\{(2R^2)^{-1}, 2C_3\}$ for $t \geq t_1$.

Now we apply the first error decomposition (3.2) for the RSK. Separate $w - w_{t+1}$ into $w - w_t + w_t - w_{t+1}$ and bound $\langle w_t - w_{t+1}, v_t - v_{t+1}\rangle$ by the Schwarz inequality as

$$\langle w_t - w_{t+1}, v_t - v_{t+1}\rangle \leq \frac{1}{2}\|w_t - w_{t+1}\|_2^2 + \frac{1}{2}\|v_t - v_{t+1}\|_2^2.$$

Combining this with the lower bound (3.1) of the Bregman distance $D_\Psi^{v_t}(w_{t+1}, w_t)$ in (3.2), we get

$$D_\Psi^{v_{t+1}}(w, w_{t+1}) - D_\Psi^{v_t}(w, w_t) \leq \langle w - w_t, v_t - v_{t+1}\rangle + \frac{1}{2}\|v_t - v_{t+1}\|_2^2.$$

Using the iteration (1.5), we find

$$D_\Psi^{v_{t+1}}(w, w_{t+1}) - D_\Psi^{v_t}(w, w_t) \leq \eta_t\langle w - w_t, (\langle w_t, x_t\rangle - y_t)x_t\rangle + 2^{-1}\eta_t^2\|(\langle w_t, x_t\rangle - y_t)x_t\|_2^2.$$

But $\|x_t\|_2 \leq R$ and the function on $\mathbb{R}^d$ mapping w to $2^{-1}(\langle w, x_t\rangle - y_t)^2$ is convex. Hence we have

$$D_\Psi^{v_{t+1}}(w, w_{t+1}) - D_\Psi^{v_t}(w, w_t) \leq \frac{\eta_t}{2}\left[(\langle w, x_t\rangle - y_t)^2 - (\langle w_t, x_t\rangle - y_t)^2\right] + \frac{\eta_t^2}{2}R^2(\langle w_t, x_t\rangle - y_t)^2. \tag{4.11}$$

Taking $w = w^*$ in the above inequality followed by taking conditional expectations on both sides (note (v_t, w_t) is independent of z_t), we derive

$$\begin{aligned}\mathbb{E}_{z_t}[D_\Psi^{v_{t+1}}(w^*, w_{t+1})] - D_\Psi^{v_t}(w^*, w_t) &\leq \eta_t[\mathcal{E}(w^*) - \mathcal{E}(w_t)] + \eta_t^2R^2\mathcal{E}(w_t)\\ &= (\eta_t - \eta_t^2R^2)[\mathcal{E}(w^*) - \mathcal{E}(w_t)] + \eta_t^2R^2\mathcal{E}(w^*).\end{aligned}$$

For $t \geq t_1$, we have $\eta_tR^2 \leq 1/2$ and know by Lemma 4 that almost surely

$$\begin{aligned}\mathbb{E}_{z_t}[D_\Psi^{v_{t+1}}(w^*, w_{t+1})] - D_\Psi^{v_t}(w^*, w_t) &\leq \frac{\eta_t}{2}[\mathcal{E}(w^*) - \mathcal{E}(w_t)] + \eta_t^2R^2\mathcal{E}(w^*)\\ &\leq -(2C_3)^{-1}\eta_t[D_\Psi^{v_t}(w^*, w_t)] + \eta_t^2R^2\mathcal{E}(w^*).\end{aligned} \tag{4.12}$$

Reformulation of the above inequality followed by taking expectations over the remaining random variables gives

$$\mathbb{E}_{z_1,\dots,z_t}[D_\Psi^{v_{t+1}}(w^*, w_{t+1})] \le (1-\tilde{a}\eta_t)\mathbb{E}_{z_1,\dots,z_{t-1}}[D_\Psi^{v_t}(w^*, w_t)] + \tilde{b}\eta_t^2, \tag{4.13}$$

where we denote the constants $\tilde{a} = (2C_3)^{-1}$ and $\tilde{b} = R^2\mathcal{E}(w^*)$. Applying this inequality iteratively for $t = T, \dots, t_1$ yields

$$\mathbb{E}_{z_1,\dots,z_T}[D_\Psi^{v_{T+1}}(w^*, w_{T+1})] \le \prod_{t=t_1}^{T}(1-\tilde{a}\eta_t)\mathbb{E}_{z_1,\dots,z_{t_1-1}}[D_\Psi^{v_{t_1}}(w^*, w_{t_1})] + \tilde{b}\sum_{t=t_1}^{T}\eta_t^2\prod_{k=t+1}^{T}(1-\tilde{a}\eta_k). \tag{4.14}$$

The first term of the above inequality can be estimated by the standard inequality $1-a \le \exp(-a)$ for $a > 0$ together with the assumption $\sum_{t=1}^{\infty}\eta_t = \infty$ as

$$\begin{aligned}\prod_{t=t_1}^{T}(1-\tilde{a}\eta_t)\mathbb{E}_{z_1,\dots,z_{t_1-1}}[D_\Psi^{v_{t_1}}(w^*, w_{t_1})] &\le \prod_{t=t_1}^{T}\exp\big(-\tilde{a}\eta_t\big)\mathbb{E}_{z_1,\dots,z_{t_1-1}}[D_\Psi^{v_{t_1}}(w^*, w_{t_1})] \\ &= \exp\left(-\tilde{a}\sum_{t=t_1}^{T}\eta_t\right)\mathbb{E}_{z_1,\dots,z_{t_1-1}}[D_\Psi^{v_{t_1}}(w^*, w_{t_1})] \to 0 \ (T\to\infty).\end{aligned}$$

Applying Lemma 9 with $a = \tilde{a}$, we get

$$\lim_{T\to\infty}\sum_{t=t_1}^{T}\eta_t^2\prod_{k=t+1}^{T}(1-\tilde{a}\eta_k) = 0.$$

Plugging the above two expressions into (4.14) shows the convergence of $\{\mathbb{E}_{z_1,\dots,z_{t-1}}[D_\Psi^{v_t}(w^*, w_t)]\}$ to 0. The proof of the proposition is complete. ∎

Remark 1. If $\eta_t \le R^{-2}\ \forall\ t \in \mathbb{N}$, one can take $w = 0$ in (4.11) to get

$$D_\Psi^{v_{t+1}}(0, w_{t+1}) - D_\Psi^{v_t}(0, w_t) \le 2^{-1}\eta_t y_t^2 + 2^{-1}\eta_t(\eta_t R^2 - 1)\big(\langle w_t, x_t\rangle - y_t\big)^2 \le 2^{-1}\eta_t y_t^2.$$

If $\mathcal{Y}$ is bounded, then there holds $D_\Psi^{v_t}(0, w_t) = O\big(\sum_{k=1}^{t}\eta_k\big)$.

If $\mathcal{E}(w^*) = 0$, then (4.13) translates to

$$\mathbb{E}_{z_1,\dots,z_t}[D_\Psi^{v_{t+1}}(w^*, w_{t+1})] \le (1-\tilde{a}\eta_t)\mathbb{E}_{z_1,\dots,z_{t-1}}[D_\Psi^{v_t}(w^*, w_t)],$$

from which one can derive linear convergence by taking a constant step size sequence $\eta_t \equiv \eta_1$.

Remark 2. There are interesting connections between (1.5) and the proximal operator in the literature of Bregman optimization (e.g., [2]). Indeed, the component w_{t+1} in (1.5) can be reformulated as

$$w_{t+1} = \arg\min_{w\in\mathbb{R}^d} D_\Psi^{v_t}(w, w_t) + \eta_t\big\langle w, \big(\langle w_t, x_t\rangle - y_t\big)x_t\big\rangle.$$

That is, w_{t+1} is the output [2] of a proximal operator associated with the Bregman distance induced by Ψ. However, from (4.13), we see that the so-called Bregman monotonicity [3] of $\{w_t\}_{t\in\mathbb{N}}$ with respect to W^* does not hold in general, due to the positive variance of the sampling process.

The RSK is also closely related to the online proximal gradient descent algorithm updating iterates [4, 10] defined as

$$w_{t+1} = S_{\tilde{\lambda}\eta_t}\big(w_t - \eta_t(\langle w_t, x_t\rangle - y_t)x_t\big), \tag{4.15}$$

which also performs a soft-thresholding operation $S_{\tilde{\lambda}}$ with a threshold parameter $\tilde{\lambda}$ after a gradient descent step. An essential difference between these two algorithms lies in the gradient descent step: online proximal gradient descent moves along the minus gradient direction from the position w_t, while the RSK moves along the minus gradient direction from the position v_t.

5. Necessity of convergence in expectation.

The proof of the necessity part of Theorem 1 is based on the following key lemma relating the one-step update on w_t to the one-step update on v_t, using special structures of the soft-thresholding operator. Recall that I is the support of w^*. Hence $\kappa := \min_{i\in I}|w^*(i)| > 0$. Denote the cardinality of the set I as $|I|$. Our idea is to estimate the expected value of the difference norm

$$\mathcal{D}_t := \|w_t(I) - v_t(I) - (w_{t+1}(I) - v_{t+1}(I))\|_2$$

according to whether w_t and w_{t+1} are close to w^* on I (satisfying $\|w_t(I) - w^*(I)\|_\infty \le \kappa/2$). Our novelty here is to bound the probability of the event $\|w_t(I) - w^*(I)\|_\infty > \kappa/2$ by the Bregman distance $D_\Psi^{v_t}(w^*, w_t)$.

Lemma 11. *Let w^* be defined by (1.7) and $\kappa = \min_{i\in I}|w^*(i)| > 0$. Then for $t \in \mathbb{N}$, there holds*

$$\begin{aligned}\mathbb{E}_{z_1,\dots,z_t}[\|v_t(I) - v_{t+1}(I)\|_2] \le{}& \mathbb{E}_{z_1,\dots,z_t}[\|w_t(I) - w_{t+1}(I)\|_2] \\ &+ \frac{16}{\kappa^2}\sqrt{|I|}\lambda\mathbb{E}_{z_1,\dots,z_t}[D_\Psi^{v_t}(w^*, w_t) + D_\Psi^{v_{t+1}}(w^*, w_{t+1})].\end{aligned} \tag{5.1}$$

Proof. Introduce a sequence of events

$$\Omega_t = \left\{\{(x_i, y_i)\}_{i=1}^{t-1} \in \mathcal{Z}^{t-1} : \|w_t(I) - w^*(I)\|_\infty \le \frac{\kappa}{2}\right\}, \qquad t \in \mathbb{N}.$$

For an event A, we denote by $\Pr(A)$ its probability and by $\bar{A}$ its complement. We express the expected value of $\mathcal{D}_t$ by the law of total expectation in terms of the event $\Omega_t \cap \Omega_{t+1}$ and its complement $\bar{\Omega}_t \cup \bar{\Omega}_{t+1}$ as

$$\mathbb{E}_{z_1,\dots,z_t}[\mathcal{D}_t] = \mathbb{E}_{z_1,\dots,z_t}\big[\mathcal{D}_t\big|\Omega_t \cap \Omega_{t+1}\big]\Pr\big(\Omega_t \cap \Omega_{t+1}\big) + \mathbb{E}_{z_1,\dots,z_t}\big[\mathcal{D}_t\big|\bar{\Omega}_t \cup \bar{\Omega}_{t+1}\big]\Pr\big(\bar{\Omega}_t \cup \bar{\Omega}_{t+1}\big).$$

Conditioned on the event $\Omega_t \cap \Omega_{t+1}$, the following identity holds for any $i \in I$:

$$\operatorname{sgn}(w_t(i)) = \operatorname{sgn}(w_{t+1}(i)) = \operatorname{sgn}(w^*(i)) \neq 0.$$

It implies by the relation $w_t = S_\lambda(v_t)$ that for $i \in I$, $w_t(i) - v_t(i) = \operatorname{sgn}(w_t(i))[|w_t(i)| - (|w_t(i)| + \lambda)]$, which together with the same equality for $w_{t+1}(i) - v_{t+1}(i)$ yields

$$w_t(i) - v_t(i) + v_{t+1}(i) - w_{t+1}(i) = 0.$$

Hence $\mathbb{E}_{z_1,\dots,z_t}\big[\mathcal{D}_t\big|\Omega_t \cap \Omega_{t+1}\big]\Pr\big(\Omega_t \cap \Omega_{t+1}\big) = 0$.

Conditioned on the event $\bar{\Omega}_t$, we have $\|w_t(I) - w^*(I)\|_\infty > 2^{-1}\kappa$. Hence

$$\begin{aligned}\Pr(\bar{\Omega}_t) &\le \Pr\left(\left\{\{(x_i, y_i)\}_{i=1}^{t-1} \in \mathcal{Z}^{t-1} : \|w_t(I) - w^*(I)\|_\infty^2/(\kappa/2)^2 > 1\right\}\right)\\ &\le \mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|w_t(I) - w^*(I)\|_\infty^2/(\kappa/2)^2\right]\\ &\le (2/\kappa)^2\,\mathbb{E}_{z_1,\dots,z_{t-1}}\left[\|w_t(I) - w^*(I)\|_2^2\right] \le \frac{8}{\kappa^2}\mathbb{E}_{z_1,\dots,z_{t-1}}[D_\Psi^{v_t}(w^*, w_t)].\end{aligned}$$

We also have $\Pr(\bar{\Omega}_{t+1}) \le \frac{8}{\kappa^2}\mathbb{E}_{z_1,\dots,z_t}[D_\Psi^{v_{t+1}}(w^*, w_{t+1})]$. Observe that $|w_t(i) - v_t(i)| \le \lambda$ for every i. So we have $\mathcal{D}_t \le \|w_t(I) - v_t(I)\|_2 + \|w_{t+1}(I) - v_{t+1}(I)\|_2 \le 2\sqrt{|I|}\lambda$ and thereby

$$\begin{aligned}\mathbb{E}_{z_1,\dots,z_t}\big[\mathcal{D}_t\big|\bar{\Omega}_t \cup \bar{\Omega}_{t+1}\big]\Pr\big(\bar{\Omega}_t \cup \bar{\Omega}_{t+1}\big) &\le 2\sqrt{|I|}\lambda[\Pr(\bar{\Omega}_t) + \Pr(\bar{\Omega}_{t+1})]\\ &\le 2\sqrt{|I|}\lambda\frac{8}{\kappa^2}\mathbb{E}_{z_1,\dots,z_t}[D_\Psi^{v_t}(w^*, w_t) + D_\Psi^{v_{t+1}}(w^*, w_{t+1})].\end{aligned}$$

Combining the above estimates for the two parts, we obtain

$$\mathbb{E}_{z_1,\dots,z_t}[\mathcal{D}_t] \le \frac{16}{\kappa^2}\sqrt{|I|}\lambda\mathbb{E}_{z_1,\dots,z_t}[D_\Psi^{v_t}(w^*, w_t) + D_\Psi^{v_{t+1}}(w^*, w_{t+1})].$$

Then the desired estimate (5.1) follows. The proof is complete. ∎

We can now state a proposition on necessity of the convergence measured by the Bregman distance, from which the necessity part of Theorem 1 follows as Corollary 13.

Proposition 12. *Let $\{(w_t, v_t)\}_t$ be the sequence generated by (1.5) and w^* defined by (1.7). Assume (2.1). If $\lim_{T\to\infty}\mathbb{E}_{z_1,\dots,z_{T-1}}[D_\Psi^{v_T}(w^*, w_T)] = 0$, then the step size sequence satisfies (2.2). In this case,*

$$\sum_{T=1}^{\infty}\sqrt{\mathbb{E}_{z_1,\dots,z_T}[D_\Psi^{v_{T+1}}(w^*, w_{T+1})]} = \infty. \tag{5.2}$$

Proof. We first show $\lim_{t\to\infty}\eta_t = 0$. Since $\lim_{t\to\infty}\mathbb{E}_{z_1,\dots,z_{t-1}}[D_\Psi^{v_t}(w^*, w_t)] = 0$, we know

$$\begin{aligned}\lim_{t\to\infty}\mathbb{E}_{z_1,\dots,z_t}[\|w_t - w_{t+1}\|_2^2] &= \lim_{t\to\infty}\mathbb{E}_{z_1,\dots,z_t}[\|w_t - w^* + w^* - w_{t+1}\|_2^2]\\ &\le 2\lim_{t\to\infty}\mathbb{E}_{z_1,\dots,z_t}\left[\|w_t - w^*\|_2^2 + \|w_{t+1} - w^*\|_2^2\right]\\ &\le 4\lim_{t\to\infty}\mathbb{E}_{z_1,\dots,z_t}\left[D_\Psi^{v_t}(w^*, w_t) + D_\Psi^{v_{t+1}}(w^*, w_{t+1})\right] = 0,\end{aligned}$$

from which and Hölder's inequality we get

$$\lim_{t\to\infty}\mathbb{E}_{z_1,\dots,z_t}[\|w_t(I) - w_{t+1}(I)\|_2] \le \lim_{t\to\infty}\mathbb{E}_{z_1,\dots,z_t}[\|w_t - w_{t+1}\|_2] \le \lim_{t\to\infty}\sqrt{\mathbb{E}_{z_1,\dots,z_t}[\|w_t - w_{t+1}\|_2^2]} = 0.$$

Combining this with (5.1) and the condition $\lim_{t\to\infty}\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi^{v_t}(w^*,w_t)]=0$ yields

$$\begin{aligned}&\lim_{t\to\infty}\eta_t\mathbb{E}_{z_1,\ldots,z_t}[\|(\langle w_t,x_t\rangle-y_t)x_t(I)\|_2]\\&=\lim_{t\to\infty}\mathbb{E}_{z_1,\ldots,z_t}[\|v_t(I)-v_{t+1}(I)\|_2]\\&\le\lim_{t\to\infty}\mathbb{E}_{z_1,\ldots,z_t}[\|w_t(I)-w_{t+1}(I)\|_2]\\&\quad+\frac{16}{\kappa^2}\sqrt{|I|}\lambda\lim_{t\to\infty}\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi^{v_t}(w^*,w_t)+D_\Psi^{v_{t+1}}(w^*,w_{t+1})]=0.\end{aligned}$$

According to the assumption $\sigma:=\inf_{w\in\mathbb{R}^d}\mathbb{E}_Z[\|(\langle w,X\rangle-Y)X(I)\|_2]>0$, we also know

$$\begin{aligned}\eta_t\mathbb{E}_{z_1,\ldots,z_t}[\|(\langle w_t,x_t\rangle-y_t)x_t(I)\|_2]&=\eta_t\mathbb{E}_{z_1,\ldots,z_{t-1}}\big[\mathbb{E}_{z_t}[\|(\langle w_t,x_t\rangle-y_t)x_t(I)\|_2]\big]\\&=\eta_t\mathbb{E}_{z_1,\ldots,z_{t-1}}\big[\mathbb{E}_Z[\|(\langle w_t,X\rangle-Y)X(I)\|_2]\big]\ge\eta_t\sigma.\end{aligned}$$

The necessary condition $\lim_{t\to\infty}\eta_t=0$ then follows from the above two inequalities.

We now turn to the necessary condition $\sum_{t=1}^\infty\eta_t=\infty$. Here we apply our second error decomposition (3.3) and use the expression for v_t-v_{t+1} in (1.5). We find

$$D_\Psi^{v_{t+1}}(w,w_{t+1})-D_\Psi^{v_t}(w,w_t)=\eta_t\langle w-w_t,(\langle w_t,x_t\rangle-y_t)x_t\rangle+D_\Psi^{v_{t+1}}(w_t,w_{t+1}).$$

Setting $w=w^*$ in the above identity, taking conditional expectations, and using $\mathbb{E}_{z_t}[x_tx_t^\top w_t-x_ty_t]=C_{\rho_X}(w_t-w^*)$ due to the definition of w^*, we get

$$\mathbb{E}_{z_t}[D_\Psi^{v_{t+1}}(w^*,w_{t+1})]-D_\Psi^{v_t}(w^*,w_t)=\eta_t\langle w^*-w_t,C_{\rho_X}(w_t-w^*)\rangle+\mathbb{E}_{z_t}[D_\Psi^{v_{t+1}}(w_t,w_{t+1})].$$

But $\langle w_t-w^*,C_{\rho_X}(w_t-w^*)\rangle\le\sigma_{\max}(C_{\rho_X})\|w_t-w^*\|_2^2\le2\sigma_{\max}(C_{\rho_X})D_\Psi^{v_t}(w^*,w_t)$ according to (3.1). It follows that

$$\mathbb{E}_{z_t}[D_\Psi^{v_{t+1}}(w^*,w_{t+1})]-D_\Psi^{v_t}(w^*,w_t)\ge-2\eta_t\sigma_{\max}(C_{\rho_X})D_\Psi^{v_t}(w^*,w_t)+\mathbb{E}_{z_t}[D_\Psi^{v_{t+1}}(w_t,w_{t+1})].$$

Taking expectations over the remaining random variables then yields

$$\begin{aligned}\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi^{v_{t+1}}(w^*,w_{t+1})]&\ge(1-2\eta_t\sigma_{\max}(C_{\rho_X}))\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi^{v_t}(w^*,w_t)]\\&\quad+\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi^{v_{t+1}}(w_t,w_{t+1})].\end{aligned}\tag{5.3}$$

Since $\lim_{t\to\infty}\eta_t=0$, there exists some integer $t_2>1$ such that $\eta_t\le(6\sigma_{\max}(C_{\rho_X}))^{-1}$ for $t\ge t_2$. Applying the standard inequality $1-\eta\ge\exp(-2\eta)$ for $\eta\in(0,1/3)$ to (5.3), we derive for $t\ge t_2$,

$$\begin{aligned}&\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi^{v_{t+1}}(w^*,w_{t+1})]\\&\quad\ge\exp\big(-4\eta_t\sigma_{\max}(C_{\rho_X})\big)\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi^{v_t}(w^*,w_t)]+\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi^{v_{t+1}}(w_t,w_{t+1})]\\&\quad\ge\exp\big(-4\eta_t\sigma_{\max}(C_{\rho_X})\big)\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi^{v_t}(w^*,w_t)].\end{aligned}$$

Applying the above inequality repeatedly from $t=T$ to $t=t_2+1$, we get

$$\begin{aligned}\mathbb{E}_{z_1,\ldots,z_T}[D_\Psi^{v_{T+1}}(w^*,w_{T+1})]&\ge\prod_{t=t_2+1}^T\exp\big(-4\eta_t\sigma_{\max}(C_{\rho_X})\big)\mathbb{E}_{z_1,\ldots,z_{t_2}}[D_\Psi^{v_{t_2+1}}(w^*,w_{t_2+1})]\\&=\exp\left(-4\sigma_{\max}(C_{\rho_X})\sum_{t=t_2+1}^T\eta_t\right)\mathbb{E}_{z_1,\ldots,z_{t_2}}[D_\Psi^{v_{t_2+1}}(w^*,w_{t_2+1})].\end{aligned}\tag{5.4}$$

We now show by contradiction that $\mathbb{E}_{z_1,\dots,z_{t_2}}[D_\Psi^{v_{t_2+1}}(w^*, w_{t_2+1})] > 0$. Suppose to the contrary that $\mathbb{E}_{z_1,\dots,z_{t_2}}[D_\Psi^{v_{t_2+1}}(w^*, w_{t_2+1})] = 0$. Then, by (5.3) and $1 - 2\eta_{t_2}\sigma_{\max}(C_{\rho_X}) > 0$, we get

$$\mathbb{E}_{z_1,\dots,z_{t_2-1}}[D_\Psi^{v_{t_2}}(w^*, w_{t_2})] = 0 \quad \text{and} \quad \mathbb{E}_{z_1,\dots,z_{t_2}}[D_\Psi^{v_{t_2+1}}(w_{t_2}, w_{t_2+1})] = 0. \tag{5.5}$$

According to (5.1), we then have

(5.6)

$$\begin{aligned}
&\mathbb{E}_{z_1,\dots,z_{t_2}}[\|w_{t_2}(I) - w_{t_2+1}(I)\|_2] \\
&\quad \geq \mathbb{E}_{z_1,\dots,z_{t_2}}[\|v_{t_2}(I) - v_{t_2+1}(I)\|_2] - \frac{16}{\kappa^2}\sqrt{|I|}\lambda\mathbb{E}_{z_1,\dots,z_{t_2}}[D_\Psi^{v_{t_2}}(w^*, w_{t_2}) + D_\Psi^{v_{t_2+1}}(w^*, w_{t_2+1})] \\
&\quad = \mathbb{E}_{z_1,\dots,z_{t_2}}[\|v_{t_2}(I) - v_{t_2+1}(I)\|_2] = \eta_{t_2}\mathbb{E}_{z_1,\dots,z_{t_2-1}}\big[\mathbb{E}_{z_{t_2}}[\|(\langle w_{t_2}, x_{t_2}\rangle - y_{t_2})x_{t_2}(I)\|_2]\big] \geq \eta_{t_2}\sigma,
\end{aligned}$$

where the first identity follows from the assumption $\mathbb{E}_{z_1,\dots,z_{t_2}}[D_\Psi^{v_{t_2+1}}(w^*, w_{t_2+1})] = 0$ and the induced identity $\mathbb{E}_{z_1,\dots,z_{t_2-1}}[D_\Psi^{v_{t_2}}(w^*, w_{t_2})] = 0$. This contradicts the second identity in (5.5). Therefore, $\mathbb{E}_{z_1,\dots,z_{t_2}}[D_\Psi^{v_{t_2+1}}(w^*, w_{t_2+1})] > 0$. Equation (5.4) can be reformulated as

$$\begin{aligned}
\sum_{t=t_2+1}^{T} \eta_t \geq \big(4\sigma_{\max}(C_{\rho_X})\big)^{-1}\big[&\log\big(\mathbb{E}_{z_1,\dots,z_{t_2}}[D_\Psi^{v_{t_2+1}}(w^*, w_{t_2+1})]\big) \\
&- \log\big(\mathbb{E}_{z_1,\dots,z_T}[D_\Psi^{v_{T+1}}(w^*, w_{T+1})]\big)\big].
\end{aligned}$$

The necessary condition $\sum_{t=1}^{\infty}\eta_t = \infty$ then follows from $\lim_{T\to\infty}\mathbb{E}_{z_1,\dots,z_T}[D_\Psi^{v_{T+1}}(w^*, w_{T+1})] = 0$ and $\mathbb{E}_{z_1,\dots,z_{t_2}}[D_\Psi^{v_{t_2+1}}(w^*, w_{t_2+1})] > 0$. This verifies (2.2).

We now turn to (5.2). Let $t \geq t_2$.

If $\mathbb{E}_{z_1,\dots,z_t}[D_\Psi^{v_t}(w^*, w_t) + D_\Psi^{v_{t+1}}(w^*, w_{t+1})] \leq \frac{\kappa^2}{32\sqrt{|I|}\lambda}\sigma\eta_t$, applying (5.1) again we have from the definition (1.5)

$$\mathbb{E}_{z_1,\dots,z_t}[\|w_t(I) - w_{t+1}(I)\|_2] \geq \mathbb{E}_{z_1,\dots,z_t}[\|v_t(I) - v_{t+1}(I)\|_2] - 2^{-1}\sigma\eta_t \geq 2^{-1}\sigma\eta_t.$$

This together with (5.3) yields

$$\begin{aligned}
\mathbb{E}_{z_1,\dots,z_t}[D_\Psi^{v_{t+1}}(w^*, w_{t+1})] &\geq \mathbb{E}_{z_1,\dots,z_t}[D_\Psi^{v_{t+1}}(w_t, w_{t+1})] \geq 2^{-1}\mathbb{E}_{z_1,\dots,z_t}[\|w_t - w_{t+1}\|_2^2] \\
&\geq 2^{-1}\big(\mathbb{E}_{z_1,\dots,z_t}[\|w_t - w_{t+1}\|_2]\big)^2 \geq 8^{-1}\sigma^2\eta_t^2.
\end{aligned}$$

This inequality can be reformulated as

$$\eta_t \leq \sigma^{-1}\sqrt{8\mathbb{E}_{z_1,\dots,z_t}[D_\Psi^{v_{t+1}}(w^*, w_{t+1})]}.$$

If $\mathbb{E}_{z_1,\dots,z_t}[D_\Psi^{v_t}(w^*, w_t) + D_\Psi^{v_{t+1}}(w^*, w_{t+1})] > \frac{\kappa^2}{32\sqrt{|I|}\lambda}\sigma\eta_t$, we immediately have

$$\eta_t \leq \frac{32\sqrt{|I|}\lambda}{\kappa^2\sigma}\mathbb{E}_{z_1,\dots,z_t}\big[D_\Psi^{v_t}(w^*, w_t) + D_\Psi^{v_{t+1}}(w^*, w_{t+1})\big].$$

Combining the above two inequalities on η_t in the two cases, we have for $t \geq t_2$,

$$\eta_t \leq 2\sigma^{-1} \max\left\{\sqrt{2\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi^{v_{t+1}}(w^*, w_{t+1})]}, \frac{16\sqrt{|I|}\lambda}{\kappa^2}\mathbb{E}_{z_1,\ldots,z_t}\left[D_\Psi^{v_t}(w^*, w_t)+D_\Psi^{v_{t+1}}(w^*, w_{t+1})\right]\right\},$$

which, coupled with the necessary condition $\sum_{t=1}^{\infty}\eta_t = \infty$, shows that

$$\sum_{t=1}^{\infty} \max\left\{\sqrt{\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi^{v_{t+1}}(w^*, w_{t+1})]}, \mathbb{E}_{z_1,\ldots,z_t}\left[D_\Psi^{v_t}(w^*, w_t) + D_\Psi^{v_{t+1}}(w^*, w_{t+1})\right]\right\} = \infty.$$

This together with $\lim_{t\to\infty}\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi^{v_t}(w^*, w_t)] = 0$ implies (5.2). The proof is complete. ∎

Corollary 13. *Let $\{(w_t, v_t)\}_t$ be the sequence generated by (1.5) and w^* defined by (1.7). Assume (2.1). If $\lim_{T\to\infty}[\|w^* - w_T\|_2^2] = 0$, then the step size sequence satisfies (2.2). In this case, (2.3) holds.*

Proof. According to (3.8) with $w = w^*$, we derive

$$\lim_{T\to\infty}\mathbb{E}_{z_1,\ldots,z_{T-1}}[D_\Psi^{v_T}(w^*, w_T)]$$
$$\leq 2\lambda\sqrt{d}\lim_{T\to\infty}\sqrt{\mathbb{E}_{z_1,\ldots,z_{T-1}}[\|w^* - w_T\|^2]} + \frac{1}{2}\lim_{T\to\infty}\mathbb{E}_{z_1,\ldots,z_{T-1}}[\|w^* - w_T\|_2^2] = 0,$$

from which and Proposition 12 we get (2.2) and (5.2). Furthermore, according to Lemma 4 and (4.6), we have almost surely

$$D_\Psi^{v_t}(w^*, w_t) \leq C_3[\mathcal{E}(w_t) - \mathcal{E}(w^*)] \leq \frac{C_3\sigma_{\max}(C_{\rho_X})}{2}\|w_t - w_t^*\|_2^2,$$

where w_t^* is defined in (4.5). But the definition of w_t^* implies $\|w_t - w_t^*\|_2 \leq \|w_t - w^*\|_2$. Therefore, together with (5.2), we have (2.3). The proof is complete. ∎

The assumption (2.1) is given in Theorem 1 in order to derive the necessary condition $\lim_{t\to\infty}\eta_t = 0$, based on the intuitive observation that the convergence of $\{w_t\}_t$ in expectation can ensure the convergence of $\{v_t(I)\}_t$ in expectation but not that of $\{v_t\}_t$. We present a proposition here to show that the necessary condition $\lim_{t\to\infty}\eta_t = 0$ is equivalent to $\lim_{t\to\infty}\mathbb{E}_{z_1,\ldots,z_t}[\|v_t - v_{t+1}\|_2] = 0$ under the assumption

$$\inf_{w\in\mathbb{R}^d}\mathbb{E}_Z[\|(\langle w, X\rangle - Y)X\|_2] > 0. \tag{5.7}$$

It would be interesting to investigate whether the condition (2.1) in Theorem 1 can be relaxed to the assumption (5.7).

Proposition 14. *Let $\{(w_t, v_t)\}_{t\in\mathbb{N}}$ be the sequence generated by (1.5) and w^* defined by (1.7). Assume (5.7) and the limit $\lim_{T\to\infty}\mathbb{E}_{z_1,\ldots,z_{T-1}}\left[\|w^* - w_T\|_2^2\right] = 0$. Then $\lim_{t\to\infty}\eta_t = 0$ if and only if $\lim_{t\to\infty}\mathbb{E}_{z_1,\ldots,z_t}[\|v_t - v_{t+1}\|_2] = 0$.*

Proof. Since $\lim_{T\to\infty}\mathbb{E}_{z_1,\ldots,z_{T-1}}\left[\|w^* - w_T\|_2^2\right] = 0$, there exists some constant $\sigma_2 > 0$ such that

$$\mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\|w^* - w_t\|_2^2\right] \leq \sigma_2 \qquad \forall t \in \mathbb{N}.$$

It follows from $\|x_t\|_2 \le R$ that

$$\begin{aligned}\eta_t \inf_{w\in\mathbb{R}^d} \mathbb{E}_Z[\|(\langle w, X\rangle - Y)X\|_2] &\le \eta_t \mathbb{E}_{z_1,\dots,z_t}\big[\|(\langle w_t, x_t\rangle - y_t)x_t\|_2\big] \\ &\le \eta_t \mathbb{E}_{z_1,\dots,z_t}\big[\|\langle w_t - w^*, x_t\rangle x_t\|_2\big] + \eta_t \mathbb{E}_{z_1,\dots,z_t}\big[\|(\langle w^*, x_t\rangle - y_t)x_t\|_2\big] \\ &\le \eta_t R^2 \mathbb{E}_{z_1,\dots,z_{t-1}}\big[\|w^* - w_t\|_2\big] + \eta_t \mathbb{E}_Z\big[\|(\langle w^*, X\rangle - Y)X\|_2\big] \\ &\le \eta_t R^2 \sqrt{\mathbb{E}_{z_1,\dots,z_{t-1}}\big[\|w^* - w_t\|_2^2\big]} + \eta_t \mathbb{E}_Z\big[\|(\langle w^*, X\rangle - Y)X\|_2\big] \\ &\le \eta_t R^2 \sqrt{\sigma_2} + \eta_t \mathbb{E}_Z\big[\|(\langle w^*, X\rangle - Y)X\|_2\big].\end{aligned}$$

But $\eta_t(\langle w_t, x_t\rangle - y_t)x_t = v_t - v_{t+1}$ by the first expression of the algorithm (1.5). So the above inequalities yield

$$\eta_t \inf_{w\in\mathbb{R}^d} \mathbb{E}_Z[\|(\langle w, X\rangle - Y)X\|_2] \le \mathbb{E}_{z_1,\dots,z_t}\big[\|v_t - v_{t+1}\|_2\big] \le \eta_t R^2\sqrt{\sigma_2} + \eta_t \mathbb{E}_Z\big[\|(\langle w^*, X\rangle - Y)X\|_2\big].$$

If $\lim_{t\to\infty} \eta_t = 0$, then the second inequality above implies $\lim_{t\to\infty} \mathbb{E}_{z_1,\dots,z_t}[\|v_t - v_{t+1}\|_2] = 0$. If $\lim_{t\to\infty} \mathbb{E}_{z_1,\dots,z_t}[\|v_t - v_{t+1}\|_2] = 0$, then (5.7) and the first inequality implies $\lim_{t\to\infty} \eta_t = 0$. The proof is complete. ∎

6. Almost sure convergence and convergence rates. The proof of Theorem 2 for almost sure convergence is based on the following Doob's forward convergence theorem (see, e.g., [9, p. 195]).

Lemma 15. *Let $\{\tilde{X}_t\}_{t\in\mathbb{N}}$ be a sequence of nonnegative random variables and let $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$ be a nested sequence of σ-algebras with $\mathcal{F}_t \subset \mathcal{F}_{t+1}$ $\forall\, t \in \mathbb{N}$. If $\mathbb{E}[\tilde{X}_{t+1}|\mathcal{F}_t] \le \tilde{X}_t$ almost surely for every $t \in \mathbb{N}$, then $\tilde{X}_t$ converges to a nonnegative random variable $\tilde{X}$ almost surely.*

Proof of Theorem 2. By (4.12) and $\mathcal{E}(w^*) \le \mathcal{E}(w_t)$, we know almost surely that for $t \ge t_1$,

$$\mathbb{E}_{z_t}[D_\Psi^{v_{t+1}}(w^*, w_{t+1})] \le D_\Psi^{v_t}(w^*, w_t) + \eta_t^2 R^2 \mathcal{E}(w^*). \tag{6.1}$$

Since $\sum_{t=1}^\infty \eta_t^2 < \infty$, we can define a random process as

$$\tilde{X}_t = D_\Psi^{v_{t+1}}(w^*, w_{t+1}) + R^2\mathcal{E}(w^*) \sum_{\tilde{t}=t+1}^\infty \eta_{\tilde{t}}^2, \qquad t \in \mathbb{N}.$$

Inequality (6.1) amounts to saying that $\mathbb{E}_{z_t}\big[\tilde{X}_t\big] \le \tilde{X}_{t-1}$ almost surely for $t \ge t_1$ and therefore the random process $\{\tilde{X}_t : t \ge t_1\}$ is a supermartingale. Also, $\tilde{X}_t \ge 0$. Lemma 15 then shows that $\tilde{X}_t$ converges almost surely, which, together with the condition $\sum_{t=1}^\infty \eta_t^2 < \infty$, implies that $D_\Psi^{v_t}(w^*, w_t)$ converges to a nonnegative random variable $\tilde{X}$ almost surely. According to Fatou's lemma and the limit $\lim_{t\to\infty} \mathbb{E}[D_\Psi^{v_t}(w^*, w_t)] = 0$ established in Proposition 10, we get

$$\mathbb{E}[\tilde{X}] = \mathbb{E}\big[\lim_{t\to\infty} D_\Psi^{v_t}(w^*, w_t)\big] \le \liminf_{t\to\infty} \mathbb{E}[D_\Psi^{v_t}(w^*, w_t)] = 0.$$

But $\tilde{X}$ is a nonnegative random variable, so we have $\tilde{X} = 0$ almost surely. It follows that $\{D_\Psi^{v_t}(w^*, w_t)\}_{t\in\mathbb{N}}$ converges to 0 almost surely. By means of (3.1), the proof is complete. ∎

To prove Theorem 3 on convergence rates in expectation, we need the following elementary inequality which can be found in [25]:

$$\sum_{i=1}^{t-1} i^{-q_2} \exp\left(-\nu \sum_{j=i+1}^{t} j^{-q_1}\right) \le \left(\frac{2^{q_1+q_2}}{\nu} + \left(\frac{1+q_2}{\nu(1-2^{q_1-1})e}\right)^{\frac{1+q_2}{1-q_1}}\right) t^{q_1-q_2}, \qquad t \in \mathbb{N}, \tag{6.2}$$

where $\nu > 0$, $q_2 \ge 0$, and $0 < q_1 < 1$.

Proof of Theorem 3. (a) Since $\eta_t \le (2R^2)^{-1}$ for $t \in \mathbb{N}$, applying $1 - a \le \exp(-a)$ for $a > 0$ to (4.14) yields

$$\begin{aligned}\mathbb{E}_{z_1,\dots,z_T}[D_\Psi^{v_{T+1}}(w^*, w_{T+1})] &\le \prod_{t=1}^{T} \exp(-\tilde{a}\eta_t) D_\Psi^{v_1}(w^*, w_1) + \tilde{b}\sum_{t=1}^{T} \eta_t^2 \prod_{k=t+1}^{T} \exp(-\tilde{a}\eta_k) \\ &= \exp\left(-\tilde{a}\eta_1 \sum_{t=1}^{T} t^{-\theta}\right) D_\Psi^{v_1}(w^*, w_1) + \tilde{b}\eta_1^2 \sum_{t=1}^{T} t^{-2\theta} \exp\left(-\tilde{a}\eta_1 \sum_{k=t+1}^{T} k^{-\theta}\right).\end{aligned}$$

Since $\sum_{t=1}^{T} t^{-\theta} \ge \frac{1}{1-\theta}[(T+1)^{1-\theta} - 1]$ for $0 < \theta < 1$, applying (6.2) with $q_1 = \theta, q_2 = 2\theta, \nu = \tilde{a}\eta_1$ gives

$$\begin{aligned}\mathbb{E}_{z_1,\dots,z_T}[D_\Psi^{v_{T+1}}(w^*, w_{T+1})] &\le \exp\left(\frac{\tilde{a}\eta_1}{1-\theta}\right) \exp\left(-\frac{\tilde{a}\eta_1}{1-\theta} T^{1-\theta}\right) D_\Psi^{v_1}(w^*, w_1) \\ &\quad + \tilde{b}\eta_1^2 T^{-2\theta} + \tilde{b}\eta_1^2 \left(\frac{2^{3\theta}}{\tilde{a}\eta_1} + \left(\frac{1+2\theta}{\tilde{a}\eta_1(1-2^{\theta-1})e}\right)^{\frac{1+2\theta}{1-\theta}}\right) T^{-\theta}.\end{aligned}$$

An elementary inequality found in [25] asserts that $\exp(-\nu x) \le \left(\alpha(\nu e x)^{-1}\right)^{\alpha}$ for $x, \nu, \alpha > 0$. Using this inequality with $\alpha = \frac{\theta}{1-\theta}, x = T^{1-\theta}$ and $\nu = \tilde{a}\eta_1(1-\theta)^{-1}$ yields

$$\begin{aligned}\mathbb{E}_{z_1,\dots,z_T}[D_\Psi^{v_{T+1}}(w^*, w_{T+1})] &\le \exp\left(\frac{\tilde{a}\eta_1}{1-\theta}\right) \left(\theta(\tilde{a}\eta_1 e)^{-1}\right)^{\frac{\theta}{1-\theta}} T^{-\theta} D_\Psi^{v_1}(w^*, w_1) \\ &\quad + \tilde{b}\eta_1^2 \left(1 + \frac{2^{3\theta}}{\tilde{a}\eta_1} + \left(\frac{1+2\theta}{\tilde{a}\eta_1(1-2^{\theta-1})e}\right)^{\frac{1+2\theta}{1-\theta}}\right) T^{-\theta}.\end{aligned}$$

Together with (3.1), this verifies (2.5) with

$$C_1 = 2\exp\left(\frac{\tilde{a}\eta_1}{1-\theta}\right) \left(\theta(\tilde{a}\eta_1 e)^{-1}\right)^{\frac{\theta}{1-\theta}} D_\Psi^{v_1}(w^*, w_1) + 2\tilde{b}\eta_1^2 \left(1 + \frac{2^{3\theta}}{\tilde{a}\eta_1} + \left(\frac{1+2\theta}{\tilde{a}\eta_1(1-2^{\theta-1})e}\right)^{\frac{1+2\theta}{1-\theta}}\right).$$

(b) Let $\tilde{t}_1$ be the smallest positive integer no less than $\frac{4R^2}{\tilde{a}}$ with $\tilde{a} = (2C_3)^{-1}$. It is clear that $\eta_t \le (2R^2)^{-1}$ for any $t \ge \tilde{t}_1$. Plugging the step size choice $\eta_t = \frac{2}{(t+1)\tilde{a}}$ into (4.13) then gives

$$\mathbb{E}_{z_1,\dots,z_t}\left[D_\Psi^{v_{t+1}}(w^*, w_{t+1})\right] \le \frac{t-1}{t+1}\mathbb{E}_{z_1,\dots,z_{t-1}}\left[D_\Psi^{v_t}(w^*, w_t)\right] + \frac{4\tilde{b}}{\tilde{a}^2(t+1)^2} \quad \forall t \ge \tilde{t}_1.$$

Multiplying both sides of the above inequality by $t(t+1)$ implies

$$t(t+1)\mathbb{E}_{z_1,\dots,z_t}\left[D_\Psi^{v_{t+1}}(w^*, w_{t+1})\right] \le t(t-1)\mathbb{E}_{z_1,\dots,z_{t-1}}\left[D_\Psi^{v_t}(w^*, w_t)\right] + 4\tilde{a}^{-2}\tilde{b} \qquad \forall t \ge \tilde{t}_1.$$

Summing the above inequality from $t = T$ to $t = \tilde{t}_1$ yields

$$T(T+1)\mathbb{E}_{z_1,\dots,z_T}\left[D_\Psi^{v_{T+1}}(w^*, w_{T+1})\right] \le (\tilde{t}_1 - 1)\tilde{t}_1\mathbb{E}_{z_1,\dots,z_{\tilde{t}_1-1}}\left[D_\Psi^{v_{\tilde{t}_1}}(w^*, w_{\tilde{t}_1})\right] + 4\tilde{a}^{-2}\tilde{b}(T - \tilde{t}_1 + 1)$$

for any $T \ge \tilde{t}_1$, from which we have

$$\mathbb{E}_{z_1,\dots,z_T}\left[D_\Psi^{v_{T+1}}(w^*, w_{T+1})\right] \le \frac{(\tilde{t}_1 - 1)\tilde{t}_1\mathbb{E}_{z_1,\dots,z_{\tilde{t}_1-1}}\left[D_\Psi^{v_{\tilde{t}_1}}(w^*, w_{\tilde{t}_1})\right]}{T(T+1)} + \frac{4\tilde{b}}{T\tilde{a}^2} \quad \forall T \ge \tilde{t}_1.$$

This together with (3.1) establishes (2.6) with

$$C_2 = 2\tilde{t}_1\mathbb{E}_{z_1,\dots,z_{\tilde{t}_1-1}}\left[D_\Psi^{v_{\tilde{t}_1}}(w^*, w_{\tilde{t}_1})\right] + 8\tilde{a}^{-2}\tilde{b}. \tag{6.3}$$

The proof of Theorem 3 is complete. ∎

Remark 3. According to C_3 defined in (4.10), we see that C_3 is bounded by $\tilde{C}_3\sigma_{\min}^{-1}(C_{\rho_X})(1+\lambda\sqrt{d})$ with a constant $\tilde{C}_3$ depending only on W^* and $\|w^*\|_2$. Therefore the constant C_2 defined in (6.3) can be asymptotically bounded by $O\big(\sigma_{\min}^{-2}(C_{\rho_X})(\lambda^2 d+1)(R^2+1)\big)$. It enjoys a linear dependency on the dimension d, a quadratic dependency on λ, and a quadratic dependency on $\sigma_{\min}^{-1}(C_{\rho_X})$. The dimension d enters into the constant C_3 by the use of (4.4), where an upper bound of ℓ_1-norm in terms of ℓ_2-norms results in a factor of $\sqrt{d}$. It would be interesting to study whether the dimension d in C_3 can be replaced by the cardinality $|I|$ of the support set of w^* reflecting the sparsity of the model. The dependency on λ in C_3 shows that λ trades off the sparsity of iterates and learning rates: a large λ would increase the sparsity level of iterates but lower the learning rates.

7. Numerical experiments.

In this section, we apply the general RSK (1.5) to the problem of tomography reconstruction in image processing. We use the AIR toolbox [14], followed with the ℓ_2-normalization of rows, to create a CT-measurement matrix $A \in \mathbb{R}^{n\times d}$ and an $N \times N$ sparse image encoded by a vector $w^\dagger \in \mathbb{R}^d$ with $d = N^2$. Each row of A corresponds to a line integral from a fan beam projection geometry. Note that the corresponding lines form a preselected set in our experiments. We consider a noisy case where the line integrals are corrupted by Gaussian noises. Our purpose is to reconstruct the image $w^\dagger$ from the matrix A and the noisy measurements by (1.5). At the tth step, we randomly choose an index i_t

from the uniform distribution over $\{1,\dots,n\}$. We then choose the transpose of A_{i_t} as x_t and $y_t = A_{i_t}w^\dagger + s_t$, where A_{i_t} is the i_tth row of A and s_t is a Gaussian random variable with mean 0 and standard deviation $0.1|A_{i_t}w^\dagger|$.

To illustrate our convergence results with step sizes, we consider three sequences of polynomially decaying step sizes $\eta_t = (1+t\sigma_{\min}(C_{\rho_X}))^{-\theta}$ with $\theta = 0, \theta = \frac{1}{2}$, and $\theta = 1$. The constant $\sigma_{\min}(C_{\rho_X})$ is chosen according to the constant $\tilde{a} = (2C_3)^{-1}$ in the proof of Theorem 3(b) and the constant C_3 given in (4.10). Note that our algorithm (1.5) with $\theta = 0$ recovers the RSK in [23] with constant step size $\eta_t \equiv 1$. To demonstrate the effectiveness of our algorithm, we also compare our algorithm with two baseline methods: the randomized Kaczmarz algorithm (RK) in [19] and the online proximal gradient descent method (OPGD) given in (4.15). We also consider different step sizes for these two baseline methods. We repeat the experiments 10 times for each case and report here the average of the results.

Figure 1 shows the experimental results for tomography reconstruction with $N = 128$ and $n = 92160$. Panel (a) displays the true image with $\|w^\dagger\|_0 = 6794$ with $\|w^\dagger\|_0$ being the number of nonzero entries of $w^\dagger$. The reconstructed image, encoded by $\hat{w} \in \mathbb{R}^{16384}$, by RSK (1.5) with linearly decaying step size sequence ($\theta = 1$) is shown in panel (b). Panels (c), (d), and (e) give the plots of relative errors $\mathrm{err}_r(w_t) := 100\|w_t - w^\dagger\|_2/\|w^\dagger\|_2$ as functions of iteration numbers for RSK with $\lambda = 1$, RK, and OPGD (4.15) with $\tilde{\lambda} = 3 \times 10^{-6}$, respectively. Panel (f) shows the number of nonzero elements of models along the iterations for the considered three models.

From panel (c), we can see that RSK with a constant step size sequence cannot reconstruct well the true image due to the existence of noise. As a comparison, (1.5) with $\theta = 1/2$ and $\theta = 1$ attain decaying relative errors along the iterations. This is consistent with our theoretical results on sufficient and necessary conditions for the convergence of (1.5). Furthermore, it can be seen clearly that (1.5) with $\theta = 1$ achieves a faster convergence rate than that with $\theta = 1/2$, which is consistent with our error rate analysis. According to panel (d) and panel (e), we know that step sizes should also diminish along the iterations for the convergence of RK and OPGD in the noisy case. From panel (f), we can see that the RSK is able to capture the sparsity of the models. Indeed, the average $\|w_t\|_0$ for the last iterate of RSK with $\theta = 1$ is 6913, while $\|w^\dagger\|_0 = 6794$. As a comparison, the sparsity of models is not preserved by either RK or OPGD. Table 2 exhibits the relative errors $\mathrm{err}_r(w_{2\times 10^6})$ with standard deviations for RSK with $\lambda = 1$, RK, and OPGD with regularization parameters $\tilde{\lambda} \in \{10^{-6}, 3\times 10^{-6}, 10^{-5}, 3\times 10^{-5}\}$, from which one can see that RSK outperforms the two baseline methods by attaining smaller relative errors in this example.

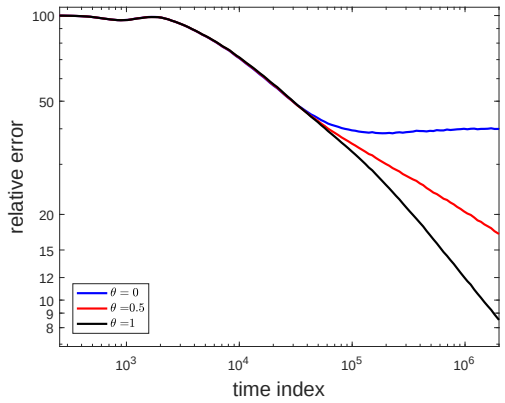
Table 2

Relative errors with standard deviations for different methods on tomography reconstruction with $N = 128, n = 92160$, $\|w^\dagger\|_0 = 6794$, and 10% relative noise. We consider models with lineary decaying step sizes, i.e., $\theta = 1$ in this table.

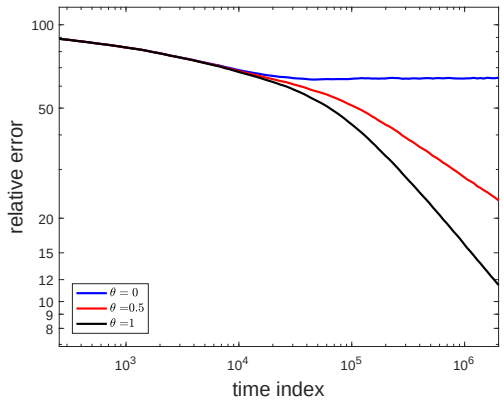
	RSK	RK	OPGD			
			$\tilde{\lambda} = 10^{-6}$	$\tilde{\lambda} = 3\times 10^{-6}$	$\tilde{\lambda} = 10^{-5}$	$\tilde{\lambda} = 3\times 10^{-5}$
$\mathrm{err}_r(w_{2\times 10^6})$	$\mathbf{8.52 \pm 0.09}$	11.44 ± 0.08	9.85 ± 0.09	9.14 ± 0.07	9.21 ± 0.08	10.68 ± 0.09

(a) True image.

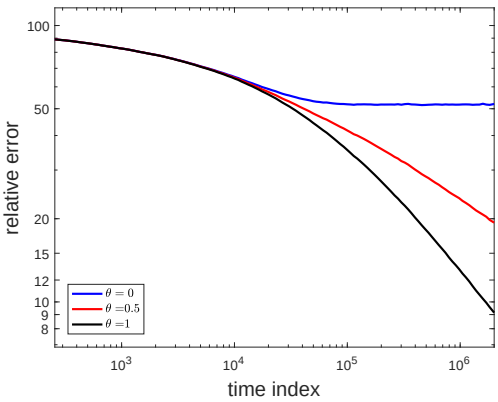
(b) Reconstructed image.



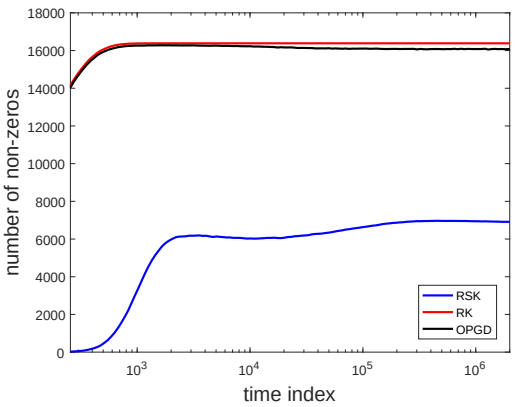
(c) Randomized sparse Kaczmarz method.



(d) Randomized Kaczmarz method.



(e) Online proximal gradient descent.



(f) Sparsity versus iteration numbers.

Figure 1. *Tomography reconstruction with* $N = 128, n = 92160$, $\|w^\dagger\|_0 = 6794$, *and* 10% *relative noise. Panel* (a) *is the true image. Panel* (b) *is the reconstructed image* $\hat{w}$ *by* (1.5) *with a linearly decaying step size sequence, i.e.,* $\theta = 1$. *Panel* (c) *shows the plots of relative errors for RSK with* $\lambda = 1$ *for polynomially decaying step sizes with* $\theta = 0$ *(blue),* $\theta = 1/2$ *(red), and* $\theta = 1$ *(black). Panel* (d) *shows the plots of relative errors for RK. Panel* (e) *shows the plots of relative errors for OPGD* (4.15) *with the regularization parameter* $\tilde{\lambda} = 3 \times 10^{-6}$. *Panel* (f) *plots* $\|w_t\|_0$ *versus iteration numbers for RSK method with* $\lambda = 1$ *(blue), RK (red), and OPGD with* $\tilde{\lambda} = 10^{-5}$ *(black).*

Appendix: Proof of Lemma 9. Since $\lim_{t\to\infty}\eta_t = 0$ and $\sum_{t=1}^{\infty}\eta_t = \infty$, for any $\epsilon > 0$ we can find some integer $t(\epsilon) \geq t_1$ such that $\eta_t \leq \frac{a\epsilon}{2}\ \forall t \geq t(\epsilon)$ and then another integer $t'(\epsilon) > t(\epsilon)$ such that

$$\sum_{k=t(\epsilon)+1}^{T}\eta_k \geq \frac{1}{a}\log\frac{2t(\epsilon)}{a^2\epsilon} \qquad \forall T \geq t'(\epsilon).$$

For $T \geq t'(\epsilon)$, we then have

$$\begin{aligned}\sum_{t=t_1}^{T}\eta_t^2\prod_{k=t+1}^{T}(1-a\eta_k) &= \sum_{t=t_1}^{t(\epsilon)}\eta_t^2\prod_{k=t+1}^{T}(1-a\eta_k) + \sum_{t=t(\epsilon)+1}^{T}\eta_t^2\prod_{k=t+1}^{T}(1-a\eta_k)\\ &\leq \sum_{t=t_1}^{t(\epsilon)}\eta_t^2\prod_{k=t+1}^{T}(1-a\eta_k) + \frac{a\epsilon}{2}\sum_{t=t(\epsilon)+1}^{T}\eta_t\prod_{k=t+1}^{T}(1-a\eta_k)\\ &\leq \sum_{t=t_1}^{t(\epsilon)}\eta_t^2\prod_{k=t+1}^{T}(1-a\eta_k) + \frac{\epsilon}{2}, \qquad \text{(A)}\end{aligned}$$

where we have used the notation $\prod_{k=t+1}^{T}(1-a\eta_k) = 1$ for $t = T$ and the inequality

$$\begin{aligned}\sum_{t=t(\epsilon)+1}^{T}a\eta_t\prod_{k=t+1}^{T}(1-a\eta_k) &= \sum_{t=t(\epsilon)+1}^{T}[1-(1-a\eta_t)]\prod_{k=t+1}^{T}(1-a\eta_k)\\ &= \sum_{t=t(\epsilon)+1}^{T}\left[\prod_{k=t+1}^{T}(1-a\eta_k) - \prod_{k=t}^{T}(1-a\eta_k)\right]\\ &= 1 - \prod_{k=t(\epsilon)+1}^{T}(1-a\eta_k) \leq 1.\end{aligned}$$

The first term on the right-hand side of (A) is bounded by

$$\begin{aligned}\sum_{t=t_1}^{t(\epsilon)}\eta_t^2\prod_{k=t+1}^{T}(1-a\eta_k) &\leq \sum_{t=t_1}^{t(\epsilon)}a^{-2}\prod_{k=t(\epsilon)+1}^{T}(1-a\eta_k) \leq a^{-2}\sum_{t=t_1}^{t(\epsilon)}\prod_{k=t(\epsilon)+1}^{T}\exp(-a\eta_k)\\ &= a^{-2}\sum_{t=t_1}^{t(\epsilon)}\exp\left(-a\sum_{k=t(\epsilon)+1}^{T}\eta_k\right) \leq a^{-2}\sum_{t=t_1}^{t(\epsilon)}\exp\left(-\log\frac{2t(\epsilon)}{a^2\epsilon}\right)\\ &\leq a^{-2}t(\epsilon)\frac{a^2\epsilon}{2t(\epsilon)} = \frac{\epsilon}{2}.\end{aligned}$$

Plugging the above inequality back into (A) then yields $\sum_{t=t_1}^{T}\eta_t^2\prod_{k=t+1}^{T}(1-a\eta_k) \leq \epsilon$ for any $T \geq t'(\epsilon)$. The proof is complete since ϵ is arbitrarily chosen.

Acknowledgments. We would like to thank the referees for their encouraging comments and constructive suggestions. We are grateful to Prof. Jian-Feng Cai for helpful discussions on the experiments. This paper was written when the corresponding author, Ding-Xuan Zhou, visited Shanghai Jiaotong University, whose hospitality is greatly appreciated.

REFERENCES

[1] H. H. Bauschke and J. M. Borwein, *On projection algorithms for solving convex feasibility problems*, SIAM Rev., 38 (1996), pp. 367–426.

[2] H. H. Bauschke, J. M. Borwein, and P. L. Combettes, *Bregman monotone optimization algorithms*, SIAM J. Control Optim., 42 (2003), pp. 596–636.

[3] H. H. Bauschke, J. M. Borwein, and W. Li, *Strong conical hull intersection property, bounded linear regularity, jamesons property (g), and error bounds in convex optimization*, Math. Program., 86 (1999), pp. 135–160.

[4] D. P. Bertsekas, *Incremental Gradient, Subgradient, and Proximal Methods for Convex Optimization: A Survey*, Report LIDS-P-2848, Laboratory for Information and Decision Systems, MIT, Cambridge, MA, 2010.

[5] L. Bottou, *On-line learning and stochastic approximations*, in On-line Learning in Neural Networks, D. Saad, ed., Cambridge University Press, New York, NY, 1998, pp. 9–42.

[6] J.-F. Cai, S. Osher, and Z. Shen, *Convergence of the linearized Bregman iteration for ℓ_1-norm minimization*, Math. Comp., 78 (2009), pp. 2127–2136.

[7] J.-F. Cai, S. Osher, and Z. Shen, *Linearized Bregman iterations for compressed sensing*, Math. Comp., 78 (2009), pp. 1515–1536.

[8] F. Cucker and D.-X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*, Cambridge Monogr. Appl. Comput. Math. 24, Cambridge University Press, Cambridge, UK, 2007.

[9] J. L. Doob, *Measure Theory*, Grad. Texts in Math. 143, Springer, New York, 1994.

[10] J. Duchi and Y. Singer, *Efficient online and batch learning using forward backward splitting*, J. Mach. Learn. Res., 10 (2009), pp. 2899–2934.

[11] J. Fan, T. Hu, Q. Wu, and D.-X. Zhou, *Consistency analysis of an empirical minimum error entropy algorithm*, Appl. Comput. Harmon. Anal., 41 (2016), pp. 164–189.

[12] T. Goldstein and S. Osher, *The split Bregman method for* l1*-regularized problems*, SIAM J. Imaging Sci., 2 (2009), pp. 323–343.

[13] Z.-C. Guo, D.-H. Xiang, X. Guo, and D.-X. Zhou, *Thresholded spectral algorithms for sparse approximations*, Anal. Appl., 15 (2017), pp. 433–455.

[14] P. C. Hansen and M. Saxild-Hansen, *AIR Tools: A MATLAB package of algebraic iterative reconstruction methods*, J. Comput. Appl. Math., 236 (2012), pp. 2167–2178.

[15] T. Hu, J. Fan, Q. Wu, and D.-X. Zhou, *Regularization schemes for minimum error entropy principle*, Anal. Appl., 13 (2015), pp. 437–455.

[16] S. Kaczmarz, *Angenäherte auflösung von systemen linearer gleichungen*, Bull. Int. Acad. Sci. Pologne A, 35 (1937), pp. 355–357.

[17] Y. Lei and D.-X. Zhou, *Analysis of online composite mirror descent algorithm*, Neural Comput., 29 (2017), pp. 825–860.

[18] Y. Lei and D.-X. Zhou, *Convergence of online mirror descent algorithms*, submitted.

[19] J. Lin and D.-X. Zhou, *Learning theory of randomized Kaczmarz algorithm*, J. Mach. Learn. Res., 16 (2015), pp. 3341–3365.

[20] D. A. Lorenz, F. Schöpfer, and S. Wenger, *The linearized Bregman method via split feasibility problems: Analysis and generalizations*, SIAM J. Imaging Sci., 7 (2014), pp. 1237–1262.

[21] D. A. Lorenz, S. Wenger, F. Schöpfer, and M. Magnor, *A sparse Kaczmarz solver and a linearized Bregman method for online compressed sensing*, in Proceedings of the IEEE International Conference on Image Processing, 2014, pp. 1347–1351.

[22] A.-S. Nemirovsky and D.-B. Yudin, *Problem Complexity and Method Efficiency in Optimization*, John Wiley & Sons, New York, 1983.

[23] F. Schöpfer and D. A. Lorenz, *Linear convergence of the randomized sparse Kaczmarz method*, Math. Program., (2018), doi:10.1007/s10107-017-1229-1.

[24] L. Shi, Y.-L. Feng, and D.-X. Zhou, *Concentration estimates for learning with ℓ_1-regularizer and data dependent hypothesis spaces*, Appl. Comput. Harmon. Anal., 31 (2011), pp. 286–302.

[25] S. Smale and D.-X. Zhou, *Online learning with markov sampling*, Anal. Appl., 7 (2009), pp. 87–113.

[26] T. Strohmer and R. Vershynin, *A randomized Kaczmarz algorithm with exponential convergence*, J. Fourier Anal. Appl., 15 (2009), pp. 262–278.

[27] P. Tarres and Y. Yao, *Online learning as stochastic approximation of regularization paths: Optimality and almost-sure convergence*, IEEE Trans. Inform. Theory, 60 (2014), pp. 5716–5735.
[28] W. Yin, *Analysis and generalizations of the linearized Bregman method*, SIAM J. Imaging Sci., 3 (2010), pp. 856–877.
[29] W. Yin, S. Osher, D. Goldfarb, and J. Darbon, *Bregman iterative algorithms for ℓ_1-minimization with applications to compressed sensing*, SIAM J. Imaging Sci., 1 (2008), pp. 143–168.
[30] Y. Ying and D.-X. Zhou, *Online regularized classification algorithms*, IEEE Trans. Inform. Theory, 52 (2006), pp. 4775–4788.
[31] X. Zhang, M. Burger, and S. Osher, *A unified primal-dual algorithm framework based on Bregman iteration*, J. Sci. Comput., 46 (2011), pp. 20–46.
[32] A. Zouzias and N. M. Freris, *Randomized extended Kaczmarz for solving least squares*, SIAM J. Matrix Anal. Appl., 34 (2013), pp. 773–793.