Stochastic Composite Mirror Descent: Optimal Bounds with High Probabilities

Yunwen Lei and **Ke Tang**[*]
Shenzhen Key Laboratory of Computational Intelligence, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
leiyw@sustc.edu.cn tangk3@sustc.edu.cn

Abstract

We study stochastic composite mirror descent, a class of scalable algorithms able to exploit the geometry and composite structure of a problem. We consider both convex and strongly convex objectives with non-smooth loss functions, for each of which we establish high-probability convergence rates optimal up to a logarithmic factor. We apply the derived computational error bounds to study the generalization performance of multi-pass stochastic gradient descent (SGD) in a non-parametric setting. Our high-probability generalization bounds enjoy a *logarithmical* dependency on the number of passes provided that the step size sequence is square-summable, which improves the existing bounds in expectation with a *polynomial* dependency and therefore gives a strong justification on the ability of multi-pass SGD to overcome overfitting. Our analysis removes boundedness assumptions on subgradients often imposed in the literature. Numerical results are reported to support our theoretical findings.

1 Introduction

Stochastic gradient descent (SGD) has found wide applications in machine learning problems due to its simplicity in implementation, low memory requirement and low computational complexity per iteration, as well as good practical behavior [2, 6, 28, 32, 41]. As an iterative method, SGD minimizes empirical errors by moving iterates along the direction of a negative gradient calculated based on a loss function on a single training example or a batch of few examples. This strategy of processing few examples per iteration makes SGD particularly suitable for large scale applications with very large data points [2, 41], which are becoming ubiquitous in the big data era.

Stochastic composite mirror descent (SCMD) is a powerful extension of SGD based on two motivations [12]. Firstly, it relaxes the Hilbert space structure of SGD by using a mirror map to capture geometric properties of data from a Banach space [4, 25]. Secondly, it exploits the problem structure by separating, at every iteration, a data-fitting term and a regularization term in structured optimization problems to obtain a desired regularization effect, which arise naturally since a regularizer is often introduced to either avoid overfitting or impose a priori information [12, 37].

Although much theoretical analysis has been performed to understand the practical behavior of SGD and SCMD, the existing theoretical results are still not quite satisfactory. Firstly, most of the existing theoretical results are stated in expectation which inevitably ignore some information on high-order moments of the random variable we are interested in. In practice, we may be more interested in high-probability bounds to understand the variability of the learned model which is also an important factor we should take into account when measuring the quality of models [32]. Secondly, the existing generalization bounds, stated in expectation, for SGD either are suboptimal

[*]Corresponding author

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

or require to impose a smoothness assumption on loss functions [13, 21]. Thirdly, a non-trivial assumption on the boundedness of subgradients is often imposed in the literature to proceed with the analysis [11, 12, 28, 32], especially in the derivation of high-probability bounds. However, this boundedness assumption may not hold if the optimization is conducted in an unbounded domain, under which scenario the derived bounds may not be intuitive.

In this paper, we aim to contribute towards a refined analysis on both convergence rates and generalization properties of SCMD. We consider both general convex and strongly convex objectives, for each of which we show that SCMD can achieve almost optimal convergence rates with high probability, which match the minimax lower rates for stochastic approximation up to a logarithmic factor [1, 25]. In particular, we identify a constraint on step sizes to guarantee the boundedness of iterates with high probability (up to a logarithmic factor). Furthermore, we apply these convergence rates related to computational errors to establish high-probability generalization bounds for the model trained by SGD through *multiple passes* over the training examples, which is a typical way of using SGD to process large datasets [20]. Our generalization bounds do not require to impose smoothness assumptions on loss functions and can be optimal up to a logarithmic factor. Surprisingly, we show that estimation errors scale logarithmically with respect to (w.r.t.) the number of passes provided that the step size sequence is square-summable, which implies that SGD may be immune to overfitting. As a contrast, estimation error bounds based on stability arguments [13] and uniform deviation arguments [21] scale polynomially w.r.t. the number of passes, which may not justify well the ability of SGD in overcoming overfitting in practice. All our theoretical results are derived without any boundedness assumptions on subgradients based on two tricks. The first trick is to use a self-bounding property of loss functions (Assumption 1) to show that a (weighted) summation of function values can be controlled by step sizes (Lemma 2). The second trick is to show that conditional variances of martingales in a one-step progress inequality of SCMD can be partially offset by some other terms in the one-step progress inequality.

The paper is organized as follows. We introduce SCMD and state convergence rates in Section 2 and Section 3, respectively. We study generalization bounds of SGD in Section 4. Discussions are given in Section 5. Simulation results and conclusions are given in Section 6 and Section 7, respectively.

2 Stochastic Composite Mirror Descent

Many machine learning problems involve optimization problems of a composite structure [12, 37]

$$\min_{w\in\mathcal{W}} \phi(w) = \mathbb{E}_z[f(w,z)] + r(w), \tag{2.1}$$

where $\mathcal{W}$ is a Banach space with a norm $\|\cdot\|$, $F(w) := \mathbb{E}_z[f(w,z)]$ is a data-fitting term and $r : \mathcal{W} \to \mathbb{R}_+$ is a simple regularizer possibly inducing sparsity. Here $f : \mathcal{W} \times \mathcal{Z} \mapsto \mathbb{R}_+$ is a function with $f(w,z)$ measuring the quality of a model indexed by $w \in \mathcal{W}$ on a random example $z = (x,y)$ drawn from a probability measure $\tilde{\rho}$ defined in a sample space $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$ with an input space $\mathcal{X} \subset \mathcal{W}^*$ and an output space $\mathcal{Y} \subset \mathbb{R}$. We denote by $\mathbb{E}_z$ the expectation w.r.t. z, and by $\mathcal{W}^*$ the dual of $\mathcal{W}$ with the dual norm $\|\cdot\|_*$. A typical choice of the data-fitting term takes the form $f(w,z) = \ell(\langle w, x\rangle, y)$, where $\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}_+$ is a loss function and $\langle w, x\rangle$ is the dual element $x \in \mathcal{W}^*$ acting on $w \in \mathcal{W}$. With specific instantiations of loss functions ℓ and regularizers r, the formulation (2.1) covers many famous machine learning problems in a unifying framework, including least squares, support vector machines, logistic regression, lasso and elastic-net, etc [12, 37].

As an extension of SGD, SCMD uses a strongly convex and Fréchet differentiable mirror map Ψ to generate an appropriate Bregman distance $D_\Psi(w, \tilde{w}) := \Psi(w) - \Psi(\tilde{w}) - \langle w - \tilde{w}, \nabla\Psi(\tilde{w})\rangle$ to capture the involved non-Euclidean geometry [4, 25], where $\nabla\Psi(\tilde{w})$ denotes the gradient of Ψ at $\tilde{w}$. Let $w_1 = 0 \in \mathcal{W}$ and $\{\eta_t\}_{t\in\mathbb{N}}$ be a positive step size sequence. Upon the arrival of z_t at the t-th iteration, SCMD calculates a subgradient $f'(w_t, z_t) \in \partial_w f(w_t, z_t)$ as an unbiased estimate of $F'(w_t) \in \partial F(w_t)$, and updates the model as follows

$$w_{t+1} = \arg\min_{w\in\mathcal{W}} \eta_t\big[\langle w - w_t, f'(w_t, z_t)\rangle + r(w)\big] + D_\Psi(w, w_t). \tag{2.2}$$

Here $\partial_w f(w_t, z_t) := \big\{g : f(w, z_t) - f(w_t, z_t) \geq \langle w - w_t, g\rangle \text{ for all } w\big\}$ denotes the subdifferential of $f(\cdot, z_t)$ at w_t. Intuitively, SCMD uses $f'(w_t, z_t)$ to form a first-order approximation of $f(\cdot, z_t)$ at w_t and uses the Bregman distance $D_\Psi(w, w_t)$ to keep w_{t+1} not far away from the current iterate. The regularizer r is kept intact here for a regularization effect [12, 37]. A typical choice of Ψ is the

p-norm divergence $\Psi_p(w) = \frac{1}{2}\|w\|_p^2$ $(1 < p \le 2)$, which works favorably for sparse problems by setting p close to 1 [12, 37]. Here $\|\cdot\|_p$ is the p-norm defined by $\|w\|_p = \big(\sum_{i=1}^d |w(i)|^p\big)^{1/p}$ for $w = (w(1), \ldots, w(d)) \in \mathbb{R}^d$. SCMD recovers SGD by taking $\Psi = \Psi_2$ and $r(w) = 0$, stochastic forward-backward splitting by taking $\Psi = \Psi_2$ [11], stochastic mirror descent by taking $r(w) = 0$ [24] and stochastic mirror descent algorithm made sparse by taking $\Psi = \Psi_p$ and $r(w) = \lambda\|w\|_1$ [30].

3 Convergence Rates

Before stating our high-probability convergence rates, we introduce some assumptions. Throughout the paper, we assume that the mirror map Ψ is Fréchet differentiable and σ_Ψ-strongly convex in the sense that $D_\Psi(w, \tilde{w}) \ge 2^{-1}\sigma_\Psi\|w - \tilde{w}\|^2$ for all $w, \tilde{w} \in \mathcal{W} \subset \mathbb{R}^d$ $(\sigma_\Psi > 0)$, and $f(w, z)$ is convex w.r.t. the first argument. We also always assume that Assumption 1 and Assumption 2 hold, the sample space $\mathcal{Z}$ is bounded and $\sup_{z\in\mathcal{Z}} f(0, z) < \infty$.

Assumption 1. We assume that there exist A and $B \ge 0$ such that the following inequalities hold for any $w \in \mathcal{W}, z \in \mathcal{Z}$ and any $f'(w, z) \in \partial f(w, z), r'(w) \in \partial r(w)$

$$\|f'(w, z)\|_*^2 \le Af(w, z) + B \quad \text{and} \quad \|r'(w)\|_*^2 \le Ar(w) + B. \tag{3.1}$$

This is a standard assumption and satisfied in many practical problems [11, 41]. For example, Lemma A.5 shows that $r(w) = \lambda\|w\|_p^p$ satisfies the second inequality of (3.1) with $\|\cdot\| = \|\cdot\|_p (1 \le p \le 2)$, $A = 2\lambda p(p-1)$ and $B = \lambda p(2-p)$. Furthermore, if $f(w, z) = \ell(\langle w, x\rangle, y)$, then Lemma A.4 shows that $\|f'(w, z)\|_*^2 = |\ell'(\langle w, x\rangle, y)|^2\|x\|_*^2$ would satisfy the first inequality of (3.1) if

$$|\ell'(a, y)|^2 \le \tilde{A}\ell(a, y) + \tilde{B}, \quad \forall a \in \mathbb{R}, y \in \mathcal{Y} \tag{3.2}$$

for some $\tilde{A}, \tilde{B} > 0$ [41], where $\ell'(a, y)$ denotes a subgradient of ℓ w.r.t. the first argument. Many popular loss functions satisfy (3.2), including the p-norm hinge loss $\ell(a, y) = \max\{0, 1 - ya\}^p$ $(1 \le p \le 2)$ [34], the logistic loss $\ell(a, y) = \log(1 + \exp(-ya))$ for classification, and the p-th power absolute distance loss $\ell(a, y) = |a - y|^p$ $(1 \le p \le 2)$, the Huber loss $\ell(a, y) = (a - y)^2$ if $|a - y| \le 1$ and $\ell(a, y) = 2|a - y| - 1$ otherwise for regression [41]. We refer the interested readers to [41] for constants $\tilde{A}, \tilde{B}$ in (3.2) with different loss functions ℓ.

Assumption 2. We assume the existence of $\sigma_F, \sigma_r \ge 0$ such that

$$\begin{aligned} F(w) - F(\tilde{w}) - \langle w - \tilde{w}, F'(\tilde{w})\rangle &\ge \sigma_F D_\Psi(w, \tilde{w}), \\ r(w) - r(\tilde{w}) - \langle w - \tilde{w}, r'(\tilde{w})\rangle &\ge \sigma_r D_\Psi(w, \tilde{w}) \end{aligned} \tag{3.3}$$

hold for all $w, \tilde{w} \in \mathcal{W}$ and any $F'(\tilde{w}) \in \partial F(\tilde{w}), r'(\tilde{w}) \in \partial r(\tilde{w})$.

The case $\sigma_\phi := \sigma_F + \sigma_r = 0$ corresponds to general convex objectives, while the case $\sigma_\phi > 0$ corresponds to strongly convex objectives. Let $w^* = \arg\min_{w\in\mathcal{W}} \phi(w)$ be the minimizer of ϕ in $\mathcal{W}$ with the minimal norm. We always assume $\|w^*\| < \infty$ in this paper.

Our theoretical analysis is based on the following lemma quantifying the one-step progress of SCMD measured by Bregman distance, which shows how $D_\Psi(w, w_t)$ would change in a single iteration.

Lemma 1. *Let $\{w_t\}_{t\in\mathbb{N}}$ be generated by* (2.2)*, then the following inequality holds for any $w \in \mathcal{W}$*

$$\begin{aligned} D_\Psi(w, w_{t+1}) - D_\Psi(w, w_t) \le{}& \eta_t\langle w - w_t, f'(w_t, z_t)\rangle + \eta_t(r(w) - r(w_t)) \\ &+ \sigma_\Psi^{-1}\eta_t^2 \underbrace{\big(Af(w_t, z_t) + Ar(w_t) + 2B\big)}_{:=\mathfrak{A}_t} - \sigma_r\eta_t D_\Psi(w, w_{t+1}). \end{aligned} \tag{3.4}$$

Existing one-step progress inequality can be found in the literature with $\mathfrak{A}_t$ replaced by $\mathfrak{B}_t := \|f'(w_t, z_t)\|_*^2 + \|r'(w_t)\|_*^2$, see, e.g., [12]. Then, a non-trivial assumption as $\mathfrak{B}_t \le G$ for all $t \in \mathbb{N}$ and a $G \in \mathbb{R}$ is imposed to control $\sum_{t=1}^T \eta_t^2\mathfrak{B}_t$ by $O(\sum_{t=1}^T \eta_t^2)$. We refine these discussions by using Assumption 1 to replace $\mathfrak{B}_t$ with $\mathfrak{A}_t$. Equation (3.6) allows us to control $\sum_{t=1}^T \eta_t^2\mathfrak{A}_t$ by $O(\sum_{t=1}^T \eta_t^2)$ without imposing any boundedness assumptions on subgradients. In our discussion for strongly convex objectives, we require to divide both sides of (3.4) by η_t^2. In this way, Eq. (3.7) plays an analogous role in removing boundedness assumptions in the strongly convex case. Both proofs of Lemma 1 and Lemma 2 are given in Supplementary Material B.

Lemma 2. *Let* $\{w_t\}_{t\in\mathbb{N}}$ *be the sequence produced by* (2.2) *with* $\eta_t \le (2A)^{-1}\sigma_\Psi$*. Then, we have*

$$\|w_{t+1}\|^2 \le 2C_1\sigma_\Psi^{-1}\sum_{k=1}^t \eta_k, \quad \forall t\in\mathbb{N}, \tag{3.5}$$

where $C_1 = \sup_{z\in\mathcal{Z}} f(0,z) + r(0) + A^{-1}B$*. Furthermore, if* $\eta_{t+1}\le\eta_t$*, then for all* $t\in\mathbb{N}$

$$\sum_{k=1}^t \eta_k^2\big(f(w_k,z_k)+r(w_k)\big) \le 2C_1\sum_{k=1}^t \eta_k^2, \tag{3.6}$$

$$\sum_{k=1}^t \big(f(w_k,z_k)+r(w_k)\big) \le 2C_1 t + 2C_1\big(\sum_{k=1}^t \eta_k\big)\eta_t^{-1}. \tag{3.7}$$

3.1 Convex Objectives

We study the behavior of SCMD for convex objectives with $\sigma_\phi = 0$. The assumption $\sum_{t=1}^\infty \eta_t^2 < \infty$ is satisfied if $\eta_t = \eta_1 t^{-\theta}$ with $\theta > 1/2$ or $\eta_t = \eta_1(t\log^\beta(et))^{-\frac{1}{2}}$ with $\beta > 1$. Our idea is to take a summation of Eq. (3.4) with $w = w^*$, and show that the conditional variance of the involved martingale $\sum_{k=1}^t \eta_k\langle w^* - w_k, f'(w_k,z_k) - \mathbb{E}_{z_k}[f'(w_k,z_k)]\rangle$ can be partially offset by some other terms. The proofs of Theorems 3 and 4 are given in Supplementary Material C.

Theorem 3. *Let* $\{w_t\}_{t\in\mathbb{N}}$ *be the sequence produced by* (2.2) *with* $\eta_t \le (2A)^{-1}\sigma_\Psi, \eta_{t+1}\le\eta_t$ *and* $\sum_{t=1}^\infty \eta_t^2 < \infty$*. Then, there exists a constant* C_2 *independent of* T *(explicitly given in the proof) such that for any* $\delta\in(0,1)$ *the following inequality holds with probability at least* $1-\delta$

$$\max_{1\le t\le T}\|w_t\|^2 \le C_2\log\frac{T}{\delta}. \tag{3.8}$$

Remark 1. Although implemented in a possibly unbounded domain, Theorem 3 shows that $\{w_t\}_{t\in\mathbb{N}}$ by (2.2) falls into a bounded ball (up to a logarithmic factor) with high probabilities. Intuitively, this suggests that SCMD is immune to overfitting if we take appropriate step sizes. In this case, we can run SCMD with many iterations without essentially harming the quality of the output model.

Based on Theorem 3, we establish high-probability convergence rates for a weighted average of iterates without any assumptions on the boundedness of iterates. In Theorem 4 and Corollary 5, we establish bounds on suboptimality of objectives w.r.t. any w and an optimal solution w^*, respectively.

Theorem 4. *Let* $w\in\mathcal{W}$ *and* $\delta\in(0,2/e)$*. Let* $\bar{w}_T^{(1)} = \big(\sum_{t=1}^T\eta_t\big)^{-1}\sum_{t=1}^T\eta_t w_t$ *be a weighted average of the first* T *iterates. Under the conditions of Theorem 3, with probability* $1-\delta$ *we have*

$$\phi(\bar{w}_T^{(1)}) - \phi(w) \le \Big(\sum_{t=1}^T\eta_t\Big)^{-1}\big(2C_3D_\Psi(w,0)+C_4\big)\log^{\frac{3}{2}}\frac{2T}{\delta}, \tag{3.9}$$

where C_3 *and* C_4 *are two constants (explicitly given in the proof) independent of* T*.*

Remark 2. A similar high-probability bound was established for SCMD in [12]. However, their discussion needs to impose an additional almost-sure boundedness assumption on iterates, i.e., $\|w_t\|_2 \le G$ for a $G > 0$ and all $t\in\mathbb{N}$. These boundedness assumptions on either subgradients or iterates are fundamental to the existing analysis but hard to check in practice. Moreover, the high-probability analysis makes these assumptions *non-trivial* to remove since one also needs to consider high-order moments of random variables.

Corollary 5. *If* $\delta\in(0,2/e)$ *and conditions of Theorem 4 are satisfied, then* (3.9) *holds with probability* $1-\delta$ *with* $w = w^*$*. Furthermore, if we choose* $\eta_t = \eta_1 t^{-\theta}$ *with* $\theta > 1/2$*, then with probability* $1-\delta$ *we have* $\phi(\bar{w}_T^{(1)}) - \phi(w^*) = O\big(T^{\theta-1}\log^{\frac{3}{2}}\frac{T}{\delta}\big)$*; if we choose* $\eta_t = \eta_1(t\log^\beta(et))^{-\frac{1}{2}}$ *with* $\beta > 1$*, then with probability* $1-\delta$ *we have* $\phi(\bar{w}_T^{(1)}) - \phi(w^*) = O\big(\big(T^{-1}\log^\beta T\big)^{\frac{1}{2}}\log^{\frac{3}{2}}\frac{T}{\delta}\big)$*.*

The convergence rate $O\big(\big(T^{-1}\log^\beta T\big)^{\frac{1}{2}}\log^{\frac{3}{2}}\frac{T}{\delta}\big)$ in Corollary 5 is optimal up to a logarithmic factor [1], which follows directly from Theorem 4 and $\sum_{t=1}^T t^{-\theta} \ge (1-\theta)^{-1}(T^{1-\theta}-1), \theta\in(0,1)$. We omit the proof for brevity.

In Theorem 6, we give sufficient conditions for the almost sure finiteness of $\lim_{t\to\infty} D_\Psi(w^*, w_t)$ and $\sum_{t=1}^\infty \eta_t\big(\phi(w_t) - \phi(w^*)\big)$. As a direct corollary, we also establish convergence rates with probability one in Corollary 7. Theorem 6 is a part of Proposition E.3 to be presented and proved in Supplementary Material E, while the proof of Corollary 7 is omitted for brevity.

Theorem 6. *Consider $\{w_t\}_{t\in\mathbb{N}}$ by (2.2) with $\sum_{t=1}^\infty \eta_t^2 < \infty$. Then $\{D_\Psi(w^*, w_t)\}_t$ converges almost surely (a.s.) to a non-negative random variable and $\lim_{t\to\infty} D_\Psi(w^*, w_t) < \infty$ a.s.. Furthermore, if $\eta_t \le (2A)^{-1}\sigma_\Psi$ and $\eta_{t+1} \le \eta_t$, then $\sum_{t=1}^\infty \eta_t\big(\phi(w_t) - \phi(w^*)\big) < \infty$ a.s..*

Corollary 7. *Let $\{w_t\}_{t\in\mathbb{N}}$ be produced by (2.2) and $\eta_1 \le (2A)^{-1}\sigma_\Psi$. If we choose $\eta_t = \eta_1 t^{-\theta}$ with $\theta > 1/2$, then $\lim_{T\to\infty} T^{1-\theta}\big(\phi(\bar{w}_T^{(1)}) - \phi(w^*)\big) < \infty$ a.s.. If we choose $\eta_t = \eta_1(t\log^\beta(et))^{-\frac{1}{2}}$ with $\beta > 1$, then $\lim_{T\to\infty}\big(\frac{T}{\log^\beta T}\big)^{\frac{1}{2}}\big(\phi(\bar{w}_T^{(1)}) - \phi(w^*)\big) < \infty$ a.s..*

3.2 Strongly Convex Objectives

We now turn to strongly convex objectives with $\sigma_\phi > 0$. In Theorem 8, we establish high-probability bounds for both $\|w_t - w^*\|^2$ and $\phi(\bar{w}_t^{(2)}) - \phi(w^*)$ with $\bar{w}_t^{(2)}$ being another weighted average of the first t iterates, for each of which we derive optimal convergence rates up to a logarithmic factor [1]. The optimality means that not only the dependency on t but also the dependency on the strong-convexity parameter σ_ϕ can not be improved up to a logarithmic factor [16, 28] (σ_ϕ is often chosen to be very small in practical learning problems [28, 31]). It should be mentioned that our analysis removes boundedness assumptions on subgradients in the literature [28]. Our idea is to take a weighted summation of (3.4) with $w = w^*$, and show that the conditional variance of an involved martingale $\sum_{k=1}^t (k + t_0 + 1)\langle w^* - w_k, f'(w_k, z_k) - \mathbb{E}_{z_k}[f'(w_k, z_k)]\rangle$ can be partially offset by another term in this weighted summation of (3.4), which is another trick to remove boundedness assumptions on subgradients. We also give a sufficient condition on the almost sure convergence of w_t to w^* in Theorem 9. The proof of Theorem 8 is given in Supplementary Material D. Theorem 9 is a part of Proposition E.3 to be presented in Supplementary Material E.

Theorem 8. *Assume $\sigma_\phi > 0$ and $\delta \in (0, e^{-\frac{1}{4}})$. Let $\{w_t\}_{t\in\mathbb{N}}$ be produced by (2.2) with $\eta_t = \frac{2}{\sigma_\phi t + 2\sigma_F + \sigma_\phi t_0}$, where $t_0 \ge \frac{16A\log\frac{T}{\delta}}{\sigma_\phi\sigma_\Psi}$. Let $\bar{w}_t^{(2)} = \big(\sum_{k=1}^t (k+t_0+1)\big)^{-1}\sum_{k=1}^t (k+t_0+1)w_k, t \in \mathbb{N}$. Then, the following inequalities hold with probability $1-\delta$ for all $t = 1, \ldots, T$*

$$\|w^* - w_t\|^2 \le \frac{C_T}{t + t_0 + 1} \quad \text{and} \quad \phi(\bar{w}_t^{(2)}) - \phi(w^*) \le \frac{\widetilde{C}_T}{t}. \tag{3.10}$$

Moreover, the dependencies of C_T and $\widetilde{C}_T$ on T/δ are logarithmic. The dependencies of C_T and $\widetilde{C}_T$ on σ_ϕ^{-1} are quadratic and linear, respectively.

Theorem 9. *Let $\{w_t\}_{t\in\mathbb{N}}$ be the sequence produced by (2.2) with $\sigma_\phi > 0$. If $\sum_{t=1}^\infty \eta_t = \infty$ and $\sum_{t=1}^\infty \eta_t^2 < \infty$, then $\lim_{t\to\infty} D_\Psi(w^*, w_t) = 0$ a.s..*

4 Generalization Error Bounds

Here we apply our high-probability convergence rates for SCMD to establish generalization error bounds for SGD. In this setting, we assume a training sample $\mathbf{z} = \{z_1, \ldots, z_n\}$ of size $n \in \mathbb{N}$ is drawn independently from a probability measure ρ defined on the sample space $\mathcal{Z}$, and our aim is to learn a hypothesis $h : \mathcal{X} \mapsto \mathbb{R}$ from a hypothesis space $\mathcal{W}$ with good generalization performance. The quality of h at (x, y) is quantified by $\ell(h(x), y)$, where $\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}_+$ is convex w.r.t. the first argument. The generalization error and empirical error of h are defined respectively by $\mathcal{E}(h) = \mathbb{E}_z\big[\ell(h(x), y)\big]$ and $\mathcal{E}_{\mathbf{z}}(h) = \frac{1}{n}\sum_{i=1}^n \ell\big(h(x_i), y_i\big)$. The best model minimizing the generalization error then becomes $h_\rho = \arg\min_h \mathcal{E}(h)$. We consider a non-parametric learning setting with $\mathcal{W}$ being a reproducing kernel Hilbert space (RKHS) associated to a Mercer kernel $K : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ which is continuous, symmetric and positive semi-definite [9, 34]. In this learning setting, the candidate models take the form $h_w(x) = \langle w, K_x\rangle$ with $w \in \mathcal{W}$. For brevity, we denote the norm in the RKHS $\mathcal{W}$ by $\|\cdot\|_2$ and introduce abbreviations $\mathcal{E}(w) = \mathcal{E}(h_w), \mathcal{E}_{\mathbf{z}}(w) = \mathcal{E}_{\mathbf{z}}(h_w)$. We assume (3.2) and apply the SGD scheme to minimize $\mathcal{E}_{\mathbf{z}}(w)$. To be specific, we let $w_1 = 0$. At the t-th iteration, we randomly choose an index j_t from the uniform distribution over $\{1, \ldots, n\}$ and produce w_{t+1} by

$$w_{t+1} = w_t - \eta_t \ell'\big(\langle w_t, K_{x_{j_t}}\rangle, y_{j_t}\big)K_{x_{j_t}}, \quad t \in \mathbb{N}. \tag{4.1}$$

It is clear that (4.1) is a specific instantiation of (2.2) with $\Psi(w) = \frac{1}{2}\|w\|_2^2, f(w,z) = \ell(\langle w, K_x\rangle, y), r(w) = 0$ and $\tilde{\rho}$ in Section 2 being the uniform distribution over $\{z_1, \ldots, z_n\}$[2]. Therefore, the objective function to which SGD is applied becomes $\phi(w) = \mathcal{E}_{\mathbf{z}}(w)$.

To state our generalization bounds, we need to introduce an assumption on a polynomial decay rate of approximation errors.

Assumption 3. We assume the approximation error $D(\lambda) := \inf_{w\in\mathcal{W}} \mathcal{E}(w) - \mathcal{E}(h_\rho) + \lambda\|w\|_2^2$ enjoys a polynomial decay with exponent $0 < \alpha \le 1$ in the sense $D(\lambda) \le c_\alpha \lambda^\alpha, \forall \lambda > 0$, where $c_\alpha > 0$.

Remark 3. Assumption 3 is standard in learning theory and satisfied under some mild conditions on the smoothness of the function h_ρ and the representation power of $\mathcal{W}$ [9, 33]. If ℓ is smooth, then $D(\lambda)$ can be controlled by $\widetilde{D}(\lambda) := \inf_{w\in\mathcal{W}} \|h_w - h_\rho\|^2_{L^2_{\rho_\mathcal{X}}} + \lambda\|w\|_2^2$, which quantifies the approximation of h_ρ by RKHS in $L^2_{\rho_\mathcal{X}}$ (square-integrable function class with marginal measure $\rho_\mathcal{X}$) and is well studied in approximation theory. $\widetilde{D}(\lambda)$ decays polynomially with $\alpha \in (0,1]$ if $h_\rho \in L_K^{\alpha/2}(L^2_{\rho_\mathcal{X}})$, where $L_K : L^2_{\rho_\mathcal{X}} \mapsto L^2_{\rho_\mathcal{X}}$ is the integral operator associated to K [9, Proposition 8.5]. Similar results hold if ℓ is Lipschitz continuous. Assumption 3 also holds if we use Gaussian kernels with flexible variances and distributions with geometric noise conditions [35]. It should be mentioned that kernels need not to be universal for Assumption 3 since it concerns the target function h_ρ, which may admit more regularity (e.g., expressed by L_K) than continuity, while universality means that $D(\lambda) \to 0$ as $\lambda \to 0$ for *all continuous* h_ρ [34].

We now establish a generalization error bound for a weighted average of iterates produced by (4.1) to be proved in Supplementary Material F, which is derived by decomposing the excess generalization error $\mathcal{E}(\bar{w}_T^{(1)}) - \mathcal{E}(h_\rho)$ into three components: an estimation error, an approximation error and a computational error. As we will see in the proof, the term $\big(\sum_{t=1}^T \eta_t\big)^{-\alpha}$ is due to the approximation and computational error, while the term $n^{-\frac{\alpha}{1+\alpha}}$ is due to the estimation and approximation error. The bound becomes $n^{-\frac{\alpha}{1+\alpha}} \log^{\frac{3}{2}} \frac{8T}{\delta}$ for sufficiently large T, which enjoys a logarithmic dependency on T and demonstrates the ability of SGD to avoid overfitting.

Theorem 10. *Let $\{w_t\}_{t\in\mathbb{N}}$ be the sequence produced by* (4.1) *with $\eta_t \le (2A)^{-1}\sigma_\Psi, \eta_{t+1} \le \eta_t$ and $\sum_{t=1}^\infty \eta_t^2 < \infty$. Suppose Assumption 3 holds. Then, for any T satisfying $\sum_{t=1}^T \eta_t \ge 1$ and $\delta \in (0, 2/e)$, the following inequality holds with probability at least $1-\delta$*

$$\mathcal{E}(\bar{w}_T^{(1)}) - \mathcal{E}(h_\rho) \le C_5 \max\left\{\Big(\sum_{t=1}^T \eta_t\Big)^{-\alpha}, n^{-\frac{\alpha}{1+\alpha}}\right\} \log^{\frac{3}{2}} \frac{8T}{\delta}, \tag{4.2}$$

where C_5 is a constant independent of T (explicitly given in the proof).

We consider specific step sizes in Theorem 10 and choose an appropriate time index to get concrete generalization bounds, as shown in Corollary 11. The bound $O\big(n^{-\frac{\alpha}{1+\alpha}} \log^{\frac{3+\alpha\beta}{2}} \frac{n}{\delta}\big)$ coincides with $O(n^{-\frac{\alpha}{1+\alpha}} \log n)$ (up to a logarithmic factor) in expectation for convex and smooth loss functions [21], and largely improves the bound $O(n^{-\frac{\alpha}{1+2\alpha}} \log n)$ in expectation for convex and non-smooth loss functions [21]. In particular, if $\alpha = 1$ we derive the optimal bound $O(n^{-\frac{1}{2}} \log^{\frac{3+\beta}{2}} \frac{n}{\delta})$ in a general case with neither Bernstein conditions on variances nor capacity assumptions on hypothesis spaces (up to a logarithmic factor). It is also clear that SGD with different step sizes can achieve similar generalization bounds. However, the computational complexity to fulfill this statistical potential can be significantly different. Corollary 11, with the proof omitted, follows directly from Theorem 10 and $\sum_{t=1}^T t^{-\theta} \ge (1-\theta)^{-1}(T^{1-\theta} - 1), \theta \in (0,1)$. Denote $\lceil a \rceil$ the least integer no less than a.

Corollary 11. *Consider $\{w_t\}_{t\in\mathbb{N}}$ by* (4.1) *and $\delta \in (0, 2/e)$. Let Assumption 3 hold and $\sum_{t=1}^T \eta_t \ge 1$.*

(a) If we take $\eta_t = \eta_1 t^{-\theta}$ with $\eta_1 \le (2A)^{-1}$ and $\theta \in (1/2, 1)$, then with probability $1-\delta$ that

$$\mathcal{E}(\bar{w}_T^{(1)}) - \mathcal{E}(h_\rho) = O\bigg(\Big(T^{-\alpha(1-\theta)} + n^{-\frac{\alpha}{1+\alpha}}\Big) \log^{\frac{3}{2}} \frac{T}{\delta}\bigg).$$

If we further take $T^* = \lceil n^{\frac{1}{(1+\alpha)(1-\theta)}} \rceil$, then we get $\mathcal{E}(\bar{w}_{T^*}^{(1)}) - \mathcal{E}(h_\rho) = O\big(n^{-\frac{\alpha}{1+\alpha}} \log^{\frac{3}{2}} \frac{n}{\delta}\big)$.

[2] ρ is related to the draw of training examples while $\tilde{\rho}$ is related to the draw of indices for SGD.

(b) If we take $\eta_t = \eta_1(t\log^\beta(et))^{-\frac{1}{2}}$ with $\eta_1 \leq (2A)^{-1}$ and $\beta > 1$, then with probability $1-\delta$ that

$$\mathcal{E}(\bar{w}_T^{(1)}) - \mathcal{E}(h_\rho) = O\bigg(\Big(T^{-\frac{\alpha}{2}}\log^{\frac{\alpha\beta}{2}} T + n^{-\frac{\alpha}{1+\alpha}}\Big)\log^{\frac{3}{2}}\frac{T}{\delta}\bigg).$$

If we further take $T^* = \lceil n^{\frac{2}{1+\alpha}}\rceil$, then we get $\mathcal{E}(\bar{w}_{T^*}^{(1)}) - \mathcal{E}(h_\rho) = O\big(n^{-\frac{\alpha}{1+\alpha}}\log^{\frac{3+\alpha\beta}{2}}\frac{n}{\delta}\big)$.

It should be noted that our discussions depend on the existence of a minimizer of $\mathcal{E}_{\mathbf{z}}(\cdot)$ over the RKHS with a finite norm. This assumption can be relaxed to the existence of a minimizer of $\mathcal{E}(\cdot)$ over the RKHS with a finite norm to derive similar generalization bounds. Indeed, one can perform deductions similar to the proof of Theorem 3 by taking w in (3.4) to be the minimizer of $\mathcal{E}(\cdot)$. However, in this case it becomes a challenge to derive estimation error bounds with a logarithmic dependency on T.

5 Related Work and Discussions

5.1 Convex Objectives

For general convex objectives, regret bounds $O(\sqrt{T})$ were established for online gradient descent with T iterations [44], from which one can directly derive convergence rates $O(T^{-\frac{1}{2}})$ for SGD with some averaging schemes. This result was extended to stochastic forward-backward splitting [11]. A convergence rate $O(T^{-\frac{1}{2}}\log T)$ was established for the T-th individual iterate of SGD [32]. All the above mentioned rates were stated in expectation and derived based on an assumption $\mathbb{E}[\|f'(w_t,z_t)\|_*^2 + \|r'(w_t)\|_*^2] \leq G$ for a $G \geq 0$ and $t \in \mathbb{N}$. This boundedness assumption was successfully removed for studying convergence rates in expectation under some smoothness assumption [23, 40, 42] or Assumption 1 [30]. As compared to these convergence rates in expectation, high-probability convergence rates were much less studied and were often based on a stronger assumption on the almost sure boundedness of subgradients. Under the assumption $\max\{D_\Psi(w^*,w_t), \sup_z \|f'(w_t,z)\|_*\} \leq G$ for a $G > 0$ and all $t \in \mathbb{N}$, it was shown with probability $1-\delta$ that $\phi(\bar{w}_T^{(1)}) - \phi(w^*) = O\big(T^{-\frac{1}{2}}\log^{\frac{1}{2}}\frac{1}{\delta}\big)$ for $\bar{w}_T^{(1)}$ defined in Theorem 4 [12, 24]. High-probability bounds were also established for stochastic dual averaging under the boundedness assumption on iterates and subgradients [37]. In our discussion, we show that the same high-probability convergence rate (up to a logarithmic factor) holds without any boundedness assumptions on either the iterates $\{w_t\}$ or the associated subgradients. In particular, we show that $\{w_t\}_{t\leq T}$ automatically falls into a ball with radius $O(\sqrt{\log T/\delta})$ with high probability. It was shown with probability $1-\delta$ that $\|w_t - w^*\|_2^2 = O(\|w^*\|_2^2\log\frac{T}{\delta})$ for the particular SGD [19]. However, the discussion in [19] requires a stronger assumption on the Hölder continuity of loss functions which excludes non-differentiable loss functions such as hinge loss and the absolute loss satisfying (3.2). Secondly, they only consider the one-pass SGD where each training example is used only once.

We also give a sufficient condition for almost sure finiteness of $\sum_{t=1}^\infty \eta_t\big(\phi(w_t) - \phi(w^*)\big)$, while most results on almost sure convergence are achieved for strongly convex objectives.

5.2 Strongly Convex Objectives

For λ-exp-concave loss functions, a regret bound $O(\lambda^{-1}\log T)$ was established for an online Newton method [15], which implies convergence rates $O\big((\lambda T)^{-1}\log T\big)$ for some average of iterates produced by the stochastic counterpart. This result was extended to online forward-backward splitting [11] and SCMD [12] applied to λ-strongly convex objectives. Optimal convergence rates $O((\lambda T)^{-1})$ for the suboptimality of objective values were derived based on a suffix averaging scheme [28], a epoch-GD scheme based on a doubling trick [14] and a weighted averaging with a weight of $t+1$ for w_t [16]. However, the above mentioned results are all associated to convergence rates in expectation and require to impose boundedness assumptions on subgradients encountered during the iterations. This boundedness assumption was relaxed as $\mathbb{E}_z[\|f'(w_t,z)\|_*^2] \leq A_1 + B_1\|F'(w_t)\|_*^2$ for SGD [6] with $A_1, B_1 \geq 0$, which was further removed for SGD [26] and stochastic mirror descent [17] by imposing smoothness assumptions on loss functions. All the above mentioned results are stated in expectation. With probability $1-\delta$, it was shown $\|w_T - w^*\|^2 = O\big((\lambda^2 T)^{-1} + (\lambda T)^{-1}\log(\delta^{-1}\log T)\big)$ for SGD [28]. High-probability convergence rates $O\big((\lambda T)^{-1}\log(\delta^{-1}\log T)\big)$ were also established for the suboptimality of objective values for the T-th iterate of the epoch-GD [14]. These two high-probability rates were derived based on an assumption on almost sure boundedness of subgradients which is more challenging to remove [14, 28]. As a comparison, we establish the same convergence rate (up to a logarithmic factor) for a more general SCMD without boundedness assumptions on subgradients. Sufficient conditions as in Theorem 9 were established for almost sure convergence of SGD [5, 26] and stochastic mirror descent [17], which were extended to SCMD in Theorem 9.

5.3 Generalization Error Bounds

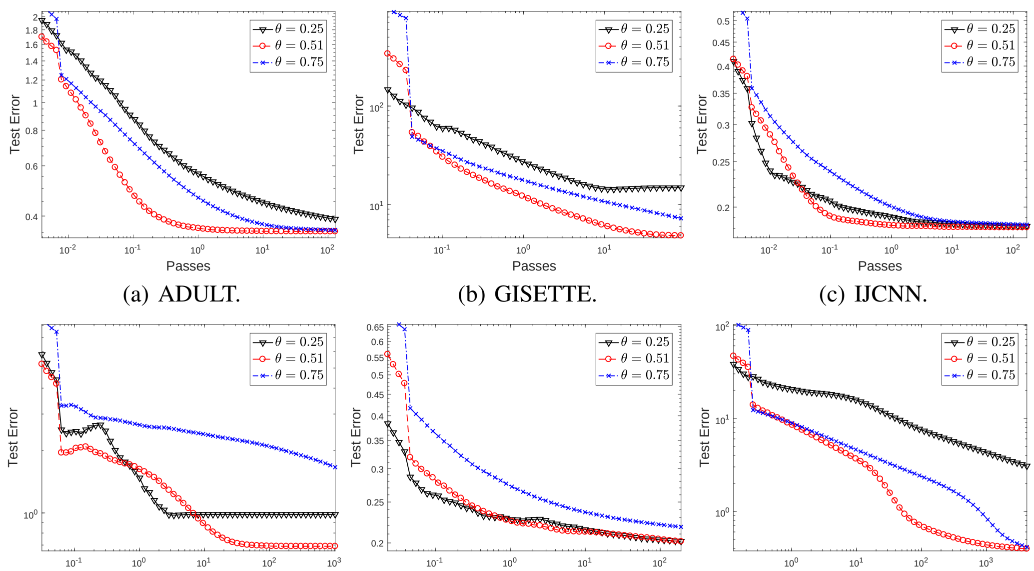
While computational complexity of SGD has been extensively studied in the optimization community, there is much less work on the generalization property of the model trained by SGD. Classical generalization bounds only hold for one-pass SGD [24, 27, 28, 32, 36, 38, 39] where each training example can be used at most once. In practice, however, multiple passes are often used to produce a model with good generalization behavior [13]. The landmark work in [7] developed a framework to analyze generalization performance of multi-pass stochastic learning algorithms by taking into account the computational complexity of learning algorithms. Under this framework, the interplay among estimation errors, computational errors and approximation errors can be studied, showing that an implicit regularization can be achieved in the absence of penalization or constraints by tuning either the step size or the number of passes (the iteration number divided by the training set size) [13, 20, 21, 29]. In a parametric setting, it was shown that SGD is algorithmically stable and the stability measure of SGD with T iterates scales as $O(n^{-1}\sum_{t=1}^{T}\eta_t)$ [13], based on which a generalization bound $\mathbb{E}[\mathcal{E}(\bar{w}_T^{(1)})]-\inf_{w\in\mathcal{W}}\mathcal{E}(w)=O(n^{-\frac{1}{2}})$ was established for $\eta_t=O(1/\sqrt{n})$ and $T=O(n)$ without considering approximation errors. The discussion in [13] requires to impose a smoothness assumption on loss functions. Generalization analysis was considered separately for smooth and non-smooth loss functions [21]. For smooth loss functions, it was shown $\mathbb{E}[\mathcal{E}(\bar{w}_T^{(1)})]-\mathcal{E}(h_\rho)=O(n^{-\frac{\alpha}{1+\alpha}}\log n)$ for $\eta_t=\eta_1/\sqrt{t}$ with $T=\lceil n^{\frac{2}{\alpha+1}}\rceil$ [21], based on the stability property of SGD established in [13]. For non-smooth loss functions, it was shown $\mathbb{E}[\mathcal{E}(\bar{w}_T^{(1)})]-\mathcal{E}(h_\rho)=O(n^{-\frac{\alpha}{2\alpha+1}}\log n)$ for $\eta_t=\eta_1/\sqrt{t}$ and $T=\lceil n^{\frac{2}{2\alpha+1}}\rceil$, by controlling estimation errors with Rademacher complexities [3, 21]. Still, the bounds in [13, 21] require to impose a boundedness assumption on subgradients and are stated in expectation. As a comparison, we establish high-probability bounds without any boundedness assumptions on subgradients. Furthermore, our generalization analysis extends the analysis in [13] to non-smooth loss functions and substantially improve the bound $O(n^{-\frac{\alpha}{2\alpha+1}}\log n)$ [21] in this setting. The generalization error bound $O\big(n^{-\frac{\alpha}{1+\alpha}}\log^{\frac{3+\alpha\beta}{2}}\frac{n}{\delta}\big)$ in Corollary 11 is optimal in the sense that it matches the best available bound for Tikhonov regularization (up to a logarithmic factor) [9, 21, 34].

We achieve this improvement by controlling better estimation errors. Specifically, estimation errors were shown to scale polynomially w.r.t. the number of passes [13, 21], which dominate the other two errors for large T. In this way, one needs to tune T to balance the estimation, approximation and computational errors. As a comparison, we show bounds scaling logarithmically w.r.t. the number of passes for $\mathcal{E}(\bar{w}_T^{(1)})-\mathcal{E}_{\mathbf{z}}(\bar{w}_T^{(1)})$ (Theorem 10). This implies that estimation errors will never essentially dominate the other two errors and one can run SGD with a sufficient number of passes with little overfitting if step sizes are square-summable, due to the key observation on the almost boundedness of iterates established in Theorem 3. Another trick in getting almost optimal bounds includes the use of Assumption 3 to control $\mathcal{E}(w_\lambda)-\mathcal{E}_{\mathbf{z}}(w_\lambda)$ with a linear (instead of quadratic) function of $\sup_z f(w_\lambda,z)$ and to select a suitable λ, where $w_\lambda=\arg\min_{w\in\mathcal{W}}\mathcal{E}(w)+\lambda\|w\|_2^2$. Optimal learning rates were given for multi-pass SGD with the least squares loss function [10, 20, 29]. However, their analysis is based on an integral operator approach and does not apply to general loss functions. Generalization bounds for SGD were also studied from a PAC-Bayesian perspective [22]. However, the high-probability bounds there require to impose Lipschitz continuity, smoothness and strong convexity assumptions on loss functions, and ignore computational and approximation errors [22].

6 Simulations

Our analysis implies that SGD can be run with a sufficient number of iterations with little overfitting if step sizes are square-summable, which meanwhile can achieve similar generalization performance with different computational complexities. In this section, we include some experimental results to validate these theoretical findings. We apply SGD (4.1) with a linear kernel $K_x=x$ and the hinge loss $\ell(a,y)=\max\{0,1-ya\}$ to several binary classification datasets (ADULT, GISETTE, IJCNN, MUSHROOMS, PHISHING and SPLICE). All these datasets, described in Supplementary Material

G, can be download from the LIBSVM website [8]. We consider polynomially decaying step sizes of the form $\eta_t = 5t^{-\theta}$ with $\theta \in \{0.25, 0.51, 0.75\}$ (we consider $\theta = 0.51$, instead of $\theta = 0.5$, since the associated step size sequence is square-summable). We repeat experiments 12 times and report the average of results. In Figure 1, we plot test errors of $\bar{w}_t^{(3)} = \big(\sum_{k=\tilde{t}+1}^{t} \eta_k\big)^{-1} \sum_{k=\tilde{t}+1}^{t} \eta_k w_k$ versus the number of passes (the iteration number divided by the training set size), where $\tilde{t} = 2^{\lfloor \log_2 t \rfloor - 1}$. Intuitively, $\bar{w}_t^{(3)}$ returns an α-suffix average of iterates [28] with $\alpha \in [1/2, 3/4]$ and one can adapt the proof of Theorem 4 to show that $\bar{w}_t^{(3)}$ enjoys similar generalization bounds as $\bar{w}_t^{(1)}$. Moreover, $\bar{w}_t^{(3)}$ is easily computable on-the-fly by storing only $\sum_{j=1}^{k} \eta_j w_j$ with $k = 2^0, 2^1, 2^2, \ldots$. From Figure 1, we see that SGD is resistant to overfitting for appropriate step sizes. For example, we observe no overfitting even if the number of passes exceeds 1000 for SGD with $\theta \in \{0.51, 0.75\}$. Moreover, SGD with $\theta \in \{0.51, 0.75\}$ can achieve similar generalization errors on ADULT, IJCNN, PHISHING and SPLICE, towards which SGD with $\theta = 0.51$ requires a significantly smaller number of passes. This is well consistent with Corollary 11.



(a) ADULT. (b) GISETTE. (c) IJCNN.

(d) MUSHROOMS. (e) PHISHING. (f) SPLICE.

Figure 1: Test errors versus the number of passes.

7 Conclusions

In this paper, we establish a rigorous theoretical foundation for SCMD by providing optimal convergence rates (up to a logarithmic factor) in the stochastic optimization setting without boundedness assumptions on either subgradients or iterates, which in turn also shed new insights on the generalization behavior of the multi-pass SGD in the statistical learning theory setting. In particular, we justify the immunity of multi-pass SGD to overfitting by giving estimation error bounds with a *logarithmic* dependency on the number of passes for square-summable step sizes, while existing bounds scale *polynomially* [13, 21]. This improvement is based on the key observation on the almost boundedness of iterates with high probability. Our generalization analysis of SGD also substantially improves learning rates in [21], removes bounded subgradient assumptions in [13, 21, 22], removes smoothness assumptions in [13, 22] and is performed in high probability instead of in expectation [13, 21]. It would be interesting to extend our results to a non-convex setting [43] and to general mirror descent algorithms with a *non-differentiable* mirror map [18].

Acknowledgments

This work is supported in part by the National Key Research and Development Program of China (Grant No. 2017YFB1003102), the National Natural Science Foundation of China (Grant

Nos. 61806091 and 61672478), the Science and Technology Innovation Committee Foundation of Shenzhen (Grant No. ZDSYS201703031748284) and Shenzhen Peacock Plan (Grant No. KQT-D2016112514355531).

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*, 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] P. L. Bartlett and S. Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

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

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

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

[7] O. Bousquet and L. Bottou. The tradeoffs of large scale learning. In *Advances in Neural Information Processing Systems*, pages 161–168, 2008.

[8] 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.

[9] F. Cucker and D.-X. Zhou. *Learning Theory: an Approximation Theory Viewpoint*. Cambridge University Press, 2007.

[10] A. Dieuleveut and F. Bach. Nonparametric stochastic approximation with large step-sizes. *Annals of Statistics*, 44(4):1363–1399, 2016.

[11] 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.

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

[13] M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *International Conference on Machine Learning*, pages 1225–1234, 2016.

[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] E. Hazan, A. Agarwal, and S. Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2):169–192, 2007.

[16] 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.

[17] 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.

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

[19] Y. Lei, L. Shi, and Z.-C. Guo. Convergence of unregularized online learning algorithms. *Journal of Machine Learning Research*, 18(171):1–33, 2018.

[20] J. Lin and L. Rosasco. Optimal learning for multi-pass stochastic gradient methods. In *Advances in Neural Information Processing Systems*, pages 4556–4564, 2016.

[21] J. Lin, R. Camoriano, and L. Rosasco. Generalization properties and implicit regularization for multiple passes SGM. In *International Conference on Machine Learning*, pages 2340–2348, 2016.

[22] B. London. A PAC-bayesian analysis of randomized learning with application to stochastic gradient descent. In *Advances in Neural Information Processing Systems*, pages 2931–2940, 2017.

[23] E. Moulines and F. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459, 2011.

[24] A. Nemirovski, A. Juditsky, G. Lan, and A. Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609, 2009.

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

[26] L. M. Nguyen, P. H. Nguyen, M. van Dijk, P. Richtárik, K. Scheinberg, and M. Takáč. SGD and hogwild! convergence without the bounded gradients assumption. *arXiv preprint arXiv:1802.03801*, 2018.

[27] F. Orabona. Simultaneous model selection and optimization through parameter-free stochastic learning. In *Advances in Neural Information Processing Systems*, pages 1116–1124, 2014.

[28] 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.

[29] L. Rosasco and S. Villa. Learning with incremental iterative regularization. In *Advances in Neural Information Processing Systems*, pages 1630–1638, 2015.

[30] S. Shalev-Shwartz and A. Tewari. Stochastic methods for ℓ_1-regularized loss minimization. *Journal of Machine Learning Research*, 12:1865–1892, 2011.

[31] S. Shalev-Shwartz, Y. Singer, and N. Srebro. Pegasos: Primal estimated sub-gradient solver for SVM. In *International Conference on Machine Learning*, pages 807–814. ACM, 2007.

[32] 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.

[33] S. Smale and D.-X. Zhou. Estimating the approximation error in learning theory. *Analysis and Applications*, 1(01):17–41, 2003.

[34] I. Steinwart and A. Christmann. *Support Vector Machines*. Springer Science & Business Media, 2008.

[35] I. Steinwart and C. Scovel. Fast rates for support vector machines using gaussian kernels. *Annals of Statistics*, 35(2):575–607, 2007.

[36] 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.

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

[38] Y. Ying and M. Pontil. Online gradient descent learning algorithms. *Foundations of Computational Mathematics*, 8(5):561–596, 2008.

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

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

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

[42] P. Zhao and T. Zhang. Stochastic optimization with importance sampling for regularized loss minimization. In *International Conference on Machine Learning*, pages 1–9, 2015.

[43] 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.

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

Supplementary Material

Yunwen Lei and **Ke Tang***
Shenzhen Key Laboratory of Computational Intelligence, Department of Computer Science and Engineering, Southern University of Science and Technology, Shenzhen 518055, China
leiyw@sustc.edu.cn tangk3@sustc.edu.cn

A Technical Lemmas

A.1 Concentration Inequalities

Our discussion on high-probability bounds is based on the following two concentration inequalities. Lemma A.1 quantifies the concentration behavior of martingales. Part (a) is the Azuma-Hoeffding inequality for martingales with bounded increments [4], and part (b) is a conditional Bernstein inequality using the conditional variance to quantify better the concentration behavior of martingales [10]. Lemma A.2 is the McDiarmid's inequality to arbitrary real-valued functions of independent random variables that satisfy a bounded increment condition [6].

Lemma A.1. *Let $z_1, \ldots, z_n$ be a sequence of random variables such that z_k may depend on the previous random variables $z_1, \ldots, z_{k-1}$ for all $k = 1, \ldots, n$. Consider a sequence of functionals $\xi_k(z_1, \ldots, z_k), k = 1, \ldots, n$. Let $\sigma_n^2 = \sum_{k=1}^n \mathbb{E}_{z_k}\big[\big(\xi_k - \mathbb{E}_{z_k}[\xi_k]\big)^2\big]$ be the conditional variance.*

(a) Assume that $|\xi_k - \mathbb{E}_{z_k}[\xi_k]| \leq b_k$ for each k. Let $\delta \in (0,1)$. With probability at least $1-\delta$ we have

$$\sum_{k=1}^n \xi_k - \sum_{k=1}^n \mathbb{E}_{z_k}[\xi_k] \leq \Big(2\sum_{k=1}^n b_k^2 \log\frac{1}{\delta}\Big)^{\frac{1}{2}}. \tag{A.1}$$

(b) Assume that $\xi_k - \mathbb{E}_{z_k}[\xi_k] \leq b$ for each k. Let $\rho \in (0,1]$ and $\delta \in (0,1)$. With probability at least $1-\delta$ we have

$$\sum_{k=1}^n \xi_k - \sum_{k=1}^n \mathbb{E}_{z_k}[\xi_k] \leq \frac{\rho\sigma_n^2}{b} + \frac{b\log\frac{1}{\delta}}{\rho}. \tag{A.2}$$

Lemma A.2. *Let $c_1, \ldots, c_n \in \mathbb{R}_+$. Let $Z_1, \ldots, Z_n$ be independent random variables taking values in a set $\mathcal{Z}$, and assume that $f : \mathcal{Z}^n \to \mathbb{R}$ satisfies*

$$\sup_{z_1,\ldots,z_n,\bar{z}_k\in\mathcal{Z}} |f(z_1,\cdots,z_n) - f(z_1,\cdots,z_{k-1},\bar{z}_k,z_{k+1},\cdots,z_n)| \leq c_k \tag{A.3}$$

for $k = 1, \ldots, n$. Then, for any $0 < \delta < 1$, with probability at least $1-\delta$ we have

$$f(Z_1,\ldots,Z_n) \leq \mathbb{E}\big[f(Z_1,\ldots,Z_n)\big] + \sqrt{\frac{\sum_{k=1}^n c_k^2 \log(1/\delta)}{2}}.$$

A.2 Behavior of Objectives

In this section, we collect some lemmas on functions g satisfying

$$\|g'(w)\|_*^2 \leq Ag(w) + B \tag{A.4}$$

for some constant $A, B \geq 0$. Lemma A.3 shows that, if g satisfies (A.4), then both $\|g'(w)\|_*^2$ and $g(w)$ can be controlled by quadratic functions of $\|w\|$.

*Corresponding author

Lemma A.3. *Let $g : \mathcal{W} \mapsto \mathbb{R}$ be a convex function. If there exist A and B such that* (A.4) *holds for all $w \in \mathcal{W}$. Then*

$$\|g'(w)\|_*^2 \leq 2A^2\|w\|^2 + 2Ag(0) + 2B \quad \textit{and} \quad g(w) \leq \big(A^2 + \frac{1}{2}\big)\|w\|^2 + (A+1)g(0) + B. \quad \text{(A.5)}$$

Proof. According to (A.4) and the convexity of g, we know

$$\begin{aligned}\|g'(w)\|_*^2 &\leq A\big(g(w) - g(0)\big) + Ag(0) + B \\ &\leq A\langle w, g'(w)\rangle + Ag(0) + B \leq A\|w\|\|g'(w)\|_* + Ag(0) + B.\end{aligned}$$

Solving the above quadratic inequality of $\|g'(w)\|_*$ shows

$$\|g'(w)\|_* \leq A\|w\| + \sqrt{Ag(0) + B},$$

from which and the elementary inequality $(a+b)^2 \leq 2(a^2 + b^2)$ we derive the first inequality.

We now turn to the second inequality. By the convexity of g and the first inequality in (A.5), we get

$$\begin{aligned}g(w) - g(0) &\leq \langle w, g'(w)\rangle \leq \|w\|\|g'(w)\|_* \\ &\leq \frac{\|w\|^2}{2} + \frac{\|g'(w)\|_*^2}{2} \leq \frac{\|w\|^2}{2} + A^2\|w\|^2 + Ag(0) + B,\end{aligned}$$

from which we derive the second inequality. The proof is complete. ◻

Lemma A.4 shows that functions of the form $f(w, z) = \ell(\langle w, x\rangle, y)$ would satisfy (3.1) if ℓ satisfies (A.6).

Lemma A.4. *Let $\ell : \mathbb{R} \times \mathcal{Y} \mapsto \mathbb{R}$ and $f(w, z) = \ell(\langle w, x\rangle, y)$ with $z = (x, y)$. If there exist $\tilde{A}, \tilde{B} \geq 0$ such that*

$$|\ell'(a, y)|^2 \leq \tilde{A}\ell(a, y) + \tilde{B}, \quad \forall a \in \mathbb{R}, y \in \mathcal{Y}. \quad \text{(A.6)}$$

Then we have $\|f'(w, z)\|_*^2 \leq Af(w, z) + B$ for any $w \in \mathcal{W}$ and $z \in \mathcal{Z}$, where $\kappa = \sup_{x\in\mathcal{X}} \|x\|_*$, $A = \tilde{A}\kappa^2$ and $B = \tilde{B}\kappa^2$.

Proof. For any $w \in \mathcal{W}$ and $z \in \mathcal{Z}$, it follows from (A.6) that

$$\|f'(w, z)\|_*^2 = \big\|\ell'(\langle w, x\rangle, y)x\big\|_*^2 \leq \kappa^2\Big(\tilde{A}\ell(\langle w, x\rangle, y) + \tilde{B}\Big) = \kappa^2\big(\tilde{A}f(w, z) + \tilde{B}\big).$$

The proof is complete. ◻

Lemma A.5 shows that regularizers $r_p(w) = \|w\|_p^p, p \in [1, 2]$ satisfy the condition (3.1). For $a \in \mathbb{R}$, denote by $\text{sgn}(a)$ the sign of a, i.e., $\text{sgn}(a) = 1$ if $a > 0$, $\text{sgn}(a) = -1$ if $a < 0$ and $\text{sgn}(a) = 0$ if $a = 0$.

Lemma A.5. *The function $r_p(w) = \|w\|_p^p$ with $1 \leq p \leq 2$ defined on $\mathcal{W}$ satisfies*

$$\|r_p'(w)\|_{p^*}^2 \leq p\big(2(p-1)\|w\|_p^p + 2 - p\big), \quad \forall w \in \mathcal{W},$$

where $p^* = \frac{p}{p-1}$ is the conjugate exponent of p.

Proof. If $p = 1$, then any $r_1'(w) \in \partial r_1(w)$ would satisfy $\|r_1'(w)\|_\infty \leq 1$, from which and $p^* = \infty$ we know $\|r_1'(w)\|_{p^*}^2 \leq 1$.

If $p > 1$, then the gradient of r_p at w can be calculated by $\nabla r_p(w) = p\big(\text{sgn}(w(i))|w(i)|^{p-1}\big)_{i=1}^d$, from which we have

$$\|\nabla r_p(w)\|_{p^*} = p\Big(\sum_{i=1}^d \big|\text{sgn}(w(i))|w(i)|^{p-1}\big|^{p^*}\Big)^{\frac{1}{p^*}} = p\Big(\sum_{i=1}^d |w(i)|^{p^*(p-1)}\Big)^{\frac{1}{p^*}} = p\|w\|_p^{p-1}.$$

It then follows from the Young's inequality

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

that

$$\|\nabla r_p(w)\|_{p^*}^2 = p^2\|w\|_p^{2(p-1)} \leq p^2\left(\frac{\|w\|_p^{2(p-1)\frac{p}{2(p-1)}}}{\frac{p}{2(p-1)}} + \frac{2-p}{p}\right) = p\big(2(p-1)\|w\|_p^p + 2 - p\big).$$

The proof is complete by combining the above two cases together. ◻

B Proofs for Lemma 1 and Lemma 2

In this section, we prove Lemma 1 quantifying the one-step progress of SCMD (2.2), and Lemma 2 which plays an important role in removing the boundedness assumptions on subgradients.

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

$$\eta_t f'(w_t, z_t) + \eta_t r'(w_{t+1}) + \nabla\Psi(w_{t+1}) - \nabla\Psi(w_t) = 0,$$

from which and the identity $D_\Psi(w, w_{t+1}) + D_\Psi(w_{t+1}, w_t) - D_\Psi(w, w_t) = \langle w - w_{t+1}, \nabla\Psi(w_t) - \nabla\Psi(w_{t+1})\rangle$, we derive

$$\begin{aligned}
D_\Psi(w, w_{t+1}) - D_\Psi(w, w_t) &= D_\Psi(w, w_{t+1}) + D_\Psi(w_{t+1}, w_t) - D_\Psi(w, w_t) - D_\Psi(w_{t+1}, w_t) \\
&= \langle w - w_{t+1}, \nabla\Psi(w_t) - \nabla\Psi(w_{t+1})\rangle - D_\Psi(w_{t+1}, w_t) \\
&= \eta_t\langle w - w_{t+1}, f'(w_t, z_t) + r'(w_{t+1})\rangle - D_\Psi(w_{t+1}, w_t) \\
&\le \eta_t\langle w - w_{t+1}, f'(w_t, z_t)\rangle + \eta_t\big[r(w) - r(w_{t+1}) - \sigma_r D_\Psi(w, w_{t+1})\big] - D_\Psi(w_{t+1}, w_t) \\
&= \eta_t\langle w - w_t, f'(w_t, z_t)\rangle + \eta_t\langle w_t - w_{t+1}, f'(w_t, z_t)\rangle + \eta_t[r(w) - r(w_t)] \\
&\qquad + \eta_t[r(w_t) - r(w_{t+1})] - \sigma_r\eta_t D_\Psi(w, w_{t+1}) - D_\Psi(w_{t+1}, w_t).
\end{aligned} \tag{B.1}$$

Here, we have used the σ_r-strong convexity of r (3.3) in the inequality. From the convexity of r, the definition of dual norm and the strong convexity of Ψ, it follows that

$$\begin{aligned}
&\eta_t\big[\langle w_t - w_{t+1}, f'(w_t, z_t)\rangle + r(w_t) - r(w_{t+1})\big] - D_\Psi(w_{t+1}, w_t) \\
&\le \eta_t\|w_t - w_{t+1}\|\|f'(w_t, z_t)\|_* + \eta_t\langle w_t - w_{t+1}, r'(w_t)\rangle - 2^{-1}\sigma_\Psi\|w_t - w_{t+1}\|^2 \\
&\le \eta_t\|w_t - w_{t+1}\|\big[\|f'(w_t, z_t)\|_* + \|r'(w_t)\|_*\big] - 2^{-1}\sigma_\Psi\|w_t - w_{t+1}\|^2 \\
&\le 2^{-1}\sigma_\Psi\|w_t - w_{t+1}\|^2 + 2^{-1}\sigma_\Psi^{-1}\eta_t^2\big[\|f'(w_t, z_t)\|_* + \|r'(w_t)\|_*\big]^2 - 2^{-1}\sigma_\Psi\|w_t - w_{t+1}\|^2 \\
&\le \sigma_\Psi^{-1}\eta_t^2\big[\|f'(w_t, z_t)\|_*^2 + \|r'(w_t)\|_*^2\big] \le \sigma_\Psi^{-1}\eta_t^2\big[Af(w_t, z_t) + Ar(w_t) + 2B\big],
\end{aligned}$$

where we have used the elementary inequality $(a+b)^2 \le 2(a^2 + b^2)$ and (3.1) in the last two inequalities. Plugging the above inequality back into (B.1), we get the stated inequality and complete the proof. □

Proof of Lemma 2. Using the convexity of f in (3.4), we derive the following inequality for any $w \in \mathcal{W}$

$$\begin{aligned}
&D_\Psi(w, w_{t+1}) - D_\Psi(w, w_t) \\
&\le \eta_t\big(f(w, z_t) - f(w_t, z_t)\big) + \eta_t(r(w) - r(w_t)) + \sigma_\Psi^{-1}\eta_t^2\big(Af(w_t, z_t) + Ar(w_t) + 2B\big) \\
&= \eta_t(f(w, z_t) + r(w)) + (\sigma_\Psi^{-1}\eta_t^2 A - \eta_t)\big(f(w_t, z_t) + r(w_t)\big) + 2\sigma_\Psi^{-1}B\eta_t^2 \qquad \text{(B.2)} \\
&\le \eta_t(f(w, z_t) + r(w)) + A^{-1}B\eta_t,
\end{aligned}$$

where the last inequality is due to the assumption $\eta_t \le (2A)^{-1}\sigma_\Psi$. Plugging $w = 0$ in the above inequality and using the definition of C_1, we derive

$$D_\Psi(0, w_{t+1}) - D_\Psi(0, w_t) \le \eta_t(f(0, z_t) + r(0)) + A^{-1}B\eta_t \le \eta_t C_1.$$

It then follows that

$$D_\Psi(0, w_{t+1}) = D_\Psi(0, w_1) + \sum_{k=1}^t \big[D_\Psi(0, w_{k+1}) - D_\Psi(0, w_k)\big] \le C_1\sum_{k=1}^t \eta_k, \tag{B.3}$$

where we have used $w_1 = 0$ in the last inequality. The stated inequality (3.5) then follows from the σ_Ψ-strong convexity of Ψ.

We now prove (3.6). Taking $w = 0$ in (B.2) and using $\eta_t \le 2^{-1}A^{-1}\sigma_\Psi$, we get

$$2^{-1}\eta_t\big(f(w_t, z_t) + r(w_t)\big) \le \eta_t\big(f(0, z_t) + r(0)\big) + 2\sigma_\Psi^{-1}B\eta_t^2 + D_\Psi(0, w_t) - D_\Psi(0, w_{t+1}). \tag{B.4}$$

Multiplying both sides by $2\eta_t$ then gives

$$\begin{aligned}\eta_t^2\big(f(w_t,z_t)+r(w_t)\big) &\le 2\eta_t^2\big(f(0,z_t)+r(0)\big)+4\sigma_\Psi^{-1}B\eta_t^3+2\eta_t\big(D_\Psi(0,w_t)-D_\Psi(0,w_{t+1})\big)\\ &\le 2\eta_t^2\big(f(0,z_t)+r(0)\big)+2A^{-1}B\eta_t^2+2\eta_t D_\Psi(0,w_t)-2\eta_{t+1}D_\Psi(0,w_{t+1})\\ &\le 2C_1\eta_t^2+2\eta_t D_\Psi(0,w_t)-2\eta_{t+1}D_\Psi(0,w_{t+1}),\end{aligned}$$

where we have used $\eta_t \le (2A)^{-1}\sigma_\Psi, \eta_{t+1}\le \eta_t$ in the second inequality and the definition of C_1 in the last inequality. Taking a summation of the above inequality further implies

$$\sum_{k=1}^t \eta_k^2\big(f(w_k,z_k)+r(w_k)\big) \le 2C_1\sum_{k=1}^t \eta_k^2 + 2\eta_1 D_\Psi(0,w_1) = 2C_1\sum_{k=1}^t \eta_k^2,$$

where the last identity is due to $w_1=0$. This proves (3.6).

We now prove (3.7). Plugging the inequality $\eta_t \le (2A)^{-1}\sigma_\Psi$ into (B.4) and multiplying both sides by $2\eta_t^{-1}$, we know

$$f(w_t,z_t)+r(w_t) \le 2\big(f(0,z_t)+r(0)\big)+2A^{-1}B+2\eta_t^{-1}\big(D_\Psi(0,w_t)-D_\Psi(0,w_{t+1})\big).$$

Taking a summation of the above inequality, we derive

$$\sum_{k=1}^t\big(f(w_k,z_k)+r(w_k)\big) \le 2\sum_{k=1}^t\big(f(0,z_k)+r(0)+A^{-1}B\big)+2\sum_{k=1}^t \eta_k^{-1}\big(D_\Psi(0,w_k)-D_\Psi(0,w_{k+1})\big).$$

The last term can be controlled by (note $w_1=0$)

$$\begin{aligned}\sum_{k=1}^t \eta_k^{-1}\big(D_\Psi(0,w_k)-D_\Psi(0,w_{k+1})\big) &= \sum_{k=2}^t D_\Psi(0,w_k)\big(\eta_k^{-1}-\eta_{k-1}^{-1}\big)+\eta_1^{-1}D_\Psi(0,w_1)-\eta_t^{-1}D_\Psi(0,w_{t+1})\\ &\le \max_{1\le\tilde{k}\le t} D_\Psi(0,w_{\tilde{k}})\sum_{k=2}^t\big(\eta_k^{-1}-\eta_{k-1}^{-1}\big) \le \max_{1\le\tilde{k}\le t} D_\Psi(0,w_{\tilde{k}})\eta_t^{-1} \le C_1\Big(\sum_{k=1}^t \eta_k\Big)\eta_t^{-1},\end{aligned}$$

where the last inequality is due to (B.3). Combining the above two inequalities together and using the definition of C_1, we derive the stated inequality (3.7). The proof is complete.

□

C Proofs for General Convex Objectives

In this section, we prove Theorem 3 and Theorem 4. We first provide a proposition to show that $\|w_{t+1}-w^*\|^2$ can be controlled by $O\big(\sum_{k=1}^t \eta_k^2\|w_k-w^*\|^2\big)$ with high probability. To this aim, we take $w=w^*$ in (3.4) to derive

$$D_\Psi(w^*,w_{t+1}) \le \sum_{k=1}^t \xi_k + \sum_{k=1}^t \eta_k\big(\phi(w^*)-\phi(w_k)\big)+\widetilde{C}_1\sum_{k=1}^t \eta_k^2, \tag{C.1}$$

where ξ_k is defined in (C.4) and $\widetilde{C}_1\in\mathbb{R}$. A key idea is to use a conditional Bernstein inequality to show $\sum_{k=1}^t \xi_k \le \sum_{k=1}^t \eta_k\big(\phi(w_k)-\phi(w^*)\big)+\widetilde{C}_2\sum_{k=1}^t\eta_k^2\|w_k-w^*\|^2$ with high probability. An interesting observation is that one can offset the term $\sum_{k=1}^t \eta_k\big(\phi(w^*)-\phi(w_k)\big)$ in (C.1) by the above bound on $\sum_{k=1}^t\xi_k$, leading to the inequality $D_\Psi(w^*,w_{t+1}) \le \widetilde{C}_1\sum_{k=1}^t\eta_k^2+\widetilde{C}_2\sum_{k=1}^t\eta_k^2\|w_k-w^*\|^2$ with high probability. In the discussion of the conditional variance, we use $\mathbb{E}_{z_k}\big[\big(\xi_k-\mathbb{E}_{z_k}[\xi_k]\big)^2\big] \le \eta_k^2\|w_k-w^*\|^2(A\phi(w_k)+B)$ and introduce the following decomposition

$$\eta_k^2\|w_k-w^*\|^2\big(A\phi(w_k)+B\big) = \eta_k^2\|w_k-w^*\|^2A\big(\phi(w_k)-\phi(w^*)\big)+\eta_k^2\|w_k-w^*\|^2\big(A\phi(w^*)+B\big).$$

We apply (3.5) to control the first $\|w_k-w^*\|^2$ on the right-hand side to show $\eta_k^2\|w_k-w^*\|^2A\big(\phi(w_k)-\phi(w^*)\big) \le \widetilde{C}_3\eta_k\big(\phi(w_k)-\phi(w^*)\big)$ for a $\widetilde{C}_3>0$. As a comparison, the second $\|w_k-w^*\|^2$ is kept intact.

Proposition C.1. *Let $\{w_t\}_{t\in\mathbb{N}}$ be the sequence produced by* (2.2) *with $\eta_t \le (2A)^{-1}\sigma_\Psi$ and $\eta_{t+1} \le \eta_t$. We assume $C_6 = \sup_{k\in\mathbb{N}} \eta_k \sum_{j=1}^{k-1}\eta_j < \infty$. Then for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\|w_{t+1}-w^*\|^2 \le \frac{A\phi(w^*)+B}{2C_1C_6A}\sum_{k=1}^t \eta_k^2\|w_k-w^*\|^2 + \frac{2D_\Psi(w^*,0)}{\sigma_\Psi} + \frac{2C_7\log\frac{1}{\delta}}{\rho_1\sigma_\Psi} + 4\sigma_\Psi^{-2}(B+AC_1)\sum_{k=1}^t \eta_k^2, \tag{C.2}$$

where $\rho_1 = \min\{1, (2A)^{-1}(\eta_1\|w^*\|^2 + 2C_1C_6\sigma_\Psi^{-1})^{-1}C_7\}$ and

$$C_7 = \eta_1\big(\sup_{z\in\mathcal{Z}} f(w^*,z) + \|w^*\|^2 + AF(0) + B\big) + 2(A^2+1)C_1\sigma_\Psi^{-1}C_6.$$

Proof. Setting $w = w^*$ in (3.4) shows

$$\begin{aligned} D_\Psi(w^*,w_{t+1}) - D_\Psi(w^*,w_t) &\le \eta_t\langle w^*-w_t, f'(w_t,z_t)\rangle + \eta_t(r(w^*)-r(w_t)) \\ &\quad + \sigma_\Psi^{-1}\eta_t^2\big(Af(w_t,z_t) + Ar(w_t) + 2B\big). \end{aligned}$$

We write

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

Combining the above equations together and using the definition of ϕ, we derive

$$\begin{aligned} D_\Psi(w^*,w_{t+1}) - D_\Psi(w^*,w_t) &\le \eta_t\langle w^*-w_t, f'(w_t,z_t) - \mathbb{E}_{z_t}[f'(w_t,z_t)]\rangle \\ &\quad + \eta_t(\phi(w^*)-\phi(w_t)) + \sigma_\Psi^{-1}\eta_t^2\big(Af(w_t,z_t) + Ar(w_t) + 2B\big). \end{aligned}$$

Together with $w_1 = 0$, it then follows that

$$\begin{aligned} D_\Psi(w^*,w_{t+1}) &= D_\Psi(w^*,w_1) + \sum_{k=1}^t \big(D_\Psi(w^*,w_{k+1}) - D_\Psi(w^*,w_k)\big) \\ &\le D_\Psi(w^*,0) + \sum_{k=1}^t \eta_k\langle w^*-w_k, f'(w_k,z_k) - \mathbb{E}_{z_k}[f'(w_k,z_k)]\rangle \\ &\quad + \sum_{k=1}^t \eta_k\big(\phi(w^*)-\phi(w_k)\big) + \sigma_\Psi^{-1}\sum_{k=1}^t \eta_k^2\big(Af(w_k,z_k) + Ar(w_k) + 2B\big). \end{aligned} \tag{C.3}$$

Introduce a sequence of random variables as follows

$$\xi_k = \eta_k\langle w^*-w_k, f'(w_k,z_k) - \mathbb{E}_{z_k}[f'(w_k,z_k)]\rangle, \quad k\in\mathbb{N}. \tag{C.4}$$

It is clear that $\mathbb{E}_{z_k}[\xi_k] = 0$ and therefore $\{\xi_k\}_k$ is a martingale difference sequence. Since $\mathbb{E}[(\xi - \mathbb{E}[\xi])^2] \le \mathbb{E}[\xi^2]$ for any real-valued random variable ξ, we know

$$\begin{aligned} \mathbb{E}_{z_k}\big[\big|\langle w^*-w_k, f'(w_k,z_k) - \mathbb{E}_{z_k}[f'(w_k,z_k)]\rangle\big|^2\big] &\le \mathbb{E}_{z_k}\big[\big|\langle w^*-w_k, f'(w_k,z_k)\rangle\big|^2\big] \\ &\le \|w^*-w_k\|^2\mathbb{E}_{z_k}\big[\|f'(w_k,z_k)\|_*^2\big] \le \|w^*-w_k\|^2\mathbb{E}_{z_k}\big[Af(w_k,z_k)+B\big] \\ &\le \|w^*-w_k\|^2\big(AF(w_k) + Ar(w_k) + B\big), \end{aligned}$$

where we have used (3.1) in the third inequality. Then, the conditional variance of ξ_k can be controlled by

$$\sum_{k=1}^{t}\mathbb{E}_{z_k}\big[\big(\xi_k-\mathbb{E}_{z_k}[\xi_k]\big)^2\big]=\sum_{k=1}^{t}\eta_k^2\mathbb{E}_{z_k}\big[\big|\langle w^*-w_k,f'(w_k,z_k)-\mathbb{E}_{z_k}[f'(w_k,z_k)]\rangle\big|^2\big]$$
$$\le\sum_{k=1}^{t}\eta_k^2\|w^*-w_k\|^2\big(A\phi(w_k)-A\phi(w^*)\big)+\sum_{k=1}^{t}\eta_k^2\|w^*-w_k\|^2\big(A\phi(w^*)+B\big)$$
$$\le 2\sum_{k=1}^{t}\eta_k^2(\|w^*\|^2+\|w_k\|^2)\big(A\phi(w_k)-A\phi(w^*)\big)+\sum_{k=1}^{t}\eta_k^2\|w^*-w_k\|^2\big(A\phi(w^*)+B\big)$$
$$\le 2A\sum_{k=1}^{t}\eta_k\Big(\eta_k\|w^*\|^2+2C_1\sigma_\Psi^{-1}\eta_k\sum_{j=1}^{k-1}\eta_j\Big)\big(\phi(w_k)-\phi(w^*)\big)+\sum_{k=1}^{t}\eta_k^2\|w_k-w^*\|^2\big(A\phi(w^*)+B\big)$$
$$\le 2A(\eta_1\|w^*\|^2+2C_1\sigma_\Psi^{-1}C_6)\sum_{k=1}^{t}\eta_k\big(\phi(w_k)-\phi(w^*)\big)+\sum_{k=1}^{t}\eta_k^2\|w_k-w^*\|^2\big(A\phi(w^*)+B\big),\tag{C.5}$$

where the last second inequality is due to (3.5) and the last inequality is due to the definition of C_6.

Furthermore, it follows from the convexity of f that

$$\begin{aligned}\xi_k-\mathbb{E}_{z_k}[\xi_k]&=\eta_k\langle w^*-w_k,f'(w_k,z_k)\rangle+\eta_k\langle w_k-w^*,\mathbb{E}_{z_k}[f'(w_k,z_k)]\rangle\\&\le\eta_k(f(w^*,z_k)-f(w_k,z_k))+\eta_k\|w_k-w^*\|\big\|\mathbb{E}_{z_k}[f'(w_k,z_k)]\big\|_*.\end{aligned}\tag{C.6}$$

By the Schwarz's inequality and Lemma A.3, we know

$$\begin{aligned}&\|w_k-w^*\|\big\|\mathbb{E}_{z_k}[f'(w_k,z_k)]\big\|_*\\&\le\frac{1}{2}\big(\|w_k-w^*\|^2+\|F'(w_k)\|_*^2\big)\le\frac{1}{2}\big(2\|w_k\|^2+2\|w^*\|^2+2A^2\|w_k\|^2+2AF(0)+2B\big)\\&\le 2(A^2+1)C_1\sigma_\Psi^{-1}\sum_{j=1}^{k-1}\eta_j+\|w^*\|^2+AF(0)+B,\end{aligned}$$

where the last inequality is due to (3.5). Plugging the above inequality back into (C.6) and using the non-negativity of $f(w_t,z_t)$ then give

$$\xi_k-\mathbb{E}_{z_k}[\xi_k]\le\eta_1\big(\sup_{z\in\mathcal{Z}}f(w^*,z)+\|w^*\|^2+AF(0)+B\big)+2(A^2+1)C_1\sigma_\Psi^{-1}\eta_k\sum_{j=1}^{k-1}\eta_j\le C_7.$$

Applying Part (b) of Lemma A.1 with the above estimates on magnitudes and variances of ξ_k, we derive the following inequality with probability at least $1-\delta$

$$\begin{aligned}\sum_{k=1}^{t}\xi_k\le\frac{\rho_1}{C_7}\Big(2A\big(\eta_1\|w^*\|^2+2C_1\sigma_\Psi^{-1}C_6\big)\sum_{k=1}^{t}\eta_k\big(\phi(w_k)-\phi(w^*)\big)+\sum_{k=1}^{t}\eta_k^2\|w_k-w^*\|^2\big(A\phi(w^*)+B\big)\Big)\\+\frac{C_7\log\frac{1}{\delta}}{\rho_1}\le\sum_{k=1}^{t}\eta_k\big(\phi(w_k)-\phi(w^*)\big)+\frac{\sigma_\Psi\big(A\phi(w^*)+B\big)}{4C_1C_6A}\sum_{k=1}^{t}\eta_k^2\|w_k-w^*\|^2+\frac{C_7\log\frac{1}{\delta}}{\rho_1},\end{aligned}$$

where we have used $2\rho_1A(\eta_1\|w^*\|^2+2C_1C_6\sigma_\Psi^{-1})\le C_7$. By (3.6) we know

$$\sum_{k=1}^{t}\eta_k^2\big(Af(w_k,z_k)+Ar(w_k)+2B\big)\le\big(2AC_1+2B\big)\sum_{k=1}^{t}\eta_k^2.$$

Plugging the above two inequalities back into (C.3) gives the following inequality with probability $1-\delta$

$$D_\Psi(w^*,w_{t+1})\le D_\Psi(w^*,0)+\frac{\sigma_\Psi(A\phi(w^*)+B)}{4C_1C_6A}\sum_{k=1}^{t}\eta_k^2\|w_k-w^*\|^2+\frac{C_7\log\frac{1}{\delta}}{\rho_1}+2\sigma_\Psi^{-1}(B+AC_1)\sum_{k=1}^{t}\eta_k^2.$$

This together with the σ_Ψ-strong convexity of Ψ gives the stated bound with probability $1-\delta$. The proof is complete.

□

We can use the assumption $\sum_{k=1}^{\infty}\eta_k^2<\infty$ to show that the right-hand side of (C.2) can be bounded by $\frac{1}{2}\max_{1\le k\le t}\|w_k-w^*\|^2+\widetilde{C}\log\frac{1}{\delta}$ for a $\widetilde{C}>0$, from which we can show the boundedness of $\|w_t\|^2$ with high probability (up to a logarithmic factor).

Proof of Theorem 3. It follows from the assumption $\sum_{t=1}^{\infty}\eta_t^2<\infty$ and $\eta_{t+1}\le\eta_t$ that $\sup_t\eta_t\sum_{j=1}^{t-1}\eta_j\le\sum_{j=1}^{\infty}\eta_j^2<\infty$. Therefore, $C_6=\sup_{k\in\mathbb{N}}\eta_k\sum_{j=1}^{k-1}\eta_j$ is well defined. We define the set Ω_T as

$$\Omega_T=\Big\{(z_1,\ldots,z_T):\|w_{t+1}-w^*\|^2\le\frac{A\phi(w^*)+B}{2C_1C_6A}\sum_{k=1}^{t}\eta_k^2\|w_k-w^*\|^2+\frac{2D_\Psi(w^*,0)}{\sigma_\Psi}+\frac{2C_7\log\frac{T}{\delta}}{\rho_1\sigma_\Psi}+4\sigma_\Psi^{-2}(B+AC_1)\sum_{k=1}^{t}\eta_k^2\text{ for all }t=1,\ldots,T\Big\},$$

where ρ_1 is defined in Proposition C.1. By Proposition C.1 and union bounds on probability of events, we know $\Pr\{\Omega_T\}\ge1-\delta$. Since $\sum_{t=1}^{\infty}\eta_t^2<\infty$, we can find a $t_1\in\mathbb{N}$ such that $\big(A\phi(w^*)+B\big)\sum_{k=t_1+1}^{\infty}\eta_k^2\le C_1C_6A$. With the occurrence of Ω_T, the following inequality holds for all $t=1,\ldots,T$

$$\begin{aligned}&\|w_{t+1}-w^*\|^2-\frac{2C_7\log\frac{T}{\delta}}{\rho_1\sigma_\Psi}-\frac{2D_\Psi(w^*,0)}{\sigma_\Psi}\\&\le\frac{A\phi(w^*)+B}{2C_1C_6A}\Big(\sum_{k=1}^{t_1}\eta_k^2\|w_k-w^*\|^2+\sum_{k=t_1+1}^{t}\eta_k^2\|w_k-w^*\|^2\Big)+\frac{4(B+AC_1)\sum_{k=1}^{t}\eta_k^2}{\sigma_\Psi^2}\\&\le\frac{A\phi(w^*)+B}{2C_1C_6A}\Big(\sum_{k=1}^{t_1}\eta_k^2\|w_k-w^*\|^2+\sum_{k=t_1+1}^{t}\eta_k^2\sup_{1\le\tilde{k}\le T}\|w_{\tilde{k}}-w^*\|^2\Big)+\frac{4(B+AC_1)\sum_{k=1}^{t}\eta_k^2}{\sigma_\Psi^2}\\&\le\frac{A\phi(w^*)+B}{C_1C_6A}\sum_{k=1}^{t_1}\big(2C_1\sigma_\Psi^{-1}\eta_k^2\sum_{j=1}^{k-1}\eta_j+\|w^*\|^2\big)+\frac{1}{2}\sup_{1\le k\le T}\|w_k-w^*\|^2+\frac{4(B+AC_1)\sum_{k=1}^{t}\eta_k^2}{\sigma_\Psi^2},\end{aligned}$$

where we have used $\|w_k-w^*\|^2\le2(\|w_k\|^2+\|w^*\|^2)$ and (3.5) in the last step. Under the event Ω_T, we then have

$$\max_{1\le t\le T}\|w_t-w^*\|^2\le\frac{A\phi(w^*)+B}{C_1C_6A}\sum_{k=1}^{t_1}\big(2C_1\sigma_\Psi^{-1}\eta_k^2\sum_{j=1}^{k-1}\eta_j+\|w^*\|^2\big)+\frac{1}{2}\sup_{1\le k\le T}\|w_k-w^*\|^2+\frac{2C_7\log\frac{T}{\delta}}{\rho_1\sigma_\Psi}+\frac{2D_\Psi(w^*,0)}{\sigma_\Psi}+\frac{4(B+AC_1)\sum_{k=1}^{t}\eta_k^2}{\sigma_\Psi^2},$$

from which and $\Pr\{\Omega_T\}\ge1-\delta$ we derive the following inequality with probability at least $1-\delta$

$$\max_{1\le t\le T}\|w_t-w^*\|^2\le\frac{2(A\phi(w^*)+B)}{C_1C_6A}\sum_{k=1}^{t_1}\big(2C_1\sigma_\Psi^{-1}\eta_k^2\sum_{j=1}^{k-1}\eta_j+\|w^*\|^2\big)+\frac{4C_7\log\frac{T}{\delta}}{\rho_1\sigma_\Psi}+\frac{4D_\Psi(w^*,0)}{\sigma_\Psi^2}+\frac{8(B+AC_1)\sum_{k=1}^{t}\eta_k^2}{\sigma_\Psi}.$$

The stated inequality holds with C_2 defined by (using $(a+b)^2\le2a^2+2b^2$)

$$C_2=\frac{4(A\phi(w^*)+B)}{C_1C_6A}\sum_{k=1}^{t_1}\big(2C_1\sigma_\Psi^{-1}\eta_k^2\sum_{j=1}^{k-1}\eta_j+\|w^*\|^2\big)+\frac{8C_7}{\rho_1\sigma_\Psi}++\frac{8D_\Psi(w^*,0)}{\sigma_\Psi^2}+\frac{16(B+AC_1)\sum_{k=1}^{t}\eta_k^2}{\sigma_\Psi^2}+2\|w^*\|^2.$$

The proof is complete. $\square$

We are now in a position to prove Theorem 4. The basic idea is to control $\sum_{t=1}^T \eta_t\big(\phi(w_t)-\phi(w^*)\big)$ in terms of a martingale, which can be further controlled by the Azuma-Hoeffding inequality. The bound of $\|w_t\|^2$ in Theorem 3 allows us to control the increments of martingale by logarithmic functions of T/δ.

Proof of Theorem 4. It follows from (3.4) that

$$\begin{aligned}
&D_\Psi(w,w_{t+1})-D_\Psi(w,w_t)\\
&\leq \eta_t\langle w-w_t, f'(w_t,z_t)\rangle+\eta_t\big(r(w)-r(w_t)\big)+\sigma_\Psi^{-1}\eta_t^2\big(Af(w_t,z_t)+Ar(w_t)+2B\big)\\
&\leq \eta_t\langle w-w_t, f'(w_t,z_t)-\mathbb{E}_{z_t}[f'(w_t,z_t)]\rangle+\eta_t\big(\phi(w)-\phi(w_t)\big)+\sigma_\Psi^{-1}\eta_t^2\big(Af(w_t,z_t)+Ar(w_t)+2B\big),
\end{aligned}$$

where we have used the inequality $\langle w-w_t,\mathbb{E}_{z_t}[f'(w_t,z_t)]\rangle\leq F(w)-F(w_t)$ and the definition of ϕ in the last inequality.

Taking a summation over $t=1,\ldots,T$ followed with a reformulation, we derive

$$\begin{aligned}
&\sum_{t=1}^T\eta_t\big(\phi(w_t)-\phi(w)\big)\\
&\leq \sum_{t=1}^T\xi_t+\sum_{t=1}^T\big(D_\Psi(w,w_t)-D_\Psi(w,w_{t+1})\big)+\sigma_\Psi^{-1}\sum_{t=1}^T\eta_t^2\big(Af(w_t,z_t)+Ar(w_t)+2B\big)\\
&\leq \sum_{t=1}^T\xi_t+D_\Psi(w,0)+2\sigma_\Psi^{-1}(AC_1+B)\sum_{t=1}^T\eta_t^2, \qquad (C.7)
\end{aligned}$$

where we have introduced a sequence of random variables

$$\xi_t=\eta_t\langle w-w_t, f'(w_t,z_t)-\mathbb{E}_{z_t}[f'(w_t,z_t)]\rangle$$

and used (3.6). Let

$$\xi_t'=\eta_t\langle w-w_t, f'(w_t,z_t)-\mathbb{E}_{z_t}[f'(w_t,z_t)]\rangle\mathbb{I}_{\{\|w_t\|^2\leq C_2\log\frac{2T}{\delta}\}},\quad t=1,\ldots,T,$$

where $\mathbb{I}_\mathcal{A}$ denotes the indicator function of an event $\mathcal{A}$, i.e., $\mathbb{I}_\mathcal{A}=1$ if $\mathcal{A}$ happens and 0 otherwise. According to the elementary inequality $(a+b)^2\leq 2(a^2+b^2)$ for $a,b\in\mathbb{R}$

$$\begin{aligned}
|\xi_t'|&\leq\frac{\eta_t}{2}\Big[\|w-w_t\|^2+\|f'(w_t,z_t)-\mathbb{E}_{z_t}[f'(w_t,z_t)]\|_*^2\Big]\mathbb{I}_{\{\|w_t\|^2\leq C_2\log\frac{2T}{\delta}\}}\\
&\leq\eta_t\Big[\|w\|^2+\|w_t\|^2+\|f'(w_t,z_t)\|_*^2+\|\mathbb{E}_{z_t}[f'(w_t,z_t)]\|_*^2\Big]\mathbb{I}_{\{\|w_t\|^2\leq C_2\log\frac{2T}{\delta}\}}.
\end{aligned}$$

It follows from Lemma A.3 that

$$\begin{aligned}
\|f'(w_t,z_t)\|_*^2+\|F'(w_t)\|_*^2&\leq 2A^2\|w_t\|^2+2Af(0,z_t)+2B+2A^2\|w_t\|^2+2AF(0)+2B\\
&\leq 4A^2\|w_t\|^2+2A\big(\sup_z f(0,z)+F(0)\big)+4B\\
&\leq 4A^2\|w_t\|^2+4AC_1. \qquad (C.8)
\end{aligned}$$

Combining the above two inequalities together, we derive

$$\begin{aligned}
|\xi_t'|&\leq\eta_t\Big[\|w\|^2+(4A^2+1)\|w_t\|^2+4AC_1\Big]\mathbb{I}_{\{\|w_t\|^2\leq C_2\log\frac{2T}{\delta}\}}\\
&\leq\eta_t\Big(\|w\|^2+4AC_1+(4A^2+1)C_2\log\frac{2T}{\delta}\Big)\leq C(w)\eta_t\log\frac{2T}{\delta},
\end{aligned}$$

where we introduce $C(w)=\|w\|^2+4AC_1+(4A^2+1)C_2$. It is clear that $\mathbb{E}_{z_t}[\xi_t']=0$ and ξ_t' depends only on $z_1,\ldots,z_t$. According to Part (a) of Lemma A.1, we can find an event $\Omega_T:=\{(z_1,\ldots,z_T): z_1,\ldots,z_T\in\mathcal{Z}\}$ with $\Pr\{\Omega_t\}\geq 1-\frac{\delta}{2}$ such that for any $(z_1,\ldots,z_T)\in\Omega_T$ the following inequality holds

$$\sum_{t=1}^T\xi_t'\leq C(w)\log\frac{2T}{\delta}\Big(2\sum_{t=1}^T\eta_t^2\log\frac{2}{\delta}\Big)^{\frac12}\leq C(w)\log^{\frac32}\frac{2T}{\delta}\Big(2\sum_{t=1}^T\eta_t^2\Big)^{\frac12}.$$

Furthermore, according to Theorem 3, there exists an event $\Omega'_T := \{(z_1,\ldots,z_T) : z_1,\ldots,z_T \in \mathcal{Z}\}$ with $\Pr\{\Omega'_t\} \geq 1-\frac{\delta}{2}$ such that for any $(z_1,\ldots,z_T) \in \Omega'_T$ the following inequality holds

$$\max_{1\leq t\leq T} \|w_t\|^2 \leq C_2 \log \frac{2T}{\delta}.$$

Under the intersection of these two events, we have $\xi_t = \xi'_t$ and therefore

$$\sum_{t=1}^T \xi_t = \sum_{t=1}^T \xi'_t \leq C(w) \log^{\frac{3}{2}} \frac{2T}{\delta} \Big(2\sum_{t=1}^T \eta_t^2\Big)^{\frac{1}{2}},$$

which, together with $\Pr\{\Omega_T \cap \Omega'_T\} \geq 1-\delta$ and (C.7), shows the following inequality with probability at least $1-\delta$

$$\sum_{t=1}^T \eta_t\big(\phi(w_t)-\phi(w)\big) \leq D_\Psi(w,0) + 2\sigma_\Psi^{-1}\big(AC_1+B\big)\sum_{t=1}^T \eta_t^2 + C(w)\log^{\frac{3}{2}} \frac{2T}{\delta}\big(2\sum_{t=1}^T \eta_t^2\big)^{\frac{1}{2}}$$
$$\leq \big(2C_3 D_\Psi(w,0) + C_4\big)\log^{\frac{3}{2}} \frac{2T}{\delta},$$

where

$$C_3 = 2^{-1} + \sigma_\Psi^{-1}\Big(2\sum_{t=1}^\infty \eta_t^2\Big)^{\frac{1}{2}}$$

and

$$C_4 := 2\sigma_\Psi^{-1}(AC_1+B)\sum_{t=1}^\infty \eta_t^2 + (4AC_1 + 4A^2C_2 + C_2)\Big(2\sum_{t=1}^\infty \eta_t^2\Big)^{\frac{1}{2}}.$$

The stated inequality then follows from the convexity of ϕ. The proof is complete. □

D Proofs for Strongly Convex Objectives

This section is devoted to proving Theorem 8. First, we take a weighted summation of (3.4) and use (3.7) to tackle $\sum_{k=1}^t \big(f(w_k,z_k)+r(w_k)\big)$ without boundedness assumptions, yielding Lemma D.2. We need the following simple lemma on step sizes in this derivation.

Lemma D.1. *Let $\eta_k = \frac{2}{\sigma_\phi k + 2\sigma_F + \sigma_\phi t_0}$, where $t_0 \in \mathbb{R}_+$. Then,*

$$\sum_{k=1}^t \eta_k \leq 2\sigma_\phi^{-1}\log(et). \tag{D.1}$$

Proof. It follows from the definition of η_t that

$$\sum_{k=1}^t \eta_k \leq 2\sigma_\phi^{-1}\sum_{k=1}^t (k+t_0)^{-1} \leq 2\sigma_\phi^{-1}\log(et).$$

The proof is complete. □

Lemma D.2. *Assume $\sigma_\phi > 0$. Let $\{w_t\}_{t\in\mathbb{N}}$ be the sequence produced by (2.2) with $\eta_t = \frac{2}{\sigma_\phi t + 2\sigma_F + \sigma_\phi t_0}$, where $t_0 \geq 4A/(\sigma_\Psi\sigma_\phi)$. Then the following inequality holds for all $t = 1,\ldots,T$*

$$2\sigma_\phi^{-1}\sum_{k=1}^t (k+t_0+1)\big(\phi(w_k)-\phi(w^*)\big) + (t+t_0+1)(t+t_0+2)D_\Psi(w^*,w_{t+1}) \leq (t_0+1)(t_0+2)D_\Psi(w^*,w_1)$$
$$+ 2\sigma_\phi^{-1}\sum_{k=1}^t (k+t_0+1)\xi_k + 16\log(eT)\sigma_\Psi^{-1}\sigma_\phi^{-2}\big(AC_1(2t+t_0+2)+Bt\big), \tag{D.2}$$

where we introduce

$$\xi_k = \langle w^*-w_k, f'(w_k,z_k) - \mathbb{E}_{z_k}[f'(w_k,z_k)]\rangle, \quad k=1,\ldots,T.$$

Proof. Since $t_0 \geq \frac{4A}{\sigma_\Psi \sigma_\phi}$, we know $\eta_t \leq (2A)^{-1}\sigma_\Psi$ and therefore Lemma 2 holds. Taking $w = w^*$ in (3.4), we derive

$$D_\Psi(w^*, w_{k+1}) - D_\Psi(w^*, w_k) \leq \eta_k \langle w^* - w_k, f'(w_k, z_k) - \mathbb{E}_{z_k}[f'(w_k, z_k)]\rangle + \eta_k \langle w^* - w_k, F'(w_k)\rangle + \eta_k\big(r(w^*) - r(w_k)\big) + \sigma_\Psi^{-1}\eta_k^2\big(Af(w_k, z_k) + Ar(w_k) + 2B\big) - \sigma_r \eta_k D_\Psi(w^*, w_{k+1}).$$

Plugging the inequality $F(w^*) - F(w_k) \geq \langle w^* - w_k, F'(w_k)\rangle + \sigma_F D_\Psi(w^*, w_k)$ into the above inequality then shows

$$D_\Psi(w^*, w_{k+1}) - D_\Psi(w^*, w_k) \leq \eta_k \xi_k + \eta_k\big(F(w^*) - F(w_k) - \sigma_F D_\Psi(w^*, w_k)\big) + \eta_k\big(r(w^*) - r(w_k)\big) + \sigma_\Psi^{-1}\eta_k^2\big(Af(w_k, z_k) + Ar(w_k) + 2B\big) - \sigma_r \eta_k D_\Psi(w^*, w_{k+1}).$$

According to the definition of ϕ, we further get

$$(1 + \sigma_r \eta_k) D_\Psi(w^*, w_{k+1}) \leq (1 - \eta_k \sigma_F) D_\Psi(w^*, w_k) + \eta_k \xi_k + \eta_k\big(\phi(w^*) - \phi(w_k)\big) + \sigma_\Psi^{-1}\eta_k^2\big(Af(w_k, z_k) + Ar(w_k) + 2B\big), \quad \text{(D.3)}$$

which can be reformulated as follows

$$\frac{\eta_k\big(\phi(w_k) - \phi(w^*)\big)}{1 + \sigma_r \eta_k} + D_\Psi(w^*, w_{k+1}) \leq \frac{1 - \eta_k \sigma_F}{1 + \eta_k \sigma_r} D_\Psi(w^*, w_k) + \frac{\eta_k \xi_k}{1 + \sigma_r \eta_k} + \frac{\sigma_\Psi^{-1}\eta_k^2\big(Af(w_k, z_k) + Ar(w_k) + 2B\big)}{1 + \sigma_r \eta_k}. \quad \text{(D.4)}$$

Since $\eta_k = \frac{2}{\sigma_\phi k + 2\sigma_F + \sigma_\phi t_0}$, we know

$$\frac{1 - \sigma_F \eta_k}{1 + \sigma_r \eta_k} = \frac{\sigma_\phi k + 2\sigma_F + \sigma_\phi t_0 - 2\sigma_F}{\sigma_\phi k + 2\sigma_F + \sigma_\phi t_0 + 2\sigma_r} = \frac{k + t_0}{k + t_0 + 2},$$
$$\frac{\eta_k}{1 + \sigma_r \eta_k} = \frac{2}{\sigma_\phi(k + t_0 + 2)}.$$

Plugging the above two equations back into (D.4), we derive

$$\frac{2\big(\phi(w_k) - \phi(w^*)\big)}{\sigma_\phi(k + t_0 + 2)} + D_\Psi(w^*, w_{k+1}) \leq \frac{k + t_0}{k + t_0 + 2} D_\Psi(w^*, w_k) + \frac{2\xi_k}{\sigma_\phi(k + t_0 + 2)} + \frac{2\eta_k\big(Af(w_k, z_k) + Ar(w_k) + 2B\big)}{\sigma_\Psi \sigma_\phi (k + t_0 + 2)}.$$

Multiplying both sides by $(k + t_0 + 1)(k + t_0 + 2)$, we get

$$\frac{2(k + t_0 + 1)}{\sigma_\phi}\big(\phi(w_k) - \phi(w^*)\big) + (k + t_0 + 1)(k + t_0 + 2) D_\Psi(w^*, w_{k+1})$$
$$\leq (k + t_0)(k + t_0 + 1) D_\Psi(w^*, w_k) + \frac{2(k + t_0 + 1)\xi_k}{\sigma_\phi} + \frac{2\eta_k (k + t_0 + 1)\big(Af(w_k, z_k) + Ar(w_k) + 2B\big)}{\sigma_\Psi \sigma_\phi}.$$

Taking a summation of the above inequality from $k = 1$ to $k = t$ and using the inequality $(k + t_0 + 1)\eta_k \leq 4\sigma_\phi^{-1}$, we derive

$$2\sigma_\phi^{-1} \sum_{k=1}^{t} (k + t_0 + 1)\big(\phi(w_k) - \phi(w^*)\big) + (t + t_0 + 1)(t + t_0 + 2) D_\Psi(w^*, w_{t+1}) \leq (t_0 + 1)(t_0 + 2) D_\Psi(w^*, w_1) + 2\sigma_\phi^{-1} \sum_{k=1}^{t} (k + t_0 + 1)\xi_k + 8\sigma_\Psi^{-1}\sigma_\phi^{-2} \sum_{k=1}^{t} \big(Af(w_k, z_k) + Ar(w_k) + 2B\big). \quad \text{(D.5)}$$

According to (3.7), (D.1) and $\eta_t^{-1} \leq 2^{-1}\sigma_\phi(t + t_0 + 2)$, we know

$$\begin{aligned}
\sum_{k=1}^{t} \big(Af(w_k, z_k) + Ar(w_k) + 2B\big) &\leq t(2AC_1 + 2B) + 2AC_1\Big(\sum_{k=1}^{t} \eta_k\Big)\eta_t^{-1} \\
&\leq 2t(AC_1 + B) + 2AC_1\Big(2\sigma_\phi^{-1}\log(et)\Big)\Big(2^{-1}\sigma_\phi(t + t_0 + 2)\Big) \\
&= 2t(AC_1 + B) + 2AC_1(t + t_0 + 2)\log(et) \\
&\leq 2\log(eT)\big(AC_1(2t + t_0 + 2) + Bt\big).
\end{aligned}$$

Plugging the above inequality into (D.5) gives the stated inequality. The proof is complete. □

In the following lemma, we establish bounds on magnitudes and conditional variances on $\{\xi_k\}_k$ defined in Lemma D.2.

Lemma D.3. *Let the assumptions of Lemma D.2 hold with $t_0 \geq \frac{4A}{\sigma_\Psi \sigma_\phi}$ and ξ_k be defined in Lemma D.2. Then for all $k \leq T$ we have*

$$|\xi_k| \leq C_8 \log(eT) \quad \textit{and} \quad \mathbb{E}_{z_k}\big[\big(\xi_k - \mathbb{E}_{z_k}[\xi_k]\big)^2\big] \leq \|w^* - w_k\|^2\big(A\phi(w_k) + B\big),$$

where

$$C_8 := (16A^2 + 4)C_1\sigma_\Psi^{-1}\sigma_\phi^{-1} + \|w^*\|^2 + 4AC_1.$$

Proof. Since $t_0 \geq \frac{4A}{\sigma_\Psi \sigma_\phi}$, we know $\eta_t \leq (2A)^{-1}\sigma_\Psi$ and therefore (3.5) holds. According to Schwarz's inequality, we have

$$\begin{aligned}\big|\langle w^* - w_k, f'(w_k, z_k) - \mathbb{E}_{z_k}[f'(w_k, z_k)]\rangle\big| &\leq \|w^* - w_k\|\big(\|f'(w_k, z_k)\|_* + \|F'(w_k)\|_*\big) \\ &\leq \frac{1}{2}\|w^* - w_k\|^2 + \frac{1}{2}\big(\|f'(w_k, z_k)\|_* + \|F'(w_k)\|_*\big)^2 \\ &\leq \|w^*\|^2 + \|w_k\|^2 + \|f'(w_k, z_k)\|_*^2 + \|F'(w_k)\|_*^2.\end{aligned}$$

Combining the above inequality and (C.8) together shows

$$\begin{aligned}&\big|\langle w^* - w_k, f'(w_k, z_k) - \mathbb{E}_{z_k}[f'(w_k, z_k)]\rangle\big| \leq (4A^2 + 1)\|w_k\|^2 + \|w^*\|^2 + 4AC_1 \\ &\leq (8A^2 + 2)C_1\sigma_\Psi^{-1}\sum_{j=1}^{k}\eta_j + \|w^*\|^2 + 4AC_1 \leq C_8 \log(ek),\end{aligned}$$

where we have used (3.5) and Lemma D.1 to control $\sum_{j=1}^k \eta_j$. This shows a bound on $|\xi_k|$.

It is clear that $\mathbb{E}_{z_k}[\xi_k] = 0$ and therefore it follows from $\mathbb{E}[(\xi - \mathbb{E}[\xi])^2] \leq \mathbb{E}[\xi^2]$ for all real-valued random variable ξ that

$$\begin{aligned}\mathbb{E}_{z_k}\big[\big(\xi_k - \mathbb{E}_{z_k}[\xi_k]\big)^2\big] &= \mathbb{E}_{z_k}[\xi_k^2] \leq \mathbb{E}_{z_k}\big[\langle w^* - w_k, f'(w_k, z_k)\rangle^2\big] \\ &\leq \|w^* - w_k\|^2\mathbb{E}_{z_k}[\|f'(w_k, z_k)\|_*^2] \leq \|w^* - w_k\|^2\big(AF(w_k) + B\big) \\ &\leq \|w^* - w_k\|^2\big(A\phi(w_k) + B\big),\end{aligned}$$

where we have used

$$\mathbb{E}_{z_k}[\|f'(w_k, z_k)\|_*^2] \leq \mathbb{E}_{z_k}[Af(w_k, z_k) + B] = AF(w_k) + B$$

due to (3.1). The proof is complete. □

Then, we apply a Bernstein inequality to show $\sum_{k=1}^t (k + t_0 + 1)\xi_k \leq \frac{1}{2}\sum_{k=1}^t (k + t_0 + 1)\big(\phi(w_k) - \phi(w^*)\big) + \mathfrak{C}_t$ with high probability, where $\mathfrak{C}_t$ is the summation of the last two terms in (D.10). An interesting observation is that $\frac{1}{2}\sum_{k=1}^t (k + t_0 + 1)\big(\phi(w_k) - \phi(w^*)\big)$ can be offset by the first term in (D.2), from which one can derive (3.10). To apply the Bernstein inequality, we use Lemma D.3 to control the conditional variance as $\mathbb{E}_{z_k}\big[\big(\xi_k - \mathbb{E}_{z_k}[\xi_k]\big)^2\big] \leq \|w^* - w_k\|^2\big(A\phi(w_k) + B\big)$, and introduce the decomposition $A\phi(w_k) + B = A\phi(w_k) - A\phi(w^*) + A\phi(w^*) + B$ to get variance partially offset by the first term in (D.2). This is a key trick for us to proceed with the discussion without boundedness assumption on subgradients.

Proof of Theorem 8. Let ξ_k be defined in Lemma D.2. Since $t_0 \geq \frac{16A\log\frac{T}{\delta}}{\sigma_\phi\sigma_\Psi}$ and $\delta \leq e^{-\frac{1}{4}}$, we know $t_0 \geq \frac{4A}{\sigma_\Psi\sigma_\phi}$ and therefore Lemma D.2 and Lemma D.3 hold. Define

$$C_T = \max\Big\{\frac{4(t_0 + 1)D_\Psi(w^*, w_1)}{\sigma_\Psi} + \frac{3t_0\big(\phi(w^*) + A^{-1}B\big)}{\sigma_\phi\sigma_\Psi} + \frac{64\log(eT)(B + 2AC_1)}{\sigma_\Psi^2\sigma_\phi^2}, \frac{C_8 t_0 \log(eT)}{2A}\Big\}. \quad \text{(D.6)}$$

Let $\rho_2 = \frac{C_8 t_0 \log(eT)}{2AC_T}$. It is clear from the definition of C_T that $\rho_2 \in (0,1]$. According to Lemma D.3, we derive the following inequalities for all $k = 1, \ldots, t (t \le T)$

$$|(k+t_0+1)\xi_k| \le C_8(t+t_0+1)\log(eT)$$

$$\mathbb{E}_{z_k}\big[\big((k+t_0+1)\xi_k - \mathbb{E}_{z_k}[(k+t_0+1)\xi_k]\big)^2\big] \le (k+t_0+1)^2\|w^*-w_k\|^2(A\phi(w_k)+B).$$

Plugging the above two inequalities back into Part (b) of Lemma A.1, we derive the following inequality with probability at least $1-\frac{\delta}{T}$

$$\sum_{k=1}^t (k+t_0+1)\xi_k \le \frac{\rho_2 \sum_{k=1}^t \Big((k+t_0+1)^2\|w^*-w_k\|^2\big(A\phi(w_k)+B\big)\Big)}{C_8(t+t_0+1)\log(eT)} + \frac{C_8(t+t_0+1)\log(eT)\log\frac{T}{\delta}}{\rho_2}. \quad \text{(D.7)}$$

Taking union bounds on probabilities of events, it is clear that (D.7) holds with probability at least $1-\delta$ simultaneously for all $t = 1, \ldots, T$. In the remainder of the proof, we always assume that (D.7) holds for all $t = 1, \ldots, T$, which happens with probability at least $1-\delta$.

Applying the σ_Ψ-strong convexity of Ψ to (D.2) and dividing both sides by $2^{-1}\sigma_\Psi(t+t_0+1)(t+t_0+2)$, we derive the following inequality with probability $1-\delta$ for all $t = 1, \ldots, T$

$$\frac{4\sum_{k=1}^t (k+t_0+1)\big(\phi(w_k)-\phi(w^*)\big)}{\sigma_\phi\sigma_\Psi(t+t_0+1)(t+t_0+2)} + \|w^*-w_{t+1}\|^2 \le \frac{2(t_0+1)(t_0+2)D_\Psi(w^*,w_1)}{\sigma_\Psi(t+t_0+1)(t+t_0+2)} + \frac{4\sum_{k=1}^t (k+t_0+1)\xi_k}{(t+t_0+1)(t+t_0+2)\sigma_\phi\sigma_\Psi} + \frac{32\log(eT)\big(AC_1(2t+t_0+2)+Bt\big)}{(t+t_0+1)(t+t_0+2)\sigma_\Psi^2\sigma_\phi^2}. \quad \text{(D.8)}$$

We now show by induction that $\|w^*-w_{\tilde{t}}\|^2 \le \frac{C_T}{\tilde{t}+t_0+1}$ for all $\tilde{t} = 1, \ldots, T$. It is clear from the definition of C_T that

$$\|w^*-w_1\|^2 \le \frac{2D_\Psi(w^*,w_1)(t_0+2)}{\sigma_\Psi(t_0+2)} \le \frac{4(t_0+1)D_\Psi(w^*,w_1)}{\sigma_\Psi(t_0+2)} \le \frac{C_T}{t_0+2}.$$

Therefore, the induction assumption holds for the case with $\tilde{t} = 1$. Suppose that $\|w^*-w_{\tilde{t}}\|^2 \le \frac{C_T}{\tilde{t}+t_0+1}$ for all $\tilde{t} \le t$. We now need to show that it also holds for $\tilde{t} = t+1$, i.e., $\|w^*-w_{t+1}\|^2 \le \frac{C_T}{t+t_0+2}$. According to (D.8) multiplied by $t+t_0+2$, it suffices to show

$$-\frac{4\sum_{k=1}^t (k+t_0+1)\big(\phi(w_k)-\phi(w^*)\big)}{\sigma_\phi\sigma_\Psi(t+t_0+1)} + \frac{2(t_0+1)(t_0+2)D_\Psi(w^*,w_1)}{\sigma_\Psi(t+t_0+1)} + \frac{4\sum_{k=1}^t (k+t_0+1)\xi_k}{\sigma_\phi\sigma_\Psi(t+t_0+1)} + \frac{32\log(eT)\big(AC_1(2t+t_0+2)+Bt\big)}{\sigma_\Psi^2\sigma_\phi^2(t+t_0+1)} \le C_T. \quad \text{(D.9)}$$

Plugging the induction assumption $\|w^*-w_{\tilde{t}}\|^2 \le C_T/(\tilde{t}+t_0+1)$ for all $\tilde{t} \le t$ back into (D.7), we derive

$$\begin{aligned}
&\sum_{k=1}^t (k+t_0+1)\xi_k \\
&\le \frac{\rho_2 C_T \sum_{k=1}^t \Big((k+t_0+1)\big(A\phi(w_k)+B\big)\Big)}{C_8(t+t_0+1)\log(eT)} + \frac{C_8(t+t_0+1)\log(eT)\log\frac{T}{\delta}}{\rho_2} \\
&= \frac{t_0 A^{-1}}{2(t+t_0+1)}\sum_{k=1}^t (k+t_0+1)\big(A\phi(w_k)-A\phi(w^*)+A\phi(w^*)+B\big) + \frac{2(t+t_0+1)AC_T\log\frac{T}{\delta}}{t_0} \\
&\le \frac{1}{2}\sum_{k=1}^t (k+t_0+1)\big(\phi(w_k)-\phi(w^*)\big) + \frac{t_0(A\phi(w^*)+B)\sum_{k=1}^t (k+t_0+1)}{2A(t+t_0+1)} + \frac{(t+t_0+1)C_T\sigma_\phi\sigma_\Psi}{8},
\end{aligned} \quad \text{(D.10)}$$

where we have used the definition of ρ_2 in the first identity and the assumption $t_0 \geq \frac{16A\log\frac{T}{\delta}}{\sigma_\phi\sigma_\Psi}$ in the last inequality. Plugging the above inequality into (D.9), it suffices to show

$$\frac{2(t_0+1)(t_0+2)D_\Psi(w^*,w_1)}{\sigma_\Psi(t+t_0+1)} + \frac{2t_0(\phi(w^*)+A^{-1}B)\sum_{k=1}^t(k+t_0+1)}{\sigma_\Psi\sigma_\phi(t+t_0+1)^2} + \frac{C_T}{2} + \frac{32(B+2AC_1)\log(eT)}{\sigma_\Psi^2\sigma_\phi^2} \leq C_T.$$

Since

$$\sum_{k=1}^t(k+t_0+1) = \frac{t(t+2t_0+3)}{2} \leq \frac{3(t+t_0+1)^2}{4}, \tag{D.11}$$

it suffices to show

$$\frac{2(t_0+1)D_\Psi(w^*,w_1)}{\sigma_\Psi} + \frac{3t_0(\phi(w^*)+A^{-1}B)}{2\sigma_\Psi\sigma_\phi} + \frac{C_T}{2} + \frac{32(B+2AC_1)\log(eT)}{\sigma_\Psi^2\sigma_\phi^2} \leq C_T.$$

which is clear from the definition of C_T in (D.6). Therefore, $\|w^*-w_{t+1}\|^2 \leq \frac{C_T}{t+t_0+2}$. This proves the first inequality in (3.10).

We now prove the second inequality in (3.10). According to (D.2), we know

$$\sum_{k=1}^t(k+t_0+1)\big(\phi(w_k)-\phi(w^*)\big) \leq \frac{\sigma_\phi(t_0+1)(t_0+2)D_\Psi(w^*,w_1)}{2} + \sum_{k=1}^t(k+t_0+1)\xi_k + \frac{8\log(eT)\big(AC_1(2t+t_0+2)+Bt\big)}{\sigma_\phi\sigma_\Psi}.$$

Plugging (D.10) into the above inequality and using (D.11), we derive the following inequality with probability at least $1-\delta$ for all $t=1,\ldots,T$

$$\frac{\sum_{k=1}^t(k+t_0+1)\big(\phi(w_k)-\phi(w^*)\big)}{2} \leq \frac{\sigma_\phi(t_0+1)(t_0+2)D_\Psi(w^*,w_1)}{2} + \frac{3t_0(A\phi(w^*)+B)(t+t_0+1)}{8A} + \frac{(t+t_0+1)C_T\sigma_\phi\sigma_\Psi}{8} + \frac{8\log(eT)\big(AC_1(2t+t_0+2)+Bt\big)}{\sigma_\phi\sigma_\Psi}.$$

With probability at least $1-\delta$, it then follows from the convexity of ϕ and the identity in (D.11) that

$$\begin{aligned}\phi(\bar{w}_t^{(2)})-\phi(w^*) &\leq \Big(\sum_{k=1}^t(k+t_0+1)\Big)^{-1}\Big(\sum_{k=1}^t(k+t_0+1)\big(\phi(w_k)-\phi(w^*)\big)\Big)\\ &\leq \frac{1}{t(t+2t_0+3)}\bigg(2\sigma_\phi(t_0+1)(t_0+2)D_\Psi(w^*,w_1) + \frac{3t_0(A\phi(w^*)+B)(t+t_0+1)}{2A}\\ &+ \frac{(t+t_0+1)C_T\sigma_\phi\sigma_\Psi}{2} + \frac{32\log(eT)\big(AC_1(2t+t_0+2)+Bt\big)}{\sigma_\phi\sigma_\Psi}\bigg), \quad \text{for all } t=1,\ldots,T.\end{aligned}$$

This establishes the second inequality in (3.10) with $\widetilde{C}_T$ defined by

$$\widetilde{C}_T = \sigma_\phi(t_0+1)D_\Psi(w^*,w_1) + \frac{3t_0(A\phi(w^*)+B)}{2A} + \frac{C_T\sigma_\phi\sigma_\Psi}{2} + \frac{32\log(eT)(2AC_1+B)}{\sigma_\phi\sigma_\Psi}.$$

The proof is complete. □

Remark 1. According to the definition of C_T and $\widetilde{C}_T$, it is clear that both C_T and $\widetilde{C}_T$ only involves logarithmic functions of T/δ. It is also clear that C_T is a quadratic function of σ_ϕ^{-1} and $\widetilde{C}_T$ is a linear function of σ_ϕ^{-1}.

E Proofs for Almost Sure Convergence

In this section, we present a proposition on almost sure convergence which covers both the general convex case (Theorem 6) and the strongly convex case (Theorem 9). To this aim, we need to introduce two lemmas. Lemma E.1 is the Doob's martingale convergence theorem [see, e.g., 2, page 195] which is a powerful tool to study almost sure convergence. We will use Lemma E.2 [9] to show that the random variable to which $D_\Psi(w^*, w_t)$ converges is zero almost surely in the strongly convex case.

Lemma E.1. *Let $\{\tilde{X}_t\}_{t\in\mathbb{N}}$ be a sequence of non-negative random variables with $\mathbb{E}[\tilde{X}_1] < \infty$ and let $\{\mathcal{F}_t\}_{t\in\mathbb{N}}$ be a nested sequence of sets of random variables with $\mathcal{F}_t \subset \mathcal{F}_{t+1}$ for all $t \in \mathbb{N}$. If $\mathbb{E}[\tilde{X}_{t+1}|\mathcal{F}_t] \leq \tilde{X}_t$ for every $t \in \mathbb{N}$, then $\tilde{X}_t$ converges to a nonnegative random variable $\tilde{X}$ almost surely. Furthermore, $\tilde{X} < \infty$ almost surely.*

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

The basic idea in proving Proposition E.3 is to construct non-negative supermartingales based on the one-step progress inequality (3.4), whose almost sure convergence based on Lemma E.1 will imply the almost sure convergence of the random variables we are interested in. We will construct different supermartingales in the general convex case and the strongly convex case.

Proposition E.3. *Let $\{w_t\}_{t\in\mathbb{N}}$ be the sequence produced by (2.2). If $\|w^*\| < \infty$ and $\sum_{t=1}^\infty \eta_t^2 < \infty$, then $\{D_\Psi(w^*, w_t)\}_t$ converges almost surely to a non-negative random variable and $\lim_{t\to\infty} D_\Psi(w^*, w_t) < \infty$ almost surely. Furthermore,*

(a) if $\eta_t \leq (2A)^{-1}\sigma_\Psi$ and $\eta_{t+1} \leq \eta_t$, then $\sum_{t=1}^\infty \eta_t\big(\phi(w_t) - \phi(w^*)\big) < \infty$ almost surely;

(b) if $\sigma_\phi > 0$ and $\sum_{t=1}^\infty \eta_t = \infty$, then $\lim_{t\to\infty} D_\Psi(w^*, w_t) = 0$ almost surely.

Proof. Since $\sum_{t=1}^\infty \eta_t^2 < \infty$, there exists a $t_2 \in \mathbb{N}$ such that $\eta_t \leq \min\{(2A)^{-1}\sigma_\Psi, 2\sigma_\phi^{-1}, \sigma_r^{-1}\}$ for all $t \geq t_2$. Taking conditional expectations w.r.t. z_t on both sides of (D.3), we derive the following inequality for all $t \geq t_2$

$$\mathbb{E}_{z_t}[D_\Psi(w^*, w_{t+1})] \leq \frac{1 - \sigma_F\eta_t}{1 + \sigma_r\eta_t} D_\Psi(w^*, w_t) + \frac{\eta_t}{1 + \sigma_r\eta_t}\big(\phi(w^*) - \phi(w_t)\big) + \sigma_\Psi^{-1}\eta_t^2\big(A\phi(w_t) - A\phi(w^*) + A\phi(w^*) + 2B\big),$$

where we have used $1 + \sigma_F\eta_t \geq 1$ and $\mathbb{E}_{z_t}[\xi_t] = 0$ for ξ_t defined in Lemma D.2. According to $\phi(w^*) \leq \phi(w_t)$ and $\eta_t \leq \min\{(2A)^{-1}\sigma_\Psi, \sigma_r^{-1}\}$, we know

$$\eta_t(1 + \sigma_r\eta_t)^{-1}\big(\phi(w^*) - \phi(w_t)\big) + \sigma_\Psi^{-1}\eta_t^2\big(A\phi(w_t) - A\phi(w^*)\big) \leq 2^{-1}\eta_t\big(\phi(w^*) - \phi(w_t)\big) + 2^{-1}\eta_t\big(\phi(w_t) - \phi(w^*)\big) = 0.$$

Combining the above two inequalities together, we derive

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

Introduce a sequence of non-negative random variables $\widetilde{X}_t$ as

$$\widetilde{X}_t = D_\Psi(w^*, w_t) + \sigma_\Psi^{-1}\big(A\phi(w^*) + 2B\big)\sum_{k=t}^\infty \eta_k^2,$$

which is well defined since $\sum_{t=1}^\infty \eta_t^2 < \infty$. By (E.1), it is clear that $\mathbb{E}_{z_t}[\widetilde{X}_{t+1}] \leq \widetilde{X}_t$ for all $t \geq t_2$. Taking $w = w^*$ and expectations on both sides of (B.2), we derive

$$\mathbb{E}[D_\Psi(w^*, w_{t+1})] \leq \mathbb{E}[D_\Psi(w^*, w_t)] + \sigma_\Psi^{-1}\eta_t^2 A\mathbb{E}[\phi(w_t)] + 2\sigma_\Psi^{-1}B\eta_t^2, \quad \text{for all } t \in \mathbb{N},$$

where we have used $\phi(w^*) \leq \phi(w_t)$. According to Lemma A.3, the term $\mathbb{E}[\phi(w_t)]$ can be controlled by $\mathbb{E}[D_\Psi(w^*, w_t)]$ and $\|w^*\|$. Therefore, we derive an upper bound on $\mathbb{E}[D_\Psi(w^*, w_{t+1})]$ in terms of $\mathbb{E}[D_\Psi(w^*,w_t)]$, $\|w^*\|$ and step sizes, from which we know $\mathbb{E}[\widetilde{X}_{t_2}]<\infty$ (t_2 is a fixed constant). Therefore, one can apply Lemma E.1 to show that $\widetilde{X}_t$ converges almost surely to a non-negative random variable, which, together with $\sum_{t=1}^\infty \eta_t^2<\infty$, further implies $\lim_{t\to\infty} D_\Psi(w^*,w_t)=\widetilde{X}$ almost surely for a non-negative random variable $\widetilde{X}$. It is clear that $\widetilde{X}<\infty$ almost surely by Lemma E.1.

We now turn to part (a). Under the assumption $\eta_t\le(2A)^{-1}\sigma_\Psi$ and $\eta_{t+1}\le\eta_t$, (C.7) holds. According to (C.7) with $w=w^*$, we know

$$\sum_{k=1}^t \eta_k\big(\phi(w_k)-\phi(w^*)\big)\le\sum_{k=1}^t\xi_k+D_\Psi(w^*,0)+2\sigma_\Psi^{-1}(AC_1+B)\sum_{k=1}^t\eta_k^2, \tag{E.2}$$

where

$$\xi_k=\eta_k\langle w^*-w_k,f'(w_k,z_k)-\mathbb{E}_{z_k}[f'(w_k,z_k)].\rangle$$

Introduce a sequence of random variables

$$\widetilde{X}'_{t+1}=\sum_{k=1}^t\xi_k+D_\Psi(w^*,0)+2\sigma_\Psi^{-1}(AC_1+B)\sum_{k=1}^\infty\eta_k^2,\quad t=0,1,\ldots,$$

which is well defined since $\sum_{t=1}^\infty\eta_t^2<\infty$. It is clear from $\mathbb{E}_{z_t}[\xi_t]=0$ that

$$\mathbb{E}_{z_t}[\widetilde{X}'_{t+1}]=\sum_{k=1}^{t-1}\xi_k+\mathbb{E}_{z_t}[\xi_t]+D_\Psi(w^*,0)+2\sigma_\Psi^{-1}(AC_1+B)\sum_{k=1}^\infty\eta_k^2=\widetilde{X}'_t.$$

Furthermore, according to the definition of w^* and (E.2), we know $\widetilde{X}'_t\ge0$ for all $t\in\mathbb{N}$. Therefore, one can apply Lemma E.1 to show that $\{\widetilde{X}'_t\}_{t\in\mathbb{N}}$ converges to a non-negative variable $\widetilde{X}'$ almost surely and $\widetilde{X}'<\infty$ almost surely. This, together with (E.2) and the definition of $\widetilde{X}'_t$, implies that $\sum_{k=1}^\infty\eta_k\big(\phi(w_k)-\phi(w^*)\big)<\infty$ almost surely. This finishes the proof of part (a).

We now turn to part (b). We have shown $\lim_{t\to\infty}D_\Psi(w^*,w_t)=\widetilde{X}$ almost surely. It suffices to show $\widetilde{X}=0$ almost surely under the condition $\sigma_\phi>0$ and $\sum_{t=1}^\infty\eta_t=\infty$. Since $\eta_t\le\sigma_r^{-1}$ for all $t\ge t_2$, we know

$$\frac{1-\sigma_F\eta_t}{1+\sigma_r\eta_t}=\frac{1+\sigma_r\eta_t-\sigma_\phi\eta_t}{1+\sigma_r\eta_t}\le1-2^{-1}\sigma_\phi\eta_t,\quad\forall t\ge t_2.$$

Plugging the above inequality back into (E.1) and taking expectations over both sides, we derive

$$\mathbb{E}[D_\Psi(w^*,w_{t+1})]\le\big(1-2^{-1}\sigma_\phi\eta_t\big)\mathbb{E}[D_\Psi(w^*,w_t)]+\sigma_\Psi^{-1}(A\phi(w^*)+2B)\eta_t^2,\quad\forall t\ge t_2.$$

Applying this inequality iteratively for $t=T,T-1,\ldots,t_2$ yields

$$\mathbb{E}[D_\Psi(w^*,w_{T+1})]\le\prod_{t=t_2}^T(1-2^{-1}\sigma_\phi\eta_t)\mathbb{E}[D_\Psi(w^*,w_{t_2})]\\ +\sigma_\Psi^{-1}(A\phi(w^*)+2B)\sum_{t=t_2}^T\eta_t^2\prod_{k=t+1}^T(1-2^{-1}\sigma_\phi\eta_k), \tag{E.3}$$

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 $1-a\le\exp(-a), a>0$ together with $\sum_{t=1}^\infty\eta_t=\infty$

$$\lim_{T\to\infty}\prod_{t=t_2}^T(1-2^{-1}\sigma_\phi\eta_t)\mathbb{E}[D_\Psi(w^*,w_{t_2})]\le\lim_{T\to\infty}\prod_{t=t_2}^T\exp\big(-2^{-1}\sigma_\phi\eta_t\big)\mathbb{E}[D_\Psi(w^*,w_{t_2})]\\ =\lim_{T\to\infty}\exp\Big(-2^{-1}\sigma_\phi\sum_{t=t_2}^T\eta_t\Big)\mathbb{E}[D_\Psi(w^*,w_{t_2})]=0.$$

Applying Lemma E.2 with $a = 2^{-1}\sigma_\phi$, we get $\lim_{T\to\infty}\sum_{t=t_2}^T \eta_t^2 \prod_{k=t+1}^T (1-2^{-1}\sigma_\phi\eta_k) = 0$. Plugging the above two expressions into (E.3) implies $\lim_{T\to\infty}\mathbb{E}[D_\Psi(w^*, w_T)] = 0$. This together with Fatou's lemma shows

$$0 \le \mathbb{E}[\widetilde{X}] = \mathbb{E}\big[\lim_{T\to\infty} D_\Psi(w^*, w_T)\big] \le \lim\inf_{T\to\infty}\mathbb{E}[D_\Psi(w^*, w_T)] = 0,$$

from which and $\widetilde{X} \ge 0$ we know $\widetilde{X} = 0$ almost surely. This finishes the proof of part (b). The proof is complete. □

F Proofs for Generalization Bounds

In this section, we prove generalization error bounds presented in Section 4. The following lemma is a standard probabilistic bound on the uniform deviation between empirical errors and generalization errors over a RKHS ball. In our case, we need to control the Lipschitz constants and the magnitudes for functions satisfying Assumption 1. According to (3.2) and Lemma A.4 we know $\|f'(w,z)\|_2^2 \le Af(w,z) + B$ with $A = \tilde{A}\kappa^2$ and $B = \tilde{B}\kappa^2$, where $\kappa = \sup_{x\in\mathcal{X}}\|K_x\|_2$. Recall that $f(w,z) = \ell(h_w(x), y)$.

Lemma F.1. *Let $R > 0$ and define $B_R = \big\{w \in \mathcal{W} : \|w\|_2 \le R\big\}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ we have*

$$\sup_{w\in B_R}\big[\mathcal{E}(w) - \mathcal{E}_{\mathbf{z}}(w)\big] \le \big(C_9R^2 + C_{10}\big)n^{-\frac{1}{2}}\log^{\frac{1}{2}}\frac{1}{\delta}, \tag{F.1}$$

where

$$C_9 = \kappa^2 + 2\tilde{A}^2\kappa^2 + \Big(\frac{A^2}{\sqrt{2}} + \frac{1}{2\sqrt{2}}\Big) \;\; \text{and} \;\; C_{10} = \Big(2\tilde{A} + \frac{A+1}{\sqrt{2}}\Big)\sup_z f(0,z) + 2\tilde{B} + \frac{B}{\sqrt{2}}.$$

Proof. We prove this lemma by McDiarmid's inequality (Lemma A.2). To this aim, we first show that the function $\mathbf{z} \mapsto \sup_{w\in B_R}\big[\mathcal{E}(w) - \mathcal{E}_{\mathbf{z}}(w)\big]$ satisfies a bounded difference property. Indeed, for any $\mathbf{z} = \{z_1,\ldots,z_{i-1},z_i,z_{i+1},\ldots,z_n\}$ and $\bar{\mathbf{z}} = \{z_1,\ldots,z_{i-1},\bar{z}_i,z_{i+1},\ldots,z_n\}$, we have

$$\begin{aligned}\Big|\sup_{w\in B_R}\big[\mathcal{E}(w) - \mathcal{E}_{\mathbf{z}}(w)\big] - \sup_{w\in B_R}\big[\mathcal{E}(w) - \mathcal{E}_{\bar{\mathbf{z}}}(w)\big]\Big| &\le \sup_{w\in B_R}\big|\mathcal{E}_{\mathbf{z}}(w) - \mathcal{E}_{\bar{\mathbf{z}}}(w)\big| \\ &\le \frac{1}{n}\sup_{w\in B_R}\big|f(w,z_i) - f(w,\bar{z}_i)\big| \le \frac{1}{n}\sup_{w\in B_R}\sup_{z\in\mathcal{Z}} f(w,z) \\ &\le \frac{1}{n}\Big(\big(A^2 + \frac{1}{2}\big)R^2 + (A+1)\sup_z f(0,z) + B\Big),\end{aligned}$$

where the third inequality is due to the non-negativity of f and the last inequality is due to (A.5) applied to the function $w \mapsto f(w,z)$. Applying McDiarmid's inequality with increments bounded above, we derive the following inequality with probability at least $1-\delta$

$$\begin{aligned}\sup_{w\in B_R}\big[\mathcal{E}(w) - \mathcal{E}_{\mathbf{z}}(w)\big] &\le \mathbb{E}_{\mathbf{z}}\Big[\sup_{w\in B_R}\big[\mathcal{E}(w) - \mathcal{E}_{\mathbf{z}}(w)\big]\Big] \\ &\quad + \sqrt{\frac{\log 1/\delta}{2n}}\Big(\big(A^2 + \frac{1}{2}\big)R^2 + (A+1)\sup_z f(0,z) + B\Big).\end{aligned} \tag{F.2}$$

We now control the term $\mathbb{E}_{\mathbf{z}}\Big[\sup_{w\in B_R}\big[\mathcal{E}(w) - \mathcal{E}_{\mathbf{z}}(w)\big]\Big]$. Let $\tilde{\mathbf{z}} = \{\tilde{z}_1,\ldots,\tilde{z}_n\}$ be training examples independently drawn from ρ and independent of $\mathbf{z}$. Let $\sigma_1,\ldots,\sigma_n$ be a sequence of independent Rademacher variables with $\Pr\{\sigma_i = 1\} = \Pr\{\sigma_i = -1\} = \frac{1}{2}$. By Jensen's inequality and the standard symmetrization technique, we get

$$\begin{aligned}&\mathbb{E}_{\mathbf{z}}\Big[\sup_{w\in B_R}\big[\mathcal{E}(w) - \mathcal{E}_{\mathbf{z}}(w)\big]\Big] = \mathbb{E}_{\mathbf{z}}\Big[\sup_{w\in B_R}\big[\mathbb{E}_{\tilde{\mathbf{z}}}[\mathcal{E}_{\tilde{\mathbf{z}}}(w)] - \mathcal{E}_{\mathbf{z}}(w)\big]\Big] \\ &\le \mathbb{E}_{\mathbf{z},\tilde{\mathbf{z}}}\Big[\sup_{w\in B_R}\big[\mathcal{E}_{\tilde{\mathbf{z}}}(w) - \mathcal{E}_{\mathbf{z}}(w)\big]\Big] = \frac{1}{n}\mathbb{E}_{\mathbf{z},\tilde{\mathbf{z}}}\Big[\sup_{w\in B_R}\sum_{i=1}^n\Big(f(w,\tilde{z}_i) - f(w,z_i)\Big)\Big] \\ &= \frac{1}{n}\mathbb{E}_{\mathbf{z},\tilde{\mathbf{z}},\sigma}\Big[\sup_{w\in B_R}\sum_{i=1}^n\sigma_i\Big(f(w,\tilde{z}_i) - f(w,z_i)\Big)\Big] \le \frac{2}{n}\mathbb{E}_{\mathbf{z},\sigma}\Big[\sup_{w\in B_R}\sum_{i=1}^n\sigma_i f(w,z_i)\Big].\end{aligned} \tag{F.3}$$

For any $w \in B_R$, it follows from Lemma A.3 that

$$\begin{aligned}\big|\ell'(\langle w, K_x\rangle, y)\big|^2 &\le 2\tilde{A}^2|\langle w, K_x\rangle|^2 + 2\tilde{A}\ell(0,y) + 2\tilde{B} \le 2\tilde{A}^2\|w\|_2^2\|K_x\|_2^2 + 2\tilde{A}\ell(0,y) + 2\tilde{B} \\ &\le 2\tilde{A}^2R^2\kappa^2 + 2\tilde{A}\sup_y \ell(0,y) + 2\tilde{B},\end{aligned}$$

from which we know

$$\big|\ell'(\langle w, K_x\rangle, y)\big| \le \sqrt{2\tilde{A}^2R^2\kappa^2 + 2\tilde{A}\sup_y \ell(0,y) + 2\tilde{B}}, \quad \forall w \in B_R.$$

Applying Talagrand's contraction lemma [5] to the last term of (F.3) together with $f(w,z) = \ell(\langle w, K_x\rangle, y)$ and the above bound on derivative of ℓ, we derive

$$\mathbb{E}_{\mathbf{z}}\Big[\sup_{w\in B_R}\big[\mathcal{E}(w) - \mathcal{E}_{\mathbf{z}}(w)\big]\Big] \le \frac{2\sqrt{2\tilde{A}^2R^2\kappa^2 + 2\tilde{A}\sup_y \ell(0,y) + 2\tilde{B}}}{n}\mathbb{E}_{\mathbf{z},\sigma}\Big[\sup_{w\in B_R}\sum_{i=1}^n \sigma_i\langle w, K_{x_i}\rangle\Big]. \tag{F.4}$$

According to the Schwarz's inequality and Jensen's inequality, we get

$$\begin{aligned}\mathbb{E}_\sigma\Big[\sup_{w\in B_R}\sum_{i=1}^n \sigma_i\langle w, K_{x_i}\rangle\Big] &= \mathbb{E}_\sigma\Big[\sup_{w\in B_R}\big\langle w, \sum_{i=1}^n \sigma_i K_{x_i}\big\rangle\Big] \le \mathbb{E}_\sigma\Big[\sup_{w\in B_R}\|w\|_2\sqrt{\Big\|\sum_{i=1}^n \sigma_i K_{x_i}\Big\|_2^2}\Big] \\ &\le R\sqrt{\mathbb{E}_\sigma\Big\langle \sum_{i=1}^n \sigma_i K_{x_i}, \sum_{i=1}^n \sigma_i K_{x_i}\Big\rangle} = R\sqrt{\sum_{i=1}^n \|K_{x_i}\|_2^2} \le R\kappa\sqrt{n}.\end{aligned}$$

Plugging the above inequality back into (F.4), we derive

$$\mathbb{E}_{\mathbf{z}}\Big[\sup_{w\in B_R}\big[\mathcal{E}(w) - \mathcal{E}_{\mathbf{z}}(w)\big]\Big] \le \frac{2R\kappa\sqrt{2\tilde{A}^2R^2\kappa^2 + 2\tilde{A}\sup_y \ell(0,y) + 2\tilde{B}}}{\sqrt{n}}.$$

Plugging the above inequality back into (F.2) and using $2ab \le a^2 + b^2$ for $a, b \in \mathbb{R}$, we derive the following inequality with probability at least $1-\delta$

$$\begin{aligned}\sup_{w\in B_R}\big[\mathcal{E}(w) - \mathcal{E}_{\mathbf{z}}(w)\big] &\le \frac{1}{\sqrt{n}}\Big(R^2\kappa^2 + 2\tilde{A}^2R^2\kappa^2 + 2\tilde{A}\sup_y \ell(0,y) + 2\tilde{B}\Big) \\ &\quad + \sqrt{\frac{\log 1/\delta}{2n}}\Big(\big(A^2 + \frac{1}{2}\big)R^2 + (A+1)\sup_z f(0,z) + B\Big),\end{aligned}$$

which can be written as (F.1) with the stated C_9 and C_{10}. The proof is complete. ☐

The following lemma aims to bound $\mathcal{E}_{\mathbf{z}}(w_\lambda) - \mathcal{E}(w_\lambda)$ with w_λ defined in (F.5). Since w_λ is a fixed element in $\mathcal{W}$, we do not need to resort to uniform deviation arguments. Instead, we can apply a Bernstein inequality to study $\mathcal{E}_{\mathbf{z}}(w_\lambda) - \mathcal{E}(w_\lambda)$, based on the observation that Assumption 3 allows us to control the variance of $f(w_\lambda, z)$ by a linear function of $\sup_z f(w_\lambda, z)$.

Lemma F.2. *Let $\lambda \in (0,1]$ and define*

$$w_\lambda = \arg\min_{w\in\mathcal{W}} \mathcal{E}(w) + \lambda\|w\|_2^2. \tag{F.5}$$

Let $\rho \in (0,1]$ and $\delta \in (0,1)$. Then, with probability at least $1-\delta$ we have

$$\mathcal{E}_{\mathbf{z}}(w_\lambda) - \mathcal{E}(w_\lambda) \le \rho\big(c_\alpha + \mathcal{E}(h_\rho)\big) + (\rho n)^{-1}\sup_z f(w_\lambda, z)\log\delta^{-1}.$$

Proof. Let $\xi_i = f(w_\lambda, z_i), i = 1, \ldots, n$. According to the definition of w_λ and Assumption 3, we know

$$\mathcal{E}(w_\lambda) - \mathcal{E}(h_\rho) + \lambda\|w_\lambda\|_2^2 \le c_\alpha\lambda^\alpha,$$

from which and $\lambda \le 1$ we derive

$$\mathcal{E}(w_\lambda) \le \mathcal{E}(h_\rho) + c_\alpha.$$

It then follows that $\xi_i - \mathbb{E}[\xi_i] \leq \sup_z f(w_\lambda, z)$ (non-negativity of ξ_i) and

$$\mathbb{E}\big[\big(\xi_i - \mathbb{E}[\xi_i]\big)^2\big] \leq \mathbb{E}[f^2(w_\lambda, z_i)] \leq \sup_z f(w_\lambda, z)\mathbb{E}[f(w_\lambda, z)] \leq \sup_z f(w_\lambda, z)\big(c_\alpha + \mathcal{E}(h_\rho)\big).$$

Applying Part (b) of Lemma A.1 with $\xi_i = f(w_\lambda, z_i)$ and the above bounds on variances and magnitudes, we derive the following inequality with probability at least $1 - \delta$

$$\mathcal{E}_{\mathbf{z}}(w_\lambda) - \mathcal{E}(w_\lambda) = \frac{1}{n}\sum_{i=1}^n \xi_i - \mathbb{E}[\xi] \leq \frac{\rho n \sup_z f(w_\lambda, z)\big(c_\alpha + \mathcal{E}(h_\rho)\big)}{n \sup_z f(w_\lambda, z)} + \frac{\sup_z f(w_\lambda, z) \log \frac{1}{\delta}}{\rho n}.$$

The stated inequality then follows directly. The proof is complete. □

We are now in a position to prove Theorem 10. Our basic idea is to use the decomposition (F.6) with w_λ and λ proportional to $n^{-\frac{\alpha}{1+\alpha}}$. The term $\mathcal{E}_{\mathbf{z}}(\bar{w}_T^{(1)}) - \mathcal{E}_{\mathbf{z}}(w_\lambda)$ is the computational error related to the optimization process. Both $\mathcal{E}(\bar{w}_T^{(1)}) - \mathcal{E}_{\mathbf{z}}(\bar{w}_T^{(1)})$ and $\mathcal{E}_{\mathbf{z}}(w_\lambda) - \mathcal{E}(w_\lambda)$ are estimation errors related to the sampling process. The term $\mathcal{E}(w_\lambda) - \mathcal{E}(h_\rho)$ is the approximation error. In the following, we apply Lemma F.1 and Lemma F.2 to control estimation errors, Theorem 4 to control the computational error and Assumption 3 to control the approximation error. Here we use three tricks to get almost optimal generalization error bounds. First, we show that $\|\bar{w}_T^{(1)}\|_2^2$ grows as a logarithmic function of T, which allows us to get $\mathcal{E}(\bar{w}_T^{(1)}) - \mathcal{E}_{\mathbf{z}}(\bar{w}_T^{(1)}) = O(n^{-\frac{1}{2}} \log T)$ (we omit the dependency on $1/\delta$ for brevity). Second, in the analysis of $\mathcal{E}_{\mathbf{z}}(w_\lambda) - \mathcal{E}(w_\lambda)$, we show the variance of $f(w_\lambda, z)$ grows as a linear function of $\sup_z f(w_\lambda, z)$ instead of a quadratic function of $\sup_z f(w_\lambda, z)$ by exploiting Assumption 3, which allows us to get a bound with a mild dependency on $\|w_\lambda\|_2^2$. As a comparison, if we use $\|w_\lambda\|_2^2 = O(\lambda^{\alpha-1})$ due to Assumption 3 and the Azuma-Hoeffding inequality we will get $\mathcal{E}_{\mathbf{z}}(w_\lambda) - \mathcal{E}(w_\lambda) = O(\lambda^{\alpha-1}n^{-\frac{1}{2}})$, which is suboptimal since λ is chosen to be very small to trade the estimation, computational and approximation errors. Indeed, if one plug $\mathcal{E}_{\mathbf{z}}(w_\lambda) - \mathcal{E}(w_\lambda) = O(\lambda^{\alpha-1}n^{-\frac{1}{2}})$ into (F.6), one can only derive the suboptimal bound $\mathcal{E}(\bar{w}_T^{(1)}) - \mathcal{E}(h_\rho) = O(n^{-\frac{\alpha}{2}} \log^{\frac{3}{2}} T)$ worse than $O(n^{-\frac{\alpha}{1+\alpha}} \log^{\frac{3}{2}} T)$ in Theorem 10. The third trick is to choose w_λ with an appropriate λ in (F.6) to fully exploit Assumption 3.

Proof of Theorem 10. Let $\lambda, \rho \in (0, 1]$ be real numbers to be fixed later and $w = w_\lambda$ defined by (F.5). We use the following error decomposition w.r.t. w_λ to study the excess generalization error $\mathcal{E}(\bar{w}_T^{(1)}) - \mathcal{E}(h_\rho)$

$$\begin{aligned}\mathcal{E}(\bar{w}_T^{(1)}) - \mathcal{E}(h_\rho) = \big(\mathcal{E}(\bar{w}_T^{(1)}) - \mathcal{E}_{\mathbf{z}}(\bar{w}_T^{(1)})\big) + \big(\mathcal{E}_{\mathbf{z}}(\bar{w}_T^{(1)}) - \mathcal{E}_{\mathbf{z}}(w_\lambda)\big) \\ + \big(\mathcal{E}_{\mathbf{z}}(w_\lambda) - \mathcal{E}(w_\lambda)\big) + \big(\mathcal{E}(w_\lambda) - \mathcal{E}(h_\rho)\big). \quad \text{(F.6)}\end{aligned}$$

It is clear that (4.1) is a specific instantiation of (2.2) with $f(w, z) = \ell\big(\langle w, K_x\rangle, y\big), \Psi(w) = \frac{1}{2}\|w\|_2^2, r(w) = 0$ and $\tilde{\rho}$ being the uniform distribution over $\{z_1, \ldots, z_n\}$. During the iteration of (4.1), the training sample $\mathbf{z} = \{z_1, \ldots, z_n\}$ is fixed and the randomness comes from the index sequence $\{j_t\}_{t\in\mathbb{N}}$. Since j_t is drawn from a uniform distribution over $\{1, \ldots, n\}$, the objective function minimized by the SGD scheme (4.1) is the empirical error $\phi(w) = \mathbb{E}_{j_t}[f(w, z_{j_t})] = \mathcal{E}_{\mathbf{z}}(w)$. An application of Theorem 4 to the SGD scheme (4.1) with $w = w_\lambda$ then gives the following inequality with probability $1 - \delta/4$

$$\mathcal{E}_{\mathbf{z}}(\bar{w}_T^{(1)}) - \mathcal{E}_{\mathbf{z}}(w_\lambda) \leq \Big(\sum_{t=1}^T \eta_t\Big)^{-1}\big(C_3\|w_\lambda\|_2^2 + C_4\big) \log^{\frac{3}{2}} \frac{8T}{\delta}. \quad \text{(F.7)}$$

We can apply Lemma F.2 to derive the following inequality with probability at least $1 - \delta/4$

$$\begin{aligned}\mathcal{E}_{\mathbf{z}}(w_\lambda) - \mathcal{E}(w_\lambda) &\leq \rho\big(c_\alpha + \mathcal{E}(h_\rho)\big) + (\rho n)^{-1} \sup_z f(w_\lambda, z) \log \frac{4}{\delta} \\ &\leq \rho\big(c_\alpha + \mathcal{E}(h_\rho)\big) + (\rho n)^{-1}\Big(\Big(A^2 + \frac{1}{2}\Big)\|w_\lambda\|_2^2 + (A+1)\sup_z f(0, z) + B\Big) \log \frac{4}{\delta},\end{aligned} \quad \text{(F.8)}$$

where the last inequality is due to Lemma A.3.

According to Theorem 3, with probability at least $1-\delta/4$ we have $\max_{1\le t\le T}\|w_t\|_2 \le \sqrt{C_2\log\frac{4T}{\delta}}$, from which and the convexity of norm we derive the following inequality with probability $1-\delta/4$

$$\|\bar{w}_T^{(1)}\|_2 \le \sqrt{C_2\log\frac{4T}{\delta}}. \tag{F.9}$$

Furthermore, an application of Lemma F.1 with $\widetilde{R}=\sqrt{C_2\log\frac{4T}{\delta}}$ shows the following inequality with probability $1-\delta/4$

$$\sup_{w\in B_{\widetilde{R}}}\Big[\mathcal{E}(w)-\mathcal{E}_{\mathbf{z}}(w)\Big] \le \Big(C_9C_2\log\frac{4T}{\delta}+C_{10}\Big)n^{-\frac{1}{2}}\log^{\frac{1}{2}}\frac{4}{\delta}.$$

Combining the above inequality and (F.9) together, we derive the following inequality with probability $1-\delta/2$

$$\Big[\mathcal{E}(\bar{w}_T^{(1)})-\mathcal{E}_{\mathbf{z}}(\bar{w}_T^{(1)})\Big] \le \Big(C_9C_2+C_{10}\Big)n^{-\frac{1}{2}}\log^{\frac{3}{2}}\frac{4T}{\delta}. \tag{F.10}$$

Plugging (F.7), (F.8) and (F.10) into (F.6), we derive the following inequality with probability at least $1-\delta$

$$\begin{aligned}\mathcal{E}(\bar{w}_T^{(1)})-\mathcal{E}(h_\rho) \le \mathcal{E}(w_\lambda)-\mathcal{E}(h_\rho)+\|w_\lambda\|_2^2\Big(C_3\big(\sum_{t=1}^T\eta_t\big)^{-1}+(\rho n)^{-1}\big(A^2+2^{-1}\big)\Big)\log^{\frac{3}{2}}\frac{8T}{\delta}\\ +C_4\big(\sum_{t=1}^T\eta_t\big)^{-1}\log^{\frac{3}{2}}\frac{8T}{\delta}+\Big(C_9C_2+C_{10}\Big)n^{-\frac{1}{2}}\log^{\frac{3}{2}}\frac{4T}{\delta}\\ +\rho\big(c_\alpha+\mathcal{E}(h_\rho)\big)+(\rho n)^{-1}\Big((A+1)\sup_z f(0,z)+B\Big)\log\frac{4}{\delta}.\end{aligned}$$

We choose $\lambda=\max\big\{\big(\sum_{t=1}^T\eta_t\big)^{-1},(\rho n)^{-1}\big\}$ in the above inequality and derive the following inequality with probability $1-\delta$

$$\begin{aligned}\mathcal{E}(\bar{w}_T^{(1)})-\mathcal{E}(h_\rho) \le \big(C_3+A^2+2^{-1}\big)D\Big(\max\Big\{\big(\sum_{t=1}^T\eta_t\big)^{-1},(\rho n)^{-1}\Big\}\Big)\log^{\frac{3}{2}}\frac{8T}{\delta}+\Big(C_4\big(\sum_{t=1}^T\eta_t\big)^{-1}+\\ \Big(C_9C_2+C_{10}\Big)n^{-\frac{1}{2}}\Big)\log^{\frac{3}{2}}\frac{8T}{\delta}+\rho\big(c_\alpha+\mathcal{E}(h_\rho)\big)+(\rho n)^{-1}\Big((A+1)\sup_z f(0,z)+B\Big)\log\frac{4}{\delta},\end{aligned}$$

where in the first inequality we have used $C_3+A^2+2^{-1}\ge 1$ and

$$\begin{aligned}&\mathcal{E}(w_\lambda)-\mathcal{E}(h_\rho)+\|w_\lambda\|_2^2\Big(C_3\big(\sum_{t=1}^T\eta_t\big)^{-1}+(\rho n)^{-1}\big(A^2+2^{-1}\big)\Big)\log^{\frac{3}{2}}\frac{8T}{\delta}\\ &\le \big(C_3+A^2+2^{-1}\big)\Big(\mathcal{E}(w_\lambda)-\mathcal{E}(h_\rho)+\lambda\|w_\lambda\|_2^2\Big)\log^{\frac{3}{2}}\frac{8T}{\delta}=\big(C_3+A^2+2^{-1}\big)D(\lambda)\log^{\frac{3}{2}}\frac{8T}{\delta}.\end{aligned}$$

Since the above inequality holds for any $\rho\in(0,1]$, we can take $\rho=n^{-\frac{\alpha}{1+\alpha}}$ to derive the following inequality with probability at least $1-\delta$

$$\begin{aligned}\mathcal{E}(\bar{w}_T^{(1)})-\mathcal{E}(h_\rho) \le c_\alpha\big(C_3+A^2+2^{-1}\big)\max\Big\{\big(\sum_{t=1}^T\eta_t\big)^{-\alpha},n^{-\frac{\alpha}{1+\alpha}}\Big\}\log^{\frac{3}{2}}\frac{8T}{\delta}+\Big(C_4\big(\sum_{t=1}^T\eta_t\big)^{-1}+\\ \Big(C_9C_2+C_{10}\Big)n^{-\frac{1}{2}}\Big)\log^{\frac{3}{2}}\frac{8T}{\delta}+n^{-\frac{\alpha}{1+\alpha}}\big(c_\alpha+\mathcal{E}(h_\rho)+(A+1)\sup_z f(0,z)+B\big)\log\frac{4}{\delta},\end{aligned}$$

from which it follows directly the stated inequality (4.2) with C_5 defined by

$$C_5=c_\alpha(C_3+A^2+2^{-1})+C_4+C_9C_2+C_{10}+c_\alpha+\mathcal{E}(h_\rho)+(A+1)\sup_z f(0,z)+B.$$

It is clear both ρ and λ defined above satisfy $\rho,\lambda\in(0,1]$. The proof is complete. □

Dataset	No. of Training Examples	No. of Test Examples	No. of Attributes	Source
ADULT	32,561	16,281	123	[7]
GISETTE	6,000	1,000	5,000	[3]
IJCNN1	49,990	91,701	22	[8]
MUSHROOMS	4,062	4,062	112	[1]
PHISHING	5,527	5,528	68	[1]
SPLICE	1,000	2,175	60	[1]

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

G Additional Information on Simulation

We present a detailed description of datasets, used in Section 6, in Table G.1.

References

[1] D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

[2] J. L. Doob. *Measure Theory, Graduate Texts in Mathematics*. Springer, 1994.

[3] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror. Result analysis of the NIPS 2003 feature selection challenge. In *Advances in neural information processing systems*, pages 545–552, 2005.

[4] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[5] M. Ledoux and M. Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer, Berlin, 1991.

[6] C. McDiarmid. On the method of bounded differences. In J. Siemous, editor, *Surveys in combinatorics*, pages 148–188. Cambridge Univ. Press, Cambridge, 1989.

[7] J. Platt. Fast training of support vector machines using sequential minimal optimization. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods—Support Vector Learning*, pages 185–208. MIT Press, Cambridge, MA, 1999.

[8] D. Prokhorov. IJCNN 2001 neural network competition. *Slide presentation in IJCNN*, 1:97, 2001.

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

[10] T. Zhang. Data dependent concentration bounds for sequential prediction algorithms. In *Conference on Learning Theory*, pages 173–187, 2005.