Convergence of Online Mirror Descent

Yunwen Lei, Ding-Xuan Zhou*

Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong

Abstract

In this paper we consider online mirror descent (OMD), a class of scalable online learning algorithms exploiting data geometric structures through mirror maps. Necessary and sufficient conditions are presented in terms of the step size sequence $\{\eta_t\}_t$ for the convergence of OMD with respect to the expected Bregman distance induced by the mirror map. The condition is $\lim_{t\to\infty}\eta_t = 0, \sum_{t=1}^{\infty}\eta_t = \infty$ in the case of positive variances. It is reduced to $\sum_{t=1}^{\infty}\eta_t = \infty$ in the case of zero variance for which linear convergence may be achieved by taking a constant step size sequence. A sufficient condition on the almost sure convergence is also given. We establish tight error bounds under mild conditions on the mirror map, the loss function, and the regularizer. Our results are achieved by some novel analysis on the one-step progress of OMD using smoothness and strong convexity of the mirror map and the loss function.

Keywords: Mirror descent, Online learning, Bregman distance, Convergence analysis, Learning theory

1. Introduction

Analyzing and processing big data in various applications has raised the need of scalable learning algorithms using geometric structures of data. One approach for scalability in learning theory is stochastic gradient descent and online learning. In this paper we are interested in online mirror descent, a class

*Corresponding author

Email addresses: yunwelei@cityu.edu.hk (Yunwen Lei), mazhou@cityu.edu.hk (Ding-Xuan Zhou)

Preprint submitted to Applied and Computational Harmonic Analysis *April 30, 2018*

of scalable learning algorithms exploiting possible data geometric structures such as sparsity.

Mirror descent is a powerful extension of the classical gradient descent [3] by relaxing the Hilbert space structure and using a mirror map $\Psi : \mathcal{W} \to \mathbb{R}$ to capture geometric properties of data from a Banach space $\mathcal{W}$. In this paper we consider $\mathcal{W} = \mathbb{R}^d$ endowed with a norm $\|\cdot\|$ which might be a non-Euclidean norm, allowing us to capture non-Euclidean geometric structures of data from $\mathbb{R}^d$. To introduce the mirror descent and online mirror descent, we assume that the mirror map Ψ is Fréchet differentiable and strongly convex. The Fréchet differentiability means the existence of a bounded linear operator $\nabla\Psi(w) : \mathcal{W} \to \mathbb{R}$ at every $w \in \mathcal{W}$ satisfying $\Psi(w+x) - \Psi(w) - \nabla\Psi(w)x = o(\|x\|)$. The strong convexity of Ψ means the existence of some $\sigma_\Psi > 0$ such that

$$D_\Psi(\tilde{w}, w) := \Psi(\tilde{w}) - \Psi(w) - \langle \tilde{w} - w, \nabla\Psi(w)\rangle \geq \frac{\sigma_\Psi}{2}\|\tilde{w} - w\|^2, \quad \forall \tilde{w}, w \in \mathcal{W},$$

where $\langle \tilde{w} - w, \nabla\Psi(w)\rangle$ is the linear operator $\nabla\Psi(w)$ acting on $\tilde{w} - w \in \mathcal{W}$. With this number σ_Ψ, we say Ψ is σ_Ψ-strongly convex (with respect to the norm $\|\cdot\|$), which we assume throughout the paper. The quantity $D_\Psi(\tilde{w}, w)$ is called the Bregman distance between $\tilde{w}$ and w.

Given a differentiable and convex objective function $F : \mathcal{W} \to \mathbb{R}$, a mirror descent algorithm approximates a minimizer of F by a sequence $\{w_t\}_{t\in\mathbb{N}} \subset \mathcal{W}$ defined with an initial vector $w_1 \in \mathcal{W}$ and the gradient descent method in terms of the gradient ∇F of F as

$$\nabla\Psi(w_{t+1}) = \nabla\Psi(w_t) - \eta_t \nabla F(w_t), \qquad t \in \mathbb{N}, \tag{1.1}$$

where $\{\eta_t\}_t$ is a sequence of positive numbers called the step size sequence. Here the gradient descent is performed in the dual $(\mathcal{W}^* = \mathbb{R}^d, \|\cdot\|_*)$ of the primal space $(\mathcal{W}, \|\cdot\|)$ since the map $\nabla\Psi : \mathcal{W} \to \mathcal{W}^*$ is well-defined, and invertible due to the strong convexity of Ψ. Useful instantiations [11] of the mirror map Ψ include the choice of p**-norm divergence** $\Psi = \Psi_p$ with $1 < p \leq 2$ defined by $\Psi_p(w) = \frac{1}{2}\|w\|_p^2$ where $\|\cdot\|_p$ is the p-norm defined by $\|w\|_p = \left(\sum_{i=1}^d |w(i)|^p\right)^{1/p}$ for $w = (w(1), \ldots, w(d)) \in \mathbb{R}^d$. The mirror descent algorithm with $\Psi = \Psi_2$

recovers the gradient descent.

In machine learning, the objective function F is often the regularized risk $F(w) = \mathbb{E}_Z[f(w, Z)]$ of the linear function $x \to \langle w, x\rangle$ induced by the action of $x \in \mathcal{W}^*$ on $w \in \mathcal{W}$, where $f(w, Z) = \phi(\langle w, X\rangle, Y) + r(w)$ is the regularized loss function induced by a loss function $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ and a convex regularizer $r : \mathcal{W} \to \mathbb{R}_+$, and $\mathbb{E}_Z$ denotes the expectation with respect to the random sample $Z = (X, Y)$ drawn from a Borel probability measure ρ on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$ with an input space $\mathcal{X} \subset \mathcal{W}^*$ and an output space $\mathcal{Y} \subset \mathbb{R}$. In the remainder of this paper, we focus on F of the form $F(w) = \mathbb{E}_Z[f(w, Z)]$ with f given in terms of ϕ and r.

In many machine learning applications, training examples $\{z_t = (x_t, y_t) \in \mathcal{Z}\}_t$ become available in a sequential manner. In such situations, instead of computing $F(w)$, we use the sample z_t at the t-th iteration of the mirror descent to compute the gradient $\nabla_w[f(w_t, z_t)]$ of $f(w, z_t)$ with respect to the variable w at w_t. This leads to the **online mirror descent** (OMD) which extends the classical online gradient descent algorithm by replacing Ψ_2 with a mirror map Ψ to capture data geometric structures beyond Hilbert spaces. It generates a sequence $\{w_t\}_t \subset \mathcal{W}$ with an initial vector $w_1 \in \mathcal{W}$ by performing the stochastic mirror descent in the dual space as

$$\nabla\Psi(w_{t+1}) = \nabla\Psi(w_t) - \eta_t \nabla_w[f(w_t, z_t)], \qquad t \in \mathbb{N}. \tag{1.2}$$

We always assume that the loss function ϕ is convex and differentiable with respect to the first variable (with the partial derivative ϕ'). When $\Psi = \Psi_2$ and $r(w) = \lambda\|w\|_2^2$ with $\lambda \geq 0$, the OMD (1.2) becomes the classical online learning algorithm with the iteration $w_{t+1} = w_t - \eta_t[\phi'(\langle w_t, x_t\rangle, y_t)x_t + 2\lambda w_t]$ generated by the stochastic gradient descent method in the Hilbert space $\mathcal{W}^* = \mathcal{W}$. The special choice $\phi(a, y) = \frac{1}{2}(a-y)^2$ of the unregularized least squares loss function with $r = 0$ corresponds to the general randomized Kaczmarz algorithm [9] given by

$$w_{t+1} = w_t - \eta_t[\langle w_t, x_t\rangle - y_t]x_t, \qquad t \in \mathbb{N}. \tag{1.3}$$

It was shown in [22] that when $\inf_{w\in\mathcal{W}} \mathbb{E}_Z\left[(Y - \langle w, X\rangle)^2\right] > 0$, the randomized

Kaczmarz algorithm (1.3) converges in expectation if and only if $\lim_{t\to\infty}\eta_t = 0$ and $\sum_{t=1}^{\infty}\eta_t = \infty$.

This paper presents **necessary and sufficient conditions** for the convergence of the OMD (1.2) with respect to the **Bregman distance** D_Ψ. It extends the results in [22, 29] from Ψ_2 to a general mirror map Ψ beyond the Hilbert space framework. Our conditions are stated in terms of the step size sequence $\{\eta_t\}_t$, under some mild assumptions on the mirror map Ψ, the regularized loss function f, and the probability measure ρ. Throughout the paper, we assume that the training examples $\{z_t\}_t$ are sampled independently from the probability measure ρ on $\mathcal{Z}$.

We illustrate our main results to be stated in the next section by presenting an example corresponding to the special choice of the unregularized least squares loss and a strongly smooth mirror map or the p-norm divergence Ψ_p (which, as shown in Proposition 7, is not strongly smooth). Here we say that Ψ is L_Ψ-strongly smooth (with respect to the norm $\|\cdot\|$) with $L_\Psi > 0$ if $D_\Psi(\tilde{w}, w) \leq \frac{L_\Psi}{2}\|\tilde{w} - w\|^2$ for any $w, \tilde{w} \in \mathcal{W}$. Examples of strongly smooth mirror maps include Ψ_2 and a mirror map $\Psi^{(\epsilon,\lambda)}$ with parameters $\epsilon > 0, \lambda > 0$ defined in the literature of compressed sensing [7] as $\Psi^{(\epsilon,\lambda)}(w) = \lambda\sum_{i=1}^{d} g_\epsilon(w(i)) + \frac{1}{2}\|w\|_2^2$, where $g_\epsilon(\xi) = \frac{\xi^2}{2\epsilon}$ for $|\xi| \leq \epsilon$ and $|\xi| - \frac{\epsilon}{2}$ for $|\xi| > \epsilon$. The mirror map Ψ_p plays an important role in the mirror descent method and it can be applied to capturing geometric structures of data for learning problems in huge dimensions. For example, the specific choice with $p = 1 + \frac{1}{\log d}$ gives convergence bounds with only a logarithmic dependence on the dimension d, see [11]. The mirror map Ψ_p is strongly convex with $\sigma_{\Psi_p} = p - 1$ when the norm of $\mathcal{W}$ takes the p-norm $\|\cdot\| = \|\cdot\|_p$ (see [2]), and by the norm equivalence, $\sigma_{\Psi_p} > 0$ for other norms.

With the special choice of the unregularized least squares loss $f(w, z) = \frac{1}{2}(\langle w, x\rangle - y)^2$, the OMD (1.2) takes a special form

$$\nabla\Psi(w_{t+1}) = \nabla\Psi(w_t) - \eta_t[\langle w_t, x_t\rangle - y_t]x_t, \qquad t \in \mathbb{N}. \tag{1.4}$$

The following result for this example will be proved in Section 6. Denote by $X^\top$ the transpose of $X \in \mathcal{W}^*$.

Theorem 1. *Assume* $\sup_{x\in\mathcal{X}}\|x\|_* < \infty$, $\mathbb{E}_Z[Y^2] < \infty$, *and that the covariance matrix* $\mathcal{C}_X = \mathbb{E}_Z[XX^\top]$ *is positive definite. Consider the OMD (1.4) and denote* $w_\rho = \mathcal{C}_X^{-1}\mathbb{E}_Z[XY]$. *Let* Ψ *be either some p-norm divergence* $\Psi = \Psi_p$ *with* $1 < p \le 2$ *or a strongly smooth mirror map.*

(a) Assume $\inf_{w\in\mathcal{W}}\mathbb{E}_Z\left[|Y - \langle w, X\rangle|\,\|X\|_*\right] > 0$. *Then* $\lim_{t\to\infty}\mathbb{E}_{z_1,\ldots,z_{t-1}}[\|w_\rho - w_t\|^2] = 0$ *if and only if*

$$\lim_{t\to\infty}\eta_t = 0 \quad \textit{and} \quad \sum_{t=1}^{\infty}\eta_t = \infty. \tag{1.5}$$

Furthermore, if Ψ *is strongly smooth and* $\lim_{t\to\infty}\eta_t = 0$, *then there exist some* $\tilde{T}_1 \in \mathbb{N}$ *and* $\tilde{C} > 0$ *such that* $\mathbb{E}_{z_1,\ldots,z_{T-1}}[\|w_\rho - w_T\|^2] \ge \tilde{C}T^{-1}$ *for* $T \ge \tilde{T}_1$. *If we take* $\eta_t = \frac{4}{(t+1)\sigma}$ *for some appropriate* $\sigma > 0$ *(given in the proof), then* $\mathbb{E}_{z_1,\ldots,z_{T-1}}[\|w_\rho - w_T\|^2] = O\left(T^{-1}\right)$.

(b) Assume $w_\rho \ne w_1, \mathbb{E}_Z\left[|Y - \langle w_\rho, X\rangle|\,\|X\|_*\right] = 0$ *and for some* $\kappa > 0$, $\eta_t \le \frac{\sigma_\Psi}{(2+\kappa)R^2}$. *Then* $\lim_{t\to\infty}\mathbb{E}_{z_1,\ldots,z_{t-1}}[\|w_\rho - w_t\|^2] = 0$ *if and only if* $\sum_{t=1}^{\infty}\eta_t = \infty$. *Furthermore, if* Ψ *is strongly smooth and* $\eta_t \equiv \eta_1 < \frac{\sigma_\Psi}{2R^2}$, *then there exist* $\tilde{c}_1, \tilde{c}_2 \in (0,1)$ *such that*

$$\left(\tilde{c}_1\right)^T\|w_\rho - w_1\|^2 \le \mathbb{E}_{z_1,\ldots,z_{T-1}}[\|w_\rho - w_T\|^2] \le \left(\tilde{c}_2\right)^T\|w_\rho - w_1\|^2, \qquad \forall T\in\mathbb{N}. \tag{1.6}$$

(c) If the step size sequence satisfies

$$\sum_{t=1}^{\infty}\eta_t = \infty \quad \textit{and} \quad \sum_{t=1}^{\infty}\eta_t^2 < \infty, \tag{1.7}$$

then $\{\|w_\rho - w_t\|^2\}_{t\in\mathbb{N}}$ *converges to* 0 *almost surely.*

Part (b) of Theorem 1 is for the case of zero variance with $y = \langle w_\rho, x\rangle$ almost surely, meaning that the sampling process has no noise and the target function (conditional mean) is linear. It asserts that the OMD with a strongly smooth mirror map and a constant step size sequence may converge linearly in this case. Part (a) asserts that for the case of positive variances (either the sampling process has noise or the target function is nonlinear) the OMD with a strongly smooth mirror map can converge of at most order $O(\frac{1}{T})$ and this order may be achieved. This solves a conjecture raised in [22, page 3346] that a convergence rate of order $O(T^{-\theta})$ with $1 < \theta \leq 2$ is impossible for the randomized Kaczmarz algorithm (with $\Psi = \Psi_2$) in the noisy case. Theorem 1 also characterizes the convergence in expectation by means of the step size condition $\sum_{t=1}^{\infty} \eta_t = \infty$ for the case of zero variance and the condition $\lim_{t\to\infty} \eta_t = 0$ and $\sum_{t=1}^{\infty} \eta_t = \infty$ for the case of positive variances.

Our analysis is based on a key identity on measuring the one-step progress of OMD by excess Bregman distances, from which lower and upper bounds on the one-step progress are established by using strong smoothness and convexity of the associated regularized loss functions as well as properties of the mirror map. These lower and upper bounds are then used to build necessary and sufficient conditions, as well as tight convergence rates.

This paper is organized as follows. In Section 2 we introduce some mild assumptions on the mirror map and the regularized risk. General results on convergence of the OMD for the cases with positive variances and zero variance are stated in subsection 2.1, and then exemplified with specific mirror maps and loss functions in subsections 2.2 and 2.3. We give some discussion and comparison with related work in subsection 2.4. In Section 3, we present a key identity on the one-step progress of the OMD and sketch the basic idea of our analysis. We prove the convergence results in the case of positive variances in Section 4, and results in the case of zero variance together with the almost sure convergence in Section 5. In Section 6, we prove the explicit results stated in Section 1, subsection 2.2 and subsection 2.3. Some simulations are given in Section 7 to validate our theoretical results.

2. Main Results

In this section we state our main results on necessary and sufficient conditions for the convergence of OMD (1.2) to a minimizer $w^* = \arg\min_{w\in\mathcal{W}} F(w)$ of the regularized risk F which is assumed to exist throughout the paper.

Our discussion requires some mild assumptions on the mirror map Ψ and the regularized risk F. On the mirror map, for necessary conditions, we shall assume that $\nabla\Psi$ is continuous at w^* and satisfies the following incremental condition at infinity.

Definition 1. We say that $\nabla\Psi$ satisfies an incremental condition (of order 1) at infinity if there exists a constant $C_\Psi > 0$ such that

$$\|\nabla\Psi(w)\|_* \leq C_\Psi(1 + \|w\|), \qquad \forall w \in \mathcal{W}. \tag{2.1}$$

We shall show later that the p-norm divergence Ψ_p with $1 < p \leq 2$ and strongly smooth mirror maps satisfy this mild condition.

For the pair (Ψ, F), we shall also assume the following condition measuring how the convexity of Ψ is controlled by that of F around w^* with a convex function Ω. Recall that w^* is a minimizer of F on $\mathcal{W}$.

Definition 2. We say that the convexity of Ψ is controlled by that of F around w^* with a convex function $\Omega : [0, \infty) \to \mathbb{R}_+$ satisfying $\Omega(0) = 0$ and $\Omega(u) > 0$ for $u > 0$ if the pair (Ψ, F) satisfies

$$\langle w^* - w, \nabla F(w^*) - \nabla F(w)\rangle \geq \Omega\left(D_\Psi(w^*, w)\right), \qquad \forall w \in \mathcal{W}. \tag{2.2}$$

Typical choices of the convex function Ω include $\Omega(u) = Cu^\alpha$ with $\alpha \geq 1$ and $C > 0$. In particular, when F is strongly convex and Ψ is strongly smooth, condition (2.2) is satisfied with a linear (convex) function $\Omega(u) = Cu$ for some $C > 0$. To see this, we notice from the definition of the Bregman distance that for a Fréchet differentiable and convex function $g : \mathbb{R}^d \to \mathbb{R}$, there holds

$$D_g(w, \tilde{w}) + D_g(\tilde{w}, w) = \langle w - \tilde{w}, \nabla g(w) - \nabla g(\tilde{w})\rangle, \qquad \forall w, \tilde{w} \in \mathcal{W}. \tag{2.3}$$

So when F is σ_F-strongly convex with $\sigma_F > 0$, we have $\langle w^* - w, \nabla F(w^*) - \nabla F(w)\rangle \geq \sigma_F\|w^* - w\|^2$. It follows that (2.2) with $\Omega(u) = \frac{2\sigma_F}{L_\Psi}u$ is satisfied when Ψ is L_Ψ-strongly smooth.

2.1. Statements of general results

Our first main result, Theorem 2, states a necessary and sufficient condition for the convergence of the OMD for the case of positive variances meaning that $\inf_{w \in \mathcal{W}} \mathbb{E}_Z [\|\nabla_w [f(w, Z)]\|_*] > 0$. It also states in Parts (a) and (b) respectively that in this case, the OMD cannot achieve convergence rates faster than $O(T^{-1})$ after T iterates, while the best rate $O(T^{-1})$ may be achieved when $\Omega(u) = Cu$ in (2.2). This theorem is a consequence of Propositions 11 and 13 to be presented in Section 4.

Theorem 2. *Assume* $\inf_{w \in \mathcal{W}} \mathbb{E}_Z [\|\nabla_w [f(w, Z)]\|_*] > 0$ *and that for some constant* $L > 0$, $f(\cdot, z)$ *is* L*-strongly smooth for almost every* $z \in Z$. *Suppose that* $\nabla \Psi$ *is continuous at* w^* *and satisfies the incremental condition (2.1) at infinity, and that the pair* (Ψ, F) *satisfies (2.2) around* w^* *with a convex function* $\Omega : [0, \infty) \to \mathbb{R}_+$ *satisfying* $\Omega(0) = 0$ *and* $\Omega(u) > 0$ *for* $u > 0$. *Then for OMD* (1.2), $\lim_{t \to \infty} \mathbb{E}_{z_1, \ldots, z_{t-1}} [D_\Psi(w^*, w_t)] = 0$ *if and only if the step size sequence satisfies* (1.5).

(a) If Ψ *is strongly smooth and* $\lim_{t \to \infty} \eta_t = 0$, *then there exist some constants* $t_0 \in \mathbb{N}$ *and* $\tilde{C} > 0$ *such that*

$$\mathbb{E}_{z_1, \ldots, z_{T-1}} [D_\Psi(w^*, w_T)] \geq \frac{\tilde{C}}{T - t_0 + 1}, \qquad \forall T \geq t_0. \tag{2.4}$$

(b) If there exists an $\sigma_F > 0$ *such that*

$$\langle w^* - w, \nabla F(w^*) - \nabla F(w) \rangle \geq \sigma_F D_\Psi(w^*, w), \qquad \forall w \in \mathcal{W}. \tag{2.5}$$

and the step size sequence takes the form $\eta_t = \frac{4}{(t+1)\sigma_F}$, *then*

$$\mathbb{E}_{z_1, \ldots, z_{T-1}} [D_\Psi(w^*, w_T)] = O\left(\frac{1}{T}\right). \tag{2.6}$$

We shall see from the proof of Proposition 11 given in Section 4 that the continuity of $\nabla \Psi$ at w^* and the incremental condition (2.1) are only required for proving $\lim_{t \to \infty} \eta_t = 0$ of the necessity, they are not required for the sufficiency or for proving $\sum_{t \to \infty} \eta_t = \infty$ of the necessity. These conditions are satisfied when Ψ is strongly smooth, as shown in Proposition 5 below.

Our second main result, Theorem 3 to be proved in Section 5, states a necessary and sufficient condition for the convergence of the OMD for the case of zero variance in the sense that $\mathbb{E}_Z\left[\|\nabla_w[f(w^*,Z)]\|_*\right]=0$.

Theorem 3. *Assume $\mathbb{E}_Z\left[\|\nabla_w[f(w^*,Z)]\|_*\right]=0$ and that for some constant $L>0$, $f(\cdot,z)$ is L-strongly smooth for almost every $z\in Z$. Suppose that the pair (Ψ,F) satisfies (2.2) around w^* with a convex function $\Omega:[0,\infty)\to\mathbb{R}_+$ satisfying $\Omega(0)=0$ and $\Omega(u)>0$ for $u>0$. Assume also $w_1\neq w^*$ and that for some $\kappa>0$, $\eta_t\leq\frac{\sigma_\Psi}{(2+\kappa)L}$ for every $t\in\mathbb{N}$. Then $\lim_{t\to\infty}\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*,w_t)]=0$ if and only if $\sum_{t=1}^\infty\eta_t=\infty$. Furthermore, if (2.5) holds and $\eta_t\equiv\eta_1<\frac{\sigma_\Psi}{2L}$, then*

$$D_\Psi(w^*,w_1)\left(1-\frac{2L\eta_1}{\sigma_\Psi}\right)^T\leq\mathbb{E}_{z_1,\ldots,z_{T-1}}[D_\Psi(w^*,w_T)]\leq D_\Psi(w^*,w_1)\left(1-\frac{\sigma_F\eta_1}{2}\right)^T. \tag{2.7}$$

Remark 1. Our results in Theorems 2 and 3 can be extended to the minibatch setting where a batch of examples $\{z_{t,1},\ldots,z_{t,m}\}$ are independently drawn from the probability measure ρ at the t-th iteration. The associated OMD then takes the form

$$\nabla\Psi(w_{t+1})=\nabla\Psi(w_t)-\frac{\eta_t}{m}\sum_{i=1}^m\nabla_w\big[f(w_t,z_{t,i})\big],\quad\forall t\in\mathbb{N}.$$

In this setting, the variance of the stochastic gradients will decrease by a factor of m. The necessary and sufficient conditions in Theorem 2 and Theorem 3 also apply. For the case with positive variances, the right-hand side of both (2.4) and (2.6) are required to be divided by m due to the variance reduction effect. For the case with zero-variances, the inequality (2.7) remains the same since the stochastic gradient at w^* does not change in the mini-batch setting.

Remark 2. The variance condition $\inf_{w\in\mathcal{W}}\mathbb{E}_Z\left[\|\nabla_w[f(w,Z)]\|_*\right]>0$ is almost complementary to the variance condition $\mathbb{E}_Z\left[\|\nabla_w[f(w^*,Z)]\|_*\right]=0$. Indeed, if $\inf_{w\in\mathcal{W}}\mathbb{E}_Z\left[\|\nabla_w[f(w,Z)]\|_*\right]=0$ and we assume the infimum can be achieved at a point $\bar{w}\in\mathcal{W}$, meaning that $\mathbb{E}_Z\left[\|\nabla_w[f(\bar{w},Z)]\|_*\right]=0$. Then we have $\nabla_w[f(\bar{w},z)]=0$ almost surely and therefore $\bar{w}$ is a minimizer of F. To see clearly these variance conditions, suppose the data are drawn according to the equation

$y_t = \langle w^*, x_t \rangle + \epsilon$ with $w^* \in \mathcal{W}$ and ϵ following the normal distribution $N(0, \sigma^2)$. Consider the loss function $f(w, z) = \frac{1}{2}\big(\langle w, x \rangle - y\big)^2$. We assume $\mathbb{E}_X[\|X\|_*] > 0$. It is clear that $\mathbb{E}_Z[XX^\top w^* - XY] = 0$ and therefore $w^* = \arg\min_{w \in \mathcal{W}} F(w)$. If $\sigma = 0$, then it is clear that

$$\mathbb{E}_Z[\|\nabla_w[f(w^*, Z)]\|_*] = \mathbb{E}_Z[|\langle w^*, X \rangle - Y|\|X\|_*] = 0,$$

which corresponds to the case with zero variance. On the other hand, if $\sigma > 0$, then for any $w \in \mathcal{W}$ and $x \in \mathcal{X}$ we have

$$\begin{aligned}
\mathbb{E}_{Y|X=x}\big[\|\nabla_w[f(w, Z)]\|_*\big] &= \|x\|_* \mathbb{E}_{Y|X=x}\big[|\langle w, X \rangle - Y|\big] \\
&= \|x\|_* \mathbb{E}_{Y|X=x}\big[|\langle w - w^*, X \rangle - \epsilon|\big] \\
&\geq \sigma \|x\|_* \Pr\big\{\big|\langle w - w^*, X \rangle - \epsilon\big| \geq \sigma \big| X = x\big\} \\
&= \sigma \|x\|_* \Big[1 - \Pr\big\{\big|\langle w - w^*, X \rangle - \epsilon\big| \leq \sigma \big| X = x\big\}\Big] \\
&\geq \sigma \|x\|_* \big[1 - \sqrt{2/\pi}\big],
\end{aligned}$$

where the first inequality is due to the Markov's inequality and the last inequality is due to following inequality (the density function of the normal distribution $N(0, \sigma^2)$ takes values in the interval $[0, \frac{1}{\sqrt{2\pi}\sigma}]$)

$$\Pr\big\{|\epsilon - a| \leq \sigma\big\} \leq \sqrt{2/\pi}, \quad \forall a \in \mathbb{R}.$$

It then follows that

$$\mathbb{E}_Z\big[\|\nabla_w[f(w, Z)]\|_*\big] \geq \sigma\big[1 - \sqrt{2/\pi}\big]\mathbb{E}_X[\|X\|_*] > 0, \quad \forall w \in \mathcal{W}.$$

That is, the case $\sigma > 0$ corresponds to exactly the case with positive variances.

Our last main result, Theorem 4 to be proved in Section 5, provides a sufficient condition for the almost sure convergence of the OMD by imposing a stronger condition with $\sum_{t=1}^\infty \eta_t^2 < \infty$.

Theorem 4. *Assume that for some constant $L > 0$, $f(\cdot, z)$ is L-strongly smooth for almost every $z \in Z$. Suppose that the pair (Ψ, F) satisfies (2.2) around w^* with a convex function $\Omega : [0, \infty) \to \mathbb{R}_+$ satisfying $\Omega(0) = 0$ and $\Omega(u) > 0$ for $u > 0$. If the step size sequence satisfies the condition* (1.7)*, then we have* $\lim_{t \to \infty} D_\Psi(w^*, w_t) = 0$ *almost surely.*

2.2. Results with strongly smooth mirror maps and p-norm divergence

In this subsection, for two classes of mirror maps Ψ and strongly convex objective functions F, we state some results to be proved in Section 6 on the continuity of $\nabla\Psi$ at w^* and the incremental condition (2.1) at infinity for $\nabla\Psi$, and the convexity condition (2.2) of (Ψ, F).

The first class of mirror maps are strongly smooth ones.

Proposition 5. *If Ψ is strongly smooth, then $\nabla\Psi$ is continuous everywhere and satisfies the incremental condition* (2.1) *at infinity. Furthermore, if F is strongly convex,* (2.2) *is satisfied for a linear convex function $\Omega(u) = C_{\Psi,L}u$ with some $C_{\Psi,L} > 0$.*

The second class of mirror maps are the p-norm divergence $\Psi = \Psi_p$ with $1 < p \leq 2$. For the case $p = 2$, we have $\nabla\Psi_2(w) = w$, $D_{\Psi_2}(\tilde{w}, w) = \frac{1}{2}\|w - \tilde{w}\|_2^2$ for $w, \tilde{w} \in \mathcal{W}$ and Ψ_2 is strongly smooth. So Proposition 5 applies.

Proposition 6. *Consider the p-norm divergence $\Psi = \Psi_p$ with $1 < p < 2$. Then $\nabla\Psi_p$ is continuous everywhere and satisfies the incremental condition* (2.1) *with $C_{\Psi_p} = 1$. Moreover, we have*

$$\|\nabla\Psi_p(w)\|_* = \|w\|_p, \quad \forall w \in \mathcal{W} \tag{2.8}$$

and for any $\tilde{w}, w \in \mathcal{W}$, there holds

$$D_{\Psi_p}(\tilde{w}, w) \leq \left((2\|\tilde{w}\|_p)^{2-p} + \|\tilde{w}\|_p^{p-1} + 1\right)\left(\|\tilde{w} - w\|_p^2 + \|\tilde{w} - w\|_p^{\min\{p,3-p\}}\right). \tag{2.9}$$

Denote $\tau_p = \frac{2}{\min\{p,3-p\}} \in (1, 2]$. For any $\tilde{w} \in \mathcal{W}$, we have

$$\|\tilde{w} - w\|_p^2 \geq B_p\Omega_p\left(D_{\Psi_p}(\tilde{w}, w)\right), \qquad \forall w \in \mathcal{W}, \tag{2.10}$$

where $\Omega_p : [0, \infty) \to [0, \infty)$ is the convex function depending on p defined by

$$\Omega_p(u) = \begin{cases} u + \frac{1}{\tau_p} - 1, & \text{if } u \geq 1, \\ \frac{1}{\tau_p}u^{\tau_p}, & \text{if } 0 \leq u < 1, \end{cases} \tag{2.11}$$

and B_p is the constant depending on $\|\tilde{w}\|_p$ and p given by

$$B_p = \quad \min\left\{ \left(2\left(2\|\tilde{w}\|_p\right)^{2-p} + 2\left\|\tilde{w}\right\|_p^{p-1} + 2\right)^{-1}, \right.$$
$$\left. \left(2\left(2\|\tilde{w}\|_p\right)^{2-p} + 2\left\|\tilde{w}\right\|_p^{p-1} + 2\right)^{-\tau_p} \right\}.$$

If F is σ_F-strongly convex with respect to the norm $\|\cdot\|_p$, then the pair (Ψ_p, F) satisfies (2.2) *around w^* with the convex function $\Omega : \mathbb{R}_+ \to \mathbb{R}_+$ given by*

$$\Omega(u) = \sigma_F B_p \Omega_p(u), \qquad u \in [0, \infty).$$

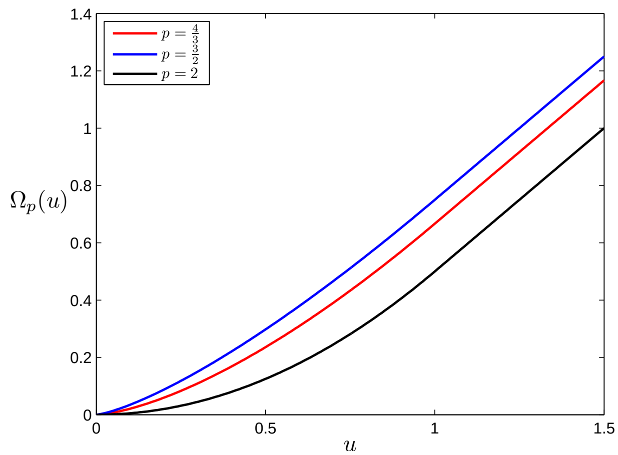


Figure 1: Plots of the convex function Ω_p with $p = \frac{4}{3}$ (red line), $p = \frac{3}{2}$ (blue line) and $p = 2$ (black line).

We remark that the convex function Ω_2 defined by (2.11) with $p = 2$ is a Huber loss [17]. Figure 1 gives the plots of the function Ω_p with $p = \frac{4}{3}, p = \frac{3}{2}$ and $p = 2$.

Following Proposition 6, a natural question to ask is whether the p-norm divergence is strongly smooth (that is, whether (2.10) holds with $\Omega_p(u) = Cu$ for some $C > 0$). When $d = 1$, $\Psi_p(w) = \frac{1}{2}w^2 = \Psi_2(w)$ is strongly smooth. When $d > 1$, the answer is negative, as shown in the following proposition to be proved in the appendix.

Proposition 7. *For $d > 1$, the p-norm divergence $\Psi = \Psi_p$ with $1 < p < 2$ is not strongly smooth.*

2.3. Explicit results with special loss functions for learning

In this subsection we state explicit results on the convergence of the OMD associated with the regularized loss function $f(w,z) = \phi(\langle w, x\rangle, y) + \lambda \|w\|_2^2$ with $\lambda > 0$ and the norm $\|\cdot\| = \|\cdot\|_2$ when the loss function ϕ has a Lipschitz continuous derivative. Common examples of such loss functions [17, 8, 30] include the least squares loss $\phi(a,y) = \frac{1}{2}(a-y)^2$, the logistic loss $\phi(a,y) = \log(1+\exp(-ay))$ or $\phi(a,y) = 1/(1+e^{ay})$, the 2-norm hinge loss $\phi(a,y) = (\max\{0, 1-ay\})^2$, and the Huber loss Ω_2 defined by (2.11) with $p = 2$.

The following explicit result will be proved in Section 6.

Theorem 8. *Assume* $\sup_{x\in\mathcal{X}} \|x\|_* < \infty$, $\|\cdot\| = \|\cdot\|_2$, *and the derivative* ϕ' *of the convex loss function* $\phi : \mathbb{R} \times \mathbb{R} \to \mathbb{R}_+$ *satisfies the Lipschitz condition*

$$\ell_\phi := \sup_{u \neq v \in \mathbb{R}, y \in \mathcal{Y}} \frac{|\phi'(u,y) - \phi'(v,y)|}{|u-v|} < \infty. \tag{2.12}$$

Then the regularized loss function $f(w,z) = \phi(\langle w,x\rangle, y) + \lambda\|w\|_2^2$ *with some* $\lambda > 0$ *is* $2(\ell_\phi R^2 + \lambda)$*-strongly smooth for every* $z \in \mathcal{Z}$*. The objective function* F *is also* $2(\ell_\phi R^2 + \lambda)$*-strongly smooth, and is* 2λ*-strongly convex. The conclusion of Theorem 1 with* w_ρ *replaced by* w^* *holds for the OMD* (1.2) *with* Ψ *being either some* p*-norm divergence* $\Psi = \Psi_p$ *with* $1 < p \leq 2$ *or a strongly smooth mirror map.*

2.4. Comparison and discussion

In the special Hilbert space setting with $\Psi = \Psi_2$, there is a large learning theory literature on the convergence of stochastic gradient descent (SGD) or online gradient descent (OGD). We first review some related work on *conditions for the convergence in expectation.* Convergence of SGD/OGD in reproducing kernel Hilbert spaces (RKHSs) was discussed in [28, 32] for regression and [33, 34] for classification. Under uniform boundedness assumptions of $\{w_t\}_t$, it was shown in [33] that a sufficient condition for the convergence of regularized SGD/OGD in expectation is the step size condition (1.5). Such a result was recently established for online regularized pairwise learning in [14]. For unregularized SGD/OGD applied to non-strongly convex and strongly smooth objective functions, it was shown in [34] that $\lim_{T\to\infty} \mathbb{E}_{z_1,\ldots,z_{T-1}}[F(w_T)] = F(w^*)$ if the step size satisfies the condition (1.7). All the above mentioned discussions on SGD/OGD considered sufficient conditions for the convergence in expectation. As a comparison, we give necessary and sufficient conditions for the convergence of a more general OMD in the strongly convex setting. We then review some related work on *convergence rates in expectation* in the strongly convex setting. Under boundedness assumptions $\mathbb{E}_Z[\|\nabla_w[f(w_t, Z)]\|_2^2] \le B$ for a constant $B > 0$, it was shown in [19, 26] that the T-th iterate of SGD/OGD satisfies $\mathbb{E}_{z_1,\ldots,z_{T-1}}[\|w_T - w^*\|_2^2] = O(1/T)$. This convergence rate was also derived in [6] under a relaxed assumption on gradients as $\mathbb{E}_Z[\|\nabla_w[f(w_t, Z)]\|_2^2] \le A + B\|\nabla F(w_t)\|_2^2$. As a comparison, we show that the same convergence rate can be achieved for the general OMD without any boundedness assumptions on gradients. Furthermore, we show this convergence rate is tight by presenting a matching lower bound up to a constant factor, which has not been established in the literature to our best knowledge. It should be mentioned that lower bounds for minimax errors were discussed for stochastic convex optimization [1], which consider the error rates of any stochastic convex optimization methods in the *worst case*. We now review some related work on the *almost sure convergence*. For SGD/OGD, under the assumption that the objective function F with a single minimizer w^* satisfies

$$\inf_{\|w-w^*\|_2^2 > \epsilon} \langle w - w^*, \nabla F(w)\rangle > 0, \quad \forall \epsilon > 0$$

and

$$\mathbb{E}_Z[\|\nabla f(w, Z)\|_*^2] \le A + B\|w - w^*\|_2^2, \quad \forall w \in \mathcal{W}$$

for some constants $A, B \ge 0$, it was shown [5] that $\{w_t\}_t$ converges to w^* almost surely if the step sizes satisfy (1.7). For regularized OGD in RKHSs associated with the specific least squares loss function, it was shown in [31] that $\{w_t\}_t$ converges to w^* almost surely for polynomially decaying step sizes $\eta_t = \eta_1 t^{-\theta}$

with $\theta \in (0,1)$. We extend these results on the almost sure convergence to the OMD.

We remark that the SGD has also been well studied in the literature of optimization (see, e.g., [27, 24]) under some conditions on the noise sequence instead of conditions on the step size sequence. For the randomized Kaczmarz algorithm (1.3), the convergence in expectation has been studied in the literature of non-uniform sampling and compressed sensing, including the characterization of the convergence [22] by (1.5) in the noisy case with $\inf_{w\in\mathcal{W}} \mathbb{E}_Z[(\langle w, X\rangle - Y)^2] > 0$, and the linear convergence [29] with a constant step size sequence in the noiseless case with $y = \langle w^*, x\rangle$ almost surely. Our work on the convergence of the OMD (1.2) with a general mirror map Ψ is motivated by these results on the randomized Kaczmarz algorithm (1.3) with the special mirror map Ψ_2.

For the OMD (1.2) with a general mirror map Ψ, the only existing work to our best knowledge is some regret bounds in [11] and some convergence rates in [25]. In this paper we characterize the convergence in expectation by the step size condition (1.5) in the noisy case and by $\sum_{t=1}^{\infty} \eta_t = \infty$ in the noiseless case, derive the linear convergence with a constant step size sequence in the noiseless case, and verify the almost sure convergence by the step size condition (1.7). The main difficulty with the general mirror map Ψ is the lack of analysis for the one-step progress $\|w_{t+1} - w^*\|_2^2 - \|w_t - w^*\|_2^2$ which was carried out in [22] by exploiting the Hilbert space structure and the special linearity caused by the least squares loss function. To overcome this difficulty due to the Banach space structure and the nonlinearity, we use the Bregman distance D_Ψ induced by the mirror map Ψ, which has been used in our recent work [20]. Our novelty here is a key identity (3.1) measuring the one-step progress of the OMD with the general mirror map Ψ. Our analysis is then conducted by extensively using properties of the Bregman distance, the smoothness and convexity of regularized loss functions, and the convexity condition (2.2) involving a related convex function Ω.

Our contribution of this paper includes not only the novel convergence analysis for the OMD (1.2) with a general mirror map Ψ, but also some improvements of our earlier work [22] on the randomized Kaczmarz algorithm (1.3) with the special mirror map Ψ_2. In particular, we confirm a conjecture raised in [22] on high order convergence rates for the randomized Kaczmarz algorithm. Furthermore, the analysis in [22] was carried out under the restriction $0 < \eta_t < 2$ on the step size sequence which is removed here. It would be interesting to get explicit convergence rates when the mirror map is Ψ_p, and to extend our analysis to other learning frameworks [12, 16, 23, 13].

3. A Key Identity and Idea of Analysis

Our analysis for the convergence of the OMD (1.2) will be carried out based on the following key identity which measures the one-step progress of the algorithm in terms of the excess Bregman distance $D_\Psi(w^*, w_{t+1}) - D_\Psi(w^*, w_t)$.

Lemma 9. *The following identity holds for $t \in \mathbb{N}$*

$$\mathbb{E}_{z_t}[D_\Psi(w^*, w_{t+1})] - D_\Psi(w^*, w_t) = \eta_t \langle w^* - w_t, \nabla F(w_t)\rangle + \mathbb{E}_{z_t}\big[D_\Psi(w_t, w_{t+1})\big]. \tag{3.1}$$

Proof. By the definition of the Bregman distance, we see the following identity

$$D_\Psi(w, v) + D_\Psi(v, u) - D_\Psi(w, u) = \langle w - v, \nabla\Psi(u) - \nabla\Psi(v)\rangle, \quad \forall u, v, w \in \mathcal{W}. \tag{3.2}$$

Choosing $v = w_{t+1}$ and $u = w_t$ yields

$$D_\Psi(w, w_{t+1}) - D_\Psi(w, w_t) = -D_\Psi(w_{t+1}, w_t) + \langle w - w_{t+1}, \nabla\Psi(w_t) - \nabla\Psi(w_{t+1})\rangle.$$

We now separate $w - w_{t+1}$ into $w - w_t$ and $w_t - w_{t+1}$, use the iteration relation (1.2) of the OMD and apply (2.3) with $g = \Psi$ to derive

$$\begin{aligned}
&D_\Psi(w, w_{t+1}) - D_\Psi(w, w_t) \\
&= -D_\Psi(w_{t+1}, w_t) + \langle w - w_t, \nabla\Psi(w_t) - \nabla\Psi(w_{t+1})\rangle + \langle w_t - w_{t+1}, \nabla\Psi(w_t) - \nabla\Psi(w_{t+1})\rangle \\
&= -D_\Psi(w_{t+1}, w_t) + \eta_t\langle w - w_t, \nabla_w[f(w_t, z_t)]\rangle + \langle w_t - w_{t+1}, \nabla\Psi(w_t) - \nabla\Psi(w_{t+1})\rangle \\
&= D_\Psi(w_t, w_{t+1}) + \eta_t\langle w - w_t, \nabla_w[f(w_t, z_t)]\rangle.
\end{aligned}$$

Taking expectations $\mathbb{E}_{z_t}$ on both sides, setting $w = w^*$ and noting that w_t is independent of z_t, we see the stated identity (3.1). The proof is complete. $\square$

The necessity of the convergence will be derived by using the strong smoothness of F and the strong convexity of Ψ to bound $\langle w_t - w^*, \nabla F(w_t)\rangle = \langle w_t - w^*, \nabla F(w_t) - \nabla F(w^*)\rangle$ by $O(1)D_\Psi(w^*, w_t)$, from which we can apply the identity (3.1) to get necessary conditions by the following inequality

$$\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi(w^*, w_{t+1})] \geq (1-O(\eta_t))\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)]+\mathbb{E}_{z_1,\ldots,z_t}\big[D_\Psi(w_t, w_{t+1})\big].$$

The sufficiency will be derived by using the strong smoothness of f and the duality $D_\Psi(w_t, w_{t+1}) = D_{\Psi^*}(\nabla\Psi(w_{t+1}), \nabla\Psi(w_t))$ to bound $\mathbb{E}_{z_t}\big[D_\Psi(w_t, w_{t+1})\big]$ in terms of $\langle w^* - w_t, \nabla F(w^*) - \nabla F(w_t)\rangle$ and $\mathbb{E}_{z_t}[\|\nabla f(w^*, z_t)\|_*^2]$, from which we can apply the identity (3.1) again to get

$$\begin{aligned}\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi(w^*, w_{t+1})] &\leq \mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)] \\ &\quad - \frac{\eta_t}{2}\mathbb{E}_{z_1,\ldots,z_t}[\langle w^* - w_t, \nabla F(w^*) - \nabla F(w_t)\rangle] + O(\eta_t^2)\end{aligned}$$

and then use (2.2) for bounding $-\langle w^* - w_t, \nabla F(w^*) - \nabla F(w_t)\rangle$ by $-\Omega\left(D_\Psi(w^*, w_t)]\right)$ to obtain

$$\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi(w^*, w_{t+1})] \leq \mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)] - \frac{\eta_t}{2}\Omega\left(\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)]\right) + O(\eta_t^2).$$

Here for a continuous convex function $g : \mathbb{R}^d \to \mathbb{R}$, the Fenchel-conjugate g^* is defined by

$$g^*(v) = \sup_{w\in\mathcal{W}}[\langle w, v\rangle - g(w)], \qquad v \in \mathbb{R}^d$$

and the duality (3.3) on the Bregman distances is stated (see, e.g., [4]) in the following lemma together with the duality between strong convexity and strong smoothness [18].

Lemma 10. *Let $g : \mathbb{R}^d \to \mathbb{R}$ be continuous and convex. Let $\beta > 0$. Then g is β-strongly convex with respect to the norm $\|\cdot\|$ if and only if g^* is $\frac{1}{\beta}$-strongly smooth with respect to the dual norm $\|\cdot\|_*$.*

If g is Fréchet differentiable and strongly convex, then there holds

$$D_g(w, \tilde{w}) = D_{g^*}(\nabla g(\tilde{w}), \nabla g(w)), \qquad \forall w, \tilde{w} \in \mathcal{W}. \tag{3.3}$$

4. Convergence in the Case of Positive Variances

In this section we prove Theorem 2 by deriving the necessary and sufficient condition from two propositions given below.

4.1. Necessary condition for convergence

The first proposition gives the necessity for the convergence of the OMD (1.2).

Proposition 11. *Assume* $\inf_{w\in\mathcal{W}} \mathbb{E}_Z\left[\|\nabla_w[f(w,Z)]\|_*\right] > 0$ *and that* F *is strongly smooth. Assume also that* $\nabla\Psi$ *satisfies the incremental condition (2.1) at infinity. If* $\lim_{t\to\infty} \mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)] = 0$ *for some* w^* *where* $\nabla\Psi$ *is continuous, then the step size sequence satisfies* (1.5)*.*

Furthermore, if Ψ *is strongly smooth, then* (2.4) *holds with some constants* $t_0 \in \mathbb{N}$ *and* $\tilde{C} > 0$*.*

Proof. We first show $\lim_{t\to\infty} \eta_t = 0$.

By the σ_Ψ-strong convexity of Ψ, we have $\|w^* - w_t\|^2 \le \frac{2}{\sigma_\Psi} D_\Psi(w^*, w_t)$. So the condition $\lim_{t\to\infty} \mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)] = 0$ implies $\lim_{t\to\infty} \mathbb{E}_{z_1,\ldots,z_{t-1}}[\|w^* - w_t\|^2] = 0$. Then we claim that

$$\lim_{t\to\infty} \mathbb{E}_{z_1,\ldots,z_{t-1}}[\|\nabla\Psi(w_t) - \nabla\Psi(w^*)\|_*] = 0. \tag{4.1}$$

To prove our claim, we use the continuity of $\nabla\Psi$ at w^* and know that for any $\varepsilon > 0$, there exists some $0 < \delta \le 1$ such that $\|\nabla\Psi(w) - \nabla\Psi(w^*)\|_* < \varepsilon$ whenever $\|w - w^*\| < \delta$.

When $\|w - w^*\| \ge \delta$, we apply the incremental condition (2.1) and $\|w\| \le \|w - w^*\| + \|w^*\|$ to find

$$\|\nabla\Psi(w) - \nabla\Psi(w^*)\|_* \le C_\Psi(1 + \|w\|) + \|\nabla\Psi(w^*)\|_* \le C_{\Psi,w^*,\delta}\|w - w^*\|,$$

where $C_{\Psi,w^*,\delta}$ is the constant given by

$$C_{\Psi,w^*,\delta} = C_\Psi + \frac{C_\Psi + C_\Psi\|w^*\| + \|\nabla\Psi(w^*)\|_*}{\delta}.$$

Combining the above two cases, we know that

$$\mathbb{E}_{z_1,\ldots,z_{t-1}}[\|\nabla\Psi(w_t)-\nabla\Psi(w^*)\|_*]\le\varepsilon+C_{\Psi,w^*,\delta}\mathbb{E}_{z_1,\ldots,z_{t-1}}[\|w_t-w^*\|].$$

But $\lim_{t\to\infty}\mathbb{E}_{z_1,\ldots,z_{t-1}}[\|w^*-w_t\|^2]=0$ ensures the existence of some $t_{\varepsilon,\delta}\in\mathbb{N}$ such that for $t>t_{\varepsilon,\delta}$, there holds $\mathbb{E}_{z_1,\ldots,z_{t-1}}[\|w_t-w^*\|^2]<\frac{\varepsilon^2}{C^2_{\Psi,w^*,\delta}}$ which implies $\mathbb{E}_{z_1,\ldots,z_{t-1}}[\|w_t-w^*\|]<\frac{\varepsilon}{C_{\Psi,w^*,\delta}}$ by the Schwarz inequality. So we have $\mathbb{E}_{z_1,\ldots,z_{t-1}}[\|\nabla\Psi(w_t)-\nabla\Psi(w^*)\|_*]<2\varepsilon$ for $t>t_{\varepsilon,\delta}$, which verifies our claim (4.1).

Denote $\sigma=\inf_{w\in\mathcal{W}}\mathbb{E}_Z\left[\|\nabla_w[f(w,Z)]\|_*\right]>0$. From the iteration relation (1.2) of the OMD, we have $\eta_t\|\nabla_w[f(w_t,z_t)]\|_*=\|\nabla\Psi(w_t)-\nabla\Psi(w_{t+1})\|_*$. Taking expectations on both sides with respect to z_t yields

$$\eta_t\sigma\le\eta_t\mathbb{E}_{z_t}\left[\|\nabla_w[f(w_t,z_t)]\|_*\right]\le\|\nabla\Psi(w_t)-\nabla\Psi(w^*)\|_*+\mathbb{E}_{z_t}\left[\|\nabla\Psi(w_{t+1})-\nabla\Psi(w^*)\|_*\right]$$

and

$$\eta_t\sigma\le\mathbb{E}_{z_1,\ldots,z_{t-1}}[\|\nabla\Psi(w_t)-\nabla\Psi(w^*)\|_*]+\mathbb{E}_{z_1,\ldots,z_t}[\|\nabla\Psi(w_{t+1})-\nabla\Psi(w^*)\|_*].$$

Hence (4.1) confirms our first limit $\lim_{t\to\infty}\eta_t=0$.

We now show $\sum_{t=1}^{\infty}\eta_t=\infty$. Assume that F is L_F-strongly smooth for some $L_F>0$. From the identity (2.3) and the optimality condition $\nabla F(w^*)=0$, we have

$$D_F(w^*,w_t)+D_F(w_t,w^*)=-\langle w^*-w_t,\nabla F(w_t)\rangle.$$

This is bounded by $L_F\|w^*-w_t\|^2$ by the L_F-strong smoothness of F. But the σ_Ψ-strong convexity of Ψ implies $D_\Psi(w^*,w_t)\ge\frac{\sigma_\Psi}{2}\|w^*-w_t\|^2$. Hence

$$\langle w^*-w_t,\nabla F(w_t)\rangle\ge-L_F\|w^*-w_t\|^2\ge-\frac{2L_F}{\sigma_\Psi}D_\Psi(w^*,w_t).$$

Plugging this inequality into (3.1) and taking expectations on both sides give

$$\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi(w^*,w_{t+1})]\ge(1-a\eta_t)\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*,w_t)]+\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi(w_t,w_{t+1})],\tag{4.2}$$

where a is the constant $a=2L_F\sigma_\Psi^{-1}$.

Since $\lim_{t\to\infty} \eta_t = 0$, we can find some integer $t_0 \in \mathbb{N}$ such that $\eta_t \leq (3a)^{-1}$ for $t \geq t_0$. Applying the elementary inequality $1 - \eta \geq \exp(-2\eta)$ valid for $\eta \in (0, 1/3]$, we know by noting $\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi(w_t, w_{t+1})] \geq 0$ in (4.2) that

$$\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi(w^*, w_{t+1})] \geq \exp(-2a\eta_t)\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)], \qquad \forall t \geq t_0. \tag{4.3}$$

Applying this inequality iteratively for $t = T, \ldots, t_0 + 1$ then yields

$$\begin{aligned}\mathbb{E}_{z_1,\ldots,z_T}[D_\Psi(w^*, w_{T+1})] &\geq \prod_{t=t_0+1}^{T} \exp(-2a\eta_t)\mathbb{E}_{z_1,\ldots,z_{t_0}}[D_\Psi(w^*, w_{t_0+1})] \\ &= \exp\Big(-2a\sum_{t=t_0+1}^{T} \eta_t\Big)\mathbb{E}_{z_1,\ldots,z_{t_0}}[D_\Psi(w^*, w_{t_0+1})].\end{aligned} \tag{4.4}$$

We claim that $\mathbb{E}_{z_1,\ldots,z_{t_0}}[D_\Psi(w^*, w_{t_0+1})] > 0$. Otherwise, we would have

$$\mathbb{E}_{z_1,\ldots,z_{t_0-1}}[D_\Psi(w^*, w_{t_0})] = \mathbb{E}_{z_1,\ldots,z_{t_0}}[D_\Psi(w^*, w_{t_0+1})] = 0$$

by (4.3), leading to $\mathbb{E}_{z_1,\ldots,z_{t_0-1}}[\|w^* - w_{t_0}\|^2] = \mathbb{E}_{z_1,\ldots,z_{t_0}}[\|w^* - w_{t_0+1}\|^2] = 0$ according to the strong convexity of Ψ. This would imply $w_{t_0+1} = w_{t_0} = w^*$ almost surely and thereby $\nabla_w[f(w^*, z_{t_0})] = 0$ almost surely by (1.2), leading to $\mathbb{E}_Z[\|\nabla_w[f(w^*, Z)]\|_*] = 0$, a contradiction to the assumption $\inf_{w\in\mathcal{W}} \mathbb{E}_Z[\|\nabla_w[f(w, Z)]\|_*] > 0$.

By $\mathbb{E}_{z_1,\ldots,z_{t_0}}[D_\Psi(w^*, w_{t_0+1})] > 0$ and $\lim_{T\to\infty} \mathbb{E}_{z_1,\ldots,z_T}[D_\Psi(w^*, w_{T+1})] = 0$, we see from (4.4) that $\sum_{t=1}^{\infty} \eta_t = \infty$. This proves the necessary condition for the convergence of the OMD.

We now prove (2.4) under the L_Ψ-strong smoothness of Ψ for some $L_\Psi > 0$. Since Ψ is σ_Ψ-strongly convex and L_Ψ-strongly smooth with respect to $\|\cdot\|$, we know from Lemma 10 that Ψ^* is σ_Ψ^{-1}-strongly smooth and L_Ψ^{-1}-strongly convex with respect to $\|\cdot\|_*$ (note $\Psi^{**} = \Psi$ since Ψ is convex and differentiable). We also know from Lemma 10 that the duality relation (3.3) between Bregman distances holds for $g = \Psi$, which yields

$$D_\Psi(w_t, w_{t+1}) = D_{\Psi^*}(\nabla\Psi(w_{t+1}), \nabla\Psi(w_t)), \quad \forall t \in \mathbb{N}.$$

Combining this with the L_Ψ^{-1}-strong convexity of Ψ^* and (4.2), we know from the bound $\eta_t \leq (3a)^{-1}$ that for $t \geq t_0$,

$$\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi(w^*, w_{t+1})] \geq (1 - a\eta_t)\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)] \\ + (2L_\Psi)^{-1}\mathbb{E}_{z_1,\ldots,z_t}\big[\|\nabla\Psi(w_t) - \nabla\Psi(w_{t+1})\|_*^2\big].$$

But $\nabla\Psi(w_t) - \nabla\Psi(w_{t+1}) = \eta_t \nabla_w[f(w_t, z_t)]$ by the definition (1.2) of the OMD. So for $t \geq t_0$ we have

$$\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi(w^*, w_{t+1})] \geq (1 - a\eta_t)\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)] \\ + (2L_\Psi)^{-1}\eta_t^2\mathbb{E}_{z_1,\ldots,z_t}\big[\|\nabla_w[f(w_t, z_t)]\|_*^2\big].$$

By the Schwarz inequality,

$$\mathbb{E}_{z_1,\ldots,z_t}\big[\|\nabla_w[f(w_t, z_t)]\|_*\big] \leq \big\{\mathbb{E}_{z_1,\ldots,z_t}\big[\|\nabla_w[f(w_t, z_t)]\|_*^2\big]\big\}^{1/2}.$$

Hence

$$\mathbb{E}_{z_1,\ldots,z_t}\big[\|\nabla_w[f(w_t, z_t)]\|_*^2\big] \geq \big\{\mathbb{E}_{z_1,\ldots,z_t}\big[\|\nabla_w[f(w_t, z_t)]\|_*\big]\big\}^2 \geq \sigma^2$$

and thereby

$$\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi(w^*, w_{t+1})] \geq (1 - a\eta_t)\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)] + (2L_\Psi)^{-1}\eta_t^2\sigma^2, \quad \forall t \geq t_0.$$

Applying this inequality iteratively from $t = T \geq t_0$ to $t = t_0$ yields (denote $\prod_{k=T+1}^{T}(1 - a\eta_k) = 1$)

$$\begin{aligned} &\mathbb{E}_{z_1,\ldots,z_T}[D_\Psi(w^*, w_{T+1})] \\ &\geq \mathbb{E}_{z_1,\ldots,z_{t_0-1}}[D_\Psi(w^*, w_{t_0})]\prod_{t=t_0}^{T}(1 - a\eta_t) + (2L_\Psi)^{-1}\sigma^2\sum_{t=t_0}^{T}\eta_t^2\prod_{k=t+1}^{T}(1 - a\eta_k) \\ &\geq (2L_\Psi)^{-1}\sigma^2\sum_{t=t_0}^{T}\eta_t^2\prod_{k=t+1}^{T}(1 - a\eta_k). \end{aligned}$$

By the Schwarz inequality and the bound $0 < 1 - a\eta_k \leq 1$ for $k \geq t_0$, we have

$$\sum_{t=t_0}^{T}\eta_t\prod_{k=t+1}^{T}(1 - a\eta_k) \leq \left\{\sum_{t=t_0}^{T}\eta_t^2\prod_{k=t+1}^{T}(1 - a\eta_k)\right\}^{1/2}(T - t_0 + 1)^{1/2}.$$

Hence

$$\begin{aligned}
\sum_{t=t_0}^{T} \eta_t^2 \prod_{k=t+1}^{T} (1 - a\eta_k) &\geq \frac{1}{a^2(T - t_0 + 1)} \left(\sum_{t=t_0}^{T} a\eta_t \prod_{k=t+1}^{T} (1 - a\eta_k) \right)^2 \\
&= \frac{1}{a^2(T - t_0 + 1)} \left(\sum_{t=t_0}^{T} \big(1 - (1 - a\eta_t)\big) \prod_{k=t+1}^{T} (1 - a\eta_k) \right)^2 \\
&= \frac{1}{a^2(T - t_0 + 1)} \left(\sum_{t=t_0}^{T} \left[\prod_{k=t+1}^{T} (1 - a\eta_k) - \prod_{k=t}^{T} (1 - a\eta_k) \right] \right)^2 \\
&= \frac{1}{a^2(T - t_0 + 1)} \left(1 - \prod_{k=t_0}^{T} (1 - a\eta_k) \right)^2 \\
&\geq \frac{1}{a^2(T - t_0 + 1)} \left(1 - (1 - a\eta_{t_0})\right)^2 = \frac{\eta_{t_0}^2}{T - t_0 + 1}.
\end{aligned}$$

Therefore,

$$\mathbb{E}_{z_1,\ldots,z_T}[D_\Psi(w^*, w_{T+1})] \geq \frac{\eta_{t_0}^2 (2L_\Psi)^{-1} \sigma^2}{T - t_0 + 1}, \qquad \forall T \geq t_0.$$

This verifies (2.4) with $\tilde{C} = \eta_{t_0}^2 (2L_\Psi)^{-1} \sigma^2$ and completes the proof. □

4.2. Sufficient condition for convergence

We now turn to the second proposition giving the sufficiency for the convergence of the OMD (1.2). We need the following lemma, to be proved in appendix by some ideas from [34], which establishes the co-coercivity of gradients for convex functions enjoying some smoothness condition.

Lemma 12. *Let $\alpha \in (0, 1]$ and $g : \mathcal{W} \to \mathbb{R}$ be a Fréchet differentiable and convex function. If there exists some constant $L > 0$ such that*

$$D_g(w, \tilde{w}) \leq \frac{L}{1 + \alpha} \|w - \tilde{w}\|^{1+\alpha}, \quad \forall w, \tilde{w} \in \mathcal{W},$$

then we have

$$\frac{2L^{-\frac{1}{\alpha}} \alpha}{1 + \alpha} \|\nabla g(w) - \nabla g(\tilde{w})\|_*^{\frac{1+\alpha}{\alpha}} \leq \langle w - \tilde{w}, \nabla g(w) - \nabla g(\tilde{w}) \rangle, \qquad \forall w, \tilde{w} \in \mathcal{W}. \quad (4.5)$$

Proposition 13. *Assume that for some constant $L > 0$, $f(\cdot, z)$ is L-strongly smooth for almost every $z \in Z$. Suppose that the pair (Ψ, F) satisfies (2.2)* *around w^* with a convex function $\Omega : [0,\infty) \to \mathbb{R}_+$ satisfying $\Omega(0) = 0$ and $\Omega(u) > 0$ for $u > 0$. If the step size sequence satisfies* (1.5)*, then we have* $\lim_{t\to\infty} \mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)] = 0$.

Furthermore, if (2.5) *holds with some $\sigma_F > 0$ and the step size takes the form $\eta_t = \frac{4}{(t+1)\sigma_F}$, then* (2.6) *holds.*

Proof. According to the key identity (3.1) for the one-step progress of the OMD and the duality relation (3.3) of the Bregman distances, we have

$$\begin{aligned}
&\mathbb{E}_{z_t}[D_\Psi(w^*, w_{t+1})] - D_\Psi(w^*, w_t) \\
= \;& \eta_t\langle w^* - w_t, \nabla F(w_t)\rangle + \mathbb{E}_{z_t}\big[D_{\Psi^*}(\nabla\Psi(w_{t+1}), \nabla\Psi(w_t))\big]. \qquad (4.6)
\end{aligned}$$

By Lemma 10, the σ_Ψ-strong convexity of Ψ implies the σ_Ψ^{-1}-strong smoothness of Ψ^*. It follows from the definition (1.2) of the OMD that

$$\begin{aligned}
\mathbb{E}_{z_t}\big[D_{\Psi^*}(\nabla\Psi(w_{t+1}), \nabla\Psi(w_t))\big] &\le \frac{1}{2\sigma_\Psi}\mathbb{E}_{z_t}\big[\|\nabla\Psi(w_{t+1}) - \nabla\Psi(w_t)\|_*^2\big] \\
&= \frac{\eta_t^2}{2\sigma_\Psi}\mathbb{E}_{z_t}\big[\|\nabla_w[f(w_t, z_t)]\|_*^2\big]. \qquad (4.7)
\end{aligned}$$

We bound $\big[\|\nabla_w[f(w_t, z_t)]\|_*^2\big]$ by $2\big[\|\nabla_w[f(w_t, z_t)] - \nabla_w[f(w^*, z_t)]\|_*^2\big] + 2\big[\|\nabla_w[f(w^*, z_t)]\|_*^2\big]$. Then we apply Lemma 12 with $w = w^*, \tilde{w} = w_t, g = f(\cdot, z_t)$ and $\alpha = 1$. By the L-strong smoothness of $f(\cdot, z)$, we know that

$$\begin{aligned}
&\mathbb{E}_{z_t}\Big[\|\nabla_w[f(w_t, z_t)] - \nabla_w[f(w^*, z_t)]\|_*^2\Big] \\
\le \;& L\mathbb{E}_{z_t}\Big[\Big\langle w_t - w^*, \nabla_w[f(w_t, z_t)] - \nabla_w[f(w^*, z_t)]\Big\rangle\Big] \\
= \;& L\langle w^* - w_t, \nabla F(w^*) - \nabla F(w_t)\rangle, \qquad (4.8)
\end{aligned}$$

where the interchange of the expectation and the gradient is valid due to the strong smoothness. Then we have

$$\begin{aligned}
&\mathbb{E}_{z_t}[D_\Psi(w^*, w_{t+1})] - D_\Psi(w^*, w_t) \le \\
&-\left(1 - \frac{L\eta_t}{\sigma_\Psi}\right)\eta_t\langle w^* - w_t, \nabla F(w^*) - \nabla F(w_t)\rangle + \frac{\eta_t^2}{\sigma_\Psi}\mathbb{E}_{z_t}\big[\|\nabla_w[f(w^*, z_t)]\|_*^2\big].
\end{aligned}$$

Since $\lim_{t\to\infty}\eta_t = 0$, there exists some $t_1 \in \mathbb{N}$ such that $\frac{L}{\sigma_\Psi}\eta_t \le \frac{1}{2}$ for $t \ge t_1$

which implies

$$\mathbb{E}_{z_t}[D_\Psi(w^*, w_{t+1})] - D_\Psi(w^*, w_t) \leq$$
$$- \frac{\eta_t}{2}\langle w^* - w_t, \nabla F(w^*) - \nabla F(w_t)\rangle + \frac{\eta_t^2}{\sigma_\Psi}\mathbb{E}_{z_t}\left[\|\nabla_w[f(w^*, z_t)]\|_*^2\right]. \quad (4.9)$$

Now we apply the relation (2.2) on the convexity to obtain

$$-\langle w^* - w_t, \nabla F(w^*) - \nabla F(w_t)\rangle \leq -\Omega\left(D_\Psi(w^*, w_t)\right). \quad (4.10)$$

It follows that

$$\mathbb{E}_{z_t}[D_\Psi(w^*, w_{t+1})] \leq D_\Psi(w^*, w_t) - \frac{\eta_t}{2}\Omega\left(D_\Psi(w^*, w_t)\right) + b\eta_t^2, \quad (4.11)$$

where b is the constant $b = \frac{1}{\sigma_\Psi}\mathbb{E}_Z\left[\|\nabla_w[f(w^*, Z)]\|_*^2\right]$. Since Ω is convex, by Jensen's inequality, we have

$$\Omega\left(\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)]\right) \leq \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[\Omega\left(D_\Psi(w^*, w_t)\right)\right].$$

Therefore, by taking expectations over $z_1, \ldots, z_{t-1}$ and denoting a sequence $\{A_t\}_t$ by

$$A_t = \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[D_\Psi(w^*, w_t)\right],$$

we have

$$A_{t+1} \leq A_t - \frac{\eta_t}{2}\Omega\left(A_t\right) + b\eta_t^2, \qquad \forall t \geq t_1. \quad (4.12)$$

To prove $\lim_{t\to\infty} A_t = 0$, we let $0 < \gamma < 1$ be an arbitrarily chosen number. The convexity of $\Omega : [0, \infty) \to \mathbb{R}_+$ tells us that for $u \geq \gamma$, there holds

$$\Omega(\gamma) = \Omega\left((1 - \frac{\gamma}{u}) \cdot 0 + \frac{\gamma}{u}u\right) \leq (1 - \frac{\gamma}{u})\Omega\left(0\right) + \frac{\gamma}{u}\Omega(u) = \frac{\gamma}{u}\Omega(u)$$

which yields

$$\Omega(u) \geq \frac{\Omega(\gamma)}{\gamma}u, \qquad \forall u \geq \gamma. \quad (4.13)$$

Since $\lim_{t\to\infty} \eta_t = 0$, we know that there exists some integer $t_\gamma \geq t_1$ such that

$$\eta_t \leq \min\left\{\frac{\Omega(\gamma)}{4b}, \frac{\Omega(\gamma)}{4\gamma b}, \sqrt{\gamma}\right\}, \qquad \forall t \geq t_\gamma. \quad (4.14)$$

We claim that

$$\sup\left\{t \in \mathbb{N} : A_t \leq \gamma\right\} = \infty. \quad (4.15)$$

If (4.15) is not true, we can find some $t'_\gamma \geq t_\gamma$ such that

$$A_t > \gamma, \qquad \forall t \geq t'_\gamma.$$

Combining this with (4.13), (4.14) and (4.12) tells us that for $t \geq t'_\gamma$,

$$A_{t+1} \leq A_t - \eta_t \frac{\Omega(\gamma)}{2\gamma} A_t + b\eta_t^2 \leq A_t - \frac{\Omega(\gamma)}{2\gamma}\eta_t A_t + \frac{\Omega(\gamma)}{4\gamma}\eta_t A_t = A_t - \frac{\Omega(\gamma)}{4\gamma}\eta_t A_t \leq A_t - \frac{\Omega(\gamma)}{4}\eta_t,$$

which implies by iteration

$$A_{t+1} \leq A_{t'_\gamma} - \frac{\Omega(\gamma)}{4} \sum_{k=t'_\gamma}^{t} \eta_k \to -\infty \text{ (as } t \to \infty).$$

This is a contradiction, which verifies our claim (4.15).

By (4.15) there exists some positive integer $t''_\gamma > t_\gamma$ such that $A_{t''_\gamma} \leq \gamma$. We now show by induction that

$$A_t \leq \gamma + b \max_{t''_\gamma \leq \ell \leq t-1} \eta_\ell^2, \qquad \forall t \geq t''_\gamma. \tag{4.16}$$

The case $t = t''_\gamma$ is true (where we denote $\max_{t''_\gamma \leq \ell \leq t''_\gamma - 1} \eta_\ell^2 = 0$) since $A_{t''_\gamma} \leq \gamma$. Supposes the statement (4.16) holds for $t = k \geq t''_\gamma$. Note that $t''_\gamma > t_\gamma$ and $\gamma < 1$. To prove the statement for $t = k+1$, we discuss in two cases. If $A_k \leq \gamma$, we see directly from (4.12) that

$$A_{k+1} \leq \gamma + b\eta_k^2 \leq \gamma + b \max_{t''_\gamma \leq \ell \leq k} \eta_\ell^2.$$

If $A_k > \gamma$, we apply (4.13), (4.14) and (4.12) again and find

$$A_{k+1} \leq A_k - \eta_k \frac{\Omega(\gamma)}{2\gamma} A_k + b\eta_k^2 \leq A_k - \frac{\Omega(\gamma)}{4\gamma}\eta_k A_k \leq A_k \leq \gamma + b \max_{t''_\gamma \leq \ell \leq k-1} \eta_\ell^2,$$

where we have used the induction hypothesis in the last inequality. This verifies the statement (4.16) for $t = k+1$ and completes the induction procedure.

Applying (4.14), (4.16) and noting $t''_\gamma > t_\gamma$, we know that

$$A_t \leq (1+b)\gamma, \qquad \forall t \geq t''_\gamma.$$

Since γ is an arbitrary number on $(0,1)$, this proves

$$\lim_{t\to\infty} A_t = \lim_{t\to\infty} \mathbb{E}_{z_1,\ldots,z_{t-1}} \left[D_\Psi(w^*, w_t)\right] = 0.$$

We now prove (2.6) under condition (2.5) and the choice $\eta_t = \frac{4}{(t+1)\sigma_F}$ of the step size sequence. Eq. (2.5) implies that (2.2) holds with $\Omega(u) = \sigma_F u$. The estimate (4.12) then becomes

$$A_{t+1} \le A_t - \frac{2}{t+1} A_t + \frac{16b}{(t+1)^2 \sigma_F^2}, \qquad \forall t \ge t_1.$$

Multiplying both sides by $t(t+1)$ gives

$$t(t+1)A_{t+1} \le (t-1)tA_t + \frac{16b}{\sigma_F^2}, \qquad \forall t \ge t_1.$$

Applying this relation iteratively, we obtain

$$(T-1)TA_T \le (t_1-1)t_1 A_{t_1} + \frac{16b(T-t_1)}{\sigma_F^2}, \qquad \forall T \ge t_1,$$

from which we see

$$\mathbb{E}_{z_1,\ldots,z_{T-1}}[D_\Psi(w^*, w_T)] \le \frac{(t_1-1)t_1 \mathbb{E}_{z_1,\ldots,z_{t_1-1}}[D_\Psi(w^*, w_{t_1})]}{(T-1)T} + \frac{16b}{T\sigma_F^2}, \qquad \forall T \ge t_1.$$

This yields (2.6). The proof is complete. □

Remark 3. Equation (2.6) gives convergence rates for $\mathbb{E}_{z_1,\ldots,z_{T-1}}[D_\Psi(w^*, w_T)]$ under an assumption on the strong convexity of F measured by the Bregman distance. It should be noticed that $D_\Psi(w^*, w_T)$ provides different geometric distance measures between w^* and w_T for different mirror maps. For example, if $\Psi = \Psi_p$, then Equation (2.6) together with the $(p-1)$-strong convexity of Ψ_p w.r.t. $\|\cdot\|_p$ implies the rate $\mathbb{E}_{z_1,\ldots,z_{T-1}}[\|w_T - w^*\|_p^2] = O(1/T)$ for the $\|\cdot\|_p$ convergence. The case $p = 2$ corresponds to the Euclidean distance while the case $1 < p < 2$ corresponds to a distance in a Banach space. Furthermore, if w^* is sparse and admits small $\|w^*\|_1$, then we can choose p to be close to 1 to make sure w_T also attains a small ℓ_1-norm: $\mathbb{E}_{z_1,\ldots,z_{T-1}}[\|w_T\|_1] \le \mathbb{E}_{z_1,\ldots,z_{T-1}}[\|w_T - w^*\|_1] + \|w^*\|_1$. In this case, w_T also enjoys some sparsity.

Let us clarify the role of the mirror map in the case when (2.2) around w^* is not imposed for the pair (Ψ, F). Take $w_1 = 0$ and $\eta_t \le \sigma_\Psi/(2L)$ for all $t \in \mathbb{N}$ (in this case t_1 for (4.9) can be taken as 1). Since the derivation of (4.9) does not depend on (2.2), we use the convexity of F and $\nabla F(w^*) = 0$ in (4.9) to derive

$$\mathbb{E}_{z_t}[D_\Psi(w^*, w_{t+1})] - D_\Psi(w^*, w_t) \le \frac{\eta_t\big[F(w^*) - F(w_t)\big]}{2} + \frac{\mathbb{E}_Z\big[\|\nabla_w[f(w^*, Z)]\|_*^2\big]\eta_t^2}{\sigma_\Psi}.$$

Taking a summation from $t = 1$ to T, we derive

$$\begin{aligned}\mathbb{E}_{z_1,\ldots,z_T}[D_\Psi(w^*, w_{T+1})] - D_\Psi(w^*, w_1) &\le \frac{1}{2}\sum_{t=1}^{T}\eta_t\big[F(w^*) - F(w_t)\big] \\ &\quad + \frac{\mathbb{E}_Z\big[\|\nabla_w[f(w^*, Z)]\|_*^2\big]\sum_{t=1}^T \eta_t^2}{\sigma_\Psi}.\end{aligned}$$

According to the convexity of F, it further follows that

$$F(\bar{w}_T) - F(w^*) \le \frac{2D_\Psi(w^*, w_1)}{\sum_{t=1}^T \eta_t} + \frac{2\big[\mathbb{E}_Z\|\nabla_w[f(w^*, Z)]\|_*^2\big]\sum_{t=1}^T \eta_t^2}{\sigma_\Psi\sum_{t=1}^T \eta_t},$$

where $\bar{w}_T = \frac{\sum_{t=1}^T \eta_t w_t}{\sum_{t=1}^T \eta_t}$ is a weighted average of the first T iterates. If we consider the mirror map $\Psi = \Psi_p$ and $\eta_t = \eta_1 t^{-\frac{1}{2}}$ with $\eta_1 = \sigma_\Psi/(2L)$, then from $w_1 = 0$ we get

$$\begin{aligned}F(\bar{w}_T) - F(w^*) &\le \frac{\|w^*\|_p^2}{\eta_1\sum_{t=1}^T t^{-\frac{1}{2}}} + \frac{2\eta_1\mathbb{E}_Z\big[\|\nabla_w[f(w^*, Z)]\|_*^2\big]\sum_{t=1}^T t^{-1}}{\sigma_\Psi\sum_{t=1}^T t^{-\frac{1}{2}}} \\ &= O\Big(\frac{\|w^*\|_p^2}{(p-1)\sqrt{T}} + \frac{\mathbb{E}_Z\big[\|\nabla_w[f(w^*, Z)]\|_*^2\big]\log T}{\sqrt{T}}\Big),\end{aligned}$$

where we have used the $(p-1)$-strong convexity of Ψ_p w.r.t. $\|\cdot\|_p$. If we choose $p = 1 + \frac{1}{\log d}$, then it follows from $\|\nabla_w[f(w^*, Z)]\|_* = \|\nabla_w[f(w^*, Z)]\|_{1+\log d} \le e\|\nabla_w[f(w^*, Z)]\|_\infty$ that

$$F(\bar{w}_T) - F(w^*) = O\Big(\frac{\|w^*\|_1^2\log d + \mathbb{E}_Z\big[\|\nabla_w[f(w^*, Z)]\|_\infty^2\big]\log T}{\sqrt{T}}\Big). \tag{4.17}$$

As a comparison, if we choose $p = 2$, the expression takes the form

$$F(\bar{w}_T) - F(w^*) = O\Big(\frac{\|w^*\|_2^2 + \mathbb{E}_Z\big[\|\nabla_w[f(w^*, Z)]\|_2^2\big]\log T}{\sqrt{T}}\Big). \tag{4.18}$$

The bound in (4.17) would be significantly smaller than that in (4.18) in the case when w^* is sparse and $\|\nabla_w[f(w^*, z)]\|_2$ is close to $\sqrt{d}\|\nabla_w[f(w^*, z)]\|_\infty$ (meaning $\nabla_w[f(w^*, z)]$ is dense). In this case, the bound (4.17) enjoys a logarithmic dependency on the dimension [11], while the bound (4.18) enjoys a square-root dependency. It should be noticed that the discussion in [11] requires a nontrivial assumption $\|\nabla_w[f(w^*,z)]\|_* \le G$ with a constant $G>0$, which is removed in this remark.

Remark 4. Some of our results can be extended to projected OMD applied to non-differentiable objective functions. For any convex function $g:\mathbb{R}^d\to\mathbb{R}$, we use $g'(w)$ to denote a subgradient of g at w satisfying $g(\tilde{w}) \ge g(w) + \langle \tilde{w} - w, g'(w)\rangle$ for all $\tilde{w}$. We assume that there exist A and $B>0$ such that

$$\|f'(w,z)]\|_*^2 \le Af(w,z)+B, \quad \forall w\in\mathcal{W}, z\in\mathcal{Z}. \tag{4.19}$$

This assumption was considered in the literature [35], and is satisfied by many (nondifferentiable) regularized loss functions wisely used in the machine learning community, including hinge loss and all strongly smooth loss functions. Let $\widetilde{W}\subset\mathcal{W}$ and $\eta_t\le\sigma_\Psi/A$. We consider the following projected OMD where a mirror descent step is followed by a Bregman projection at each iteration:

$$\begin{cases}\nabla\Psi(w_{t+\frac{1}{2}}) = \nabla\Psi(w_t) - \eta_t f'(w_t,z_t),\\ w_{t+1} = \arg\min_{w\in\widetilde{W}} D_\Psi(w, w_{t+\frac{1}{2}}).\end{cases}$$

We can replace w_{t+1} with $w_{t+\frac{1}{2}}$ in (3.1) to get (by definition one can show $F'(w_t) =: \mathbb{E}_Z[f'(w_t,Z)]$ is a subgradient of F at w_t)

$$\begin{aligned} E_{z_t}[D_\Psi(w^*, w_{t+\frac{1}{2}})] - D_\Psi(w^*,w_t) &= \eta_t\langle w^*-w_t, F'(w_t)\rangle + \mathbb{E}_{z_t}[D_\Psi(w_t, w_{t+\frac{1}{2}})]\\ &= \eta_t\langle w^*-w_t, F'(w_t)\rangle + \mathbb{E}_{z_t}[D_{\Psi^*}(\nabla\Psi(w_{t+\frac{1}{2}}), \nabla\Psi(w_t))]\\ &\le \eta_t\langle w^*-w_t, F'(w_t)\rangle + \frac{\eta_t^2}{2\sigma_\Psi}\mathbb{E}_{z_t}[\|f'(w_t,z_t)\|_*^2], \end{aligned} \tag{4.20}$$

where the second identity is due to (3.3) and the last inequality is due to the σ_Ψ^{-1}-strong smoothness of Ψ^*. By the first-order condition in the definition w_{t+1} above, we derive

$$\langle w^* - w_{t+1}, \nabla\Psi(w_{t+1}) - \nabla\Psi(w_{t+\frac{1}{2}})\rangle \ge 0,$$

from which and (3.2) we derive

$$\begin{aligned} &D_\Psi(w^*, w_{t+1}) - D_\Psi(w^*, w_{t+\frac{1}{2}}) =\\ &\qquad - D_\Psi(w_{t+1}, w_{t+\frac{1}{2}}) - \langle w^*-w_{t+1}, \nabla\Psi(w_{t+1}) - \nabla\Psi(w_{t+\frac{1}{2}})\rangle \le 0. \end{aligned}$$

Plugging the above inequality back into (4.20) and using (4.19), we derive

$$\mathbb{E}_{z_t}[D_\Psi(w^*,w_{t+1})]-D_\Psi(w^*,w_t) \le \eta_t\langle w^*-w_t,F'(w_t)\rangle+\frac{\eta_t^2}{2\sigma_\Psi}\big[A\mathbb{E}_{z_t}[f(w_t,z_t)]+B\big]. \tag{4.21}$$

According to the definition of subgradient, we know

$$\mathbb{E}_{z_t}[f(w_t,z_t)] = F(w_t)-F(w^*)+F(w^*) \le \langle w_t-w^*,F'(w_t)\rangle + F(w^*).$$

This together with (4.21) gives

$$\begin{aligned}
&\mathbb{E}_{z_t}[D_\Psi(w^*,w_{t+1})]-D_\Psi(w^*,w_t)\\
\le\;& \eta_t\langle w^*-w_t,F'(w_t)\rangle\Big(1-\frac{\eta_t A}{2\sigma_\Psi}\Big)+\frac{\eta_t^2[AF(w^*)+B]}{2\sigma_\Psi}\\
\le\;& \eta_t\langle w^*-w_t,F'(w_t)-F'(w^*)\rangle\Big(1-\frac{\eta_t A}{2\sigma_\Psi}\Big)+\frac{\eta_t^2[AF(w^*)+B]}{2\sigma_\Psi},
\end{aligned}$$

where in the last step we have used $\langle w^*-w_t,-F'(w^*)\rangle \ge 0$ due to the first-order condition in the definition of w^*. If we impose an assumption similar to (2.2) as $\langle w^*-w,F'(w^*)-F'(w)\rangle \ge \Omega(D_\Psi(w^*,w))$ for all $w\in\mathcal{W}$ and use $\eta_t \le \sigma_\Psi/A$, then we derive

$$\mathbb{E}_{z_t}[D_\Psi(w^*,w_{t+1})] \le D_\Psi(w^*,w_t)-\frac{\eta_t}{2}\Omega\left(D_\Psi(w^*,w_t)\right)+b'\eta_t^2,$$

where $b'=\frac{AF(w^*)+B}{2\sigma_\Psi}$. The above inequality takes the same form as (4.11), from which we can derive exactly the same sufficient condition for the convergence and upper bounds on convergence rates. Our analysis may not be used to get necessary conditions or lower bounds for either projected OMD or non-differentiable objective functions. Indeed, the derivation of (4.2) is based on an identity on the one-step progress which may not hold for the projected algorithm, and the L_F-strong smoothness of F which does not hold for non-differentiable loss functions.

5. Convergence in the Case of Zero Variance and Almost Sure Convergence

In this section we prove Theorem 3 for the convergence in the case of zero variance and Theorem 4 for the almost sure convergence.

Proof of Theorem 3. Necessity. For any $w, \tilde{w} \in \mathcal{W}$, we know

$$\begin{aligned} D_F(w, \tilde{w}) &= F(w) - F(\tilde{w}) - \langle w - \tilde{w}, \nabla F(\tilde{w}) \rangle \\ &= \mathbb{E}_z \Big[f(w, z) - f(\tilde{w}, z) - \langle w - \tilde{w}, \nabla f(\tilde{w}, z) \rangle \Big] \\ &\leq \frac{L \mathbb{E}_z \left[\|w - \tilde{w}\|^2 \right]}{2} = \frac{L\|w - \tilde{w}\|^2}{2}, \end{aligned}$$

where the inequality follows from the L-strong smoothness of $f(\cdot, z)$ for almost every $z \in \mathcal{Z}$. Hence F is L-strongly smooth w.r.t. $\|\cdot\|$. Notice that we do not require the increment condition (2.1) nor the variance condition in the derivation of (4.2). Indeed, we only use the L_F-strong smoothness of F and σ_Ψ-strong convexity of Ψ there. Therefore, (4.2) holds, from which we derive

$$\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi(w^*, w_{t+1})] \geq (1 - 2L\sigma_\Psi^{-1}\eta_t)\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)]. \tag{5.1}$$

We now need the assumption $0 < \eta_t \leq \frac{\sigma_\Psi}{(2+\kappa)L}$ with $\kappa > 0$ on the step size sequence. Denote the constant $\tilde{a} = \frac{2+\kappa}{2} \log \frac{2+\kappa}{\kappa}$ and apply the elementary inequality (see e.g., [20])

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

We know from (5.1) that

$$\mathbb{E}_{z_1,\ldots,z_t}[D_\Psi(w^*, w_{t+1})] \geq \exp\big(-2\tilde{a}L\sigma_\Psi^{-1}\eta_t\big)\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)].$$

Applying this inequality iteratively for $t = 1, \ldots, T$ then gives

$$\begin{aligned} \mathbb{E}_{z_1,\ldots,z_T}[D_\Psi(w^*, w_{T+1})] &\geq \prod_{t=1}^{T} \exp\big(-2\tilde{a}L\sigma_\Psi^{-1}\eta_t\big) D_\Psi(w^*, w_1) \\ &= \exp\left\{-2\tilde{a}L\sigma_\Psi^{-1}\sum_{t=1}^{T}\eta_t\right\} D_\Psi(w^*, w_1). \end{aligned}$$

From the assumption $w^* \neq w_1$, we have $D_\Psi(w^*, w_1) > 0$. The convergence $\lim_{t\to\infty} \mathbb{E}_{z_1,\ldots,z_{t-1}}[D_\Psi(w^*, w_t)] = 0$ then implies $\sum_{t=1}^{\infty} \eta_t = \infty$.

Sufficiency. Here we use the estimate (4.12) derived in the proof of Proposition 13. But in our case of zero variance, $b = \frac{1}{\sigma_\Psi}\mathbb{E}_Z\left[\|\nabla_w[f(w^*, Z)]\|_*^2\right] = 0$. So (4.12) takes the form (note that we can choose $t_1 = 1$ in deriving (4.9))

$$A_{t+1} \leq A_t - \frac{\eta_t}{2}\Omega\left(A_t\right), \qquad \forall t \in \mathbb{N}. \tag{5.2}$$

This implies that for any $0 < \gamma < 1$, there must exist some integer $\tilde{t}_\gamma \in \mathbb{N}$ such that $A_{\tilde{t}_\gamma} \leq \gamma$, since otherwise $A_t > \gamma$ for every $t \in \mathbb{N}$, which by (4.13) and (5.2) leads to a contradiction:

$$A_{t+1} \leq A_t - \frac{\eta_t \Omega(\gamma)}{2\gamma} A_t \leq A_t - \frac{\eta_t}{2}\Omega(\gamma) \leq A_{\tilde{t}_\gamma} - \frac{\Omega(\gamma)}{2} \sum_{k=\tilde{t}_\gamma}^{t} \eta_k \to -\infty \text{ (as } t \to \infty).$$

But (5.2) also tells us that the sequence $\{A_t\}_{t\in\mathbb{N}}$ of nonnegative numbers is decreasing. Hence $A_{\tilde{t}_\gamma} \leq \gamma$ for every $t \geq \tilde{t}_\gamma$. This proves the limit

$$\lim_{t\to\infty} \mathbb{E}_{z_1,\ldots,z_{t-1}}\left[D_\Psi(w^*, w_t)\right] = \lim_{t\to\infty} A_t = 0.$$

We now turn to prove (2.7) under the special choice of the constant step size sequence $\eta_t \equiv \eta_1$. It follows from (5.1) that $A_{T+1} \geq (1 - 2L\sigma_\Psi^{-1}\eta_1)^T A_1$. Furthermore, assumption (2.5) means that (2.2) holds with $\Omega(u) = \sigma_F u$. So (5.2) translates to

$$A_{t+1} \leq (1 - 2^{-1}\eta_1\sigma_F)A_t,$$

from which we find $A_{T+1} \leq (1 - 2^{-1}\eta_1\sigma_F)^T A_1$ by iteration. This verifies (2.7) and completes the proof of Theorem 3. □

The proof of Theorem 4 for the almost sure convergence is based on the following Doob's forward convergence theorem (see, e.g., [10] on page 195).

Lemma 14. *Let $\{\tilde{X}_t\}_{t\in\mathbb{N}}$ be sequences of nonnegative random variables and let $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$ be a sequence of random variable sets with $\mathcal{F}_t \subset \mathcal{F}_{t+1}$ for every $t \in \mathbb{N}$. Suppose that $\mathbb{E}[\tilde{X}_{t+1}|\mathcal{F}_t] \leq \tilde{X}_t$ almost surely for every $t \in \mathbb{N}$. Then the sequence $\{\tilde{X}_t\}$ converges to a nonnegative random variable $\tilde{X}$ almost surely.*

Proof of Theorem 4. We follow the proof of Proposition 13 and apply (4.9). Since $\langle w^* - w_t, \nabla F(w^*) - \nabla F(w_t)\rangle \geq 0$, (4.9) implies

$$\mathbb{E}_{z_t}[D_\Psi(w^*, w_{t+1})] \leq D_\Psi(w^*, w_t) + \frac{\eta_t^2}{\sigma_\Psi}\mathbb{E}_Z\left[\|\nabla_w[f(w^*, Z)]\|_*^2\right], \qquad \forall t \geq t_1. \tag{5.3}$$

The condition $\sum_{t=1}^\infty \eta_t^2 < \infty$ enables us to define a stochastic process $\{\tilde{X}_t\}_t$ by

$$\tilde{X}_t = D_\Psi(w^*, w_t) + \frac{1}{\sigma_\Psi}\mathbb{E}_Z\left[\|\nabla_w[f(w^*, Z)]\|_*^2\right] \sum_{\ell=t}^{\infty} \eta_\ell^2.$$

By (5.3), we know that $\mathbb{E}_{z_t}[\tilde{X}_{t+1}] \leq \tilde{X}_t$ for $t \geq t_1$. Also, $\tilde{X}_t \geq 0$. So the stochastic process $\{\tilde{X}_t\}_{t \geq t_1}$ is a supermartingale. Then by the supermartingale convergence theorem, Lemma 14, we know that the sequence $\{\tilde{X}_t\}_{t \geq t_1}$ converges to a non-negative random variable $\tilde{X}$ almost surely. According to Fatou's Lemma and the limit $\lim_{t\to\infty} \mathbb{E}[D_\Psi(w^*, w_t)] = 0$ proved by Proposition 13, we get

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

But $\tilde{X}$ is a non-negative random variable, so we have $\tilde{X} = 0$ almost surely. It follows that $\{D_\Psi(w^*, w_t)\}_{t\in\mathbb{N}}$ converges to 0 almost surely. The proof of Theorem 4 is complete. □

6. Proving Explicit Results

In this section we prove the propositions stated in subsection 2.2 on some properties of special mirror maps, and Theorems 1 and 8 on necessary and sufficient conditions for the convergence, as well as tight convergence rates.

Proof of Proposition 5. If Ψ is L_Ψ-strongly smooth, then the condition in Lemma 12 is satisfied with $g = \Psi, L = L_\Psi$ and $\alpha = 1$. So by Lemma 12, there holds

$$\|\nabla\Psi(w) - \nabla\Psi(\tilde{w})\|_*^2 \leq L_\Psi \langle w - \tilde{w}, \nabla\Psi(w) - \nabla\Psi(\tilde{w})\rangle, \qquad \forall w, \tilde{w} \in \mathcal{W}.$$

By the Schwarz inequality $\langle w - \tilde{w}, \nabla\Psi(w) - \nabla\Psi(\tilde{w})\rangle \leq \|w - \tilde{w}\|\|\nabla\Psi(w) - \nabla\Psi(\tilde{w})\|_*$, this implies

$$\|\nabla\Psi(w) - \nabla\Psi(\tilde{w})\|_* \leq L_\Psi\|w - \tilde{w}\|, \quad \forall w, \tilde{w} \in \mathcal{W}. \tag{6.1}$$

So the function $\nabla\Psi$ is Lipschitz, and hence is continuous everywhere.

Setting $\tilde{w} = 0$ in (6.1) also yields

$$\|\nabla\Psi(w)\|_* \leq \|\nabla\Psi(0)\|_* + L_\Psi\|w\| \leq (\|\nabla\Psi(0)\|_* + L_\Psi)(1 + \|w\|), \qquad \forall w \in \mathcal{W}.$$

This establishes the incremental conditional (2.1) at infinity with $C_\Psi = \|\nabla\Psi(0)\|_* + L_\Psi$.

If F is σ_F-strongly convex, by the identity (2.3), we have

$$\langle w-\tilde{w}, \nabla F(w)-\nabla F(\tilde{w})\rangle = D_F(w,\tilde{w})+D_F(\tilde{w},w) \geq \sigma_F\|w-\tilde{w}\|^2, \qquad \forall w,\tilde{w} \in \mathcal{W}.$$

But $D_\Psi(\tilde{w},w) \leq \frac{L_\Psi}{2}\|w-\tilde{w}\|^2$. So we have

$$\langle w-\tilde{w}, \nabla F(w)-\nabla F(\tilde{w})\rangle \geq \sigma_F\|w-\tilde{w}\|^2 \geq \frac{2\sigma_F}{L_\Psi} D_\Psi(\tilde{w},w), \qquad \forall w,\tilde{w} \in \mathcal{W}.$$

Hence (2.2) is satisfied for a linear convex function $\Omega(u) = \frac{2\sigma_F}{L_\Psi}u$. This proves Proposition 5. □

For proving Proposition 6, we need the following inequalities which follow easily from the elementary inequalities

$$|a^\beta - b^\beta| \leq |a-b|^\beta, \quad (a+b)^\beta \leq a^\beta + b^\beta \leq 2^{1-\beta}(a+b)^\beta, \qquad \forall a,b \geq 0, \beta \in (0,1].$$

Lemma 15. *Let $0 < \beta \leq 1$. Then we have*

$$|sgn(a)|a|^\beta - sgn(b)|b|^\beta| \leq 2^{1-\beta}|a-b|^\beta, \qquad \forall a,b \in \mathbb{R}, \tag{6.2}$$

$$\left|\|\tilde{w}\|_p^\beta - \|w\|_p^\beta\right| \leq \left|\|\tilde{w}\|_p - \|w\|_p\right|^\beta \leq \|\tilde{w}-w\|_p^\beta, \qquad \forall w,\tilde{w} \in \mathcal{W}, \tag{6.3}$$

where we denote the sign of $a \in \mathbb{R}$ by $sgn(a) = 1$ if $a > 0$, -1 if $a < 0$, and 0 if $a = 0$.

Proof of Proposition 6. Let $p^* = \frac{p}{p-1} > 2$ be the dual number of p satisfying $\frac{1}{p} + \frac{1}{p^*} = 1$. Then the dual norm $\|\cdot\|_*$ is exactly the p^*-norm $\|\cdot\|_{p^*}$, and the gradient of Ψ_p at $w \in \mathcal{W}$ equals

$$\nabla\Psi_p(w) = \|w\|_p^{2-p}\hat{w}, \tag{6.4}$$

where $\hat{w} \in \mathcal{W}^*$ is the vector depending on w given by

$$\hat{w} = \left(\mathrm{sgn}(w(j))|w(j)|^{p-1}\right)_{j=1}^d.$$

It follows that $\nabla\Psi_p$ is continuous everywhere, and by calculating the norm $\left\|\hat{w}\right\|_{p^*}$ directly that

$$\|\nabla\Psi_p(w)\|_* = \|w\|_p^{2-p}\left\|\hat{w}\right\|_{p^*} = \|w\|_p^{2-p+\frac{p}{p^*}} = \|w\|_p.$$

This proves the identity (2.8) and the incremental condition (2.1) with $C_{\Psi_p} = 1$.

To bound the Bregman distance $D_{\Psi_p}(\tilde{w}, w)$, we apply the identity (2.3) and find that for any $w, \tilde{w} \in \mathcal{W}$,

$$D_{\Psi_p}(\tilde{w}, w) \leq D_{\Psi_p}(\tilde{w}, w) + D_{\Psi_p}(w, \tilde{w}) \leq \|\tilde{w} - w\|_p \left\| \nabla\Psi_p(\tilde{w}) - \nabla\Psi_p(w) \right\|_{p^*}. \tag{6.5}$$

We use the expression (6.4) and write $\nabla\Psi_p(\tilde{w}) - \nabla\Psi_p(w)$ as

$$\nabla\Psi_p(\tilde{w}) - \nabla\Psi_p(w) = \|\tilde{w}\|_p^{2-p} \hat{\tilde{w}} - \|w\|_p^{2-p} \hat{w} = \|\tilde{w}\|_p^{2-p} \left(\hat{\tilde{w}} - \hat{w} \right) + \left(\|\tilde{w}\|_p^{2-p} - \|w\|_p^{2-p} \right) \hat{w}.$$

Applying (6.2) to the j-th components of $\hat{\tilde{w}} - \hat{w}$ and $\beta = p - 1 \in (0, 1)$, we have

$$\left| \mathrm{sgn}(\tilde{w}(j)) |\tilde{w}(j)|^{p-1} - \mathrm{sgn}(w(j)) |w(j)|^{p-1} \right| \leq 2^{2-p} \left| \tilde{w}(j) - w(j) \right|^{p-1}, \qquad j = 1, \ldots, d.$$

So for the first term, we have

$$\begin{aligned} \left\| \hat{\tilde{w}} - \hat{w} \right\|_{p^*} &\leq \left\{ \sum_{j=1}^{d} 2^{p^*(2-p)} \left| \tilde{w}(j) - w(j) \right|^{p^*(p-1)} \right\}^{1/p^*} \\ &= 2^{2-p} \|\tilde{w} - w\|_p^{\frac{p}{p^*}} = 2^{2-p} \|\tilde{w} - w\|_p^{p-1}. \end{aligned} \tag{6.6}$$

For the second term, we apply (6.3) with $\beta = 2 - p$ and find

$$\left\| \left(\|\tilde{w}\|_p^{2-p} - \|w\|_p^{2-p} \right) \hat{w} \right\|_{p^*} \leq \|\tilde{w} - w\|_p^{2-p} \|\hat{w}\|_{p^*} = \|\tilde{w} - w\|_p^{2-p} \|w\|_p^{p-1}.$$

Applying (6.3) with $\beta = p - 1$ yields

$$\|w\|_p^{p-1} \leq \|\tilde{w}\|_p^{p-1} + \|\tilde{w} - w\|_p^{p-1}.$$

Hence

$$\left\| \left(\|\tilde{w}\|_p^{2-p} - \|w\|_p^{2-p} \right) \hat{w} \right\|_{p^*} \leq \|\tilde{w}\|_p^{p-1} \|\tilde{w} - w\|_p^{2-p} + \|\tilde{w} - w\|_p.$$

Combining this with (6.6) gives

$$\left\| \nabla\Psi_p(\tilde{w}) - \nabla\Psi_p(w) \right\|_{p^*} \leq \left(2\|\tilde{w}\|_p \right)^{2-p} \|\tilde{w} - w\|_p^{p-1} + \|\tilde{w}\|_p^{p-1} \|\tilde{w} - w\|_p^{2-p} + \|\tilde{w} - w\|_p.$$

Putting this bound into (6.5), we obtain

$$D_{\Psi_p}(\tilde{w}, w) \leq \left(2\|\tilde{w}\|_p \right)^{2-p} \|\tilde{w} - w\|_p^{p} + \|\tilde{w}\|_p^{p-1} \|\tilde{w} - w\|_p^{3-p} + \|\tilde{w} - w\|_p^{2}.$$

Since $1 < 3-p < 2$, we have

$$D_{\Psi_p}(\tilde{w}, w) \le \begin{cases} \left((2\|\tilde{w}\|_p)^{2-p} + \|\tilde{w}\|_p^{p-1} + 1\right)\|\tilde{w} - w\|_p^2, & \text{when } \|\tilde{w} - w\|_p \ge 1, \\ \left((2\|\tilde{w}\|_p)^{2-p} + \|\tilde{w}\|_p^{p-1} + 1\right)\|\tilde{w} - w\|_p^{\min\{p,3-p\}}, & \text{when } \|\tilde{w} - w\|_p < 1. \end{cases}$$

Then our desired estimate (2.9) for $D_{\Psi_p}(\tilde{w}, w)$ follows.

Let $\tilde{w} \in \mathcal{W}$ and denote the constant $C_{\|\tilde{w}\|_p,p} = \left((2\|\tilde{w}\|_p)^{2-p} + \|\tilde{w}\|_p^{p-1} + 1\right)^{-1}$. We know from (2.9)

$$\|\tilde{w} - w\|_p^2 + \|\tilde{w} - w\|_p^{\min\{p,3-p\}} \ge C_{\|\tilde{w}\|_p,p} D_{\Psi_p}(\tilde{w}, w). \tag{6.7}$$

When $D_{\Psi_p}(\tilde{w}, w) \ge 1$, we have $\Omega_p\left(D_{\Psi_p}(\tilde{w}, w)\right) = D_{\Psi_p}(\tilde{w}, w) + \frac{1}{\tau_p} - 1 \le D_{\Psi_p}(\tilde{w}, w)$ and see from (6.7) that either

$$\|\tilde{w} - w\|_p^2 \ge 1 \Longrightarrow \|\tilde{w} - w\|_p^2 \ge \frac{1}{2}\left(\|\tilde{w} - w\|_p^2 + \|\tilde{w} - w\|_p^{\min\{p,3-p\}}\right) \ge \frac{C_{\|\tilde{w}\|_p,p}}{2}\Omega_p\left(D_{\Psi_p}(\tilde{w}, w)\right)$$

or $\|\tilde{w} - w\|_p^2 < 1$ which implies

$$\|\tilde{w} - w\|_p^{\min\{p,3-p\}} \ge \frac{C_{\|\tilde{w}\|_p,p}}{2} D_{\Psi_p}(\tilde{w}, w) \ge \frac{C_{\|\tilde{w}\|_p,p}}{2}$$

by our assumption $D_{\Psi_p}(\tilde{w}, w) \ge 1$, and thereby

$$\begin{aligned} \|\tilde{w} - w\|_p^2 &= \|\tilde{w} - w\|_p^{\min\{p,3-p\}}\|\tilde{w} - w\|_p^{2-\min\{p,3-p\}} \\ &\ge \left\{\frac{C_{\|\tilde{w}\|_p,p}}{2} D_{\Psi_p}(\tilde{w}, w)\right\}\left(\frac{C_{\|\tilde{w}\|_p,p}}{2}\right)^{\frac{2-\min\{p,3-p\}}{\min\{p,3-p\}}}. \end{aligned}$$

Hence

$$\|\tilde{w} - w\|_p^2 \ge \min\left\{\frac{C_{\|\tilde{w}\|_p,p}}{2}, \left(\frac{C_{\|\tilde{w}\|_p,p}}{2}\right)^{\tau_p}\right\}\Omega_p\left(D_{\Psi_p}(\tilde{w}, w)\right).$$

When $D_{\Psi_p}(\tilde{w}, w) < 1$, we have $\Omega_p\left(D_{\Psi_p}(\tilde{w}, w)\right) = \frac{1}{\tau_p}\left(D_{\Psi_p}(\tilde{w}, w)\right)^{\tau_p}$. Again, from (6.7), we have either

$$\begin{aligned} \|\tilde{w} - w\|_p^2 < 1 &\Longrightarrow \|\tilde{w} - w\|_p^{\min\{p,3-p\}} \ge \frac{C_{\|\tilde{w}\|_p,p}}{2} D_{\Psi_p}(\tilde{w}, w) \\ &\Longrightarrow \|\tilde{w} - w\|_p^2 \ge \tau_p\left(\frac{C_{\|\tilde{w}\|_p,p}}{2}\right)^{\tau_p}\Omega_p\left(D_{\Psi_p}(\tilde{w}, w)\right) \end{aligned}$$

or $\|\tilde{w} - w\|_p^2 \ge 1$ which implies

$$\|\tilde{w} - w\|_p^2 \ge \frac{C_{\|\tilde{w}\|_p,p}}{2} D_{\Psi_p}(\tilde{w}, w) \ge \frac{\tau_p C_{\|\tilde{w}\|_p,p}}{2}\Omega_p\left(D_{\Psi_p}(\tilde{w}, w)\right)$$

by our assumption $D_{\Psi_p}(\tilde{w}, w) < 1$. Therefore,

$$\|\tilde{w} - w\|_p^2 \geq \min\left\{\tau_p \frac{C_{\|\tilde{w}\|_p,p}}{2}, \tau_p \left(\frac{C_{\|\tilde{w}\|_p,p}}{2}\right)^{\tau_p}\right\} \Omega_p\left(D_{\Psi_p}(\tilde{w}, w)\right).$$

Combining the above two cases and noting $\tau_p > 1$, we see (2.10) holds.

The last statement follows immediately from the identity (2.3), the definition of σ_F-strong convexity, and (2.10). The proof is complete. □

Proof of Theorem 1. Denote $\sup_{x\in\mathcal{X}} \|x\|_* = R > 0$. The Hessian matrix of $f(\cdot, z) = \frac{1}{2}\left(\langle \cdot, x\rangle - y\right)^2$ for every z is $\nabla_w^2[f(w, z)] = xx^\top$, from which we know that $f(\cdot, z)$ and F are R^2-strongly smooth. Moreover, we have

$$\nabla F(w) = \mathbb{E}_Z[XX^\top w - XY] = \mathcal{C}_X w - \mathbb{E}_Z[XY].$$

So we know from the positive definiteness of the covariance matrix $\mathcal{C}_X$ that the only minimizer w^* is $w^* = w_\rho$. For any $w, \tilde{w} \in \mathcal{W}$, there holds

$$\begin{aligned}
D_F(w, \tilde{w}) &= \frac{1}{2}\mathbb{E}_Z\left[\left(\langle w, X\rangle - \langle \tilde{w}, X\rangle + \langle \tilde{w}, X\rangle - Y\right)^2\right] \\
&\quad - \frac{1}{2}\mathbb{E}_Z\left[\left(\langle \tilde{w}, X\rangle - Y\right)^2\right] - \langle w - \tilde{w}, \nabla F(\tilde{w})\rangle \\
&= \frac{1}{2}\mathbb{E}_Z\left[\left(\langle w - \tilde{w}, X\rangle\right)^2\right] + \mathbb{E}_Z\left[\left\langle w - \tilde{w}, \langle \tilde{w}, X\rangle X - XY\right\rangle\right] \\
&\quad - \langle w - \tilde{w}, \nabla F(\tilde{w})\rangle \\
&= \frac{1}{2}(w - \tilde{w})^\top \mathcal{C}_X (w - \tilde{w}) \geq \frac{\lambda_{min}}{2}\|w - \tilde{w}\|_2^2,
\end{aligned}$$

where $\lambda_{min} > 0$ is the smallest eigenvalue of the positive definite covariance matrix $\mathcal{C}_X$. But the norms $\|\cdot\|_2$ and $\|\cdot\|$ on $\mathbb{R}^d$ are equivalent. So there exist two positive numbers $b_1 \leq b_2$ such that $b_1\|w\|^2 \leq \|w\|_2^2 \leq b_2\|w\|^2$ for $w \in \mathbb{R}^d$. It follows that

$$D_F(w, \tilde{w}) \geq \frac{\lambda_{min} b_1}{2}\|w - \tilde{w}\|^2, \qquad \forall w, \tilde{w} \in \mathcal{W}.$$

This verifies the $\lambda_{min} b_1$-strong convexity of F. So by Propositions 5 and 6, the conditions of Theorems 2, 3 and 4 are satisfied. Moreover,

$$\mathbb{E}_Z\left[\|\nabla_w[f(w, Z)]\|_*\right] = \mathbb{E}_Z\left[\|(Y - \langle w, X\rangle)X\|_*\right] = \mathbb{E}_Z\left[|Y - \langle w, X\rangle|\, \|X\|_*\right].$$

So the assumption $\inf_{w\in\mathcal{W}} \mathbb{E}_Z\left[\|\nabla_w[f(w,Z)]\|_*\right] > 0$ in Theorem 2 is the same as the assumption $\inf_{w\in\mathcal{W}} \mathbb{E}_Z\left[|Y - \langle w, X\rangle| \, \|X\|_*\right] > 0$ in Theorem 1, and from Theorem 2 we know that if we replace $\|w_\rho - w_t\|^2$ by $D_\Psi(w_\rho, w_t)$, our statement (a) holds true and the constant σ can be taken as $\sigma = \frac{2\lambda_{min} b_1}{L_\Psi}$ in the case of an L_Ψ-strongly smooth mirror map Ψ. To get the statement for the norm square $\|w_\rho - w_t\|^2$, we notice first from the strong convexity of Ψ that $\frac{\sigma_\Psi}{2}\|w_\rho - w_t\|^2 \leq D_\Psi(w_\rho, w_t)$.

When Ψ is strongly smooth satisfying $D_\Psi(w_\rho, w_t) \leq \frac{L_\Psi}{2}\|w_\rho - w_t\|^2$, we know that our statement (a) holds true. When $\Psi = \Psi_p$ for some $1 < p \leq 2$, we use (2.10) with $\tilde{w} = w_\rho$ and Jensen's inequality to get from the convexity of Ω

$$\mathbb{E}_{z_1,\ldots,z_{t-1}}[\|w_\rho - w_t\|^2] \geq B_p'\Omega_p\left(\mathbb{E}_{z_1,\ldots,z_{t-1}}[D_{\Psi_p}(w_\rho, w_t)]\right),$$

where B_p' is a constant depending on $p, \|w_\rho\|$, and a constant c_p such that $c_p\|w\|_p \leq \|w\|$ holds for every $w \in \mathcal{W}$. Combining this relation with the explicit formula (2.11) for Ω_p, we know that $\lim_{t\to\infty} \mathbb{E}_{z_1,\ldots,z_{t-1}}[\|w_\rho - w_t\|^2] = 0$ implies $\lim_{t\to\infty} \mathbb{E}_{z_1,\ldots,z_{t-1}}[D_{\Psi_p}(w_\rho, w_t)] = 0$. Hence our statement (a) also holds true for $\Psi = \Psi_p$.

Note that the assumption $\mathbb{E}_Z\left[\|\nabla_w[f(w^*,Z)]\|_*\right] = 0$ in our statement (b) of Theorem 3 is the same as the the assumption $\mathbb{E}_Z\left[|Y - \langle w_\rho, X\rangle| \, \|X\|_*\right] = 0$ in Theorem 1. So our statement (b) can be proved from Theorem 3 by the same argument for dealing with the norm square $\|w_\rho - w_t\|^2$ from $D_\Psi(w_\rho, w_t)$ as we did for our statement (a).

Our statement (c) follows from Theorem 4 and the strong convexity of Ψ. The proof of Theorem 1 is complete. □

Proof of Theorem 8. Recall that for the regularizer r given by $r(w) = \lambda\|w\|_2^2$, there holds $D_r(\tilde{w}, w) = \lambda\|\tilde{w} - w\|_2^2$ for $\tilde{w}, w \in \mathcal{W}$. So we know that F is 2λ-strongly convex for every $z \in \mathcal{Z}$.

For the Bregman distance induced by the loss function

$$D_{\phi(\langle\cdot,x\rangle,y)}(\tilde{w}, w) = \phi(\langle\tilde{w}, x\rangle, y) - \phi(\langle w, x\rangle, y) - \langle\tilde{w} - w, \phi'(\langle w, x\rangle, y)x\rangle,$$

we apply the mean value theorem to find

$$\phi(\langle\tilde{w},x\rangle,y)-\phi(\langle w,x\rangle,y)=\phi'(\xi,y)\left(\langle\tilde{w},x\rangle-\langle w,x\rangle\right)=\langle\tilde{w}-w,\phi'(\xi,y)x\rangle,$$

where ξ is a number between $\langle\tilde{w},x\rangle$ and $\langle w,x\rangle$. We can write

$$\xi=(1-\theta)\langle\tilde{w},x\rangle+\theta\langle w,x\rangle=\langle(1-\theta)\tilde{w}+\theta w,x\rangle$$

for some $\theta\in(0,1)$. It follows that

$$D_{\phi(\langle\cdot,x\rangle,y)}(\tilde{w},w)=\langle\tilde{w}-w,\left(\phi'(\langle(1-\theta)\tilde{w}+\theta w,x\rangle,y)-\phi'(\langle w,x\rangle,y)\right)x\rangle$$

and

$$D_{\phi(\langle\cdot,x\rangle,y)}(\tilde{w},w)\leq\|\tilde{w}-w\|\|x\|_*\left|\phi'(\langle(1-\theta)\tilde{w}+\theta w,x\rangle,y)-\phi'(\langle w,x\rangle,y)\right|.$$

Then we apply the Lipschitz condition (2.12) and obtain

$$D_{\phi(\langle\cdot,x\rangle,y)}(\tilde{w},w)\leq\|\tilde{w}-w\|\|x\|_*\ell_\phi\left|\langle(1-\theta)\tilde{w}+\theta w,x\rangle-\langle w,x\rangle\right|\leq\|\tilde{w}-w\|^2\|x\|_*^2\ell_\phi.$$

If we denote $\sup_{x\in\mathcal{X}}\|x\|_*=R>0$, then we have

$$D_{\phi(\langle\cdot,x\rangle,y)}(\tilde{w},w)\leq\ell_\phi R^2\|\tilde{w}-w\|^2,\qquad\forall\tilde{w},w\in\mathcal{W}.$$

Therefore, $f(\cdot,z)$ is $2(\ell_\phi R^2+\lambda)$-strongly smooth for every $z\in\mathcal{Z}$, and the statements on the strong smoothness of F follows. Our desired statement on the convergence follows from Theorems 2, 3 and 4, as we have done in the proof of Theorem 1. The proof of Theorem 8 is complete. □

7. Simulations

In this section, we present some numerical simulations to validate our theoretical results. We use the AIR toolbox [15] to create a CT-measurement matrix $A\in\mathbb{R}^{n\times d}$ and an $N\times N$ sparse image represented by a vector $w^\dagger\in\mathbb{R}^d$ with $d=N^2$. Our objective is to recover the image $w^\dagger$ based on a sequence of noisy measurements $\{(x_t,y_t)\}_{t\in\mathbb{N}}$. In our experiment, we consider the measurement vector $x_t=\frac{A_{i_t}^\top}{\|A_{i_t}\|_2}$ and $y_t=\langle w^\dagger,x_t\rangle+s_t$, where A_{i_t} is the i_t-th row of A with the index i_t randomly drawn from the uniform distribution over $\{1,\dots,n\}$ and s_t is a Gaussian random variable with mean 0 and standard deviation $\sigma|\langle w^\dagger, x_t\rangle|$. We set $N = 128$ and $n = 92160$.

We apply the following online version of a modified linearized Bregman iteration [7] to recover the image $w^\dagger$ from noisy measurements $\{(x_t, y_t)\}_{t\in\mathbb{N}}$

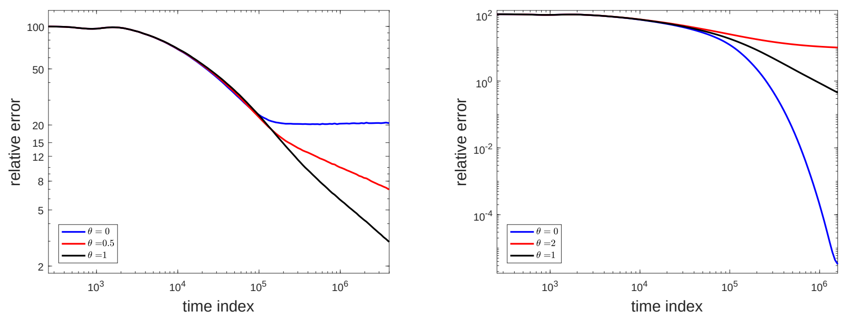
$$\begin{cases} v_{t+1} &= v_t - \eta_t\big(\langle w_t, x_t\rangle - y_t\big)x_t, \\ w_{t+1} &= T_{\lambda,\epsilon}(v_{t+1}), \end{cases} \tag{7.1}$$

where $T_{\lambda,\epsilon} : \mathbb{R}^d \to \mathbb{R}^d$ is defined component-wisely in terms of the function $T_{\lambda,\epsilon} : \mathbb{R} \to \mathbb{R}$ given by

$$T_{\lambda,\epsilon}(v) = \begin{cases} \frac{v\epsilon}{\lambda+\epsilon}, & \text{if } |v| \le \lambda + \epsilon, \\ \operatorname{sgn}(v)(|v| - \lambda), & \text{otherwise.} \end{cases}$$

Here we set $w_1 = v_1 = 0 \in \mathbb{R}^d$. This is a specific instantiation of the OMD with $f(w,z) = \frac{1}{2}\big(\langle w, x\rangle - y\big)^2$ and $\Psi = \Psi^{(\epsilon,\lambda)}$ defined [21] in Section 1. We choose $\lambda = 1$ and, as suggested in [7], $\epsilon = 10^{-8}$ here. We consider several step size sequences of the form $\eta_t = \big(1 + t\sigma_{\min}(\mathcal{C}_X)\big)^{-\theta}$ with $\theta \ge 0$, where $\sigma_{\min}(\mathcal{C}_X)$ is the smallest positive eigenvalue of the covariance matrix $\mathcal{C}_X$. We repeat the experiments 8 times and report the average of experimental results in this section.

We first consider the noisy case with $\sigma > 0$, which, as suggested in Remark 2, corresponds to the case with positive variances. We plot in panel (a) of Figure 2, the relative error $\mathrm{err}_r(w_t) := 100\|w_t - w^\dagger\|_2/\|w^\dagger\|_2$ versus the number of iterations for polynomially decaying step sizes with exponents $\theta \in \{0, \frac{1}{2}, 1\}$. The blue line is a plot for $\theta = 0$, which verifies the divergence of the algorithm since the step sizes do not satisfy the necessary condition $\lim_{t\to\infty}\eta_t = 0$ for the convergence of (7.1). The red and black lines are the plots for $\theta = \frac{1}{2}$ and $\theta = 1$, respectively. It is clear that both of these step size sequences satisfy the sufficient condition (1.5) for the convergence of the algorithm, which explains the convergence of (7.1) in the setting with positive variances. It can also be



(a) Positive variances (b) Zero variances

Figure 2: Relative error of algorithm (7.1) with different step sizes. Panel (a) shows the relative error in the case with *positive variances* for the polynomially decaying step sizes with $\theta = 0$ (blue line), $\theta = \frac{1}{2}$ (red line) and $\theta = 1$ (black line). Panel (b) shows the relative error in the case with *zero variance* for the polynomially decaying step sizes with $\theta = 0$ (blue line), $\theta = 2$ (red line) and $\theta = 1$ (black line).

seen that a faster convergence rate is achieved by setting $\theta = 1$ as compared to $\theta = 1/2$, which verifies Theorem 2 on tight convergence rates with $\theta = 1$.

We now consider the noiseless case with $\sigma = 0$, which, as clarified in Remark 2, corresponds to the case with zero variance. In panel (b) of Figure 2, we report the relative error as a function of the number of iterations for the step size sequences with $\theta = 0$ (blue line), $\theta = 2$ (red line) and $\theta = 1$ (black line). The step size sequence with $\theta = 2$ does not satisfy the necessary condition $\sum_{t=1}^{\infty} \eta_t = \infty$ for the convergence, which is well consistent with the divergence behavior of the algorithm as shown in panel (b). Both the step size sequences with $\theta = 1$ and $\theta = 0$ satisfy the sufficient condition $\sum_{t=1}^{\infty} \eta_t = \infty$, implying the convergence behavior of the algorithm (7.1). It is also clear that (7.1) with $\theta = 0$ achieves a faster convergence rate than that with $\theta = 1$, which is also consistent with the linear convergence rate established in (2.7) corresponding to $\theta = 0$.

Acknowledgments

We would like to thank the referees for their constructive comments. The work described in this paper is partially supported by the Research Grants Council of Hong Kong [Project No. CityU 11338616] and by National Natural Science Foundation of China under Grants 11461161006 and 11471292. This paper was written when the corresponding author, Ding-Xuan Zhou, visited Shanghai Jiaotong University (SJTU). The hospitality and sponsorships from SJTU and the Ministry of Education are greatly appreciated.

Appendix

This appendix provides the proofs of the co-coercivity of gradients stated in Lemma 12 and Proposition 7 together with a remark on variances involving stochastic gradients.

To prove Lemma 12, we need the following lemma on the Fenchel-conjugate of some norm power functions which is of independent interest.

Lemma 16. *Let $\kappa > 1$. The Fenchel-conjugate of $f = \frac{1}{\kappa}\|\cdot\|^{\kappa}$ is given by $f^*(v) = \frac{\kappa-1}{\kappa}\|v\|_*^{\frac{\kappa}{\kappa-1}}$.*

Proof. According to Young's inequality $ab \leq \frac{1}{\kappa}a^{\kappa} + \frac{\kappa-1}{\kappa}a^{\frac{\kappa}{\kappa-1}}$, we have for $v \in \mathcal{W}^*$,

$$\begin{aligned} f^*(v) &= \sup_{w\in\mathcal{W}}\left[\langle w, v\rangle - \frac{1}{\kappa}\|w\|^{\kappa}\right] \leq \sup_{w\in\mathcal{W}}\left[\|w\|\|v\|_* - \frac{1}{\kappa}\|w\|^{\kappa}\right] \\ &\leq \sup_{w\in\mathcal{W}}\left[\frac{1}{\kappa}\|w\|^{\kappa} + \frac{\kappa-1}{\kappa}\|v\|_*^{\frac{\kappa}{\kappa-1}} - \frac{1}{\kappa}\|w\|^{\kappa}\right] \\ &= \frac{\kappa-1}{\kappa}\|v\|_*^{\frac{\kappa}{\kappa-1}}. \end{aligned}$$

Since $\mathcal{W} = \mathcal{W}^{**}$, for $v \in \mathcal{W}^*$, there exists some $w \in \mathcal{W} = \mathcal{W}^{**}$ such that $\langle w, v\rangle = \|v\|_*$ and $\|w\| = 1$. Taking the vector $\|v\|_*^{\frac{1}{\kappa-1}}w$ in the definition of f^* gives

$$f^*(v) \geq \langle \|v\|_*^{\frac{1}{\kappa-1}}w, v\rangle - \frac{1}{\kappa}\|w\|^{\kappa}\|v\|_*^{\frac{\kappa}{\kappa-1}} = \|v\|_*^{\frac{1}{\kappa-1}}\|v\|_* - \frac{1}{\kappa}\|v\|_*^{\frac{\kappa}{\kappa-1}} = \frac{\kappa-1}{\kappa}\|v\|_*^{\frac{\kappa}{\kappa-1}}.$$

Combining the above two inequalities yields the stated result. □

Proof of Lemma 12. We use some ideas from [34]. Fix a $w \in \mathcal{W}$. Define $h : \mathcal{W} \to \mathbb{R}$ by $h(\bar{w}) = g(\bar{w}) - \langle \bar{w}, \nabla g(w) \rangle$. It is clear that h satisfies the condition

$$D_h(\bar{w}, \tilde{w}) = D_g(\bar{w}, \tilde{w}) \le \frac{L}{1+\alpha} \|\bar{w} - \tilde{w}\|^{1+\alpha}, \quad \forall \bar{w}, \tilde{w} \in \mathcal{W}.$$

Since h is convex and $\nabla h(w) = 0$, we know that h attains its minimum at w. So for $\tilde{w} \in \mathcal{W}$, we have

$$\begin{aligned} h(w) = \min_{\bar{w} \in \mathcal{W}} h(\bar{w}) &\le \min_{\bar{w} \in \mathcal{W}} \Big[h(\tilde{w}) + \langle \bar{w} - \tilde{w}, \nabla h(\tilde{w}) \rangle + \frac{L}{1+\alpha} \|\tilde{w} - \bar{w}\|^{\alpha+1} \Big] \\ &= h(\tilde{w}) - L \max_{\bar{w} \in \mathcal{W}} \Big[\langle \tilde{w} - \bar{w}, L^{-1} \nabla h(\tilde{w}) \rangle - \frac{1}{1+\alpha} \|\tilde{w} - \bar{w}\|^{\alpha+1} \Big] \\ &= h(\tilde{w}) - L \max_{\bar{w} \in \mathcal{W}} \Big[\langle \bar{w}, L^{-1} \nabla h(\tilde{w}) \rangle - \frac{1}{1+\alpha} \|\bar{w}\|^{\alpha+1} \Big]. \end{aligned}$$

According to the definition of Fenchel-conjugate and Lemma 16 with $\kappa = \alpha + 1$, we know

$$\begin{aligned} \max_{\bar{w} \in \mathcal{W}} \Big[\langle \bar{w}, L^{-1} \nabla h(\tilde{w}) \rangle - \frac{1}{1+\alpha} \|\bar{w}\|^{\alpha+1} \Big] &= \Big(\frac{1}{1+\alpha} \| \cdot \|^{\alpha+1} \Big)^* (L^{-1} \nabla h(\tilde{w})) \\ &= \frac{\alpha}{1+\alpha} \big\| L^{-1} \nabla h(\tilde{w}) \big\|_*^{\frac{1+\alpha}{\alpha}}. \end{aligned}$$

Combining the above discussions yields

$$h(w) \le h(\tilde{w}) - \frac{L^{-\frac{1}{\alpha}} \alpha}{1+\alpha} \big\| \nabla h(\tilde{w}) \big\|_*^{\frac{1+\alpha}{\alpha}}, \qquad \forall \tilde{w} \in \mathcal{W}.$$

The above inequality can be equivalently written as

$$g(\tilde{w}) \ge g(w) + \langle \tilde{w} - w, \nabla g(w) \rangle + \frac{L^{-\frac{1}{\alpha}} \alpha}{1+\alpha} \|\nabla g(\tilde{w}) - \nabla g(w)\|_*^{\frac{1+\alpha}{\alpha}}.$$

Switching w and $\tilde{w}$ also shows

$$g(w) \ge g(\tilde{w}) + \langle w - \tilde{w}, \nabla g(\tilde{w}) \rangle + \frac{L^{-\frac{1}{\alpha}} \alpha}{1+\alpha} \|\nabla g(w) - \nabla g(\tilde{w})\|_*^{\frac{1+\alpha}{\alpha}}.$$

Summing up the above two inequalities gives the stated inequality (4.5) and completes the proof. □

Now we turn to the proof of Proposition 7.

Proof of Proposition 7. Recall the dual number $p^* = \frac{p}{p-1} > 2$ of p given in the proof of Proposition 6 satisfying $\frac{1}{p} + \frac{1}{p^*} = 1$. Take the norm $\| \cdot \| = \| \cdot \|_p$.

Suppose to the contrary that Ψ_p is L-strong smooth for some $L > 0$. Then we know from the inequality (6.1) derived in the proof of Proposition 5 that

$$\|\nabla\Psi_p(w) - \nabla\Psi_p(\tilde{w})\|_* \leq L\|w - \tilde{w}\|, \qquad \forall w, \tilde{w} \in \mathcal{W}. \tag{7.2}$$

Let $a \geq 1$ and define two vectors $w, \tilde{w} \in \mathbb{R}^d$ as

$$w = \begin{cases} (a+1, a-1, \ldots, a+1, a-1), & \text{if } d \text{ is even,} \\ (a+1, a-1, \ldots, a+1, a-1, a), & \text{if } d \text{ is odd,} \end{cases}$$

and

$$\tilde{w} = \begin{cases} (a-1, a+1, \ldots, a-1, a+1), & \text{if } d \text{ is even,} \\ (a-1, a+1, \ldots, a-1, a+1, a), & \text{if } d \text{ is odd.} \end{cases}$$

By the elementary inequality $(a+1)^p + (a-1)^p \geq 2a^p$, we find

$$\|w\|_p = \|\tilde{w}\|_p = \begin{cases} \left[\frac{d}{2}(a+1)^p + \frac{d}{2}(a-1)^p\right]^{\frac{1}{p}} \geq d^{\frac{1}{p}}a, & \text{if } d \text{ is even,} \\ \left[\frac{d-1}{2}(a+1)^p + \frac{d-1}{2}(a-1)^p + a^p\right]^{\frac{1}{p}} \geq d^{\frac{1}{p}}a, & \text{if } d \text{ is odd.} \end{cases}$$

Combining this with the expression of $\nabla\Psi_p$ given in (6.4) yields

$$\begin{aligned} \|\nabla\Psi_p(w) - \nabla\Psi_p(\tilde{w})\|_* &= \|w\|_p^{2-p}\left\|\left(|w(j)|^{p-1} - |\tilde{w}(j)|^{p-1}\right)_{j=1}^d\right\|_* \\ &\geq \|w\|_p^{2-p}[(a+1)^{p-1} - (a-1)^{p-1}](d-1)^{\frac{1}{p^*}} \\ &\geq (d-1)^{\frac{1}{p}}a^{2-p}[(a+1)^{p-1} - (a-1)^{p-1}]. \end{aligned}$$

But

$$\|w - \tilde{w}\| = \begin{cases} 2d^{1/p}, & \text{if } d \text{ is even,} \\ 2(d-1)^{1/p} < 2d^{1/p}, & \text{if } d \text{ is odd.} \end{cases}$$

It follows that

$$\|\nabla\Psi_p(w) - \nabla\Psi_p(\tilde{w})\|_* \geq \frac{1}{2}\left(\frac{d-1}{d}\right)^{\frac{1}{p}} a^{2-p}[(a+1)^{p-1} - (a-1)^{p-1}]\|w - \tilde{w}\|.$$

Since $d \geq 2$, we have $\frac{d-1}{d} \geq \frac{1}{2}$. Therefore we apply the inequality (7.2) to obtain

$$L\|w - \tilde{w}\| \geq \frac{1}{4}a^{2-p}[(a+1)^{p-1} - (a-1)^{p-1}]\|w - \tilde{w}\|.$$

This is a contradiction to the limit $\lim_{a\to\infty} a^{2-p}[(a+1)^{p-1} - (a-1)^{p-1}] = \infty$. So Ψ_p is not strongly smooth. The proof of Proposition 7 is complete. $\square$

At the end, we give the following remark on the conditions on the variances.

Proposition 17. *If F is Fréchet differentiable, then the following two statements hold.*

(a) If there exists a $w^* \in \mathcal{W}$ with $\mathbb{E}_Z[\|\nabla_w[f(w^*, Z)]\|_*] = 0$, then we have $\mathbb{E}_Z[\|\nabla_w[f(w^*, Z)] - \nabla F(w^*)\|_*^2] = 0$.

(b) If $\inf_{w \in \mathcal{W}} \mathbb{E}_Z[\|\nabla_w[f(w, Z)]\|_*] > 0$, then we have $\mathbb{E}_Z[\|\nabla_w[f(w^*, Z)] - \nabla F(w^*)\|_*^2] > 0$ for any minimizer w^* of F.

Proof. For the statement (a), the condition $\mathbb{E}_Z[\|\nabla_w[f(w^*, Z)]\|_*] = 0$ amounts to saying that $\nabla_w[f(w^*, Z)] = 0$ holds almost surely, from which it follows that $\nabla F(w^*) = 0$ and therefore $\mathbb{E}_Z[\|\nabla_w[f(w^*, Z)] - \nabla F(w^*)\|_*^2] = 0$.

The statement (b) follows from the optimality condition $\nabla F(w^*) = 0$ and the Schwarz inequality $\mathbb{E}_Z[\|\nabla_w[f(w^*, Z)]\|_*] \le \left\{\mathbb{E}_Z[\|\nabla_w[f(w^*, Z)]\|_*^2]\right\}^{1/2}$. $\square$

References

[1] A. Agarwal, P. L. Bartlett, P. Ravikumar, M. J. Wainwright, Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization, IEEE Transactions on Information Theory 5 (2012) 3235–3249.

[2] K. Ball, E. A. Carlen, E. H. Lieb, Sharp uniform convexity and smoothness inequalities for trace norms, Inventiones Mathematicae 115 (1994) 463–482.

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

[4] J. M. Borwein, A. S. Lewis, Convex Analysis and Nonlinear Optimization: Theory and Examples, Springer Science & Business Media, 2010.

[5] L. Bottou, On-line learning in neural networks, Cambridge University Press, New York, NY, USA, 1998, pp. 9–42.

[6] L. Bottou, F. E. Curtis, J. Nocedal, Optimization methods for large-scale machine learning, arXiv:1606.04838 (2016).

[7] J.-F. Cai, S. Osher, Z. Shen, Linearized Bregman iterations for compressed sensing, Mathematics of Computation 78 (2009) 1515–1536.

[8] D.-R. Chen, Q. Wu, Y. Ying, D.-X. Zhou, Support vector machine soft margin classifiers: error analysis, Journal of Machine Learning Research 5 (2004) 1143–1175.

[9] X. Chen, A. M. Powell, Almost sure convergence of the Kaczmarz algorithm with random measurements, Journal of Fourier Analysis and Applications 18 (2012) 1195–1214.

[10] J. L. Doob, Measure Theory, Graduate Texts in Mathematics, volume 143, Springer, 1994.

[11] J. C. Duchi, S. Shalev-Shwartz, Y. Singer, A. Tewari, Composite objective mirror descent, in: Annual Conference on Learning Theory, Citeseer, 2010, pp. 14–26.

[12] Z.-C. Guo, S.-B. Lin, D.-X. Zhou, Learning theory of distributed spectral algorithms, Inverse Problems 33 (2017) 074009.

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

[14] Z.-C. Guo, Y. Ying, D.-X. Zhou, Online regularized learning with pairwise loss functions, Advances in Computational Mathematics 43 (2017) 127–150.

[15] P. C. Hansen, M. Saxild-Hansen, Air toolsa matlab package of algebraic iterative reconstruction methods, Journal of Computational and Applied Mathematics 236 (2012) 2167–2178.

[16] T. Hu, J. Fan, Q. Wu, D.-X. Zhou, Regularization schemes for minimum error entropy principle, Analysis and Applications 13 (2015) 437–455.

[17] P. J. Huber, Robust estimation of a location parameter, The Annals of Mathematical Statistics 35 (1964) 73–101.

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

[19] S. Lacoste-Julien, M. Schmidt, F. Bach, A simpler approach to obtaining an $o(1/t)$ convergence rate for the projected stochastic subgradient method, arXiv preprint arXiv:1212.2002 (2012).

[20] Y. Lei, D.-X. Zhou, Analysis of singular value thresholding algorithm for matrix completion, Preprint (2016).

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

[22] J. Lin, D.-X. Zhou, Learning theory of randomized Kaczmarz algorithm, Journal

of Machine Learning Research 16 (2015) 3341–3365.

[23] S.-B. Lin, X. Guo, D.-X. Zhou, Distributed learning with regularized least squares, Journal of Machine Learning Research 18 (2017) 3202–3232.

[24] A. Nedic, D. P. Bertsekas, Incremental subgradient methods for nondifferentiable optimization, SIAM Journal on Optimization 12 (2001) 109–138.

[25] A. Nemirovski, A. Juditsky, G. Lan, A. Shapiro, Robust stochastic approximation approach to stochastic programming, SIAM Journal on Optimization 19 (2009) 1574–1609.

[26] A. Rakhlin, O. Shamir, K. Sridharan, Making gradient descent optimal for strongly convex stochastic optimization, in: Proceedings of the 29th International Conference on Machine Learning, 2012, pp. 449–456.

[27] H. Robbins, S. Monro, A stochastic approximation method, The Annals of Mathematical Statistics (1951) 400–407.

[28] S. Smale, Y. Yao, Online learning algorithms, Foundations of Computational Mathematics 6 (2006) 145–170.

[29] T. Strohmer, R. Vershynin, A randomized Kaczmarz algorithm with exponential convergence, Journal of Fourier Analysis and Applications 15 (2009) 262–278.

[30] Q. Wu, Y. Ying, D.-X. Zhou, Multi-kernel regularized classifiers, Journal of Complexity 23 (2007) 108–134.

[31] Y. Yao, On complexity issues of online learning algorithms, IEEE Transactions on Information Theory 56 (2010) 6470–6481.

[32] Y. Ying, M. Pontil, Online gradient descent learning algorithms, Foundations of Computational Mathematics 8 (2008) 561–596.

[33] Y. Ying, D.-X. Zhou, Online regularized classification algorithms, IEEE Transactions on Information Theory 52 (2006) 4775–4788.

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

[35] T. Zhang, Solving large scale linear prediction problems using stochastic gradient descent algorithms, in: International Conference on Machine Learning, 2004, pp. 919–926.