Optimal Stochastic and Online Learning with Individual Iterates

Yunwen Lei[1,2]   **Peng Yang**[1]   **Ke Tang**[1]*   **Ding-Xuan Zhou**[3]

[1]University Key Laboratory of Evolving Intelligent Systems of Guangdong Province,
Department of Computer Science and Engineering,
Southern University of Science and Technology, Shenzhen 518055, China

[2]Department of Computer Science,
Technical University of Kaiserslautern, Kaiserslautern 67653, Germany

[3]School of Data Science and Department of Mathematics,
City University of Hong Kong, Kowloon, Hong Kong, China

{leiyw, yangp, tangk3}@sustech.edu.cn   mazhou@cityu.edu.hk

Abstract

Stochastic composite mirror descent (SCMD) is a simple and efficient method able to capture both geometric and composite structures of optimization problems in machine learning. Existing strategies require to take either an average or a random selection of iterates to achieve optimal convergence rates, which, however, can either destroy the sparsity of solutions or slow down the practical training speed. In this paper, we propose a theoretically sound strategy to select an *individual* iterate of the vanilla SCMD, which is able to achieve optimal rates for both convex and strongly convex problems in a non-smooth learning setting. This strategy of outputting an individual iterate can preserve the sparsity of solutions which is crucial for a proper interpretation in sparse learning problems. We report experimental comparisons with several baseline methods to show the effectiveness of our method in achieving a fast training speed as well as in outputting sparse solutions.

1   Introduction

Gradient-based methods have found wide applications to solve various optimization problems. A basic and representative method of this type is the gradient descent, which iteratively moves iterates along the minus gradient direction of the current iterate. However, gradient descent applied to machine learning problems requires to go through all training examples at each iteration, which is not efficient when the sample size is large. Stochastic gradient descent (SGD) relieves this computational burden by approximating the true gradient of the objective function with an unbiased estimation based on a randomly selected training example. With this strategy, SGD can achieve sample-size independent computational cost per iteration and therefore can be successfully applied to very large datasets which are becoming ubiquitous in the big data era [2, 4, 44].

From different viewpoints, SGD has been extended in various ways. For example, the trick of variance-reduction has been introduced to exploit the finite summation structure of objective functions for reducing the inherent variance [16, 33, 40, 43]. Adaptive step sizes were proposed to dynamically incorporate the knowledge of the geometry of the data observed [9]. Decreasing step sizes in a stagewise manner is used as a common trick in practice [14, 19, 41]. The trick of momentum is widely used to accelerate SGD by choosing an appropriate direction to pursue per iteration [18, 26, 29]. Proximal operators have been introduced to capture a structure of optimization problems [28], which

---

*Corresponding author

separates the regularizer and data-fitting terms to achieve a desired regularization effect. The concept of mirror map has been introduced to induce a Bregman distance for reflecting the geometry of the associated optimization problem [3, 25]. Empirical studies have shown that these variants can further improve the practical performance of SGD [4]. In this paper, we consider stochastic composite mirror descent (SCMD), which combines the technique of mirror map and proximal operator to capture both the composite and geometry structure of the associated optimization problem.

Theoretical studies of SCMD have attracted much attention since its appearance and many results have been derived to understand its promising behavior in practice. If the objective function is convex, then the expected suboptimality of the uniform average of iterates decays with the rate $O(T^{-\frac{1}{2}})$ after T iterations [6–8, 30, 46]. If the objective function is strongly convex, then the expected suboptimality of the uniform average of iterates decays with the rate $O(T^{-1} \log T)$ [17, 35], which is unimprovable if the objective function is not smooth [31]. Surprisingly, some averaging schemes have been proposed to achieve the optimal rate $O(T^{-1})$ [14, 20, 31, 37]. Although taking averages of previous iterates can achieve theoretically minimax optimal rates [1], it can have some practical side effects [31, 36, 37][2]. For example, it can destroy the sparsity of the solution which is often crucial for a proper interpretation of models in many applications. Moreover, averaging can affect the practical training speed due to the existence of some possibly poor iterates in the iterate sequence produced by SCMD [31, 36]. The side effect of destroying sparsity can be resolved by randomly drawing an individual iterate with probabilities proportional to the weights in the optimal averaging scheme. However, this strategy of selecting a random iterate would introduces new variances, which may further slow down the training speed due to the possibility of selecting a poor iterate. Outputting the last SGD iterate is a natural strategy in practice, which would not affect sparsity and practical training speed. However, the suboptimality for the last SGD iterate converges with suboptimal rates $O(T^{-\frac{1}{2}} \log T)$ and $O(T^{-1} \log T)$ in the convex and strongly convex cases [37], respectively. Very recently, it was shown that the last SGD iterate can attain at most a rate $O(T^{-1} \log T)$ with high probabilities [13]. These facts motivate us to ask a natural question: can we develop an algorithm which can inherit the optimal convergence rate from the averaging strategies as well as the sparsity-preservation and fast practical training speed from the last SGD iterate?

In this paper, we aim to give an affirmative answer to the above question by proposing a novel strategy to select an individual iterate able to achieve theoretically optimal rates and fast training speed in practice. We first consider the case with the number of iterations known to us, which allows us to divide the implementation into two stages. The first stage produces an output with some averaging scheme, which is then used in the second stage to select an individual iterate according to the one-step progress evolution of SCMD. We then extend this algorithm to the online learning setting where training examples arrive in a sequential way. We show that our algorithm can achieve the optimal rates $O(T^{-\frac{1}{2}})$ and $O(T^{-1})$ in the convex and strongly convex case, respectively. Rooted on a careful one-step progress analysis, our selection of iterates is based on the difference of two successive iterates, which shares some spirits with the common heuristic trick of terminating an algorithm if the distance of two successive iterates is lower than a threshold. Our analysis also removes bounded subgradient assumptions in the existing discussions of optimal learning rates [14, 20, 31, 37]. We report experimental results to confirm the effectiveness of the proposed algorithm in both attaining a fast training speed and producing a sparse solution.

We present algorithms with motivation in Section 2. Theoretical results and discussions are given in Sections 3 and 4. Experimental results and conclusions are presented in Sections 5 and 6.

2 Algorithms with Motivations

2.1 Background

In supervised learning, we aim to infer an unknown relationship between input and output variables from a sequence of training examples $\{z_t = (x_t, y_t)\}_{t \in \mathbb{N}}$ drawn from a probability measure defined in a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with an input space $\mathcal{X} \subset \mathbb{R}^d$ and an output space $\mathcal{Y} \subset \mathbb{R}$, where $d \in \mathbb{N}$ is the dimension. A very elementary and powerful approximation of this relationship is a linear model of the form $x \mapsto \langle \mathbf{w}, x \rangle$ with $\mathbf{w} \in \mathbb{R}^d$, whose behavior at a single training example (x, y) can

---

[2]When mentioning averaging, individual iterates or last iterate, our attention is with respect to the iterate sequence produced by (2.2) below.

be quantified by a function $f(\mathbf{w}, z) = \ell(\langle \mathbf{w}, x \rangle, y)$, where the loss function $\ell : \mathbb{R}^2 \mapsto \mathbb{R}_+$ is convex with respect to (w.r.t.) the first argument and $\langle \mathbf{w}, x \rangle$ denotes the inner product between $\mathbf{w}$ and x. The learning process can be often formulated as an optimization problem of a composite structure

$$\min_{\mathbf{w} \in \mathbb{R}^d} \phi(\mathbf{w}) = \mathbb{E}_z[f(\mathbf{w}, z)] + r(\mathbf{w}), \tag{2.1}$$

where $\mathbb{E}_z[\cdot]$ denotes the expectation w.r.t. z and $r : \mathbb{R}^d \mapsto \mathbb{R}_+$ is a regularizer possibly inducing sparsity. With specific instantiations of ℓ and r, Eq. (2.1) covers many famous learning problems in a unifying framework, including least squares, SVMs, logistic regression, lasso and elastic-net [8, 39].

SCMD provides an efficient first-order method to exploit the composite and geometry structure of the problem (2.1) [8, 24]. It extends SGD by using a strongly convex and differentiable mirror map Ψ to generate an appropriate Bregman distance [3, 25] $D_\Psi(\mathbf{w}, \tilde{\mathbf{w}}) = \Psi(\mathbf{w}) - \Psi(\tilde{\mathbf{w}}) - \langle \mathbf{w} - \tilde{\mathbf{w}}, \nabla \Psi(\tilde{\mathbf{w}}) \rangle$, where $\nabla \Psi(\tilde{\mathbf{w}})$ denotes the gradient of Ψ at $\tilde{\mathbf{w}}$. Let $\mathbf{w}_1 \in \mathbb{R}^d$ be an initial point and $\{\eta_t\}_{t \in \mathbb{N}}$ be a step size sequence. Up on the arrival of z_t at the t-th iteration, we calculate a subgradient $f'(\mathbf{w}_t, z_t) \in \partial_w f(\mathbf{w}_t, z_t)$ as an unbiased estimate of $F'(\mathbf{w}_t) \in \partial F(\mathbf{w}_t)$, where $\partial_w f(\mathbf{w}_t, z_t)$ denotes the subdifferential of $f(\cdot, z_t)$ at $\mathbf{w}_t$ and $F(\mathbf{w}) = \mathbb{E}_z[f(\mathbf{w}, z)]$. SCMD updates the model by

$$\mathbf{w}_{t+1} = \arg \min_{\mathbf{w} \in \mathbb{R}^d} D_\Psi(\mathbf{w}, \mathbf{w}_t) + \eta_t\big(\langle \mathbf{w} - \mathbf{w}_t, f'(\mathbf{w}_t, z_t) \rangle + r(\mathbf{w})\big). \tag{2.2}$$

Intuitively, SCMD uses $f'(\mathbf{w}_t, z_t)$ to form a first-order approximation of $f(\cdot, z_t)$ at $\mathbf{w}_t$ and uses the Bregman distance $D_\Psi(\mathbf{w}, \mathbf{w}_t)$ to keep $\mathbf{w}_{t+1}$ not far away from the current iterate. The regularizer r is kept intact here to preserve a regularization effect. Typical choices of mirror maps include the p-norm divergence $\Psi_p(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_p^2$ $(1 < p \leq 2)$, which works favorably when the solution of (2.1) is sparse by setting p close to 1 [8, 39]. Here $\|\cdot\|_p$ is the p-norm defined by $\|\mathbf{w}\|_p = \big(\sum_{i=1}^d |w(i)|^p\big)^{1/p}$ for $\mathbf{w} = (w(1), \ldots, w(d))^\top \in \mathbb{R}^d$. SCMD recovers the stochastic proximal gradient descent by taking $\Psi = \Psi_2$ [7] and stochastic mirror descent by taking $r(\mathbf{w}) = 0$ [14]. It should be mentioned that ρ is not necessarily to be an empirical distribution over training examples, and therefore the setting we consider here is more general than stochastic learning to minimize an empirical risk.

Algorithm 1: SCMDI

Input: $\{\eta_t\}_t, \sigma_\phi, \mathbf{w}_1$ and T.
Output: an approximate solution of (2.1)
1 **if** $\sigma_\phi == 0$ **then**
2      $s_w \leftarrow \mathbf{w}_1, \ s \leftarrow 1$
3 **else**
4      $s_w \leftarrow 6\eta_1 \mathbf{w}_1, \ s \leftarrow 6\eta_1$

5 **for** $t = 1, 2$ **to** $T - 1$ **do**
6   calculate $\mathbf{w}_{t+1}$ by (2.2)
7   **if** $\sigma_\phi == 0$ **then**
8         $s_w \leftarrow s_w + \mathbf{w}_{t+1}, \ s \leftarrow s + 1$
9   **else**
10        $s_w \leftarrow s_w + (t+2)(t+3)\eta_{t+1}\mathbf{w}_{t+1}$
11        $s \leftarrow s + (t+2)(t+3)\eta_{t+1}$

12 $\bar{\mathbf{w}}_T \leftarrow s_w/s$
13 **for** $t = T, T + 1$ **to** $2T - 1$ **do**
14   calculate $\mathbf{w}_{t+1}$ by (2.2)
15     $\triangle \leftarrow D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_t) - D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{t+1})$
16   **if** $\triangle \leq T^{-1} D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)$ **then**
17        $T^* \leftarrow t, \ \mathbf{w}_{T^*} \leftarrow \mathbf{w}_t$

18 **return** $\mathbf{w}_{T^*}$

Algorithm 2: OCMDI

Input: $\{\eta_t\}_t, \sigma_\phi$ and $\mathbf{w}_1$.
Output: an approximate solution of (2.1)
1 **if** $\sigma_\phi == 0$ **then**
2      $s_w \leftarrow \mathbf{w}_1, \ s \leftarrow 1$
3 **else**
4      $s_w \leftarrow 6\eta_1 \mathbf{w}_1, \ s \leftarrow 6\eta_1$

5 $\bar{\mathbf{w}} \leftarrow \mathbf{w}_1, \ \hat{\mathbf{w}} \leftarrow \mathbf{w}_1, \ k \leftarrow 1$
6 **for** $t = 1, 2, \cdots$ **do**
7   calculate $\mathbf{w}_{t+1}$ by (2.2)
8      $\triangle \leftarrow D_\Psi(\bar{\mathbf{w}}, \mathbf{w}_t) - D_\Psi(\bar{\mathbf{w}}, \mathbf{w}_{t+1})$
9   **if** $\triangle \leq 2^{1-k} D_\Psi(\bar{\mathbf{w}}, \hat{\mathbf{w}})$ **then**
10        $\tilde{\mathbf{w}} \leftarrow \mathbf{w}_t$
11   **if** $\sigma_\phi == 0$ **then**
12        $s_w \leftarrow s_w + \mathbf{w}_{t+1}, \ s \leftarrow s + 1$
13   **else**
14        $s_w \leftarrow s_w + (t+2)(t+3)\eta_{t+1}\mathbf{w}_{t+1}$
15        $s \leftarrow s + (t+2)(t+3)\eta_{t+1}$
16   **if** $t == 2^k - 1$ **then**
17        $k \leftarrow k + 1, \ \bar{\mathbf{w}} \leftarrow s_w/s, \ \hat{\mathbf{w}} \leftarrow \mathbf{w}_t$

18 **return** $\tilde{\mathbf{w}}$

2.2 Algorithms

We now present our first algorithm (Algorithm 1) which performs SCMD in two stages. In the first stage, we simply perform updates according to (2.2). An output $\bar{\mathbf{w}}_T$ is then produced by taking some

averages of the previous iterates. It is clear that $\bar{\mathbf{w}}_T$ is a simple average with uniform weights in the convex case, while the construction of $\bar{\mathbf{w}}_T$ in the strongly convex case is motivated by theoretical analysis (see Theorem 4) [20]. In the second stage, other than updating $\{\mathbf{w}_t\}$ with (2.2), we also search a time index (line 17 of Algorithm 1) at which the difference between two consecutive Bregman distance is less than a small number. We refer to Algorithm 1 as Stochastic Composite Mirror Descent with Individual iterates (SCMDI) since it outputs an individual iterate of the vanilla SCMD.

Remark 1. For any t, denote $A_t := D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_t) - D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{t+1})$. It should be mentioned that the function $t \mapsto A_t$ is not monotonically w.r.t. t. Therefore, not all $t \geq T^*$ can necessarily satisfy the condition in line 16 of Algorithm 1, which is a requirement to achieve optimal convergence rates in our analysis.

In Algorithm 1, we update T^* once we find a new t satisfying $A_t \leq T^{-1} D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)$. Another feasible strategy is to set $T^* = \arg\min_{t \in \{T, T+1, \dots, 2T-1\}} A_t$. However, experiments show that this strategy does not work as well as Algorithm 1.

Algorithm 2 is an extension of Algorithm 1 by removing the information on T. It differs from the vanilla SCMD in two aspects. Firstly, it computes some average $\bar{\mathbf{w}}_T$ of previous iterates at the (2^k-1)-th iteration, $k \in \mathbb{N}$. Secondly, it searches an index t such that $D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_t) - D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{t+1}) \leq 2^{1-k} D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)$ and outputs this individual iterate. Algorithm 2 with $2^k \leq t < 2^{k+1}$ recovers Algorithm 1 with $T = 2^k$ or $T = 2^{k-1}$, depending on whether an index $t \geq 2^k$ satisfying $A_t \leq 2^{1-k} D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)$ is found or not. Therefore, Algorithm 2 can achieve the same convergence rates as Algorithm 1 in both convex and strongly convex cases. It does not need the information on T and therefore applies to the online learning setting. We refer to Algorithm 2 as Online Composite Mirror Descent with Individual iterates (OCMDI).

2.3 Motivation

Before giving theoretical results, we sketch the key **idea** underlying the design of Algorithm 1 which also forms a key foundation of our theoretical analysis. Let $\mathbf{w}^* = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \phi(\mathbf{w})$. Typically we can get the following one-step error bounds for $\{\mathbf{w}_t\}_{t \in \mathbb{N}}$ produced by (2.2) (Lemma A.1 in Appendix)

$$\eta_t \mathbb{E}[\phi(\mathbf{w}_t) - \phi(\mathbf{w})] \leq \mathbb{E}[D_\Psi(\mathbf{w}, \mathbf{w}_t) - D_\Psi(\mathbf{w}, \mathbf{w}_{t+1})] + \eta_t^2 \widetilde{C} \tag{2.3}$$

for any $\mathbf{w} \in \mathbb{R}^d$ independent of z_t. Here $\widetilde{C}$ is a constant independent of t. If we choose $\mathbf{w} = \mathbf{w}^*$ and can show $\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_t) - D_\Psi(\mathbf{w}^*, \mathbf{w}_{t+1})] = O(\eta_t^2)$, then we would immediately obtain $\mathbb{E}[\phi(\mathbf{w}_t)] - \phi(\mathbf{w}^*) = O(\eta_t)$. This would immediately imply the rate $O(1/\sqrt{t})$ in the convex case and the rate $O(1/t)$ in the strongly convex case since in these two cases the typical step size choices are $\eta_t = 1/\sqrt{t}$ and $\eta_t = 1/t$ (ignoring constant factors), respectively [8]. Since $\mathbf{w}^*$ is unknown, any algorithm requiring an access to $\mathbf{w}^*$ is not implementable. A good surrogate $\bar{\mathbf{w}}_T$ of $\mathbf{w}^*$ should satisfy $\mathbb{E}[\phi(\bar{\mathbf{w}}_T)] - \phi(\mathbf{w}^*) = O(\eta_T)$ to enjoy a tight rate, for which a natural choice should be some average of $\{\mathbf{w}_t\}_{t \leq T}$. Then, we take $\mathbf{w} = \bar{\mathbf{w}}_T$ in (2.3) and need to find an index $T^* \in \{T, T+1, \dots, 2T-1\}$ such that $\mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{T^*}) - D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{T^*+1})] = O(\eta_T^2)$. By the non-negativity of Bregman distance, there always exists a $T^* \in \{T, T+1, \dots, 2T-1\}$ satisfying (see Lemma A.2 in Appendix)

$$D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{T^*}) - D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{T^*+1}) \leq T^{-1} D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T).$$

This motivates us to search the time index T^* by Algorithm 1 (line 17). It is clear from (2.3) that

$$\mathbb{E}[\phi(\mathbf{w}_{T^*}) - \phi(\bar{\mathbf{w}}_T)] \leq (T\eta_{T^*})^{-1} \mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)] + \eta_{T^*} \widetilde{C}. \tag{2.4}$$

The term $\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*)$ then can be estimated by the following error decomposition

$$\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*) = \mathbb{E}[\phi(\mathbf{w}_{T^*}) - \phi(\bar{\mathbf{w}}_T)] + \mathbb{E}[\phi(\bar{\mathbf{w}}_T) - \phi(\mathbf{w}^*)] \tag{2.5}$$

$$\leq (T\eta_{T^*})^{-1} \mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)] + \eta_{T^*} \widetilde{C} + \mathbb{E}[\phi(\bar{\mathbf{w}}_T) - \phi(\mathbf{w}^*)].$$

To derive the desired bound $\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*) = O(\eta_T)$, it suffices to show

$$(T\eta_{T^*})^{-1} \mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)] = O(\eta_T) \quad \text{and} \quad \mathbb{E}[\phi(\bar{\mathbf{w}}_T) - \phi(\mathbf{w}^*)] = O(\eta_T).$$

More specifically, we need to show $\mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)] = O(1), \mathbb{E}[\phi(\bar{\mathbf{w}}_T) - \phi(\mathbf{w}^*)] = O(T^{-\frac{1}{2}})$ in the convex case with $\eta_t = 1/\sqrt{t}$, and $\mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)] = O(T^{-1}), \mathbb{E}[\phi(\bar{\mathbf{w}}_T) - \phi(\mathbf{w}^*)] = O(T^{-1})$ in the strongly convex case with $\eta_t = 1/t$. We will show this is possible by choosing T^* in Algorithm 1.

3 Optimal Convergence Rates

We present here optimal convergence rates for SCMDI for both convex and strongly convex objectives. To this aim, we need to impose some standard assumptions. We assume that the mirror map Ψ is σ_Ψ-strongly convex w.r.t. a norm $\|\cdot\|$ in the sense $D_\Psi(\mathbf{w}, \tilde{\mathbf{w}}) \geq 2^{-1}\sigma_\Psi\|\mathbf{w} - \tilde{\mathbf{w}}\|^2$ for all $\mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^d$ ($\sigma_\Psi > 0$). We always assume the existence of A and $B > 0$ such that ($\|\cdot\|_*$ is the dual norm of $\|\cdot\|$)

$$\|f'(\mathbf{w}, z)\|_*^2 \leq Af(\mathbf{w}, z) + B \text{ and } \|r'(\mathbf{w})\|_*^2 \leq Ar(\mathbf{w}) + B \qquad (3.1)$$

for any $\mathbf{w} \in \mathbb{R}^d, z \in \mathcal{Z}$ and any $f'(\mathbf{w}, z) \in \partial f(\mathbf{w}, z), r'(\mathbf{w}) \in \partial r(\mathbf{w})$. Many popular learning methods use loss functions and regularizers satisfying (3.1) [7, 44]. For example, if $|\ell'(a, y)|^2 \leq \tilde{A}\ell(a, y) + \tilde{B}$ for some $\tilde{A}, \tilde{B} > 0$ and all a, y, then $f(\mathbf{w}, z) = \ell(\langle \mathbf{w}, x \rangle, y)$ would satisfy the first inequality of (3.1) if $\mathcal{X}$ is bounded. Here $\ell'(a, y)$ denotes a subgradient of ℓ w.r.t. the first argument. Examples of such ℓ include all smooth functions and all Lipschitz continuous functions widely used in machine learning [42, 44]. Examples of r satisfying (3.1) include $r(\mathbf{w}) = \lambda\|\mathbf{w}\|_p^p$ with $p \in [1, 2]$. We say Ψ is L_Ψ-smooth w.r.t. $\|\cdot\|$ if $D_\Psi(\mathbf{w}, \tilde{\mathbf{w}}) \leq \frac{L_\Psi}{2}\|\mathbf{w} - \tilde{\mathbf{w}}\|^2$ for all $\mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^d$.

We always assume the existence of $\sigma_F, \sigma_r \geq 0$ such that for all $\mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^d$

$$F(\mathbf{w}) - F(\tilde{\mathbf{w}}) - \langle \mathbf{w} - \tilde{\mathbf{w}}, F'(\tilde{\mathbf{w}}) \rangle \geq \sigma_F D_\Psi(\mathbf{w}, \tilde{\mathbf{w}}), \ r(\mathbf{w}) - r(\tilde{\mathbf{w}}) - \langle \mathbf{w} - \tilde{\mathbf{w}}, r'(\tilde{\mathbf{w}}) \rangle \geq \sigma_r D_\Psi(\mathbf{w}, \tilde{\mathbf{w}}) \qquad (3.2)$$

For $\sigma_\phi := \sigma_F + \sigma_r$, the cases $\sigma_\phi = 0$ and $\sigma_\phi > 0$ correspond to convex and strongly convex objectives, respectively.

3.1 Convex objectives

Our first result is an optimal rate $O(T^{-\frac{1}{2}})$ for convex objectives under a boundedness assumption of iterates. The boundedness assumptions $\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_t)] \leq D, \forall t \in \mathbb{N}$, imposed also in the literature [7, 8, 37], always hold for the regularizer of the form $r(\mathbf{w}) = I_{\mathcal{W}_0}(\mathbf{w}) + \tilde{r}(\mathbf{w})$, where $\mathcal{W}_0$ is a convex and compact domain, $I_{\mathcal{W}_0}$ is an indicator function with $I_{\mathcal{W}_0}(\mathbf{w}) = 0$ if $\mathbf{w} \in \mathcal{W}_0$ and ∞ otherwise, and $\tilde{r} : \mathbb{R}^d \to \mathbb{R}_+$ is convex. Proofs of theoretical results in this subsection are given in Appendix B.

Theorem 1. *Let $D > 0$. Assume $\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_t)] \leq D$ for all $t \in \mathbb{N}$ and $\mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)] \leq D$. Let $\mathbf{w}_{T^*}$ be defined by Algorithm 1 with $\eta_t = \mu/\sqrt{t}$ satisfying $\mu \leq \sigma_\Psi(2A)^{-1}$. Then, $\mathbb{E}[\phi(\mathbf{w}_{T^*}) - \phi(\mathbf{w}^*)] \leq \frac{\widetilde{C}_1}{\sqrt{T}}$, where $\widetilde{C}_1 = (4 + 2\sqrt{2})\mu^{-1}D + 10\mu\sigma_\Psi^{-1}(A\phi(\mathbf{w}^*) + 2B)$.*

Theorem 1 requires to impose a boundedness assumption on iterates, which is removed in the following theorem on convergence rates. The assumption $D_\Psi(\mathbf{w}, \tilde{\mathbf{w}}) \leq L_\Psi\|\mathbf{w} - \tilde{\mathbf{w}}\|^\alpha$ is milder than assuming a smoothness of Ψ, the latter of which is satisfied if $\Psi = \Psi_2$.

Theorem 2. *Let $\mathbf{w}_{T^*}$ be produced by Algorithm 1 with a non-increasing step size sequence and $\eta_1 \leq \sigma_\Psi(2A)^{-1}$. Assume that there exists $\alpha \in [0, 2]$ and $L_\Psi > 0$ such that $D_\Psi(\mathbf{w}, \tilde{\mathbf{w}}) \leq L_\Psi\|\mathbf{w} - \tilde{\mathbf{w}}\|^\alpha$ for all $\mathbf{w}, \tilde{\mathbf{w}} \in \mathbb{R}^d$. Let $D_T = D_\Psi(\mathbf{w}^*, \mathbf{w}_1) + \sigma_\Psi^{-1}(A\phi(\mathbf{w}^*) + 2B)\sum_{t=1}^T \eta_t^2$. Then,*

$$\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*) \leq \frac{2L_\Psi[4\sigma_\Psi^{-1}\alpha D_T + 1] + 4D_T}{T\eta_{2T}} + \frac{2A\phi(\mathbf{w}^*) + 4B}{\sigma_\Psi}\Big[\eta_T + \frac{2}{T}\sum_{t=1}^T \eta_t\Big]. \quad (3.3)$$

Corollary 3. *Suppose the assumptions in Theorem 2 hold. (a) If $\lim_{t\to\infty} \frac{\sum_{\tilde{t}=1}^t \eta_{\tilde{t}}^2}{t\eta_t} = 0$, then $\lim_{T\to\infty} \mathbb{E}[\phi(\mathbf{w}_{T^*})] = \phi(\mathbf{w}^*)$. (b) If $\eta_t = \frac{\mu}{\sqrt{T}}$ for $t = 1, \ldots, 2T$, then $\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*) = O(T^{-\frac{1}{2}})$.*

3.2 Strongly convex objectives

We now turn to Algorithm 1 for strongly convex objectives, towards which a first step is the following theorem on the performance of vanilla SCMD. Theorem 4 shows that both the suboptimality of some weighted averaged iterates and the Bregman distance between the last iterate and $\mathbf{w}^*$ converge with the rate $O(T^{-1})$. Theorem 4 is an extension of the results in [20] on the projected SGD to SCMD. Discussions in [20] require to impose a boundedness assumption on subgradients as $\mathbb{E}_Z[\|f'(\mathbf{w}_t, Z)\|_2^2] \leq \widetilde{D}$ for some $\widetilde{D} > 0$ and all $t \in \mathbb{N}$. Indeed, an independent section was included in [20] to show that this boundedness assumption holds for an SVM-like objective. Theorem 4 shows that we can derive the same convergence rate without this boundedness assumption and

5

is therefore more applicable. Theorem 5 gives a sufficient condition on step size sequences for the convergence. Upper and lower bounds on convergence rates are established in Theorem 6 and Theorem 7, respectively. The proofs of theoretical results in this subsection are given in Appendix C. If $\sigma_r = 0$, then $\bar{\mathbf{w}}_T$ defined in Theorem 4 can be simplified as $\bar{\mathbf{w}}_T = \left[\sum_{t=1}^{T} (t+1) \right]^{-1} \sum_{t=1}^{T} (t+1)\mathbf{w}_t$.

Theorem 4. *Assume* (3.2) *holds with* $\sigma_\phi > 0$. *Let* $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ *be generated by* (2.2) *with* $\eta_t = \frac{2}{\sigma_\phi t + 2\sigma_F}$. *Define*

$$\bar{\mathbf{w}}_T = \Big[\sum_{t=1}^{T} (t+1)(t+2)\eta_t \Big]^{-1} \sum_{t=1}^{T} (t+1)(t+2)\eta_t \mathbf{w}_t.$$

Then, there exists a constant $\widetilde{C}_2 > 0$ *independent of* T *and* σ_ϕ *(explicitly given in the proof) such that*

$$\mathbb{E}[\phi(\bar{\mathbf{w}}_T) - \phi(\mathbf{w}^*)] \leq \frac{4\widetilde{C}_2}{\sigma_\phi(T+1)} \quad and \quad \mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_{T+1})] \leq \frac{\widetilde{C}_2}{\sigma_\phi^2(T+2)}. \quad (3.4)$$

Theorem 5. *Assume* (3.2) *holds with* $\sigma_\phi > 0$. *Let* $\{\mathbf{w}_t\}_{t\in\mathbb{N}}$ *be generated by* (2.2). *If* $\lim_{t\to\infty} \eta_t = 0$ *and* $\sum_{t=1}^{\infty} \eta_t = \infty$, *then* $\lim_{T\to\infty} \mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_T)] = 0$.

With Theorem 4 at hand, we can provide optimal rates for the output produced by Algorithm 1.

Theorem 6. *Assume* (3.2) *holds with* $\sigma_\phi > 0$. *Let* $\mathbf{w}_{T^*}$ *be generated by Algorithm 1 with* $\eta_t = \frac{2}{\sigma_\phi t + 2\sigma_F}$. *Assume that* Ψ *is* L_Ψ*-strongly smooth* $(L_\Psi > 0)$. *If* $T \geq \frac{4A}{\sigma_\Psi \sigma_\phi}$, *then there exists a constant* $\widetilde{C}_3 > 0$ *independent of* T *and* σ_ϕ *(explicitly given in the proof) such that* $\mathbb{E}\big[\phi(\mathbf{w}_{T^*}) - \phi(\mathbf{w}^*)\big] \leq \frac{\widetilde{C}_3}{T\sigma_\phi}$.

Theorem 7 presents lower bounds matching the above upper bounds up to a constant factor, which shows that the selected iterate has achieved the best possible rate and there are no stronger results. The matching upper and lower bounds here apply to the specific SGD and specific optimization problems, while minimax bounds is related to the existence of a problem for any algorithm [1]. Similar results were derived in [21] for online mirror descent. Here we give a different and simpler proof.

Theorem 7. *Let* $\{\mathbf{w}_t\}_t$ *be the sequence produced by* (2.2) *with* $\Psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ *and* $r(\mathbf{w}) = 0$. *Suppose* f *is differentiable,* ϕ *is* L_ϕ*-smooth w.r.t.* $\|\cdot\|_2$ *and* $\eta_t \leq 1/(2L_\phi)$. *If for any* $\mathbf{w} \in \mathbb{R}^d$, $\mathbb{E}[\|\nabla f(\mathbf{w}, z) - \nabla F(\mathbf{w})\|_2^2] \geq \sigma^2$ *for some* $\sigma > 0$, *then*
$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2] \geq \min\{\|\mathbf{w}_1 - \mathbf{w}^*\|_2^2, \eta_1\sigma^2/(2L_\phi), \ldots, \eta_t\sigma^2/(2L_\phi)\}.$$

4 Discussions

We discuss related work on stochastic/online learning algorithms with different averaging schemes.

A very common scheme is to output an average of all iterates with uniform weights (UNI-AVE) [44, 46], which is able to attain the optimal rate $O(1/\sqrt{T})$ and the suboptimal rate $O(T^{-1}\log T)$ in the convex and strongly convex case, respectively [8]. A counterpart is established to show that the rate $O(T^{-1}\log T)$ is the best possible one for this simple averaging scheme [31].

For strongly convex objectives, the first algorithm able to attain the optimal rate $O(T^{-1})$ is the epoch-GD algorithm proposed in [14], which performs stochastic mirror descent in each epoch with a fixed step size. The step size decreases exponentially in each consecutive epoch and the averaged iterate of the last epoch is outputted as the solution. Since then, several other averaging schemes were developed to attain optimal rates, including a suffix-averaging scheme (SUFFIX) returning the uniform average of the last half of SGD iterates [31], a weighted averaging scheme (WEI-AVE) with a weight of $t + 1$ for $\mathbf{w}_t$ [20] and a polynomial-decay averaging scheme [37].

Motivated by the side effects of averaging in either destroying the sparsity of the solution or slowing down the training speed [31], the property of the last iterate (LAST) has also received a lot of attention. For smooth objective functions, SGD with the last iterate can achieve the optimal rate [31]. The analysis of last iterate for non-smooth objective functions is much more challenging, for which an interesting technique to relate the last iterate with suffix-average of iterates was developed [37, 44]. Based on this, rates $O(T^{-\frac{1}{2}}\log T)$ and $O(T^{-1}\log T)$ were established in the convex and strongly-convex cases [23, 37], respectively. These results motivate a natural question of whether the last iterate achieves the optimal rate in the non-smooth scenario. This open question was resolved in [13]

by constructing a problem for which SGD with the last iterate achieves at most the rate $\Omega(T^{-1} \log T)$. An interesting probabilistic rate $O(T^{-1} \log T)$ was also developed for SGD with the last iterate [13].

All the above mentioned methods require an averaging scheme to attain the optimal convergence rates, which, however, may destroy the sparsity of solutions and slow down the practical training speed. A simple scheme to preserve the sparsity of solutions while still enjoying the optimal rate is to draw an individual iterate *randomly* from the iterate sequence by (2.2) [10, 16]. Indeed, one can randomly draw the output according to a distribution over the iterate sequence with the probability mass function determined by any optimal weighted averaging scheme [20, 31, 37]. However, this scheme introduces new variances due to the choice of a random iterate as the output. Moreover, as we will verify in experiments, this scheme can slow down the practical training speed since the randomly selected iterate is not necessarily to be a favorable one in the iterate sequence. Furthermore, it requires to either store all the iterate sequence or determine the number of iterations beforehand, which does not apply to the online learning scenario. After the acceptance of the paper, we noticed an interesting step-size modification scheme to get optimal convergence rates [15]. However, the step-size scheme there requires to use the information of T, and therefore is not applicable to the online learning scenario. Furthermore, the algorithm and optimal convergence rates there are developed for the particular T-th iterate and not for other iterates.

In this paper, we propose a novel stochastic/online learning algorithm able to achieve optimal rates. As compared to the averaging scheme in [14, 20, 31, 37], our scheme of outputting an individual iterate is able to preserve the sparsity structure of solutions. As compared to the scheme of randomly choosing an individual iterate, our method avoids the added variance as well as the drawback of slowing down the training speed due to the random iterate index. Our idea is not to output the last iterate but to select an individual iterate based on a careful analysis on the one-step progress bound sketched in Section 2.3. Indeed, by (2.3), the quality of $\mathbf{w}_t$ depends on $A_t = D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_t) - D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{t+1})$. The smaller the A_t, the better the quality. This motivates us to select a T^* with small A_{T^*}. Intuitively, A_t is related to the distance between $\mathbf{w}_t$ and $\mathbf{w}_{t+1}$. Therefore, our selection of the iterate shares some spirit with the widely used heuristic of terminating the algorithm when the successive iterates are close. Our analysis also refines existing studies of optimal rates of SGD by removing the boundedness assumption on subgradients [20, 31, 37]. It should be mentioned that the boundedness assumption was relaxed as $\mathbb{E}_z[\|f'(\mathbf{w}_t, z)\|_*^2] \leq \widetilde{A} + \widetilde{B}\|F'(\mathbf{w}_t)\|_*^2$ in [4] for $\widetilde{A}, \widetilde{B} > 0$ and removed in [27], the latter of which, however, requires the objective function to be smooth.

5 Experimental Results

In this section, we justify the effectiveness of our algorithm by presenting experimental comparisons with the following averaging strategies: WEI-AVE [20], UNI-AVE, LAST, SUFFIX [31] and RAND (outputting a random iterate chosen from the uniform distribution over the last half of iterates). We consider two applications: binary classification and tomography reconstruction in image processing.

5.1 Binary classification

We first consider SVM models with linear kernels. We use 16 real-world datasets whose information is summarized in Table C.1 in Appendix D.1[3]. For each dataset, we use 80 percents of the data for training and reserve the remaining 20 percents for testing. The objective function we consider for a training dataset $\{(x_1, y_1), \ldots, (x_n, y_n)\}$ is

$$\phi(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \frac{1}{n}\sum_{i=1}^{n}\max\{0, 1 - y_i\langle\mathbf{w}, x_i\rangle\},$$

which is λ-strongly convex w.r.t. $\|\cdot\|_2$. Analogous to [20], we set λ to the reciprocal of training sample size. For datasets with multiple class labels, we group the first half of labels into the positive label, and the second half into the negative label. We consider step sizes of the form $\eta_t = \mu/(\lambda t)$ and tune the parameter μ in the set $2^{\{-12, -11, \ldots, 4\}}$ by 10-fold cross validation. We repeat the experiment 40 times and report the average of experimental results. We consider two approaches to optimize the objective function ϕ by whether to separate the regularizer and the data-fitting term. In the first approach, we apply (2.2) with $\Psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2, f(\mathbf{w}, z) = \frac{\lambda}{2}\|\mathbf{w}\|_2^2 + \max\{0, 1 - y\langle\mathbf{w}, x\rangle\}$ and $r(\mathbf{w}) = 0$,

---

[3]We display experimental results for 4 datasets here due to space limit. Complete results are in Appendix.

which is just the SGD. In the second approach, we apply (2.2) with $\Psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2, f(\mathbf{w}, z) = \max\{0, 1 - y\langle\mathbf{w}, x\rangle\}$ and $r(\mathbf{w}) = \frac{\lambda}{2}\|\mathbf{w}\|_2^2$, which is a stochastic proximal gradient descent (SPGD). In Figure 1 and Figure 2, we plot the objective function values on the testing dataset against iteration numbers for SGD and SPGD, respectively. It is clear that UNI-AVE is always the worst strategy in our experiments. WEI-AVE assigns more weights to recent iterates and can improve the performance, which is further outperformed by SUFFIX. RAND fluctuates a bit since the randomly selected index at the t-th iteration may not be an increasing function of t. LAST is overwritten by OCMDI due to their similar behavior, which behave best in our experiments especially in the beginning of the optimization process. The similarity between OCMDI and LAST can be explained as follows. If A_t (defined in Section 4) is small then it is likely that A_{t+1} is also small. Therefore, our algorithm is prone to select the last part of iterates. We also plot the objective function values on training datasets against T, which show similar behavior and are deferred to Figures D.4, D.3 in Appendix.
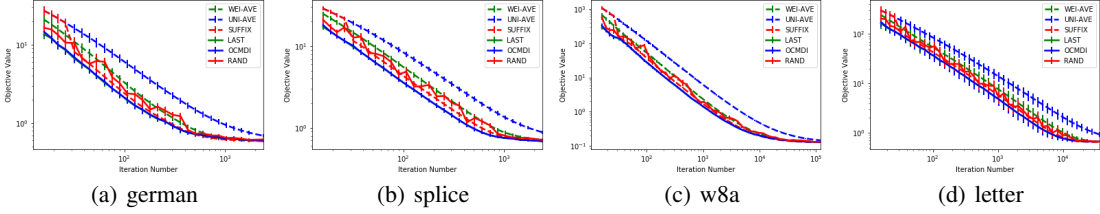

(a) german (b) splice (c) w8a (d) letter

Figure 1: Objective function values on test datasets versus iteration numbers for SGD.
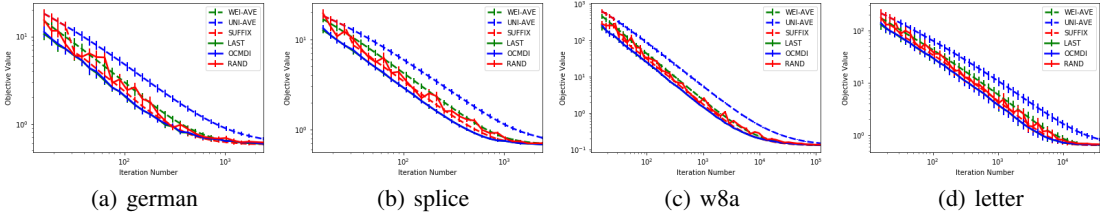

(a) german (b) splice (c) w8a (d) letter

Figure 2: Objective function values on test datasets versus iteration numbers for SPGD.

5.2 Tomography reconstruction

We now consider tomography reconstruction in image processing. We use AIR toolbox [12] to create a CT-measurement matrix $A \in \mathbb{R}^{n \times d}$, an output vector $\mathbf{y}^\dagger \in \mathbb{R}^n$ and a $N \times N$ sparse image encoded by a vector $\mathbf{w}^\dagger \in \mathbb{R}^d$ with $d = N^2$. Each row of A together with the corresponding row in $\mathbf{y}$ indicates a line integral from a fan bean projection geometry. Therefore, the true image $\mathbf{w}^\dagger$ satisfies $A\mathbf{w}^\dagger = \mathbf{y}^\dagger$. We consider a noisy case by adding Gaussian noise $N(0, (0.05|y_i^\dagger|)^2)$ to $y_i^\dagger$ and get the noisy output $\mathbf{y}$. Our aim is to reconstruct the image $\mathbf{w}^\dagger$ from the matrix A and the noisy measurements $\mathbf{y}$ by finding an approximate solution of the equation $A\mathbf{w} = \mathbf{y}$. Since many components of the true image $\mathbf{w}^\dagger$ varnish, we apply randomized sparse Kaczmarz algorithm, which is a simple and efficient algorithm to generate sparse approximate solutions to linear systems [5, 34]. Let $\epsilon > 0, \lambda > 0$ be two parameters and consider the mirror map $\Psi^{(\epsilon,\lambda)}(\mathbf{w}) = \lambda\sum_{i=1}^d g_\epsilon(w(i)) + \frac{1}{2}\|\mathbf{w}\|_2^2$, where $g_\epsilon(\xi) = \frac{\xi^2}{2\epsilon}$ for $|\xi| \leq \epsilon$ and $|\xi| - \frac{\epsilon}{2}$ for $|\xi| > \epsilon$. At each iteration t, we randomly select an index i_t from the uniform distribution over $\{1, \ldots, n\}$ and denote $x_{i_t} = A_{i_t}^\top$, where $A_{i_t}^\top$ is the transpose of A_{i_t}. Given $\mathbf{w}_1 \in \mathbb{R}^d$ and $\mathbf{v}_1 = \nabla\Psi^{(\epsilon,\lambda)}(\mathbf{w}_1)$, the randomized sparse Kaczmarz algorithm updates the model as

$$\mathbf{v}_{t+1} = \mathbf{v}_t - \eta_t(\langle\mathbf{w}_t, x_{i_t}\rangle - y_{i_t})x_{i_t}, \quad \mathbf{w}_{t+1} = S_{\lambda,\epsilon}(\mathbf{v}_{t+1}), \tag{5.1}$$

where $S_{\lambda,\epsilon} : \mathbb{R}^d \mapsto \mathbb{R}^d$ is defined component-wisely by the soft-thresholding function $S_{\lambda,\epsilon} : \mathbb{R} \mapsto \mathbb{R}$ given as [5]

$$S_{\lambda,\epsilon}(\xi) := \begin{cases} (\xi\epsilon)/(\lambda + \epsilon), & \text{if } |\xi| \leq \lambda + \epsilon, \\ \text{sgn}(\xi)\max(|\xi| - \lambda, 0), & \text{otherwise.} \end{cases}$$

Here $\text{sgn}(a)$ denotes the sign of $a \in \mathbb{R}$. Algorithm (5.1) can be equivalently formulated as [22]

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w} \in \mathbb{R}^d} \eta_t\langle\mathbf{w} - \mathbf{w}_t, (\langle\mathbf{w}_t, x_{i_t}\rangle - y_{i_t})x_{i_t}\rangle + D_{\Psi^{(\epsilon,\lambda)}}(\mathbf{w}, \mathbf{w}_t).$$

8

It is clear from the above equivalent formulation that (5.1) is an instantiation of (2.2) with $f(\mathbf{w}, z) = \frac{1}{2}(\langle \mathbf{w}, x \rangle - y)^2$, $r(\mathbf{w}) = 0$ and $\Psi(\mathbf{w}) = \Psi^{(\epsilon, \lambda)}(\mathbf{w})$ to minimize $F(\mathbf{w}) = \frac{1}{n}\|A\mathbf{w} - \mathbf{y}\|_2^2$. It was shown that $F(\mathbf{w})$ satisfies (3.2) with $\sigma_F = \sigma_{\min}(A^\top A/n)$, $\mathbf{w} = \mathbf{w}^*$ and $\tilde{\mathbf{w}} = \mathbf{w}_t, t \in \mathbb{N}$ [22], where $\sigma_{\min}(\widetilde{A})$ denotes the minimal positive eigenvalue of a matrix $\widetilde{A}$. Therefore, our theoretical analysis in Section 3.2 applies[4]. We randomly choose $\mathbf{w}_1$ from the uniform distribution in $[0, 1]^d$ and set $\lambda = 1, \epsilon = 10^{-8}$ as suggested in [5]. We repeat the experiment 40 times and report the average of results.
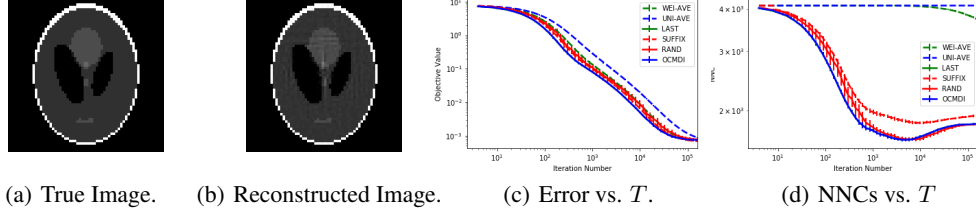

(a) True Image.	(b) Reconstructed Image.	(c) Error vs. T.	(d) NNCs vs. T

Figure 3: Tomography reconstruction with $N = 64, n = 23040$ and 5% relative noise. Panel (a) and (b) are the true image and the reconstructed image by OCMDI, respectively. Panel (c) and (d) plot the errors and NNCs versus iteration numbers.

In Figure 3, we present results with $N = 64, d = 4096$ and $n = 23040$. Panels (a) and (b) display the true image and the image output by OCMDI, respectively. In Panels (c) and (d), we plot the errors and the number of non-zero components (NNCs) for the output of algorithms with different averaging strategies. According to Figure 3, LAST is overwritten by OCMDI since they behave analogously, both of which achieve the fastest training speed among all considered methods. RAND is outperformed by SUFFIX in terms of training speed. Moreover, OCMDI, RAND and LAST can produce much more sparse solutions than other methods. Indeed, the true image $\mathbf{w}^\dagger$ has 1686 non-zero components, while UNI-AVE, WEI-AVE, SUFFIX, RAND and OCMDI produce outputs with 4096, 3727, 1943, 1823 and 1828 non-zero components, respectively. The advantage of OCMDI over LAST is that OCMDI can theoretically achieve optimal convergence rate. Similar phenomenon also appears for the case with $N = 32$ and $n = 11520$, which we plot in Figure D.5 in Appendix.

6 Conclusion

We propose a novel variant of SCMD with optimal convergence rates. An advantage of this algorithm over existing optimal learning algorithms is that it outputs an individual iterate without a random selection of index, which is able to preserve the sparsity structure without slowing down the training speed. Experimental results in both the domain of binary classification and tomography reconstruction demonstrate the ability of our algorithm in getting a fast training speed as well as in producing sparse solutions. Some interesting work include a theoretical justification on the advantage of sparsity of our method over other methods [11, 38] and an extension of our analysis to non-convex problems [10, 32, 45].

References

[1] A. Agarwal, M. J. Wainwright, P. L. Bartlett, and P. K. Ravikumar. Information-theoretic lower bounds on the oracle complexity of convex optimization. In *Advances in Neural Information Processing Systems*,

---

[4]In our analysis, we only use (3.2) for $\mathbf{w} = \mathbf{w}^*$ and $\tilde{\mathbf{w}} = \mathbf{w}_t$ (a restricted strong convexity in literature).

pages 1–9, 2009.

[2] F. Bach and E. Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate O(1/n). In *Advances in Neural Information Processing Systems*, pages 773–781, 2013.

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

[4] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

[5] J.-F. Cai, S. Osher, and Z. Shen. Linearized bregman iterations for compressed sensing. *Mathematics of Computation*, 78(267):1515–1536, 2009.

[6] N. Cesa-Bianchi, A. Conconi, and C. Gentile. On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, 50(9):2050–2057, 2004.

[7] J. Duchi and Y. Singer. Efficient online and batch learning using forward backward splitting. In *Advances in Neural Information Processing Systems*, pages 495–503, 2009.

[8] J. Duchi, S. Shalev-Shwartz, Y. Singer, and A. Tewari. Composite objective mirror descent. In *Conference on Learning Theory*, pages 14–26, 2010.

[9] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159, 2011.

[10] S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

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

[12] P. C. Hansen and M. Saxild-Hansen. Air tools—a matlab package of algebraic iterative reconstruction methods. *Journal of Computational and Applied Mathematics*, 236(8):2167–2178, 2012.

[13] N. J. A. Harvey, C. Liaw, Y. Plan, and S. Randhawa. Tight analyses for non-smooth stochastic gradient descent. In A. Beygelzimer and D. Hsu, editors, *Conference on Learning Theory*, pages 1579–1613, 2019.

[14] E. Hazan and S. Kale. Beyond the regret minimization barrier: optimal algorithms for stochastic strongly-convex optimization. *Journal of Machine Learning Research*, 15(1):2489–2512, 2014.

[15] P. Jain, D. Nagaraj, and P. Netrapalli. Making the last iterate of sgd information theoretically optimal. In A. Beygelzimer and D. Hsu, editors, *Conference on Learning Theory*, pages 1752–1755, 2019.

[16] R. Johnson and T. Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

[17] S. M. Kakade and A. Tewari. On the generalization ability of online strongly convex programming algorithms. In *Advances in Neural Information Processing Systems*, pages 801–808, 2009.

[18] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[19] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*, pages 1097–1105, 2012.

[20] S. Lacoste-Julien, M. Schmidt, and 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.

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

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

[23] J. Lin, L. Rosasco, and D.-X. Zhou. Iterative regularization for learning with convex loss functions. *Journal of Machine Learning Research*, 17(77):1–38, 2016.

[24] P.-L. Lions and B. Mercier. Splitting algorithms for the sum of two nonlinear operators. *SIAM Journal on Numerical Analysis*, 16(6):964–979, 1979.

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

[26] Y. Nesterov. *Introductory Lectures on Convex Optimization: A Basic Course*, volume 87. Springer Science & Business Media, 2013.

[27] L. M. Nguyen, P. H. Nguyen, M. van Dijk, P. Richtárik, K. Scheinberg, and M. Takác. SGD and hogwild! convergence without the bounded gradients assumption. In *International Conference on Machine Learning*, pages 3747–3755, 2018.

[28] N. Parikh and S. P. Boyd. Proximal algorithms. *Foundations and Trends in optimization*, 1(3):127–239, 2014.

[29] B. T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

[30] B. T. Polyak and A. B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, 1992.

[31] A. Rakhlin, O. Shamir, and K. Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *International Conference on Machine Learning*, pages 449–456, 2012.

[32] S. Reddi, A. Hefny, S. Sra, B. Poczos, and A. Smola. Stochastic variance reduction for nonconvex optimization. In *International Conference on Machine Learning*, pages 314–323, 2016.

[33] M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

[34] F. Schöpfer and D. A. Lorenz. Linear convergence of the randomized sparse Kaczmarz method. *Mathematical Programming*, pages 1–28, 2018.

[35] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter. Pegasos: Primal estimated sub-gradient solver for svm. *Mathematical programming*, 127(1):3–30, 2011.

[36] O. Shamir. Open problem: Is averaging needed for strongly convex stochastic gradient descent? 2012.

[37] O. Shamir and T. Zhang. Stochastic gradient descent for non-smooth optimization convergence results and optimal averaging schemes. In *International Conference on Machine Learning*, pages 71–79, 2013.

[38] J. Steinhardt, S. Wager, and P. Liang. The statistics of streaming sparse regression. *arXiv preprint arXiv:1412.4182*, 2014.

[39] L. Xiao. Dual averaging methods for regularized stochastic learning and online optimization. *Journal of Machine Learning Research*, 11:2543–2596, 2010.

[40] L. Xiao and T. Zhang. A proximal stochastic gradient method with progressive variance reduction. *SIAM Journal on Optimization*, 24(4):2057–2075, 2014.

[41] Y. Xu, Q. Lin, and T. Yang. Stochastic convex optimization: Faster local growth implies faster global convergence. In *International Conference on Machine Learning*, pages 3821–3830, 2017.

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

[43] L. Zhang, M. Mahdavi, and R. Jin. Linear convergence with condition number independent access of full gradients. In *Advances in Neural Information Processing Systems*, pages 980–988, 2013.

[44] T. Zhang. Solving large scale linear prediction problems using stochastic gradient descent algorithms. In *International Conference on Machine Learning*, pages 919–926, 2004.

[45] Z. Zhou, P. Mertikopoulos, N. Bambos, S. Boyd, and P. W. Glynn. Stochastic mirror descent in variationally coherent optimization problems. In *Advances in Neural Information Processing Systems*, pages 7043–7052, 2017.

[46] M. Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *International Conference on Machine Learning*, pages 928–936, 2003.

Supplementary Material for "Optimal Stochastic and Online Learning with Individual Iterates"

Yunwen Lei[1,2]   **Peng Yang**[1]   **Ke Tang**[1*]   **Ding-Xuan Zhou**[3]

[1]University Key Laboratory of Evolving Intelligent Systems of Guangdong Province,
Department of Computer Science and Engineering,
Southern University of Science and Technology, Shenzhen 518055, China

[2]Department of Computer Science,
Technical University of Kaiserslautern, Kaiserslautern 67653, Germany

[3]School of Data Science and Department of Mathematics,
City University of Hong Kong, Kowloon, Hong Kong, China

{leiyw, yangp, tangk3}@sustech.edu.cn   mazhou@cityu.edu.hk

A   Motivation of Algorithms

In this section, we provide rigorous theoretical results on the motivation of algorithms presented in Section 2.3. The first result (Lemma A.1) is a standard one-step progress bound which quantifies how the iterate would move towards $\mathbf{w}^*$ with a single update. The second result (Lemma A.2) shows that the time index T^* in Algorithm 1 can be always found. Therefore, the algorithm is well defined.

Lemma A.1. *Let the sequence $\{\mathbf{w}_t\}_{t \in \mathbb{N}}$ be generated by (2.2), then the following inequality holds for any $\mathbf{w} \in \mathbb{R}^d$ independent of z_t*

$$
\mathbb{E}_{z_t}[D_\Psi(\mathbf{w}, \mathbf{w}_{t+1})] - D_\Psi(\mathbf{w}, \mathbf{w}_t) \leq \eta_t\big[\phi(\mathbf{w}) - \phi(\mathbf{w}_t)\big] + \sigma_\Psi^{-1}\eta_t^2\big[A\phi(\mathbf{w}_t) + 2B\big]
$$
$$
- \eta_t\Big[\sigma_F D_\Psi(\mathbf{w}, \mathbf{w}_t) + \sigma_r \mathbb{E}_{z_t}[D_\Psi(\mathbf{w}, \mathbf{w}_{t+1})]\Big]. \quad \text{(A.1)}
$$

Proof. According to the first-order optimality condition in (2.2), there exists an $r'(\mathbf{w}_{t+1}) \in \partial r(\mathbf{w}_{t+1})$ satisfying

$$
\eta_t f'(\mathbf{w}_t, z_t) + \eta_t r'(\mathbf{w}_{t+1}) + \nabla\Psi(\mathbf{w}_{t+1}) - \nabla\Psi(\mathbf{w}_t) = 0,
$$

from which and the identity [1]

$$
D_\Psi(\mathbf{w}, \mathbf{w}_{t+1}) + D_\Psi(\mathbf{w}_{t+1}, \mathbf{w}_t) - D_\Psi(\mathbf{w}, \mathbf{w}_t) = \langle \mathbf{w} - \mathbf{w}_{t+1}, \nabla\Psi(\mathbf{w}_t) - \nabla\Psi(\mathbf{w}_{t+1})\rangle
$$

we derive

$$
D_\Psi(\mathbf{w}, \mathbf{w}_{t+1}) - D_\Psi(\mathbf{w}, \mathbf{w}_t) = D_\Psi(\mathbf{w}, \mathbf{w}_{t+1}) + D_\Psi(\mathbf{w}_{t+1}, \mathbf{w}_t) - D_\Psi(\mathbf{w}, \mathbf{w}_t) - D_\Psi(\mathbf{w}_{t+1}, \mathbf{w}_t)
$$
$$
= \langle \mathbf{w} - \mathbf{w}_{t+1}, \nabla\Psi(\mathbf{w}_t) - \nabla\Psi(\mathbf{w}_{t+1})\rangle - D_\Psi(\mathbf{w}_{t+1}, \mathbf{w}_t)
$$
$$
= \eta_t\langle \mathbf{w} - \mathbf{w}_{t+1}, f'(\mathbf{w}_t, z_t) + r'(\mathbf{w}_{t+1})\rangle - D_\Psi(\mathbf{w}_{t+1}, \mathbf{w}_t)
$$
$$
\leq \eta_t\langle \mathbf{w} - \mathbf{w}_{t+1}, f'(\mathbf{w}_t, z_t)\rangle + \eta_t\big[r(\mathbf{w}) - r(\mathbf{w}_{t+1}) - \sigma_r D_\Psi(\mathbf{w}, \mathbf{w}_{t+1})\big] - D_\Psi(\mathbf{w}_{t+1}, \mathbf{w}_t)
$$
$$
= \eta_t\langle \mathbf{w} - \mathbf{w}_t, f'(\mathbf{w}_t, z_t)\rangle + \eta_t\langle \mathbf{w}_t - \mathbf{w}_{t+1}, f'(\mathbf{w}_t, z_t)\rangle + \eta_t[r(\mathbf{w}) - r(\mathbf{w}_t)]
$$
$$
+ \eta_t[r(\mathbf{w}_t) - r(\mathbf{w}_{t+1})] - \sigma_r\eta_t D_\Psi(\mathbf{w}, \mathbf{w}_{t+1}) - D_\Psi(\mathbf{w}_{t+1}, \mathbf{w}_t). \quad \text{(A.2)}
$$

---

*Corresponding author

Here, we have used the σ_r-strong convexity of r (3.2) in the inequality. From the convexity of r, the definition of dual norm and the strong convexity of Ψ, it follows that

$$\eta_t\big[\langle \mathbf{w}_t - \mathbf{w}_{t+1}, f'(\mathbf{w}_t, z_t)\rangle + r(\mathbf{w}_t) - r(\mathbf{w}_{t+1})\big] - D_\Psi(\mathbf{w}_{t+1}, \mathbf{w}_t)$$
$$\leq \eta_t\|\mathbf{w}_t - \mathbf{w}_{t+1}\|\|f'(\mathbf{w}_t, z_t)\|_* + \eta_t\langle \mathbf{w}_t - \mathbf{w}_{t+1}, r'(\mathbf{w}_t)\rangle - 2^{-1}\sigma_\Psi\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2$$
$$\leq \eta_t\|\mathbf{w}_t - \mathbf{w}_{t+1}\|\big[\|f'(\mathbf{w}_t, z_t)\|_* + \|r'(\mathbf{w}_t)\|_*\big] - 2^{-1}\sigma_\Psi\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2$$
$$\leq 2^{-1}\sigma_\Psi\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2 + 2^{-1}\sigma_\Psi^{-1}\eta_t^2\big[\|f'(\mathbf{w}_t, z_t)\|_* + \|r'(\mathbf{w}_t)\|_*\big]^2 - 2^{-1}\sigma_\Psi\|\mathbf{w}_t - \mathbf{w}_{t+1}\|^2$$
$$\leq \sigma_\Psi^{-1}\eta_t^2\big[\|f'(\mathbf{w}_t, z_t)\|_*^2 + \|r'(\mathbf{w}_t)\|_*^2\big] \leq \sigma_\Psi^{-1}\eta_t^2\big[Af(\mathbf{w}_t, z_t) + Ar(\mathbf{w}_t) + 2B\big],$$

where we have used the elementary inequality $(a + b)^2 \leq 2(a^2 + b^2)$ and (3.1) in the last two inequalities. Plugging the above inequality back into (A.2), we get

$$D_\Psi(\mathbf{w}, \mathbf{w}_{t+1}) - D_\Psi(\mathbf{w}, \mathbf{w}_t) \leq \eta_t\langle \mathbf{w} - \mathbf{w}_t, f'(\mathbf{w}_t, z_t)\rangle + \eta_t[r(\mathbf{w}) - r(\mathbf{w}_t)]$$
$$+ \sigma_\Psi^{-1}\eta_t^2\big[Af(\mathbf{w}_t, z_t) + Ar(\mathbf{w}_t) + 2B\big] - \sigma_r\eta_t D_\Psi(\mathbf{w}, \mathbf{w}_{t+1}).$$

Taking conditional expectations with respect to z_t over both sides and using the σ_F-strong convexity of F (3.2) then give (note both $\mathbf{w}$ and $\mathbf{w}_t$ are independent of z_t)

$$\mathbb{E}_{z_t}[D_\Psi(\mathbf{w}, \mathbf{w}_{t+1})] - D_\Psi(\mathbf{w}, \mathbf{w}_t) + \sigma_r\eta_t\mathbb{E}_{z_t}[D_\Psi(\mathbf{w}, \mathbf{w}_{t+1})]$$
$$\leq \eta_t\langle \mathbf{w} - \mathbf{w}_t, F'(\mathbf{w}_t)\rangle + \eta_t[r(\mathbf{w}) - r(\mathbf{w}_t)] + \sigma_\Psi^{-1}\eta_t^2\big[AF(\mathbf{w}_t) + Ar(\mathbf{w}_t) + 2B\big]$$
$$\leq \eta_t\big[F(\mathbf{w}) - F(\mathbf{w}_t) - \sigma_F D_\Psi(\mathbf{w}, \mathbf{w}_t)\big] + \eta_t[r(\mathbf{w}) - r(\mathbf{w}_t)] + \sigma_\Psi^{-1}\eta_t^2\big[A\phi(\mathbf{w}_t) + 2B\big].$$

This gives the stated inequality and completes the proof. □

Lemma A.2. *There exists an $t \in \{T, T+1, \ldots, 2T-1\}$ such that*

$$D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_t) - D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{t+1}) \leq T^{-1}D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T).$$

Proof. We prove this lemma by contradiction. Suppose that

$$D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_t) - D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{t+1}) > T^{-1}D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T), \quad \forall t \in \{T, T+1, \ldots, 2T-1\}.$$

Taking a summation of the above inequality from $t = T$ to $t = 2T - 1$ gives

$$D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T) - D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{2T}) = \sum_{t=T}^{2T-1}\big[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_t) - D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{t+1})\big]$$
$$> T^{-1}\sum_{t=T}^{2T-1}D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T) = D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T),$$

which contradicts with the non-negativity of the Bregman distance. The proof is complete. □

B Proofs of Convergence Rates for Convex Objectives

In this section, we give proofs of Theorem 1, Theorem 2 and Corollary 3 on the performance of Algorithm 1 for convex objectives.

Proof of Theorem 1. Choosing $\mathbf{w} = \mathbf{w}^*$ in (A.1) and using $\sigma_r, \sigma_F \geq 0$, we derive

$$\mathbb{E}_{z_t}[D_\Psi(\mathbf{w}^*, \mathbf{w}_{t+1})] - D_\Psi(\mathbf{w}^*, \mathbf{w}_t) \leq \eta_t\big[\phi(\mathbf{w}^*) - \phi(\mathbf{w}_t)\big] + \sigma_\Psi^{-1}\eta_t^2\big[A\phi(\mathbf{w}_t) + 2B\big].$$

Taking expectations over both sides then gives

$$\mathbb{E}\big[D_\Psi(\mathbf{w}^*, \mathbf{w}_{t+1}) - D_\Psi(\mathbf{w}^*, \mathbf{w}_t)\big] \leq \eta_t\mathbb{E}\big[\phi(\mathbf{w}^*) - \phi(\mathbf{w}_t)\big] + \sigma_\Psi^{-1}\eta_t^2\big[A\mathbb{E}[\phi(\mathbf{w}_t)] + 2B\big]$$
$$= \eta_t\mathbb{E}\big[\phi(\mathbf{w}^*) - \phi(\mathbf{w}_t)\big] + \sigma_\Psi^{-1}\eta_t^2\big[A\mathbb{E}[\phi(\mathbf{w}_t)] - A\phi(\mathbf{w}^*) + A\phi(\mathbf{w}^*) + 2B\big]$$
$$= (\eta_t - \sigma_\Psi^{-1}\eta_t^2 A)\mathbb{E}\big[\phi(\mathbf{w}^*) - \phi(\mathbf{w}_t)\big] + \sigma_\Psi^{-1}\eta_t^2\big[A\phi(\mathbf{w}^*) + 2B\big]$$
$$\leq 2^{-1}\eta_t\mathbb{E}\big[\phi(\mathbf{w}^*) - \phi(\mathbf{w}_t)\big] + \sigma_\Psi^{-1}\eta_t^2\big[A\phi(\mathbf{w}^*) + 2B\big], \tag{B.1}$$

2

where we have used the inequality $\eta_t \leq \sigma_\Psi(2A)^{-1}$ and $\phi(\mathbf{w}^*) \leq \phi(\mathbf{w}_t)$. The above inequality can be rewritten as

$$2^{-1}\mathbb{E}\big[\phi(\mathbf{w}_t) - \phi(\mathbf{w}^*)\big] \leq \eta_t^{-1}\mathbb{E}\big[D_\Psi(\mathbf{w}^*, \mathbf{w}_t) - D_\Psi(\mathbf{w}^*, \mathbf{w}_{t+1})\big] + \sigma_\Psi^{-1}\eta_t(A\phi(\mathbf{w}^*) + 2B).$$

Taking a summation of the above inequality from $t = 1$ to $t = T$ gives

$$2^{-1}\sum_{t=1}^{T}\mathbb{E}\big[\phi(\mathbf{w}_t) - \phi(\mathbf{w}^*)\big]$$

$$\leq \sum_{t=1}^{T}\eta_t^{-1}\mathbb{E}\big[D_\Psi(\mathbf{w}^*, \mathbf{w}_t) - D_\Psi(\mathbf{w}^*, \mathbf{w}_{t+1})\big] + \sigma_\Psi^{-1}(A\phi(\mathbf{w}^*) + 2B)\sum_{t=1}^{T}\eta_t$$

$$\leq \sum_{t=2}^{T}\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_t)][\eta_t^{-1} - \eta_{t-1}^{-1}] + \eta_1^{-1}D_\Psi(\mathbf{w}^*, \mathbf{w}_1) + \sigma_\Psi^{-1}(A\phi(\mathbf{w}^*) + 2B)\sum_{t=1}^{T}\eta_t \quad \text{(B.2)}$$

$$\leq D\eta_T^{-1} + 2\sigma_\Psi^{-1}\mu(A\phi(\mathbf{w}^*) + 2B)\sqrt{T},$$

where we have used $\eta_t \leq \eta_{t-1}$ and the elementary inequality $\sum_{t=1}^{T} t^{-\frac{1}{2}} \leq 2\sqrt{T}$ in the last step. Let $\bar{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^{T}\mathbf{w}_t$ be defined in Algorithm 1 with $\sigma_\phi = 0$. The convexity of ϕ then gives

$$\mathbb{E}[\phi(\bar{\mathbf{w}}_T)] - \phi(\mathbf{w}^*) \leq \frac{2\mu^{-1}D + 4\sigma_\Psi^{-1}\mu(A\phi(\mathbf{w}^*) + 2B)}{\sqrt{T}}. \quad \text{(B.3)}$$

According to Lemma A.2 and the definition of T^* in Algorithm 1, we know

$$D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{T^*}) - D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{T^*+1}) \leq T^{-1}D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T). \quad \text{(B.4)}$$

Choosing $\mathbf{w} = \bar{\mathbf{w}}_T$ in (A.1) followed with expectations over both sides implies ($\bar{\mathbf{w}}_T$ is independent of z_t for any $t \geq T$)

$$\eta_t\mathbb{E}\big[\phi(\mathbf{w}_t) - \phi(\bar{\mathbf{w}}_T)\big] \leq \mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_t) - D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{t+1})] + \sigma_\Psi^{-1}\eta_t^2\big[A\mathbb{E}[\phi(\mathbf{w}_t)] + 2B\big], \quad \forall t \geq T.$$

Choosing $t = T^*$ in the above inequality and using (B.4) then give

$$\eta_{T^*}\mathbb{E}\big[\phi(\mathbf{w}_{T^*}) - \phi(\bar{\mathbf{w}}_T)\big] \leq \mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{T^*}) - D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_{T^*+1})] + \sigma_\Psi^{-1}\eta_{T^*}^2\big[A\mathbb{E}[\phi(\mathbf{w}_{T^*})] + 2B\big]$$

$$\leq T^{-1}\mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)] + \sigma_\Psi^{-1}\eta_{T^*}^2\big[A\mathbb{E}[\phi(\mathbf{w}_{T^*})] + 2B\big]. \quad \text{(B.5)}$$

Plugging the step size $\eta_t = \mu/\sqrt{t}$ into the above inequality and using $T \leq T^* \leq 2T$, we derive

$$\mathbb{E}\big[\phi(\mathbf{w}_{T^*}) - \phi(\bar{\mathbf{w}}_T)\big] \leq \frac{\sqrt{2}\mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)]}{\mu\sqrt{T}} + \frac{\mu\big[A\mathbb{E}[\phi(\mathbf{w}_{T^*})] + 2B\big]}{\sigma_\Psi\sqrt{T}},$$

from which and (B.3) it further follows that

$$\mathbb{E}\big[\phi(\mathbf{w}_{T^*}) - \phi(\mathbf{w}^*)\big] = \mathbb{E}\big[\phi(\mathbf{w}_{T^*}) - \phi(\bar{\mathbf{w}}_T)\big] + \mathbb{E}\big[\phi(\bar{\mathbf{w}}_T) - \phi(\mathbf{w}^*)\big] \leq \frac{\sqrt{2}\mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)]}{\mu\sqrt{T}}$$

$$+ \frac{\mu A\big(\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*)\big)}{\sigma_\Psi\sqrt{T}} + \frac{\mu[A\phi(\mathbf{w}^*) + 2B]}{\sigma_\Psi\sqrt{T}} + \frac{2\mu^{-1}D + 4\sigma_\Psi^{-1}\mu(A\phi(\mathbf{w}^*) + 2B)}{\sqrt{T}} \quad \text{(B.6)}$$

$$\leq \frac{\sqrt{2}D}{\mu\sqrt{T}} + \frac{\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*)}{2} + \frac{\mu[A\phi(\mathbf{w}^*) + 2B]}{\sigma_\Psi\sqrt{T}} + \frac{2\mu^{-1}D + 4\sigma_\Psi^{-1}\mu(A\phi(\mathbf{w}^*) + 2B)}{\sqrt{T}},$$

where the last inequality is due to $\mu \leq \sigma_\Psi(2A)^{-1}$ and $\phi(\mathbf{w}_{T^*}) \geq \phi(\mathbf{w}^*)$. The above inequality can be written as the stated inequality. The proof is complete. $\qquad\square$

Proof of Theorem 2. Taking a summation of (B.1) from $t = 1$ to $t = \tilde{t}$, we derive the following inequality for any $\tilde{t} = 0, \dots, T - 1$

$$\mathbb{E}\big[D_\Psi(\mathbf{w}^*, \mathbf{w}_{\tilde{t}+1})\big] = D_\Psi(\mathbf{w}^*, \mathbf{w}_1) + \sum_{t=1}^{\tilde{t}}\mathbb{E}\big[D_\Psi(\mathbf{w}^*, \mathbf{w}_{t+1}) - D_\Psi(\mathbf{w}^*, \mathbf{w}_t)\big]$$

$$\leq D_\Psi(\mathbf{w}^*, \mathbf{w}_1) + \frac{1}{2}\sum_{t=1}^{\tilde{t}}\eta_t\mathbb{E}\big[\phi(\mathbf{w}^*) - \phi(\mathbf{w}_t)\big] + \sigma_\Psi^{-1}\big[A\phi(\mathbf{w}^*) + 2B\big]\sum_{t=1}^{\tilde{t}}\eta_t^2$$

$$\leq D_{\tilde{t}}, \quad \text{(B.7)}$$

where we have used $\phi(\mathbf{w}^*) \leq \phi(\mathbf{w}_t)$ and $D_{\tilde{t}}$ is defined in Theorem 2. According to the Young's inequality

$$ab \leq \frac{a^s}{s} + \frac{b^{\tilde{s}}}{\tilde{s}} \quad \text{for all } a, b, s, \tilde{s} > 0 \text{ satisfying } \frac{1}{s} + \frac{1}{\tilde{s}} = 1,$$

the following inequality holds for all $t = 1, 2, \ldots, T$

$$
\begin{aligned}
D_\Psi(\mathbf{w}_t, \mathbf{w}_T) \leq L_\Psi \|\mathbf{w}_t - \mathbf{w}_T\|^\alpha &\leq L_\Psi \left[2^{-1}\alpha \|\mathbf{w}_t - \mathbf{w}_T\|^{\alpha \cdot \frac{2}{\alpha}} + 1 - 2^{-1}\alpha \right] \\
&= L_\Psi \left[2^{-1}\alpha \|\mathbf{w}_t - \mathbf{w}^* + \mathbf{w}^* - \mathbf{w}_T\|^2 + 1 - 2^{-1}\alpha \right] \\
&\leq L_\Psi \left[\alpha \|\mathbf{w}_t - \mathbf{w}^*\|^2 + \alpha \|\mathbf{w}^* - \mathbf{w}_T\|^2 + 1 - 2^{-1}\alpha \right] \\
&\leq L_\Psi \left[2\sigma_\Psi^{-1}\alpha D_\Psi(\mathbf{w}^*, \mathbf{w}_t) + 2\sigma_\Psi^{-1}\alpha D_\Psi(\mathbf{w}^*, \mathbf{w}_T) + 1 - 2^{-1}\alpha \right],
\end{aligned}
$$

which, coupled with the convexity of $g(\cdot) := D_\Psi(\cdot, \mathbf{w}_T)$ as a function on $\mathbb{R}^d$ and (B.7), then implies

$$\mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)] \leq \frac{1}{T} \sum_{t=1}^{T} \mathbb{E}[D_\Psi(\mathbf{w}_t, \mathbf{w}_T)] \leq L_\Psi \left[4\sigma_\Psi^{-1}\alpha D_T + 1 - 2^{-1}\alpha \right], \tag{B.8}$$

where $\bar{\mathbf{w}}_T$ is defined in Algorithm 1 with $\sigma_\phi = 0$. Analyzing analogously to (B.2) excepting using $\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_t)] \leq D_T$ for all $t \leq T$ (η_t is a non-increasing sequence), we derive

$$\mathbb{E}[\phi(\bar{\mathbf{w}}_T)] - \phi(\mathbf{w}^*) \leq \frac{2(A\phi(\mathbf{w}^*) + 2B) \sum_{t=1}^{T} \eta_t}{T\sigma_\Psi} + \frac{2D_T}{T\eta_T}. \tag{B.9}$$

Eq. (B.5) implies further

$$\mathbb{E}\left[\phi(\mathbf{w}_{T^*}) - \phi(\bar{\mathbf{w}}_T)\right] \leq (T\eta_{T^*})^{-1}\mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)] + \sigma_\Psi^{-1}\eta_{T^*}\left[A\mathbb{E}[\phi(\mathbf{w}_{T^*})] + 2B\right].$$

Combining the above inequality, (B.8), (B.9) together and using $T \leq T^* \leq 2T$, we get

$$
\begin{aligned}
\mathbb{E}\left[\phi(\mathbf{w}_{T^*}) - \phi(\mathbf{w}^*)\right] = \mathbb{E}\left[\phi(\mathbf{w}_{T^*}) - \phi(\bar{\mathbf{w}}_T)\right] + \mathbb{E}\left[\phi(\bar{\mathbf{w}}_T) - \phi(\mathbf{w}^*)\right] &\leq \frac{\mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)]}{T\eta_{2T}} \\
+ \frac{A\eta_T\left(\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*)\right)}{\sigma_\Psi} + \frac{\eta_T(A\phi(\mathbf{w}^*) + 2B)}{\sigma_\Psi} + \frac{2(A\phi(\mathbf{w}^*) + 2B)\sum_{t=1}^{T} \eta_t}{T\sigma_\Psi} &+ \frac{2D_T}{T\eta_T} \\
\leq \frac{L_\Psi\left(4\sigma_\Psi^{-1}\alpha D_T + 1\right) + 2D_T}{T\eta_{2T}} + \frac{\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*)}{2} + \frac{A\phi(\mathbf{w}^*) + 2B}{\sigma_\Psi}&\left(\eta_T + \frac{2}{T}\sum_{t=1}^{T} \eta_t\right),
\end{aligned}
$$

where the last inequality is due to $2A\eta_T \leq \sigma_\Psi$. The above inequality can be written as (3.3). The proof is complete. $\qquad\square$

Proof of Corollary 3. (a) Since $\{\eta_t\}$ is a non-increasing sequence, we know

$$0 \leq \lim_{t\to\infty} \eta_t = \lim_{t\to\infty} \frac{t\eta_t^2}{t\eta_t} \leq \lim_{t\to\infty} \frac{1}{t\eta_t} \sum_{\tilde{t}=1}^{t} \eta_{\tilde{t}}^2 = 0.$$

This in turn shows that the sequence of arithmetic mean also converges to 0, i.e., $\lim_{t\to\infty} \frac{1}{t}\sum_{\tilde{t}=1}^{t} \eta_{\tilde{t}} = 0$. Also, $\lim_{t\to\infty} \frac{1}{t\eta_t}\sum_{\tilde{t}=1}^{t} \eta_{\tilde{t}}^2 = 0$ immediately implies $\lim_{t\to\infty} t\eta_t = \infty$. Then, all the terms on the right-hand side of (3.3) converges to zero as T tends to ∞ and therefore $\lim_{T\to\infty} \mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*) = 0$.

(b) For the step size sequence $\eta_t = \mu/\sqrt{T}$, (3.3) translates to

$$\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*) \leq \frac{1}{\sqrt{T}}\left[2L_\Psi\mu^{-1}(4\sigma_\Psi^{-1}\alpha D_T + 1) + 6\mu\sigma_\Psi^{-1}(A\phi(\mathbf{w}^*) + 2B) + 4\mu^{-1}D_T\right],$$

which is of the stated form since in this case $D_T = D_\Psi(\mathbf{w}^*, \mathbf{w}_1) + \sigma_\Psi^{-1}[A\phi(\mathbf{w}^*) + 2B]\mu^2$ is a constant independent of T.

The proof is complete. $\qquad\square$

4

Remark 1. If we consider polynomially decaying step sizes $\eta_t = \frac{\mu}{\sqrt{t}}$ for all $t = 1, 2, \ldots, 2T$, then there exists a constant $\widetilde{C} > 0$ independent of T such that

$$\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*) \leq \frac{\widetilde{C}\log(eT)}{\sqrt{T}}.$$

Indeed, for the step size sequence $\eta_t = \mu/\sqrt{t}$, we have $\sum_{t=1}^{T}\eta_t \leq 2\mu\sqrt{T}$ and $\sum_{t=1}^{T}\eta_t^2 \leq \mu^2 \log(eT)$. Plugging these inequalities back into (3.3) gives

$$\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*) \leq \frac{1}{\sqrt{T}}\Big[2\sqrt{2}L_\Psi\mu^{-1}(4\sigma_\Psi^{-1}\alpha D_T + 1) + 10\mu\sigma_\Psi^{-1}(A\phi(\mathbf{w}^*) + 2B) + 4\sqrt{2}\mu^{-1}D_T\Big],$$

which is of the stated form since in this case $D_T \leq D_\Psi(\mathbf{w}^*, \mathbf{w}_1) + \sigma_\Psi^{-1}[A\phi(\mathbf{w}^*) + 2B]\mu^2 \log(eT)$.

C   Proofs of Convergence Rates for Strongly Convex Objectives

In this section, we present proofs of convergence rates (Theorem 4, Theorem 5, Theorem 6 and Theorem 7) for strongly convex objectives.

Proof of Theorem 4. Choosing $\mathbf{w} = \mathbf{w}^*$ in (A.1) and taking expectation over both sides, we derive the following inequality for any $t \in \mathbb{N}$

$$(1 + \sigma_r\eta_t)\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_{t+1})] \leq$$
$$(1 - \sigma_F\eta_t)\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_t)] + \eta_t\mathbb{E}[\phi(\mathbf{w}^*) - \phi(\mathbf{w}_t)] + \sigma_\Psi^{-1}\eta_t^2[A\mathbb{E}[\phi(\mathbf{w}_t)] + 2B]. \quad \text{(C.1)}$$

For any $t \geq t_0 := \max\{\lceil \frac{8A - 2\sigma_F\sigma_\Psi}{\sigma_\Psi\sigma_\phi}\rceil, 1\}$ ($\lceil a \rceil$ denotes the smallest integer not less than a), we have $\eta_t \leq \sigma_\Psi(4A)^{-1}$ and therefore derive

$$(1 + \sigma_r\eta_t)\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_{t+1})]$$
$$\leq (1 - \sigma_F\eta_t)\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_t)] + \big(\eta_t - \sigma_\Psi^{-1}\eta_t^2 A\big)\mathbb{E}[\phi(\mathbf{w}^*) - \phi(\mathbf{w}_t)] + \sigma_\Psi^{-1}\eta_t^2[A\phi(\mathbf{w}^*) + 2B]$$
$$\leq (1 - \sigma_F\eta_t)\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_t)] + \frac{3\eta_t}{4}\mathbb{E}[\phi(\mathbf{w}^*) - \phi(\mathbf{w}_t)] + \sigma_\Psi^{-1}\eta_t^2[A\phi(\mathbf{w}^*) + 2B], \quad \forall t \geq t_0.$$

Combining the above inequality and (C.1) together, we derive

$$(1 + \sigma_r\eta_t)\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_{t+1})] \leq (1 - \sigma_F\eta_t)\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_t)] + \frac{3\eta_t}{4}\mathbb{E}[\phi(\mathbf{w}^*) - \phi(\mathbf{w}_t)] + \eta_t^2\widetilde{C}_4 \ \forall t \in \mathbb{N}, \tag{C.2}$$

where

$$\widetilde{C}_4 = \sigma_\Psi^{-1}\Big[A\max_{t \leq t_0}\mathbb{E}[\phi(\mathbf{w}_t)] + 2B\Big]$$

is a constant independent of T. From (C.2), it then follows that

$$\frac{3\eta_t}{4(1 + \sigma_r\eta_t)}\mathbb{E}[\phi(\mathbf{w}_t) - \phi(\mathbf{w}^*)] + \mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_{t+1})] \leq \frac{1 - \sigma_F\eta_t}{1 + \sigma_r\eta_t}\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_t)] + \eta_t^2\widetilde{C}_4, \ \forall t \in \mathbb{N}. \tag{C.3}$$

Since $\eta_t = \frac{2}{\sigma_\phi t + 2\sigma_F}$, we know

$$\frac{1 - \sigma_F\eta_t}{1 + \sigma_r\eta_t} = \frac{\sigma_\phi t + 2\sigma_F - 2\sigma_F}{\sigma_\phi t + 2\sigma_F + 2\sigma_r} = \frac{t}{t + 2}$$

and

$$\frac{1}{1 + \sigma_r\eta_t} \geq \frac{1}{1 + \sigma_r\eta_1} = \frac{\sigma_\phi + 2\sigma_F}{3\sigma_\phi} \geq \frac{1}{3}.$$

Plugging the above two inequalities into Eq. (C.3), we derive

$$4^{-1}\eta_t\mathbb{E}[\phi(\mathbf{w}_t) - \phi(\mathbf{w}^*)] + \mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_{t+1})] \leq \frac{t}{t + 2}\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_t)] + \eta_t^2\widetilde{C}_4, \quad \forall t \in \mathbb{N}.$$

5

Multiplying both sides by $(t+1)(t+2)$ gives

$$4^{-1}(t+1)(t+2)\eta_t\mathbb{E}[\phi(\mathbf{w}_t)-\phi(\mathbf{w}^*)]+(t+1)(t+2)\mathbb{E}[D_\Psi(\mathbf{w}^*,\mathbf{w}_{t+1})]$$
$$\leq t(t+1)\mathbb{E}[D_\Psi(\mathbf{w}^*,\mathbf{w}_t)]+(t+1)(t+2)\eta_t^2\widetilde{C}_4,\ \forall t\in\mathbb{N}.$$

Taking a summation of the above inequality from $t=1$ to $t=T$ gives

$$4^{-1}\sum_{t=1}^{T}\Big[(t+1)(t+2)\eta_t\mathbb{E}[\phi(\mathbf{w}_t)-\phi(\mathbf{w}^*)]\Big]+(T+1)(T+2)\mathbb{E}[D_\Psi(\mathbf{w}^*,\mathbf{w}_{T+1})]$$

$$\leq 2D_\Psi(\mathbf{w}^*,\mathbf{w}_1)+\widetilde{C}_4\sum_{t=1}^{T}(t+1)(t+2)\eta_t^2\leq 2D_\Psi(\mathbf{w}^*,\mathbf{w}_1)+\frac{4\widetilde{C}_4}{\sigma_\phi^2}\sum_{t=1}^{T}\frac{(t+1)(t+2)}{t^2}$$

$$\leq 2D_\Psi(\mathbf{w}^*,\mathbf{w}_1)+\frac{4\widetilde{C}_4}{\sigma_\phi^2}\big[T+3\log(eT)+4\big]$$

$$\leq \sigma_\phi^{-2}\big[4\widetilde{C}_4(T+3e^{-1}T+7)+2D_\Psi(\mathbf{w}^*,\mathbf{w}_1)\sigma_\phi^2\big]\leq\widetilde{C}_2\sigma_\phi^{-2}T,$$

where we have used the elementary inequality $\log(T)\leq e^{-1}T$ for any $T\in\mathbb{N}$ and introduced

$$\widetilde{C}_2=4\widetilde{C}_4(3e^{-1}+8)+2D_\Psi(\mathbf{w}^*,\mathbf{w}_1)\sigma_\phi^2.$$

It then follows that

$$\mathbb{E}[D_\Psi(\mathbf{w}^*,\mathbf{w}_{T+1})]\leq\widetilde{C}_2\sigma_\phi^{-2}(T+2)^{-1}$$

and

$$\mathbb{E}[\phi(\bar{\mathbf{w}}_T)-\phi(\mathbf{w}^*)]\leq\frac{\sum_{t=1}^{T}(t+1)(t+2)\eta_t\mathbb{E}[\phi(\mathbf{w}_t)-\phi(\mathbf{w}^*)]}{\sum_{t=1}^{T}(t+1)(t+2)\eta_t}$$

$$\leq\frac{4\widetilde{C}_2T}{\sigma_\phi^2\sum_{t=1}^{T}(t+1)(t+2)\eta_t}\leq\frac{4\widetilde{C}_2}{(T+1)\sigma_\phi}.$$

Here we have used the inequality

$$\sum_{t=1}^{T}(t+1)(t+2)\eta_t\geq\sum_{t=1}^{T}\frac{2(t+1)(t+2)}{\sigma_\phi t}\geq\frac{2}{\sigma_\phi}\sum_{t=1}^{T}t=\frac{T(T+1)}{\sigma_\phi}.$$

The proof is complete. $\qquad\qquad\square$

We will use the following lemma to prove sufficient conditions for the convergence of SCMD established in Theorem 5. The following lemma is known in the literature (see, e.g., [3, 5]).

Lemma C.1. *Let $\{\eta_t\}_{t\in\mathbb{N}}$ be a sequence of non-negative numbers such that $\lim_{t\to\infty}\eta_t=0$ and $\sum_{t=1}^{\infty}\eta_t=\infty$. Let $a>0$ and $t_1\in\mathbb{N}$ such that $\eta_t<a^{-1}$ for any $t\geq t_1$. Then we have*

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

Proof of Theorem 5. Since $\lim_{t\to\infty}=0$, there exists a $\tilde{t}_0\in\mathbb{N}$ such that $\eta_t\leq\sigma_\Psi(4A)^{-1}$ and $\eta_t\leq\sigma_r^{-1}$ for all $t\geq\tilde{t}_0$. From (C.2) and $\eta_t\sigma_r\leq 1$ it follows that

$$\mathbb{E}[D_\Psi(\mathbf{w}^*,\mathbf{w}_{t+1})]\leq\frac{1+\sigma_r\eta_t-\sigma_\phi\eta_t}{1+\sigma_r\eta_t}\mathbb{E}[D_\Psi(\mathbf{w}^*,\mathbf{w}_t)]+\eta_t^2\widetilde{C}_5$$

$$\leq\big(1-2^{-1}\sigma_\phi\eta_t\big)\mathbb{E}[D_\Psi(\mathbf{w}^*,\mathbf{w}_t)]+\eta_t^2\widetilde{C}_5,\quad\forall t\geq\tilde{t}_0,$$

where

$$\widetilde{C}_5=\sigma_\Psi^{-1}\big[A\max_{t\leq\tilde{t}_0}\mathbb{E}[\phi(\mathbf{w}_t)]+2B\big].$$

6

Applying this inequality iteratively for $t = T, \ldots, \tilde{t}_0$ yields

$$\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_{T+1})] \leq \prod_{t=\tilde{t}_0}^{T}(1 - 2^{-1}\sigma_\phi\eta_t)\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_{\tilde{t}_0})] + \widetilde{C}_5 \sum_{t=\tilde{t}_0}^{T}\eta_t^2 \prod_{k=t+1}^{T}(1 - 2^{-1}\sigma_\phi\eta_k), \quad \text{(C.4)}$$

where we denote $\prod_{k=t+1}^{T}(1 - 2^{-1}\sigma_\phi\eta_k) = 1$ for $t = T$. The first term of the above inequality can be controlled by the standard inequality $1 - a \leq \exp(-a), a > 0$ together with $\sum_{t=1}^{\infty}\eta_t = \infty$

$$\lim_{T\to\infty}\prod_{t=\tilde{t}_0}^{T}(1 - 2^{-1}\sigma_\phi\eta_t)\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_{\tilde{t}_0})] \leq \lim_{T\to\infty}\prod_{t=\tilde{t}_0}^{T}\exp\big(-2^{-1}\sigma_\phi\eta_t\big)\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_{\tilde{t}_0})]$$

$$= \lim_{T\to\infty}\exp\Big(-2^{-1}\sigma_\phi\sum_{t=\tilde{t}_0}^{T}\eta_t\Big)\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_{\tilde{t}_0})] = 0.$$

Applying Lemma C.1 with $a = 2^{-1}\sigma_\phi$, we get

$$\lim_{T\to\infty}\sum_{t=\tilde{t}_0}^{T}\eta_t^2\prod_{k=t+1}^{T}(1 - 2^{-1}\sigma_\phi\eta_k) = 0.$$

Plugging the above two expressions into (C.4) completes the proof. $\qquad\square$

Proof of Theorem 6. According to the strong convexity of ϕ given in (3.2) and the optimality condition $0 \in \partial\phi(\mathbf{w}^*)$, we get

$$\mathbb{E}[\phi(\bar{\mathbf{w}}_T) - \phi(\mathbf{w}^*)] = \mathbb{E}[\phi(\bar{\mathbf{w}}_T) - \phi(\mathbf{w}^*) - \langle\bar{\mathbf{w}}_T - \mathbf{w}^*, 0\rangle] \geq \sigma_\phi\mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}^*)],$$

which, together with the strong convexity of Ψ and the first inequality of (3.4), gives

$$\mathbb{E}[\|\bar{\mathbf{w}}_T - \mathbf{w}^*\|^2] \leq \frac{2\mathbb{E}[\phi(\bar{\mathbf{w}}_T) - \phi(\mathbf{w}^*)]}{\sigma_\phi\sigma_\Psi} \leq \frac{8\widetilde{C}_2}{(T+1)\sigma_\phi^2\sigma_\Psi}.$$

It then follows from the L_Ψ-strong smoothness of Ψ, $\mathbb{E}[\|\mathbf{w}^* - \mathbf{w}_T\|^2] \leq \frac{2}{\sigma_\Psi}\mathbb{E}[D_\Psi(\mathbf{w}^*, \mathbf{w}_T)]$ and the second inequality of (3.4) that

$$\mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)] \leq 2^{-1}L_\Psi\mathbb{E}[\|\bar{\mathbf{w}}_T - \mathbf{w}_T\|^2] = 2^{-1}L_\Psi\mathbb{E}[\|\bar{\mathbf{w}}_T - \mathbf{w}^* + \mathbf{w}^* - \mathbf{w}_T\|^2]$$

$$\leq L_\Psi\mathbb{E}[\|\bar{\mathbf{w}}_T - \mathbf{w}^*\|^2 + \|\mathbf{w}^* - \mathbf{w}_T\|^2] \leq \frac{10L_\Psi\widetilde{C}_2}{(T+1)\sigma_\phi^2\sigma_\Psi}.$$

Plugging the above inequality back into (B.5) and using $T \leq T^* \leq 2T - 1$ give

$$\mathbb{E}[\phi(\mathbf{w}_{T^*}) - \phi(\bar{\mathbf{w}}_T)] \leq T^{-1}\eta_{T^*}^{-1}\mathbb{E}[D_\Psi(\bar{\mathbf{w}}_T, \mathbf{w}_T)] + \sigma_\Psi^{-1}\eta_{T^*}[A\mathbb{E}[\phi(\mathbf{w}_{T^*})] + 2B]$$

$$\leq \frac{\sigma_\phi(T^* + 2)}{2T}\frac{10L_\Psi\widetilde{C}_2}{(T+1)\sigma_\phi^2\sigma_\Psi} + \frac{2[A\mathbb{E}[\phi(\mathbf{w}_{T^*})] + 2B]}{\sigma_\Psi\sigma_\phi T^*}$$

$$= \frac{10L_\Psi\widetilde{C}_2}{T\sigma_\phi\sigma_\Psi} + \frac{2[A\mathbb{E}[\phi(\mathbf{w}_{T^*})] + 2B]}{\sigma_\Psi\sigma_\phi T}. \qquad\qquad \text{(C.5)}$$

Plugging the first inequality in (3.4) and (C.5) back into the error decomposition (2.5) gives

$$\mathbb{E}[\phi(\mathbf{w}_{T^*}) - \phi(\mathbf{w}^*)] \leq \frac{10L_\Psi\widetilde{C}_2\sigma_\Psi^{-1} + 4\widetilde{C}_2}{T\sigma_\phi} + \frac{2A[\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*)]}{\sigma_\Psi\sigma_\phi T} + \frac{2[A\phi(\mathbf{w}^*) + 2B]}{\sigma_\Psi\sigma_\phi T}$$

$$\leq \frac{10L_\Psi\widetilde{C}_2\sigma_\Psi^{-1} + 4\widetilde{C}_2}{T\sigma_\phi} + \frac{\mathbb{E}[\phi(\mathbf{w}_{T^*})] - \phi(\mathbf{w}^*)}{2} + \frac{2[A\phi(\mathbf{w}^*) + 2B]}{\sigma_\Psi\sigma_\phi T},$$

where we have used the inequality $\frac{2A}{\sigma_\Psi\sigma_\phi T} \leq \frac{1}{2}$ and $\phi(\mathbf{w}_{T^*}) - \phi(\mathbf{w}^*) \geq 0$ in the last step. The above inequality can be written as stated inequality with

$$\widetilde{C}_3 = 4\widetilde{C}_2(5L_\Psi\sigma_\Psi^{-1} + 2) + 4\sigma_\Psi^{-1}[A\phi(\mathbf{w}^*) + 2B].$$

The proof is complete. $\qquad\square$

Finally, we give the proof of Theorem 7 on lower bounds of convergence rates under a lower-bound assumption on the variance of $\nabla f(\mathbf{w}, z)$ as an unbiased estimate of $\nabla F(\mathbf{w})$.

Proof of Theorem 7. If $\Psi(\mathbf{w}) = \frac{1}{2}\|\mathbf{w}\|_2^2$ and $r(\mathbf{w}) = 0$, then (2.2) becomes

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta_t \nabla f(\mathbf{w}_t, z_t),$$

from which we know

$$\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2 = \|\mathbf{w}_t - \mathbf{w}^*\|_2^2 + \eta_t^2 \|\nabla f(\mathbf{w}_t, z_t)\|_2^2 - 2\eta_t \langle \mathbf{w}_t - \mathbf{w}^*, \nabla f(\mathbf{w}_t, z_t) \rangle.$$

Taking expectations over both sides, we get

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2] = \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2] + \eta_t^2 \mathbb{E}[\|\nabla f(\mathbf{w}_t, z_t)\|_2^2] - 2\eta_t \mathbb{E}[\langle \mathbf{w}_t - \mathbf{w}^*, \nabla \phi(\mathbf{w}_t) \rangle]. \quad \text{(C.6)}$$

According to the lower bound assumption on variances, we know

$$\mathbb{E}_{z_t}[\|\nabla f(\mathbf{w}_t, z_t)\|_2^2] = \mathbb{E}_{z_t}\left[\|\nabla f(\mathbf{w}_t, z_t) - \nabla F(\mathbf{w}_t)\|_2^2\right] + \|\nabla F(\mathbf{w}_t)\|_2^2 \geq \sigma^2.$$

We can combine the above inequality and (C.6) to derive

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2] \geq \mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2] + \eta_t^2 \sigma^2 - 2\eta_t \mathbb{E}[\langle \mathbf{w}_t - \mathbf{w}^*, \nabla \phi(\mathbf{w}_t) \rangle]. \quad \text{(C.7)}$$

Since ϕ is L_ϕ-smooth, we know

$$\langle \mathbf{w}^* - \mathbf{w}_t, \nabla \phi(\mathbf{w}_t) \rangle \geq -L_\phi \|\mathbf{w}^* - \mathbf{w}_t\|_2^2,$$

which, plugged into (C.6) with the assumption $2L_\phi \eta_t \leq 1$, implies the following inequality

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2] \geq (1 - 2L_\phi \eta_t)\mathbb{E}[\|\mathbf{w}_t - \mathbf{w}^*\|_2^2] + 2L_\phi \eta_t (\eta_t \sigma^2/(2L_\phi)),$$

from which we know

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2] \geq \min\{\mathbb{E}\|\mathbf{w}_t - \mathbf{w}^*\|_2^2, \eta_t \sigma^2/(2L_\phi)\}.$$

We can apply the above inequality iteratively to show

$$\mathbb{E}[\|\mathbf{w}_{t+1} - \mathbf{w}^*\|_2^2] \geq \min\{\|\mathbf{w}_1 - \mathbf{w}^*\|_2^2, \eta_1 \sigma^2/(2L_\phi), \ldots, \eta_t \sigma^2/(2L_\phi)\}.$$

The proof is complete. □

Table C.1: Description of the datasets used in the experiments.

datasets	# inst	# feat	datasets	# inst	# feat	datasets	# inst	# feat	datasets	# inst	# feat
diabetes	768	8	german	1000	24	splice	1000	60	usps	7291	256
mnist	60000	780	w8a	49749	300	letter	15000	16	satimage	4456	36
ijcnn1	141691	22	mushrooms	8124	112	a9a	32561	123	connect	67557	126
cover	286048	10	webspam_u	350000	254	real-sim	72309	20958	rcv1	20242	47236

D Additional Experimental Results

In this section, we give the description of datasets used in the experiments and report more experimental results.

D.1 Description of Datasets

In this subsection, we provide in Table C.1 the information for the datasets used in Section 5.1. Webspam_u is a subset used in the Pascal Large Scale Learning Challenge [4] to detect malicious web pages. The remaining datasets can be downloaded from the LIBSVM homepage [2].

D.2 Testing errors versus iteration numbers

In this subsection, we compare the behavior of several variants of SGD and SPGD on testing datasets. In Figure D.1, we plot the objective function values on testing datasets versus iteration numbers for SPGD (mean±0.5std). In Figure D.2, we plot the objective function values on testing datasets versus iteration numbers for SGD applied to different datasets.
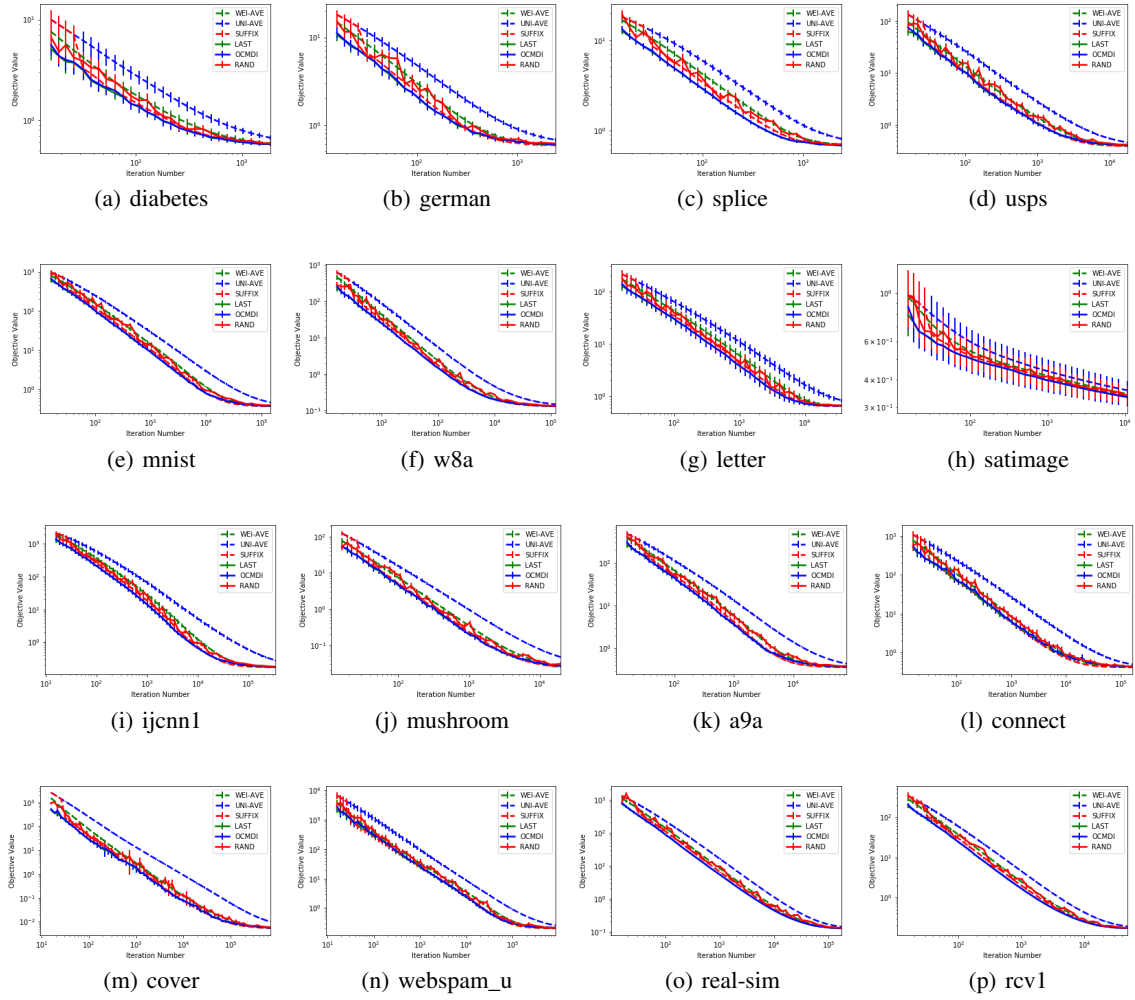
Figure D.1: Objective function values on test datasets versus iteration numbers for SPGD.

D.3 Training errors versus iteration numbers

In this section, we report more experimental results on training errors. In Figure D.3 and Figure D.4, we report the objective function values on training examples versus the number of iterations for SPGD and SGD, which behave analogously to testing errors.

D.4 Additional experimental results for tomography reconstruction

In this section, we report more experimental results for tomography reconstruction. In Figure D.5, we compare the behavior of our method with several baseline methods for the tomography reconstruction problem with $N = 32, n = 11520$ and 5% relative noise.
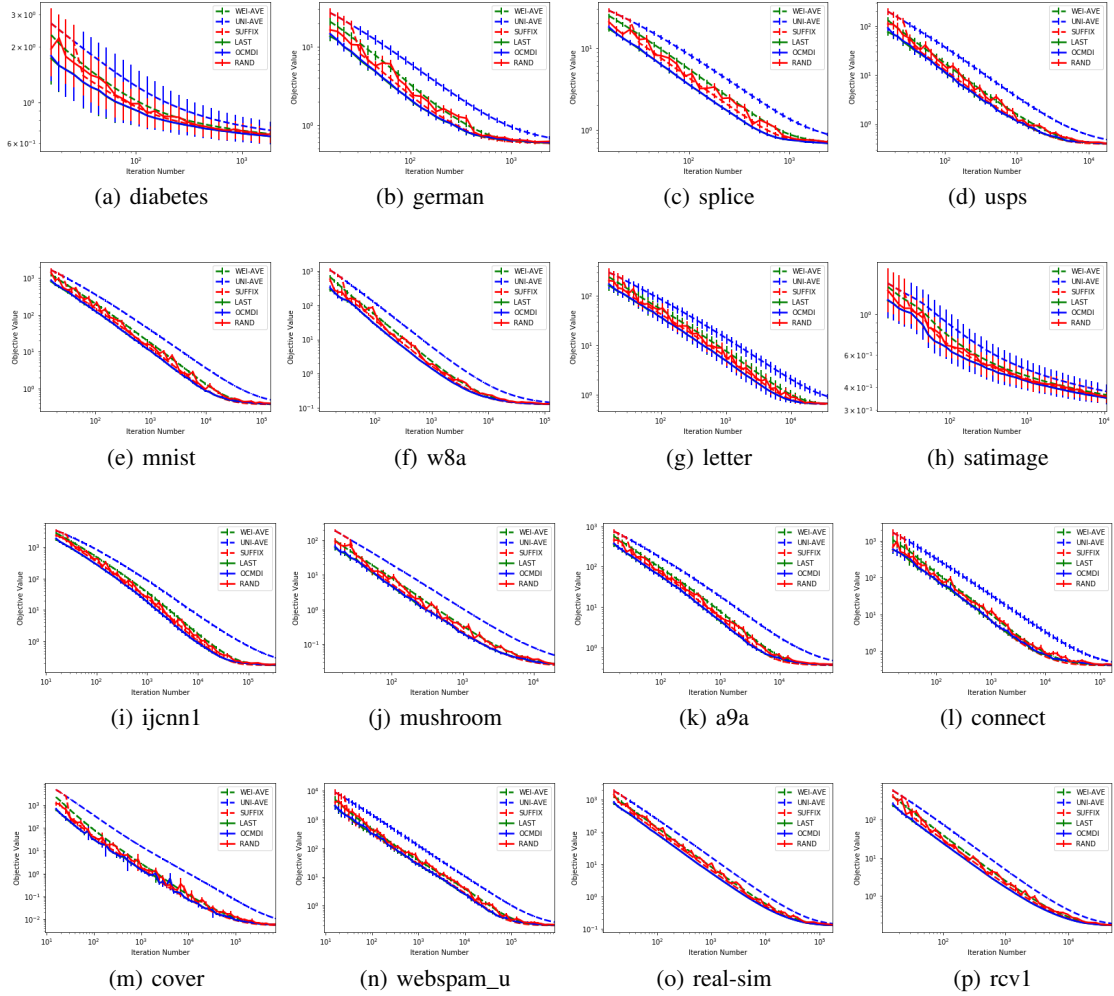
Figure D.2: Objective function values on test datasets versus iteration numbers for SGD.
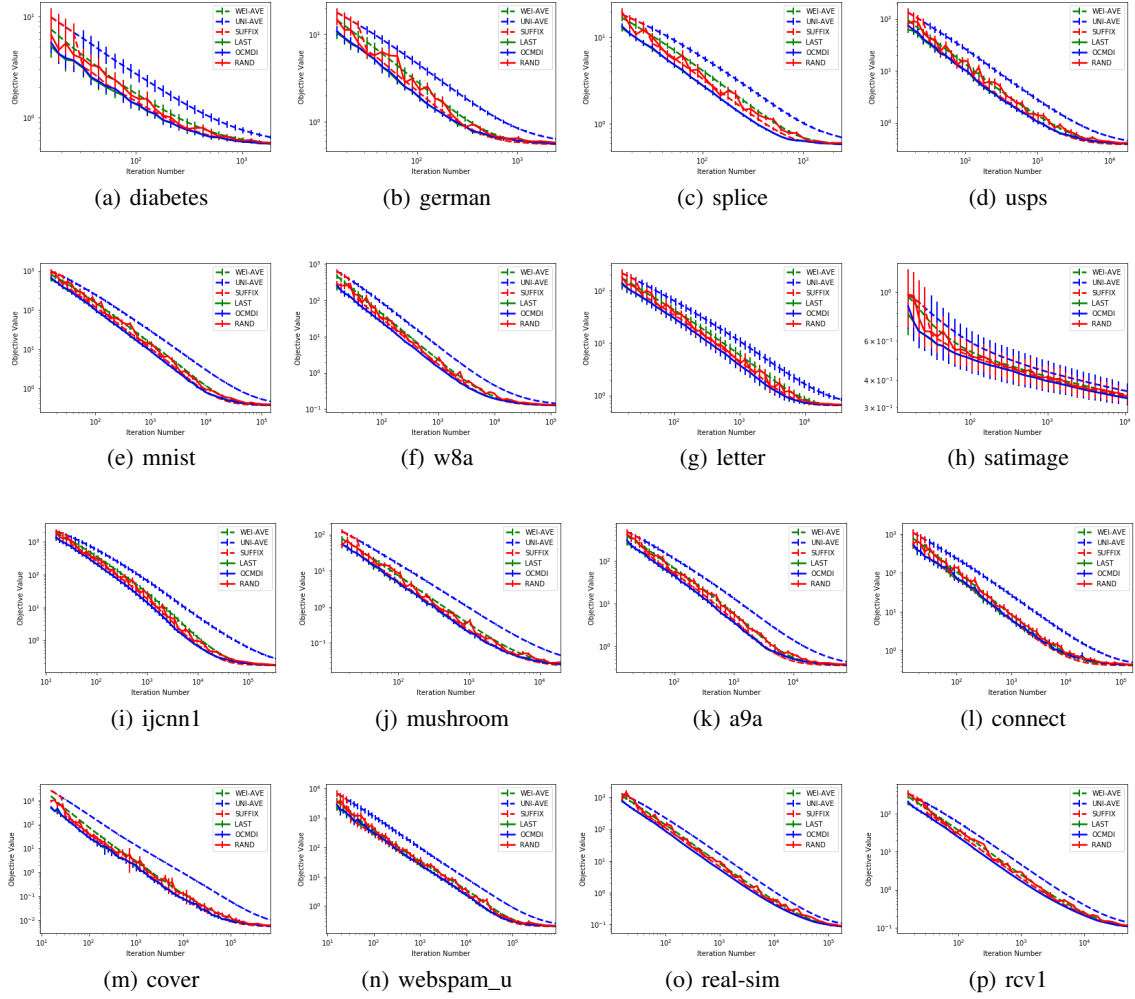
(a) diabetes   (b) german   (c) splice   (d) usps

(e) mnist   (f) w8a   (g) letter   (h) satimage

(i) ijcnn1   (j) mushroom   (k) a9a   (l) connect

(m) cover   (n) webspam_u   (o) real-sim   (p) rcv1

Figure D.3: Objective function values on training datasets versus iteration numbers for SPGD.

(a) diabetes   (b) german   (c) splice   (d) usps

(e) mnist   (f) w8a   (g) letter   (h) satimage

(i) ijcnn1   (j) mushroom   (k) a9a   (l) connect

(m) cover   (n) webspam_u   (o) real-sim   (p) rcv1
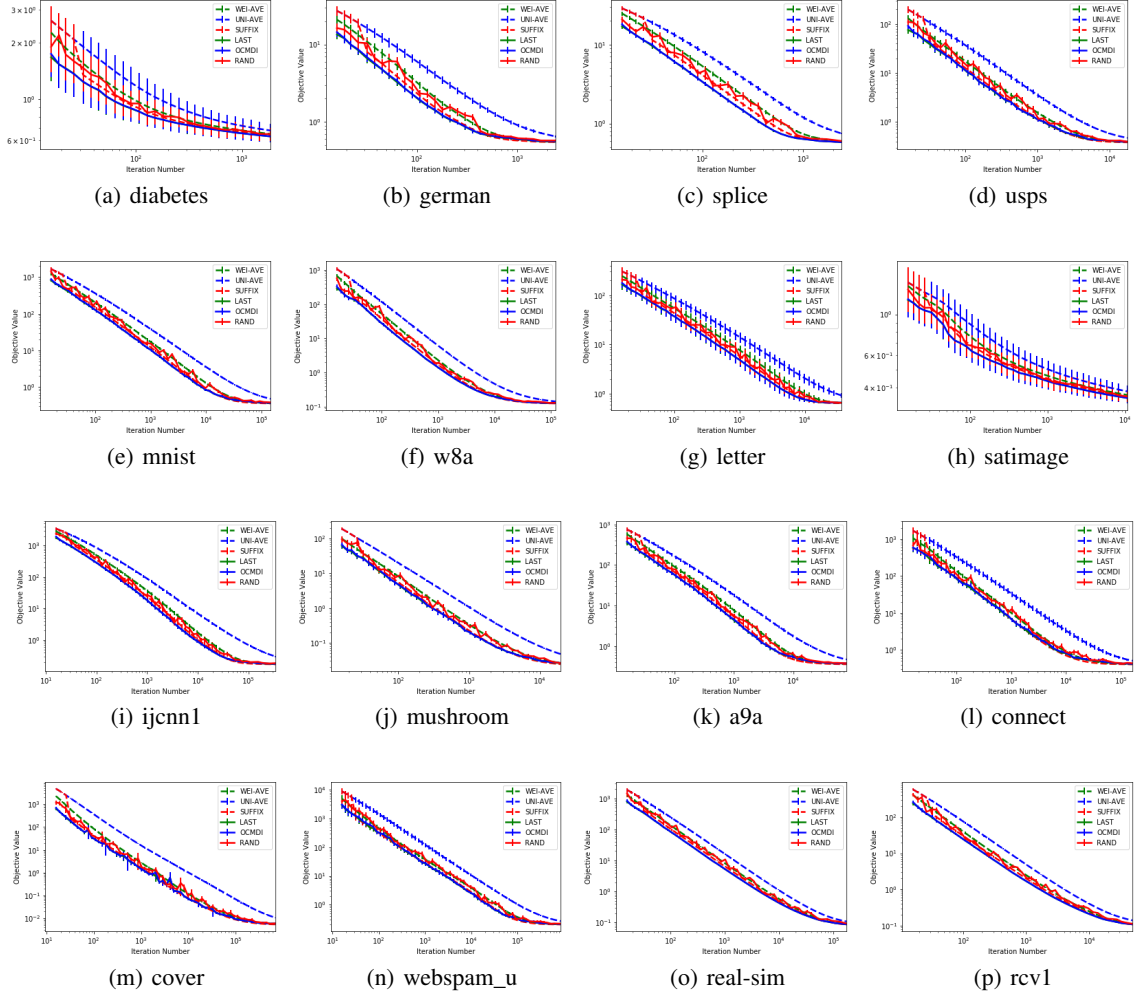
Figure D.4: Objective function values on training datasets versus iteration numbers for SGD.

(a) True Image.   (b) Reconstructed Image.   (c) Error vs. T.   (d) NNCs vs. T
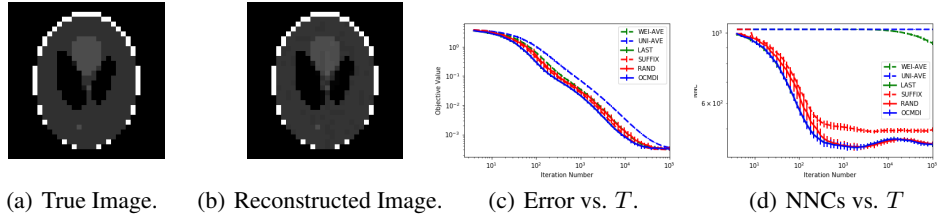
Figure D.5: Tomography reconstruction with $N = 32, n = 11520$ and 5% relative noise. Panel (a) and (b) are the true image and the reconstructed image by OCMDI, respectively. Panel (c) and (d) plot the errors and NNCs versus iteration numbers.

References

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

[2] C.-C. Chang and C.-J. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[3] P. Tarres and Y. Yao. Online learning as stochastic approximation of regularization paths: optimality and almost-sure convergence. *IEEE Transactions on Information Theory*, 60(9): 5716–5735, 2014.

[4] D. Wang, D. Irani, and C. Pu. Evolutionary study of web spam: Webb spam corpus 2011 versus webb spam corpus 2006. In *8th International Conference on Collaborative Computing: Networking, Applications and Worksharing*, pages 40–49. IEEE, 2012.

[5] Y. Ying and D.-X. Zhou. Online regularized classification algorithms. *IEEE Transactions on Information Theory*, 52(11):4775–4788, 2006.