Fast Rates of ERM and Stochastic Approximation: Adaptive to Error Bound Conditions

Mingrui Liu[†], **Xiaoxuan Zhang**[†], **Lijun Zhang**[‡], **Rong Jin**[♮], **Tianbao Yang**[†]
[†]Department of Computer Science, The University of Iowa, Iowa City, IA 52242, USA
[‡]National Key Laboratory for Novel Software Technology, Nanjing University, China
[♮] Machine Intelligence Technology, Alibaba Group, Bellevue, WA 98004, USA
mingrui-liu@uiowa.edu, zljzju@gmail.com, tianbao-yang@uiowa.edu

Abstract

Error bound conditions (EBC) are properties that characterize the growth of an objective function when a point is moved away from the optimal set. They have recently received increasing attention for developing optimization algorithms with fast convergence. However, the studies of EBC in statistical learning are hitherto still limited. The main contributions of this paper are two-fold. First, we develop fast and intermediate rates of empirical risk minimization (ERM) under EBC for risk minimization with Lipschitz continuous, and smooth convex random functions. Second, we establish fast and intermediate rates of an efficient stochastic approximation (SA) algorithm for risk minimization with Lipschitz continuous random functions, which requires only one pass of n samples and adapts to EBC. For both approaches, the convergence rates span a full spectrum between $\widetilde{O}(1/\sqrt{n})$ and $\widetilde{O}(1/n)$ depending on the power constant in EBC, and could be even faster than $O(1/n)$ in special cases for ERM. Moreover, these convergence rates are automatically adaptive without using any knowledge of EBC.

1 Introduction

In this paper, we focus on the following stochastic convex optimization problems arising in statistical learning and many other fields:

$$\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{z}\sim\mathbb{P}}[f(\mathbf{w},\mathbf{z})], \tag{1}$$

and more generally

$$\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{z}\sim\mathbb{P}}[f(\mathbf{w},\mathbf{z})] + r(\mathbf{w}), \tag{2}$$

where $f(\cdot,\mathbf{z}):\mathcal{W}\to\mathbb{R}$ is a random function depending on a random variable $\mathbf{z}\in\mathcal{Z}$ that follows a distribution $\mathbb{P}$, $r(\mathbf{w})$ is a lower semi-continuous convex function. In statistical learning [48], the problem above is also referred to as **risk minimization** where $\mathbf{z}$ is interpreted as data, $\mathbf{w}$ is interpreted as a model (or hypothesis), $f(\cdot,\cdot)$ is interpreted as a loss function, and $r(\cdot)$ is a regularization. For example, in supervised learning one can take $\mathbf{z}=(\mathbf{x},y)$ - a pair of feature vector $\mathbf{x}\in\mathcal{X}\subseteq\mathbb{R}^d$ and label $\mathbf{y}\in\mathcal{Y}$, $f(\mathbf{w},\mathbf{z})=\ell(\mathbf{w}(\mathbf{x}),y)$ - a loss function measuring the error of the prediction $\mathbf{w}(\mathbf{x}):\mathcal{X}\to\mathcal{Y}$ made by the model $\mathbf{w}$. Nonetheless, we emphasize that the risk minimization problem (1) is more general than supervised learning and could be more challenging (c.f. [35]). In this paper, we assume that $\mathcal{W}\subseteq\mathbb{R}^d$ is a compact and convex set. Let $\mathcal{W}_*=\arg\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w})$ denote the optimal set and $P_*=\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w})$ denote the optimal risk.

There are two popular approaches for solving the risk minimization problem. The first one is by empirical risk minimization that minimizes the empirical risk defined over a set of n i.i.d. samples

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

drawn from the same distribution $\mathbb{P}$ (sometimes with a regularization term on the model). The second approach is called stochastic approximation that iteratively learns the model from random samples $\mathbf{z}_t \sim \mathbb{P}, t = 1, \ldots, n$. Both approaches have been studied broadly and extensive results are available about the theoretical guarantee of the two approaches in the machine learning and optimization community. A central theme in these studies is to bound the excess risk (or optimization error) of a learned model $\widehat{\mathbf{w}}$ measured by $P(\widehat{\mathbf{w}}) - P_*$, i.e., given a set of n samples $(\mathbf{z}_1, \ldots, \mathbf{z}_n)$ how fast the learned model converges to the optimal model in terms of the excess risk.

A classical result about the excess risk bound for the considered risk minimization problem is in the order of $\widetilde{O}(\sqrt{d/n})$ [1] and $O(\sqrt{1/n})$ for ERM and SA, respectively, under appropriate conditions of the loss functions (e.g., Lipschitz continuity, convexity) [29, 35]. Various studies have attempted to establish faster rates by imposing additional conditions on the loss functions (e.g., strong convexity, smoothness, exponential concavity) [13, 42, 21], or on both the loss functions and the distribution (e.g., Tsybakov condition, Bernstein condition, central condition) [45, 3, 46]. In this paper, we will study a different family of conditions called the error bound conditions (EBC) (see Definition 1), which has a long history in the community of optimization and variational analysis [31] and recently revived for developing fast optimization algorithms without strong convexity [4, 6, 17, 28, 54]. However, the exploration of EBC in statistical learning for risk minimization is still under-explored and the connection to other conditions is not fully understood.

Definition 1. *For any $\mathbf{w} \in \mathcal{W}$, let $\mathbf{w}^* = \arg\min_{\mathbf{u} \in \mathcal{W}_*} \|\mathbf{u} - \mathbf{w}\|_2$ denote an optimal solution closest to $\mathbf{w}$, where $\mathcal{W}_*$ is the set containing all optimal solutions. Let $\theta \in (0, 1]$ and $0 < \alpha < \infty$. The problem (1) satisfies an EBC(θ, α) if for any $\mathbf{w} \in \mathcal{W}$, the following inequality holds*

$$\|\mathbf{w} - \mathbf{w}^*\|_2^2 \leq \alpha(P(\mathbf{w}) - P(\mathbf{w}^*))^\theta. \tag{3}$$

This condition has been well studied in optimization and variational analysis. Many results are available for understanding the condition for different problems. For example, it has been shown that when $P(\mathbf{w})$ is semi-algebraic and continuous, the inequality (3) is known to hold on any compact set with certain $\theta \in (0, 1]$ and $\alpha > 0$ [4] [2]. We will study both ERM and SA under the above error bound condition. In particular, we show that the benefits of exploiting EBC in statistical learning are noticeable and profound by establishing the following results.

- **Result I.** First, we show that for Lipchitz continous loss EBC implies a *relaxed* Bernstein condition, and therefore leads to intermediate rates of $\widetilde{O}((d/n)^{\frac{1}{2-\theta}})$ for Lipschitz continuous loss. Although this result does not improve over existing rates based on Bernstein condition, however, we emphasize that it provides an alternative route for establishing fast rates and brings richer results than literature to statistical learning in light of the examples provided in this paper.
- **Result II.** Second, we develop fast and optimistic rates of ERM for non-negative, Lipschitz continuous and smooth convex loss functions in the order of $\widetilde{O}(d/n + (dP_*/n)^{\frac{1}{2-\theta}})$, and in the order of $\widetilde{O}((d/n)^{\frac{2}{2-\theta}} + (dP_*/n)^{\frac{1}{2-\theta}})$ when the sample size n is sufficiently large, which imply that when the optimal risk P_* is small one can achieve a fast rate of $\widetilde{O}(d/n)$ even with $\theta < 1$ and a faster rate of $\widetilde{O}((d/n)^{\frac{2}{2-\theta}})$ when n is sufficiently large.
- **Result III.** Third, we develop an efficient SA algorithm with almost the same per-iteration cost as stochastic subgradient methods for Lipschitz continuous loss, which achieves the same order of rate $\widetilde{O}((1/n)^{\frac{1}{2-\theta}})$ as ERM without an explicit dependence on d. More importantly it is "parameter"-free with no need of prior knowledge of θ and α in EBC.

Overall, these results not only strengthen the understanding of ERM for statistical learning but also bring new fast stochastic algorithms for solving a broad range of statistical learning problems. Before ending this section, we would like to point out that all the results are automatically adaptive to the largest possible value of $\theta \in (0, 1]$ in hindsight of the problem, and the dependence on d for ERM is generally unavoidable according to the lower bounds studied in [9].

[1] $\widetilde{O}$ hides a poly-logarithmic factor of n.

[2] One may consider $\theta \in (1, 2]$, which will yield the same order of excess risk bound as $\theta = 1$ in our settings.

2 Related Work

The results for statistical learning under EBC are limited. A similar one to our **Result I** for ERM was established in [39]. However, their result requires the convexity condition of random loss functions, making it weaker than our result. Ramdas and Singh [33] and Xu et al. [50] considered SA under the EBC condition and established similar adaptive rates. Nonetheless, their stochastic algorithms require knowing the values of θ and possibly the constant α in the EBC. In contrast, the SA algorithm in this paper is "parameter"-free without the need of knowing θ and α while still achieving the adaptive rates of $O(1/n^{2-\theta})$. Fast rates under strong convexity (a special case of EBC) are well-known for ERM, online optimization and SA [35, 43, 13, 16, 36, 14]. In the presence of strong convexity of $P(\mathbf{w})$, our results of ERM and SA recover known rates (see below for more discussions).

Fast (intermediate) rates of ERM have been studied under various conditions, including Tsybakov margin condition [44, 25], Bernstein condition [3, 2, 19], exp-concavity condition [21, 11, 26, 51], mixability condition [27], central condition [46], etc. The Bernstein condition (see Definition 2) is a generalization of Tsybakov margin condition for classification. The connection between the exp-concavity condition, the Bernstein condition and the v-central condition was studied in [46]. In particular, the exp-concavity implies a v-central condition under an appropriate condition of the decision set $\mathcal{W}$ (e.g., well-specificity or convexity). With the bounded loss condition, the Bernstein condition implies the v-central condition and the v-central condition also implies a Bernstein condition.

In this work, we also study the connection between the EBC and the Bernstein condition and the v-central condition. In particular, we will develop weaker forms of the Bernstein condition and the v-central condition from the EBC for Lipschitz continuous loss functions. Building on this connection, we establish our **Result I**, which is on a par with existing results for bounded loss functions relying on the Bernstein condition or the central condition. Nevertheless, we emphasize that employing the EBC for developing fast rates has noticeable benefits: (i) it is complementary to the Bernstein condition and the central condition and enjoyed by several interesting problems whose fast rates are not exhibited yet; (ii) it can be leveraged for developing fast and intermediate optimistic rates for non-negative and smooth loss functions; (iii) it can be leveraged to develop efficient SA algorithms with intermediate and fast convergence rates.

Sebro et al. [42] established an optimistic rate of $O(1/n + \sqrt{P_*/n})$ of both ERM and SA for supervised learning with generalized linear loss functions. However, their SA algorithm requires knowing the value of P_*. Recently, Zhang et al. [55] considered the general stochastic optimization problem (1) with non-negative and smooth loss functions and achieved a series of optimistic results. It is worth mentioning that their excess risk bounds for both convex problems and strongly convex problems are special cases of our **Result II** when $\theta = 0$ and $\theta = 1$, respectively. However, the intermediate optimistic rates for $\theta \in (0, 1)$ are first shown in this paper. Importantly, our **Result II** under the EBC with $\theta = 1$ is more general than the result in [55] under strong convexity assumption.

Finally, we discuss about stochastic approximation algorithms with fast and intermediate rates to understand the significance of our **Result III**. Different variants of stochastic gradient methods have been analyzed for stochastic strongly convex optimization [14, 32, 38] with a fast rate of $O(1/n)$. But these stochastic algorithms require knowing the strong convexity modulus. A recent work established adaptive regret bounds $O(n^{\frac{1-\theta}{2-\theta}})$ for online learning with a total of n rounds under the Bernstein condition [20]. However, their methods are based on the second-order methods and therefore are not as efficient as our stochastic approximation algorithm. For example, for online convex optimization they employed the MetaGrad algorithm [47], which needs to maintain $\log(n)$ copies of the online Newton step (ONS) [13] with different learning rates. Notice that the per-iteration cost of ONS is usually $O(d^4)$ even for very simple domain $\mathcal{W}$ [21], while that of our SA algorithm is dominated by the Euclidean projection onto $\mathcal{W}$ that is as fast as $O(d)$ for a simple domain.

3 Empirical Risk Minimization (ERM)

We first formally state the minimal assumptions that are made throughout the paper. Additional assumptions will be made in the sequel for developing fast rates for different families of the random functions $f(\mathbf{w}, \mathbf{z})$.

Assumption 1. *For the stochastic optimization problems (1) and (2), we assume: (i) $P(\mathbf{w})$ is a convex function, $\mathcal{W}$ is a closed and bounded convex set, i.e., there exists $R > 0$ such that $\|\mathbf{w}\|_2 \leq R$ for any $\mathbf{w} \in \mathcal{W}$, and $r(\mathbf{w})$ is a Lipschitz continuous convex function. (ii) the problem (1) and (2) satisfy an EBC(θ, α), i.e., there exist $\theta \in (0, 1]$ and $0 < \alpha < \infty$ such that the inequality (3) hold.*

In this section, we focus on the development of theory of ERM for risk minimization. In particular, we learn a model $\widehat{\mathbf{w}}$ by solving the following ERM problem corresponding to (1):

$$\widehat{\mathbf{w}} \in \arg\min_{\mathbf{w}\in\mathcal{W}} P_n(\mathbf{w}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}, \mathbf{z}_i), \tag{4}$$

where $\mathbf{z}_1, \ldots, \mathbf{z}_n$ are i.i.d samples following the distribution $\mathbb{P}$. A similar ERM problem can be formulated for (2). This section is divided into two subsections. First, we establish intermediate rates of ERM under EBC when the random function is Lipschitz continuous. Second, we develop intermediate rates of ERM under EBC when the random function is smooth. In the sequel and the supplement, we use $\vee$ to denote the max operation and use $\wedge$ to denote the min operation.

3.1 ERM for Lipschitz continuous random functions

In this subsection, w.l.o.g we restrict our attention to (1) since we make the following assumption besides Assumption 1. If $r(\mathbf{w})$ is present, it can be absorbed into $f(\mathbf{w}, \mathbf{z})$.

Assumption 2. *For the stochastic optimization problem (1), we assume that $f(\mathbf{w}, \mathbf{z})$ is a G-Lipschitz continuous function w.r.t $\mathbf{w}$ for any $\mathbf{z} \in \mathcal{Z}$.*

It is notable that we do not assume $f(\mathbf{w}, \mathbf{z})$ is convex in terms of $\mathbf{w}$ or any $\mathbf{z}$. First, we compare EBC with two very important conditions considered in literature for developing fast rates of ERM, namely the Bernstein condition and the central condition. We first give the definitions of these two conditions.

Definition 2. *(Bernstein Condition) Let $\beta \in (0, 1]$ and $B \geq 1$. Then $(f, \mathbb{P}, \mathcal{W})$ satisfies the (β, B)-Bernstein condition if there exists a $\mathbf{w}_* \in \mathcal{W}$ such that for any $\mathbf{w} \in \mathcal{W}$*

$$\mathbb{E}_{\mathbf{z}}[(f(\mathbf{w}, \mathbf{z}) - f(\mathbf{w}_*, \mathbf{z}))^2] \leq B(\mathbb{E}_{\mathbf{z}}[f(\mathbf{w}, \mathbf{z}) - f(\mathbf{w}_*, \mathbf{z})])^{\beta}. \tag{5}$$

It is clear that if such an $\mathbf{w}_*$ exists it has to be the minimizer of the risk.

Definition 3. *(v-Central Condition) Let $v : [0, \infty) \to [0, \infty)$ be a bounded, non-decreasing function satisfying $v(x) > 0$ for all $x > 0$. We say that $(f, \mathbb{P}, \mathcal{W})$ satisfies the v-central condition if for all $\varepsilon \geq 0$, there exists $\mathbf{w}_* \in \mathcal{W}$ such that for any $\mathbf{w} \in \mathcal{W}$ the following holds with $\eta = v(\varepsilon)$.*

$$\mathbb{E}_{\mathbf{z}\sim\mathbb{P}}\left[e^{\eta(f(\mathbf{w}_*, \mathbf{z}) - f(\mathbf{w}, \mathbf{z}))}\right] \leq e^{\eta\varepsilon}. \tag{6}$$

If $v(\varepsilon)$ is a constant for all $\varepsilon \geq 0$, the v-central condition reduces to the strong η-central condition, which implies the $O(1/n)$ fast rate [46]. The connection between the Bernstein condition or v-central condition has been studied in [46]. For example, if the random functions $f(\mathbf{w}, \mathbf{z})$ take values in $[0, a]$, then (β, B)-Bernstein condition implies v-central condition with $v(x) \propto x^{1-\beta}$.

The following lemma shows that for Lipchitz continuous function, EBC condition implies a relaxed Bernstein condition and a relaxed v-central condition.

Lemma 1. ***(Relaxed Bernstein condition and v-central condition)** Suppose Assumptions 1, 2 hold. For any $\mathbf{w} \in \mathcal{W}$, there exists $\mathbf{w}^* \in \mathcal{W}_*$ (which is actually the one closest to $\mathbf{w}$), such that*

$$\mathbb{E}_{\mathbf{z}}[(f(\mathbf{w}, \mathbf{z}) - f(\mathbf{w}^*, \mathbf{z}))^2] \leq B(\mathbb{E}_{\mathbf{z}}[f(\mathbf{w}, \mathbf{z}) - f(\mathbf{w}^*, \mathbf{z})])^{\theta},$$

where $B = G^2\alpha$, and $\mathbb{E}_{\mathbf{z}\sim\mathbb{P}}\left[e^{\eta(f(\mathbf{w}^*, \mathbf{z}) - f(\mathbf{w}, \mathbf{z}))}\right] \leq e^{\eta\varepsilon}$, where $\eta = v(\varepsilon) := c\varepsilon^{1-\theta} \wedge b$. Additionally, for any $\varepsilon > 0$ if $P(\mathbf{w}) - P(\mathbf{w}^*) \geq \varepsilon$, we have $\mathbb{E}_{\mathbf{z}\sim\mathbb{P}}\left[e^{v(\varepsilon)(f(\mathbf{w}^*, \mathbf{z}) - f(\mathbf{w}, \mathbf{z}))}\right] \leq 1$, where $b > 0$ is any constant and $c = 1/(\alpha G^2\kappa(4GRb))$, where $\kappa(x) = (e^x - x - 1)/x^2$.

Remark: There is a subtle difference between the above relaxed Bernstein condition and v-central condition and their original definitions in Definitions 2 and 3. The difference is that in Definitions 2 and 3, it requires there exists a universal $\mathbf{w}_*$ for all $\mathbf{w} \in \mathcal{W}$ such that (5) and (6) hold. In Lemma 1

it only requires for every $\mathbf{w} \in \mathcal{W}$ there exists one $\mathbf{w}^*$ that could be different for different $\mathbf{w}$ such that (5) and (6) hold. This relaxation enables us to establish richer results by exploring EBC than the Bernstein condition and v-central condition, which are postponed to Section 5.

Next, we present the main result of this subsection.

Theorem 1 (Result I). *Suppose Assumptions 1, 2 hold. For any $n \geq aC$, with probability at least $1-\delta$ we have*

$$P(\widehat{\mathbf{w}}) - P_* \leq O\left(\frac{d\log n + \log(1/\delta)}{n}\right)^{\frac{1}{2-\theta}}, \tag{7}$$

where $a = 3(d\log(32GRn^{1/(2-\theta)}) + \log(1/\delta))/c + 1$ and $C > 0$ is some constant.

Remark: The proof utilizes Lemma 1 and follows similarly as the proofs in previous studies [46, 26] based on v-central condition. Our analysis essentially shows that relaxed Bernstein condition and relaxed v-central condition with non-universal $\mathbf{w}^*$ suffice to establish the intermediate rates. Although the rate in Theorem 1 does not improve that in previous works [46], the relaxation brought by EBC allows us to establish fast rates for interesting problems that were unknown before. More details are postponed into Section 5. For example, under the condition that the input data $\mathbf{x}, y$ are bounded, ERM for hinge loss minimization with ℓ_1, ℓ_∞ norm constraints, and for minimizing a quadratic function and an ℓ_1 norm regularization enjoys an $\widetilde{O}(1/n)$ fast rate. To the best of our knowledge, such a fast rate of ERM for these problems has not been shown in literature using other conditions or theories.

3.2 ERM for non-negative, Lipschitz continuous and smooth convex random functions

Below we will present improved optimistic rates of ERM for non-negative smooth loss functions expanding the results in [55]. To be general, we consider (2) and the following ERM problem:

$$\widehat{\mathbf{w}} \in \arg\min_{\mathbf{w}\in\mathcal{W}} P_n(\mathbf{w}) \triangleq \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w}, \mathbf{z}_i) + r(\mathbf{w}). \tag{8}$$

Besides Assumptions 1, 2, we further make the following assumption for developing faster rates.

Assumption 3. *For the stochastic optimization problem (1), we assume $f(\mathbf{w}, \mathbf{z})$ is a non-negative and L-smooth convex function w.r.t $\mathbf{w}$ for any $\mathbf{z} \in \mathcal{Z}$.*

It is notable that we do not assume that $r(\mathbf{w})$ is smooth. Our main result in this subsection is presented in the following theorem.

Theorem 2 (Result II). *Under Assumptions 1, 2, and 3, with probability at least $1-\delta$ we have*

$$P(\widehat{\mathbf{w}}) - P_* \leq O\left(\frac{d\log n + \log(1/\delta)}{n} + \left[\frac{(d\log n + \log(1/\delta))P_*}{n}\right]^{\frac{1}{2-\theta}}\right).$$

When $n \geq \Omega\left(\left(\alpha^{1/\theta} d\log n\right)^{2-\theta}\right)$, with probability at least $1-\delta$,

$$P(\widehat{\mathbf{w}}) - P_* \leq O\left(\left[\frac{d\log n + \log(1/\delta)}{n}\right]^{\frac{2}{2-\theta}} + \left[\frac{(d\log n + \log(1/\delta))P_*}{n}\right]^{\frac{1}{2-\theta}}\right).$$

Remark: The constant in big O and Ω can be seen from the proof, which is tedious and included in the supplement. Here we focus on the understanding of the results. First, the above results are optimistic rates that are no worse than that in Theorem 1. Second, the first result implies that when the optimal risk P_* is less than $O((d\log n/n)^{1-\theta})$, the excess risk bound is in the order of $O(d\log n/n)$. Third, when the number of samples n is sufficiently large and the optimal risk is sufficiently small, the second result can imply a faster rate than $O(d\log n/n)$. Considering smooth functions presented in Section 5 with $\theta = 1$, when $n \geq \Omega(\alpha d\log n)$ and $P_* \leq O(d\log n/n)$ (large-sample and small optimal risk), the excess risk can be bounded by $O((d\log n/n)^2)$. In another word, the sample complexity for achieving an ϵ-excess risk bound is given by $\widetilde{O}(d/\sqrt{\epsilon})$. To the best of our knowledge, the sample complexity of ERM in the order of $1/\sqrt{\epsilon}$ for these examples is the first result appearing in the literature.

Algorithm 1 SSG($\mathbf{w}_1, \gamma, T, \mathcal{W}$)

Input: $\mathbf{w}_1 \in \mathcal{W}$, $\gamma > 0$ and T

for $t = 1, \ldots, T$ **do**
  $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}_t - \gamma g_t)$
end for
$\widehat{\mathbf{w}}_T = \frac{1}{T+1}\sum_{t=1}^{T+1} \mathbf{w}_t$
return $\widehat{\mathbf{w}}_T$

Algorithm 2 ASA($\mathbf{w}_1, n, R$)

Set $R_0 = 2R$, $\widehat{\mathbf{w}}_0 = \mathbf{w}_1$, $m = \lfloor \frac{1}{2}\log_2 \frac{2n}{\log_2 n} \rfloor - 1$, $n_0 = \lfloor \frac{n}{m} \rfloor$
for $k = 1, \ldots, m$ **do**
  Set $\gamma_k = \frac{R_{k-1}}{G\sqrt{n_0+1}}$ and $R_k = R_{k-1}/2$
  $\widehat{\mathbf{w}}_k = \text{SSG}(\widehat{\mathbf{w}}_{k-1}, \gamma_k, n_0, \mathcal{W} \cap \mathcal{B}(\widehat{\mathbf{w}}_{k-1}, R_{k-1}))$
end for
return $\widehat{\mathbf{w}}_m$

4 Efficient SA for Lipschitz continuous random functions

In this section, we will present intermediate rates of an efficient stochastic approximation algorithm for solving (1) adaptive to the EBC under the Assumption 1 and 2. Note that (2) can be considered as a special case by absorbing $r(\mathbf{w})$ into $f(\mathbf{w}, \mathbf{z})$.

Denote by $\mathbf{z}_1, \ldots \mathbf{z}_k, \ldots$ i.i.d samples drawn sequentially from the distribution $\mathbb{P}$, by $g_k \in \partial f(\mathbf{w}, \mathbf{z}_k)|_{\mathbf{w}=\mathbf{w}_k}$ a *stochastic subgradient* evaluated at $\mathbf{w}_k$ with sample $\mathbf{z}_k$, and by $\mathcal{B}(\mathbf{w}, R)$ a bounded ball centered at $\mathbf{w}$ with a radius R. By the Lipschitz continuity of f, we have $\|\partial f(\mathbf{w}, \mathbf{z})\|_2 \leq G$ for $\forall \mathbf{w} \in \mathcal{W}, \forall \mathbf{z} \in \mathcal{Z}$.

The proposed adaptive stochastic approximation algorithm is presented in Algorithm 2, which is referred to as ASA. The updates are divided into m stages, where at each stage a stochastic subgradient method (Algorithm 1) is employed for running $n_0 = \lfloor n/m \rfloor$ iterations with a constant step size γ_k. The step size γ_k will be decreased by half after each stage and the next stage will be warm-started using the solution returned from the last stage as the initial solution. The projection onto the intersection of $\mathcal{W}$ and a shrinking bounded ball at each stage is a commonly used trick for the high probability analysis [14, 15, 49]. We emphasize that the subroutine in ASA can be replaced by other SA algorithms, e.g., the proximal variant of stochastic subgradient for handling a non-smooth deterministic component such as ℓ_1 norm regularization [7], stochastic mirror descent with with a p-norm divergence function [8], and etc. Please see an example in the supplement.

It is worth mentioning that the dividing schema of ASA is due to [15], which however restricts its analysis to uniformly convex functions where uniform convexity is a stronger condition than the EBC. ASA is also similar to a recently proposed accelerated stochastic subgradient (ASSG) method under the EBC [49]. However, the key differences are that (i) ASA is developed for a fixed number of iterations while ASSG is developed for a fixed accuracy level ϵ; (ii) the adaptive iteration complexity of ASSG requires knowing the value of $\theta \in (0, 2]$ while ASA does not require the value of θ. As a trade-off, we restrict our attention to $\theta \in (0, 1]$.

Theorem 3 (Result III). *Suppose Assumptions 1 and 2 hold, and $\|\mathbf{w}_1 - \mathbf{w}^*\|_2 \leq R_0$, where $\mathbf{w}^*$ is the closest optimal solution to $\mathbf{w}_1$. Define $\bar{\alpha} = \max(\alpha G^2, (R_0 G)^{2-\theta})$. For $n \geq 100$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta$, we have*

$$P(\widehat{\mathbf{w}}_m) - P_* \leq O\left(\frac{\bar{\alpha}(\log(n)\log(\log(n)/\delta))}{n}\right)^{\frac{1}{2-\theta}}.$$

Remark: The significance of the result is that although Algorithm 2 does not utilize any knowledge about EBC, it is automatically adaptive to the EBC. As a final note, the projection onto the intersection of $\mathcal{W}$ and a bounded ball can be efficiently computed by employing the projection onto $\mathcal{W}$ and a binary search for the Lagrangian multiplier of the ball constraint. Moreover, we can replace the subroutine with a slightly different variant of SSG to get around of the projection onto the intersection of $\mathcal{W}$ and a bounded ball, which is presented in the supplement.

5 Applications

From the last two sections, we can see that $\theta = 1$ is a favorable case, which yields the fastest rate in our results. It is obvious that if $f(\mathbf{w}, \mathbf{z})$ is strongly convex or $P(\mathbf{w})$ is strongly convex, then EBC($\theta = 1, \alpha$) holds. Below we show some examples of problem (1) and (2) with $\theta = 1$ without

strong convexity, which not only recover some known results of fast rate $\widetilde{O}(d/n)$, but also induce new results of fast rates that are even faster than $\widetilde{O}(d/n)$.

Quadratic Problems (QP): $$\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w}) \triangleq \mathbf{w}^\top \mathbb{E}_{\mathbf{z}}[A(\mathbf{z})]\mathbf{w} + \mathbf{w}^\top \mathbb{E}_{\mathbf{z}'}[\mathbf{b}(\mathbf{z}')] + c, \tag{9}$$

where c is a constant. The random function can be taken as $f(\mathbf{w}, \mathbf{z}, \mathbf{z}') = \mathbf{w}^\top A(\mathbf{z})\mathbf{w} + \mathbf{w}^\top \mathbf{b}(\mathbf{z}') + c$. We have the following corollary.

Corollary 1. *If $\mathbb{E}_{\mathbf{z}}[A(\mathbf{z})]$ is a positive semi-definite matrix (not necessarily positive definite) and $\mathcal{W}$ is a bounded polyhedron, then the problem (9) satisfies EBC($\theta = 1, \alpha$). Assume that $\max(\|A(\mathbf{z})\|_2, \|b(\mathbf{z}')\|_2) \leq \sigma < \infty$, then ERM has a fast rate at least $\widetilde{O}(d/n)$. If $f(\mathbf{w}, \mathbf{z}, \mathbf{z}')$ is further non-negative, convex and smooth, then ERM has a fast rate of $\widetilde{O}((d/n)^2 + dP_*/n)$ when $n \geq \Omega(d \log n)$. ASA has a convergence rate of $\widetilde{O}(1/n)$.*

Next, we present some instances of the quadratic problem (9).
Instance 1 of QP: minimizing the expected square loss. Consider the following problem:

$$\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{x},y}[(\mathbf{w}^\top \mathbf{x} - y)^2], \tag{10}$$

where $\mathbf{x} \in \mathcal{X}, y \in \mathcal{Y}$ and $\mathcal{W}$ is a bounded polyhedron (e.g., ℓ_1-ball or ℓ_∞-ball). It is not difficult to show that it is an instance of (9) and has the property that $f(\mathbf{w}, \mathbf{z}, \mathbf{z}')$ is non-negative, smooth, convex, Lipchitz continuous over $\mathcal{W}$. The convergence results in Corollary 1 for this instance not only recover some known results of $\widetilde{O}(d/n)$ rate [22, 26], but also imply a faster rate than $\widetilde{O}(d/n)$ in a large-sample regime and an optimistic case when $n \geq \Omega(d \log n), P_* \leq O(d \log n/n)$, where the latter result is the first such result of its own.

Instance 2 of QP. Let us consider the following problem:

$$\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{z}}[\mathbf{w}^\top (S - \mathbf{z}\mathbf{z}^\top)\mathbf{w}] - \mathbf{w}^\top \mathbf{b}, \tag{11}$$

where $S - \mathbb{E}_{\mathbf{z}}[\mathbf{z}\mathbf{z}^\top] \succeq 0$. It is notable that $f(\mathbf{w}, \mathbf{z}) = \mathbf{w}^\top (S - \mathbf{z}\mathbf{z}^\top)\mathbf{w} - \mathbf{w}^\top \mathbf{b}$ might be non-convex. A similar problem as (11) could arise in computing the leading eigen-vector of $\mathbb{E}[\mathbf{z}\mathbf{z}^\top]$ by performing shifted-and-inverted power method over random samples $\mathbf{z} \sim \mathbb{P}$ [10].

Piecewise Linear Problems (PLP): $$\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w}) \triangleq \mathbb{E}[f(\mathbf{w}, \mathbf{z})], \tag{12}$$

where $\mathbb{E}[f(\mathbf{w}, \mathbf{z})]$ is a piecewise linear convex function and $\mathcal{W}$ is a bounded polyhedron. We have the following corollary.

Corollary 2. *If $\mathbb{E}[f(\mathbf{w}, \mathbf{z})]$ is piecewise linear and convex and $\mathcal{W}$ is a bounded polyhedron, then the problem (12) satisfies EBC($\theta = 1, \alpha$). If $f(\mathbf{w}, \mathbf{z})$ is Lipschitz continuous, then ERM has a fast rate at least $\widetilde{O}(d/n)$, and ASA has a convergence rate of $\widetilde{O}(1/n)$. If $f(\mathbf{w}, \mathbf{z})$ is further non-negative and linear, then ERM has a fast rate of $\widetilde{O}((d/n)^2 + dP_*/n)$ when $n \geq \Omega(d \log n)$.*

Instance 1 of PLP: minimizing the expected hinge loss for bounded data. Consider the following problem:

$$\min_{\|\mathbf{w}\|_p \leq B} P(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{x},y}[(1 - y\mathbf{w}^\top \mathbf{x})_+], \tag{13}$$

where $p = 1, \infty$ and $y \in \{1, -1\}$. Suppose that $\mathbf{x} \in \mathcal{X}$ is bounded and scaled such that $|\mathbf{w}^\top \mathbf{x}| \leq 1$. Koolen et al. [20] has considered this instance with $p = 2$ and proved that the Bernstein condition (Definition 2) holds with $\beta = 1$ for the problem (13) when $\mathbb{E}[y\mathbf{x}] \neq 0$ and $|\mathbf{w}^\top \mathbf{x}| \leq 1$. In contrast, we can show that the problem (13) with any $p = 1, 2, \infty$ norm constraint [3], the EBC($\theta = 1, \alpha$) holds since the objective $P(\mathbf{w}) = 1 - \mathbf{w}^\top \mathbb{E}[y\mathbf{x}]$ is essentially a linear function of $\mathbf{w}$. Then all results in Corollary 2 hold. To the best of our knowledge, the fast rates of ERM and SA for this instance with ℓ_1 and ℓ_∞ norm constraint are the new results. In comparison, Koolen et al.'s [20] fast rate of

[3] The case of $p = 2$ is showed later.

$\widetilde{O}(1/n)$ only applies to SA and ℓ_2 norm constraint, and their SA algorithm is not as efficient as our SA algorithm.

Instance 2 of PLP: multi-dimensional newsvendor problem. Consider a firm that manufactures p products from q resources. Suppose that a manager must decide on a resource vector $\mathbf{x} \in \mathbb{R}^q_+$ before the product demand vector $\mathbf{z} \in \mathbb{R}^p$ is observed. After the demand becomes known, the manager chooses a production vector $\mathbf{y} \in \mathbb{R}^p$ so as to maximize the operating profit. Assuming that the demand $\mathbf{z}$ is a random vector with discrete probability distribution, the problem is equivalent to

$$\min_{\mathbf{x}\in\mathbb{R}^q_+, \mathbf{x}\leq\mathbf{b}} \mathbf{c}^\top \mathbf{x} - \mathbb{E}[\Pi(\mathbf{x}; \mathbf{z})],$$

where both $\Pi(\mathbf{x}; \mathbf{z})$ and $\mathbb{E}[\Pi(\mathbf{x}; \mathbf{z})]$ are piecewise linear concave functions [18]. Then the problem fits to the setting in Corollary 2.

Risk Minimization Problems over an ℓ_2 ball. Consider the following problem

$$\min_{\|\mathbf{w}\|_2\leq B} P(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{z}}[f(\mathbf{w}, \mathbf{z})]. \tag{14}$$

Assuming that $P(\mathbf{w})$ is convex and $\min_{\mathbf{w}\in\mathbb{R}^d} P(\mathbf{w}) < \min_{\|\mathbf{w}\|_2\leq B} P(\mathbf{w})$, we can show that EBC($\theta = 1, \alpha$) holds (see supplement). Using this result, we can easily show that the considered problem (13) with $p = 2$ satisfies EBC($\theta = 1, \alpha$).

Risk Minimization with ℓ_1 Regularization Problems. For ℓ_1 regularized risk minimization:

$$\min_{\|\mathbf{w}\|_1\leq B} P(\mathbf{w}) \triangleq \mathbb{E}[f(\mathbf{w}; \mathbf{z})] + \lambda\|\mathbf{w}\|_1, \tag{15}$$

we have the following corollary.

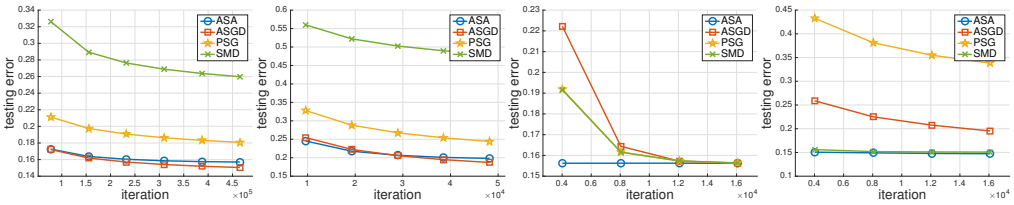
Corollary 3. *If the first component is quadratic as in (9) or is piecewise linear and convex, then the problem (15) satisfies EBC($\theta = 1, \alpha$). If the random function is Lipschitz continuous, then ERM has a fast rate at least $\widetilde{O}(d/n)$, and ASA has a convergence rate of $\widetilde{O}(1/n)$. If $f(\mathbf{w}, \mathbf{z})$ is further non-negative, convex and smooth, then ERM has a fast rate of $\widetilde{O}((d/n)^2 + dP_*/n)$ when $n \geq \Omega(d \log n)$.*

To the best of our knowledge, this above general result is the first of its kind. Next, we show some instances satisfying EBC(θ, α) with $\theta < 1$. Consider the problem $\min_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w}) \triangleq P(\mathbf{w}) + \lambda\|\mathbf{w}\|_p^p$, where $P(\mathbf{w})$ is quadratic as in (9), and $\mathcal{W}$ is a bounded polyhedron. In the supplement, we prove that EBC($\theta = 2/p, \alpha$) holds.

A Case Study for ASA. Finally, we provide some empirical evidence to support the effectiveness of the proposed ASA algorithm. In particular, we will consider solving an ℓ_1 regularized expected square loss minimization problem (15) for learning a predictive model. We compare with two baselines whose convergence rate are known as $O(1/\sqrt{n})$, namely proximal stochastic gradient (PSG) method [7], and stochastic mirror descent (SMD) method using a p-norm divergence function ($p = 2\log d$) other than the Euclidean function. For SMD, we implement the algorithm proposed in [37], which was proposed for solving (15) and could be effective for very high-dimensional data. For ASA, we implement two versions that use PSG and SMD as the subroutine and report the one that gives the best performance. The two versions differ in using the Euclidean norm or the p-norm for measuring distance. Since the comparison is focused on the testing error, we also include another strong baseline, i.e, averaged stochastic gradient (ASGD) with a constant step size, which enjoys an $O(d/n)$ rate for minimizing the expected square loss without any constraints or regularizations [1].

We use four benchmark datasets from libsvm website[4], namely, real-sim, rcv1_binary, E2006-tfidf, E2006-log1p, whose dimensionality is 20958, 47236, 150360, 4272227, respectively. We divide each dataset into three sets, respectively training, validation, and testing. For E2006-tfidf and E2006-log1p dataset, we randomly split the given testing set into half validation and half testing. For the dataset real-sim which do not explicitly provides a testing set, we randomly split the entire data into 4:1:1 for training, validation, and testing. For rcv1_binary, despite that the test set is given, the size of the

[4]http://www.csie.ntu.edu.tw/ cjlin/libsvmtools/datasets/



(a) rcv1_binary (b) real-sim (c) E2006-tfidf (d) E2006-log1p

Figure 1: Testing Error vs Iteration of ASA and other baselines for SA

training set is relatively small. Thus we first combine the training and the testing sets and then follow the above procedure to split it.

The involved parameters of each algorithm are tuned based on the validation data. With the selected parameters, we run each algorithm by passing through training examples once and evaluate intermediate models on the testing data to compute the testing error measured by square loss. The results on different data sets averaged over 5 random runs over shuffled training examples are shown in Figure 1. From the testing curves, we can see that the proposed ASA has similar convergence rate to ASGD on two relatively low-dimensional data sets. This is not surprise since both algorithms enjoy an $\widetilde{O}(1/n)$ convergence rate indicated by their theories. For the data set E2006-tfidf and E2006-log1p, we observe that ASA converges much faster than ASGD, which is due to the presence of ℓ_1 regularization. In addition, ASA converges much faster than SGD and SMD with one exception on E2006-log1p, on which ASA performs slightly better than SMD.

6 Conclusion

We have comprehensively studied statistical learning under the error bound condition for both ERM and SA. We established the connection between the error bound condition and previous conditions for developing fast rates of empirical risk minimization for Lipschitz continuous loss functions. We also developed improved rates for non-negative and smooth convex loss functions, which induce faster rates that were not achieved before. Finally, we analyzed an efficient "parameter"-free SA algorithm under the error bound condition and showed that it is automatically adaptive to the error bound condition. Applications in machine learning and other fields are considered and empirical studies corroborate the fast rate of the developed algorithms. An open question is how to develop efficient SA algorithms under the error bound condition with optimistic rates for non-negative smooth loss functions similar to the results obtained for empirical risk minimization in this paper.

Acknowledgement

The authors thank the anonymous reviewers for their helpful comments. M. Liu and T. Yang are partially supported by National Science Foundation (IIS-1545995). L. Zhang is partially supported by YESS (2017QNRC001). We thank Nishant A. Mehta for pointing out the work [12] for the proof of Theorem 1.

References

[1] Francis R. Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $\mathcal{O}(1/n)$. In *Advances in Neural Information Processing Systems (NIPS)*, pages 773–781, 2013.

[2] Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 2005.

[3] Peter L. Bartlett and Shahar Mendelson. Empirical minimization. *Probability Theory and Related Fields*, 2006.

[4] Jerome Bolte, Trong Phong Nguyen, Juan Peypouquet, and Bruce Suter. From error bounds to the complexity of first-order descent methods for convex functions. *CoRR*, abs/1510.08234, 2015.

[5] James V. Burke and Michael C. Ferris. Weak sharp minima in mathematical programming. *SIAM Journal on Control and Optimization*, 31(5):1340–1359, 1993.

[6] Dmitriy Drusvyatskiy and Adrian S. Lewis. Error bounds, quadratic growth, and linear convergence of proximal methods. *arXiv:1602.06661*, 2016.

[7] John Duchi and Yoram Singer. Efficient online and batch learning using forward backward splitting. *Journal of Machine Learning Research*, 10:2899–2934, 2009.

[8] John C. Duchi, Shai Shalev-Shwartz, Yoram Singer, and Ambuj Tewari. Composite objective mirror descent. In *COLT*, pages 14–26. Omnipress, 2010.

[9] Vitaly Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. In *NIPS*. 2016.

[10] Dan Garber, Elad Hazan, Chi Jin, Sham M. Kakade, Cameron Musco, Praneeth Netrapalli, and Aaron Sidford. Faster eigenvector computation via shift-and-invert preconditioning. In *ICML*, 2016.

[11] Alon Gonen and Shai Shalev-Shwartz. Average stability is invariant to data preconditioning: Implications to exp-concave empirical risk minimization. *J. Mach. Learn. Res.*, 18(1):8245–8257, January 2017.

[12] Peter D. Grünwald and Nishant A. Mehta. Fast rates with unbounded losses. *CoRR*, abs/1605.00252, 2016.

[13] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 2007.

[14] Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *COLT*, 2011.

[15] Anatoli Juditsky and Yuri Nesterov. Deterministic and stochastic primal-dual subgradient algorithms for uniformly convex minimization. *Stoch. Syst.*, 2014.

[16] Sham M. Kakade and Ambuj Tewari. On the generalization ability of online strongly convex programming algorithms. In *NIPS*, 2008.

[17] Hamed Karimi, Julie Nutini, and Mark W. Schmidt. Linear convergence of gradient and proximal-gradient methods under the polyak-łojasiewicz condition. In *ECML-PKDD*, 2016.

[18] Sujin Kim, Raghu Pasupathy, and Shane G. Henderson. *A Guide to Sample Average Approximation*, pages 207–243. Springer New York, New York, NY, 2015.

[19] Vladimir Koltchinskii. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 2006.

[20] Wouter M. Koolen, Peter Grünwald, and Tim van Erven. Combining adversarial guarantees and stochastic fast rates in online learning. In *NIPS*, 2016.

[21] Tomer Koren and Kfir Y. Levy. Fast rates for exp-concave empirical risk minimization. In *NIPS*, 2015.

[22] Wee Sun Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 44(5):1974–1980, 1998.

[23] Guoyin Li. Global error bounds for piecewise convex polynomials. *Math. Program.*, 2013.

[24] Guoyin Li and Ting Kei Pong. Calculus of the exponent of kurdyka- łojasiewicz inequality and its applications to linear convergence of first-order methods. *CoRR*, abs/1602.02915, 2016.

[25] Enno Mammen and Alexandre B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27(6):1808–1829, 12 1999.

[26] Nishant A. Mehta. Fast rates with high probability in exp-concave statistical learning. In *AISTATS*, pages –, 2017.

[27] Nishant A. Mehta and Robert C. Williamson. From stochastic mixability to fast rates. In *NIPS*, 2014.

[28] I. Necoara, Yu. Nesterov, and F. Glineur. Linear convergence of first order methods for non-strongly convex optimization. *CoRR*, abs/1504.06298, v4, 2015.

[29] Arkadi Nemirovski, Anatoli Juditsky, Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 2009.

[30] Yurii Nesterov. *Introductory lectures on convex optimization: a basic course*. 2004.

[31] Jong-Shi Pang. Error bounds in mathematical programming. *Math. Program.*, 1997.

[32] Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *ICML*, 2012.

[33] Aaditya Ramdas and Aarti Singh. Optimal rates for stochastic convex optimization under tsybakov noise condition. In *ICML*, 2013.

[34] R.T. Rockafellar. *Convex Analysis*. 1970.

[35] Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.

[36] Shai Shalev-Shwartz, Yoram Singer, and Nathan Srebro. Pegasos: Primal estimated sub-gradient solver for svm. In *ICML*, 2007.

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

[38] Ohad Shamir and Tong Zhang. Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *ICML*, 2013.

[39] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory, Second Edition*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2014.

[40] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 2007.

[41] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Optimistic rates for learning with a smooth loss. *ArXiv e-prints*, arXiv:1009.3896, 2010.

[42] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In *NIPS*, 2010.

[43] Karthik Sridharan, Shai Shalev-Shwartz, and Nathan Srebro. Fast rates for regularized objectives. In *NIPS*, 2008.

[44] Alexander B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 02 2004.

[45] Alexander B Tsybakov et al. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32(1):135–166, 2004.

[46] Tim van Erven, Peter D. Grünwald, Nishant A. Mehta, Mark D. Reid, and Robert C. Williamson. Fast rates in statistical and online learning. *JMLR*, 2015.

[47] Tim van Erven and Wouter M. Koolen. Metagrad: Multiple learning rates in online learning. In *NIPS*, 2016.

[48] Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, 1998.

[49] Yi Xu, Qihang Lin, and Tianbao Yang. Accelerate stochastic subgradient method by leveraging local error bound. *CoRR*, abs/1607.01027, 2016.

[50] Yi Xu, Qihang Lin, and Tianbao Yang. Stochastic convex optimization: Faster local growth implies faster global convergence. In *ICML*, pages 3821–3830, 2017.

[51] Tianbao Yang, Zhe Li, and Lijun Zhang. A simple analysis for exp-concave empirical minimization with arbitrary convex regularizer. In *AISTATS*, pages 445–453, 2018.

[52] Tianbao Yang and Qihang Lin. Rsg: Beating subgradient method without smoothness and strong convexity. *CoRR*, abs/1512.03107, 2016.

[53] W. H. Yang. Error bounds for convex polynomials. *SIAM Journal on Optimization*, 2009.

[54] Hui Zhang. New analysis of linear convergence of gradient-type methods via unifying error bound conditions. *CoRR*, abs/1606.00269, 2016.

[55] Lijun Zhang, Tianbao Yang, and Rong Jin. Empirical risk minimization for stochastic convex optimization: O(1/n)- and o(1/n^2)-type of risk bounds. *CoRR*, abs/1702.02030, 2017.

[56] Martin Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. In *ICML*, pages 928–936, 2003.

A Proof of Lemma 1

Proof. The proof follows similarly as the proof of Theorem 5.4 in [46]. Let us fix an arbitrary $\mathbf{w} \in \mathcal{W}$ and its closest optimal solution $\mathbf{w}^* \in \mathcal{W}_*$. Let $X = f(\mathbf{w}, \mathbf{z}) - f(\mathbf{w}^*, \mathbf{z})$ be a random variable due to $\mathbf{z}$. Then $|X| \leq 2GR \triangleq a$. Let $b > 0$ be any finite constant, $\kappa(x) = (e^x - x - 1)/x^2$ for $x \neq 0$ and $\kappa(0) = 1/2$, $c_1^b = 1/\kappa(2ba)$. Let $B = \alpha G^2$ and $v(x) = \frac{c_1^b}{B} x^{1-\theta} \wedge b$. Let $\varepsilon \geq 0$ and set $\eta = v(\varepsilon) \leq \frac{c_1^b}{B} \varepsilon^{1-\theta}$.

According to our analysis in the paper, we have established a similar condition to the Bernstein condition under our conditions, i.e.,

$$\mathbb{E}_{\mathbf{z}}[(f(\mathbf{w}, \mathbf{z}) - f(\mathbf{w}^*, \mathbf{z}))^2] \leq B(\mathbb{E}_{\mathbf{z}}[f(\mathbf{w}, \mathbf{z}) - f(\mathbf{w}^*, \mathbf{z})])^\theta$$

where $B = \alpha G^2$. Then

$$\text{Var}[(f(\mathbf{w}, \mathbf{z}) - f(\mathbf{w}^*, \mathbf{z}))] \leq B(\mathbb{E}_{\mathbf{z}}[f(\mathbf{w}, \mathbf{z}) - f(\mathbf{w}^*, \mathbf{z})])^\theta$$

First, when $\varepsilon = 0$ we have $\eta = 0$, the $\mathbb{E}[e^{-\eta X}] \leq e^{\eta\varepsilon}$ hold trivially. Thus we focus on the case $\varepsilon > 0$, which implies that $\eta > 0$. Then Lemma 5.6 in [46] applied to the random variable η gives

$$\mathbb{E}[X] + \frac{1}{\eta} \log \mathbb{E}[e^{-\eta X}] \leq \kappa(2ba)\eta \text{Var}(X) \leq \kappa(2ba)\eta B(\mathbb{E}[X])^\theta \leq \varepsilon^{1-\theta}(\mathbb{E}[X])^\theta.$$

If $\varepsilon \leq \mathbb{E}[X]$, then $\varepsilon^{1-\theta}(\mathbb{E}[X])^\theta \leq \mathbb{E}[X]$, which implies $\frac{1}{\eta} \log \mathbb{E}[e^{-\eta X}] \leq 0 \leq \varepsilon$. This establishes the second part and the first part for $\varepsilon \leq \mathbb{E}[X]$. For $\varepsilon \geq \mathbb{E}[X]$, we have $\varepsilon^{1-\theta}(\mathbb{E}[X])^\theta \leq \varepsilon$. Then due to $\mathbb{E}[X] \geq 0$, we have $\frac{1}{\eta} \log \mathbb{E}[e^{-\eta X}] \leq \varepsilon$. □

B Proof of Theorem 1

Proof. Let $F_{\mathbf{w}}(\mathbf{z}) = f(\mathbf{w}, \mathbf{z}) - f(\mathbf{w}^*, \mathbf{z})$, where $\mathbf{w}^*$ is the closest optimal solution to $\mathbf{w}$. Denote by $B = 2GR$. It is clear that $F_{\mathbf{w}}(\mathbf{z}) \leq B$. The goal is to show that with high probability, ERM does not select any $\mathbf{w} \in \mathcal{W}$ whose excess risk $P(\mathbf{w}) - P_* = \mathbb{E}_{\mathbf{z}}[F_{\mathbf{w}}(\mathbf{z})]$ is large than $\left(\frac{a}{n}\right)^{\frac{1}{2-\theta}}$ for some constant a. Clearly, with probability 1 ERM will never select any $\mathbf{w}$ for which both $F_{\mathbf{w}}(\mathbf{z}) > 0$ almost surely and with some positive probability $F_{\mathbf{w}}(\mathbf{z}) > 0$. These predictors are called the empirically inadmissible models. For any $\gamma_n > 0$, let $\mathcal{W}_{\geq\gamma_n}$ denote the subclass of models by starting with $\mathcal{W}$, retaining only models whose excess risk is at least γ_n, and further removing the empirically inadmissible models.

The goal now can be expressed equivalently as showing that, with high probability, ERM does not select any model $\mathbf{w} \in \mathcal{W}_{\geq\gamma_n}$, where $\gamma_n = \left(\frac{a}{n}\right)^{\frac{1}{2-\theta}}$. Let $\mathcal{W}_{\geq\gamma_n,\varepsilon}$ be the optimal proper $(\varepsilon/(2G))$-cover of $\mathcal{W}_{\geq\gamma_n}$. Note that this cover induces an ε-cover in sup norm over the function class $\{F_{\mathbf{w}} : \mathbf{w} \in \mathcal{W}_{\geq\gamma_n}\}$. To see this, for any $\mathbf{w} \in \mathcal{W}_{\geq\gamma_n}$, there exists $\widetilde{\mathbf{w}} \in \mathcal{W}_{\geq\gamma_n,\varepsilon}$ such that $\|\mathbf{w} - \widetilde{\mathbf{w}}\|_2 \leq \varepsilon/(2G)$. As a result,

$$\begin{aligned}
&\sup_{\mathbf{z}} |F_{\mathbf{w}}(\mathbf{z}) - F_{\widetilde{\mathbf{w}}}(\mathbf{z})| = \sup_{\mathbf{z}} |f(\mathbf{w}, \mathbf{z}) - f(\widetilde{\mathbf{w}}, \mathbf{z})| + \sup_{\mathbf{z}} |f(\mathbf{w}^*, \mathbf{z}) - f(\widetilde{\mathbf{w}}^*, \mathbf{z})| \\
&\leq G\|\mathbf{w} - \widetilde{\mathbf{w}}\|_2 + G\|\mathbf{w}^* - \widetilde{\mathbf{w}}^*\|_2 \leq 2G\|\mathbf{w} - \widetilde{\mathbf{w}}\|_2 \leq \varepsilon,
\end{aligned}$$

where $\mathbf{w}^*, \widetilde{\mathbf{w}}^*$ are projections of $\mathbf{w}$ and $\widetilde{\mathbf{w}}$ onto $\mathcal{W}_*$ and the last inequality uses the non-expansiveness of the projection onto $\mathcal{W}_*$, which is convex due to the convexity of $P(\mathbf{w})$ and $\mathcal{W}$. Observe that the ϵ-cover of $\mathcal{W}_{\geq\gamma_n} \subseteq \mathcal{B}^d(R)$ has cardinality at most $\left(\frac{4R}{\varepsilon}\right)^d$, and the cardinality of an optimal proper ε-cover is at most the cardinality of an optimal $\varepsilon/2$-cover. It hence follows that $|\mathcal{W}_{\geq\gamma_n,\varepsilon}| \leq \left(\frac{16GR}{\varepsilon}\right)^d$.

Let us consider a fixed $\mathbf{w} \in \mathcal{W}_{\geq\gamma_n,\varepsilon}$ and its closest optimal solution $\mathbf{w}^* \in \mathcal{W}_*$. According to Lemma 1, we have

$$\mathbb{E}_{\mathbf{z}}[e^{-v(\gamma_n)F_{\mathbf{w}}(\mathbf{z})}] \leq 1$$

Then using Theorem 13 in [12], where we set $u = B$ and $c = 1$, for all $\eta \in (0, v(\gamma_n))$ we have

$$\gamma_n \leq \mathbb{E}_{\mathbf{z}}[F_{\mathbf{w}}(\mathbf{z})] \leq -\frac{\eta B + 1}{1 - \eta/v(\gamma_n)} \frac{1}{\eta} \log \mathbb{E}_{\mathbf{z}}[e^{-\eta F_{\mathbf{w}}(\mathbf{z})}]$$

Let $\eta = v(\gamma_n)/2$, we have

$$\log \mathbb{E}_{\mathbf{z}}[e^{-(v(\gamma_n)/2)F_{\mathbf{w}}(\mathbf{z})}] \leq -\frac{0.5v(\gamma_n)}{Bv(\gamma_n)+2}\gamma_n$$

Applying Theorem 1 in [27] with $t = \frac{\gamma_n}{2}$, we have

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} F_{\mathbf{w}}(\mathbf{z}_i) \leq \frac{\gamma_n}{2}\right) \leq \exp\left(-\frac{0.5v(\gamma_n)}{Bv(\gamma_n)+2}n\gamma_n + \frac{v(\gamma_n)\gamma_n}{4}\right).$$

Assume that $\left(\frac{a}{n}\right)^{\frac{1-\theta}{2-\theta}} \leq \alpha b G^2 \kappa(4GRb)$, i.e., $n \geq a\left(\alpha b G^2 \kappa(4GRb)\right)^{(2-\theta)/(1-\theta)}$, which implies that $v(\gamma_n) = c\left(\frac{a}{n}\right)^{\frac{1-\theta}{2-\theta}} \wedge b = c\left(\frac{a}{n}\right)^{\frac{1-\theta}{2-\theta}}$ by noting the value of $c = 1/(\alpha G^2\kappa(4GRb))$ in Lemma 1. Further we assume $n \geq a(0.5Bc)^{\frac{2-\theta}{1-\theta}}$. Hence $Bv(\gamma_n) \leq 2$.

$$\begin{aligned}\Pr\left(\frac{1}{n}\sum_{i=1}^{n} F_{\mathbf{w}}(\mathbf{z}_i) \leq \frac{\gamma_n}{2}\right) &\leq \exp\left(-\frac{0.5v(\gamma_n)}{Bv(\gamma_n)+2}n\gamma_n + \frac{v(\gamma_n)\gamma_n}{4}\right) \leq \exp\left(-0.125v(\gamma_n)n\gamma_n + \frac{v(\gamma_n)\gamma_n}{4}\right) \\ &= \exp\left(-0.125ca + \frac{ca}{4n}\right) \leq \exp\left(-0.375ca\right),\end{aligned}$$

where we use $n \geq 1$.

As a result, we have

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{n} F_{\mathbf{w}}(\mathbf{z}_i) \leq \frac{\gamma_n}{2}\right) \leq \exp\left(-0.375ca\right).$$

Taking a union bound over $\mathcal{W}_{\geq\gamma_n,\varepsilon}$ we have that

$$\Pr\left(\exists \mathbf{w} \in \mathcal{W}_{\geq\gamma_n,\varepsilon}, \frac{1}{n}\sum_{i=1}^{n} F_{\mathbf{w}}(\mathbf{z}_i) \leq \frac{\gamma_n}{2}\right) \leq \left(\frac{16GR}{\varepsilon}\right)^d \exp\left(-0.375ca\right)$$

Taking $\varepsilon = \frac{1}{2n^{1/(2-\theta)}}$ and $a = \frac{3}{c}(d\log(32GRn^{1/(2-\theta)}) + \log(1/\delta))$, with probability $1-\delta$ for all $\mathbf{w} \in \mathcal{W}_{\geq\gamma_n,\varepsilon}$, we have $\frac{1}{n}\sum_{i=1}^{n} F_{\mathbf{w}}(\mathbf{z}_i) \geq \frac{a^{1/(2-\theta)}}{2n^{1/(2-\theta)}}$.

Now, since $\sup_{\mathbf{w}\in\mathcal{W}_{\geq\gamma_n}} \min_{\widetilde{\mathbf{w}}\in\mathcal{W}_{\geq\gamma_n,\varepsilon}} \|F_{\mathbf{w}} - F_{\mathbf{w}}\|_\infty \leq \varepsilon = \frac{1}{2n^{1/(2-\theta)}}$, and by increasing a by 1 to guarantee that $a > 1$, with probability $1-\delta$, for all $\mathbf{w} \in \mathcal{W}_{\geq\gamma_n}$, $\frac{1}{n}\sum_{i=1}^{n} F_{\mathbf{w}}(\mathbf{z}_i) > 0$. □

C Proof of Theorem 2

The proof follows the framework developed in [55], which converts the excess risk bound of $\widehat{\mathbf{w}}$ into large deviation of gradients. In particular, if we let $F(\mathbf{w}) = \mathbb{E}[f(\mathbf{w};\mathbf{z})]$ and $F_n(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n} f(\mathbf{w};\mathbf{z}_i)$, we will prove the following lemma.

Lemma 2. *If we let $\widehat{\mathbf{w}}^*$ be an optimal solution to $\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w})$ that is closest to $\widehat{\mathbf{w}}$, then we have*

$$\begin{aligned}P(\widehat{\mathbf{w}}) - P(\widehat{\mathbf{w}}^*) \leq &\|\nabla F(\widehat{\mathbf{w}}) - \nabla F(\widehat{\mathbf{w}}^*) - [\nabla F_n(\widehat{\mathbf{w}}) - \nabla F_n(\widehat{\mathbf{w}}^*)]\|_2 \|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2 \\ &+ \|\nabla F(\widehat{\mathbf{w}}^*) - \nabla F_n(\widehat{\mathbf{w}}^*)\|_2 \|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2\end{aligned}$$

where $P(\mathbf{w}) = F(\mathbf{w}) + r(\mathbf{w})$.

Note that [55] only proves the above result for $P(\mathbf{w}) = F(\mathbf{w})$. Then we use concentration inequalities, covering numbers, and a refined analysis leveraging the EBC to bound the excess risk, where the refined analysis leveraging the EBC is our main contribution for proving Theorem 2.

Proof. (Proof of Lemma 2)

$$\begin{aligned}&P(\widehat{\mathbf{w}}) - P(\widehat{\mathbf{w}}^*) \leq \langle \partial P(\widehat{\mathbf{w}}), \widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\rangle = \langle \partial P(\widehat{\mathbf{w}}) - \partial P(\widehat{\mathbf{w}}^*), \widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\rangle + \langle \partial P(\widehat{\mathbf{w}}^*), \widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\rangle \\ &= \langle \partial P(\widehat{\mathbf{w}}) - \partial P(\widehat{\mathbf{w}}^*) - [\partial P_n(\widehat{\mathbf{w}}) - \partial P_n(\widehat{\mathbf{w}}^*)], \widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\rangle + \langle \partial P_n(\widehat{\mathbf{w}}) - \partial P_n(\widehat{\mathbf{w}}^*) + \partial P(\widehat{\mathbf{w}}^*), \widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\rangle \\ &= \langle \partial P(\widehat{\mathbf{w}}) - \partial P(\widehat{\mathbf{w}}^*) - [\partial P_n(\widehat{\mathbf{w}}) - \partial P_n(\widehat{\mathbf{w}}^*)], \widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\rangle + \langle \partial P(\widehat{\mathbf{w}}^*) - \partial P_n(\widehat{\mathbf{w}}^*), \widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\rangle \\ &+ \langle \partial P_n(\widehat{\mathbf{w}}), \widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\rangle\end{aligned}$$

According to the optimality condition of $\widehat{\mathbf{w}}$, there exists $\mathbf{v} \in \partial r(\widehat{\mathbf{w}})$ such that $\langle \nabla F_n(\widehat{\mathbf{w}}) + \mathbf{v}, \widehat{\mathbf{w}} - \widehat{\mathbf{w}}^* \rangle \leq 0$. Let $\partial P_n(\widehat{\mathbf{w}}) = \nabla F_n(\widehat{\mathbf{w}}) + \mathbf{v}$ and $\partial P(\widehat{\mathbf{w}}) = \nabla F(\widehat{\mathbf{w}}) + \mathbf{v}$ in the above inequality, we have

$$
\begin{aligned}
&P(\widehat{\mathbf{w}}) - P(\widehat{\mathbf{w}}^*) \leq \langle \nabla F(\widehat{\mathbf{w}}) - \nabla F(\widehat{\mathbf{w}}^*) - [\nabla F_n(\widehat{\mathbf{w}}) - \nabla F_n(\widehat{\mathbf{w}}^*)], \widehat{\mathbf{w}} - \widehat{\mathbf{w}}^* \rangle \\
&+ \langle \nabla F(\widehat{\mathbf{w}}^*) - \nabla F_n(\widehat{\mathbf{w}}^*), \widehat{\mathbf{w}} - \widehat{\mathbf{w}}^* \rangle \\
&\leq (\|\nabla F(\widehat{\mathbf{w}}) - \nabla F(\widehat{\mathbf{w}}^*) - [\nabla F_n(\widehat{\mathbf{w}}) - \nabla F_n(\widehat{\mathbf{w}}^*)]\|_2 + \|\nabla F(\widehat{\mathbf{w}}^*) - \nabla F_n(\widehat{\mathbf{w}}^*)\|_2) \cdot \|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2
\end{aligned}
$$

□

The proof below uses the L-smoothness and convexity of $F(\mathbf{w})$, i.e., there exists $L \geq 0$ such that for any $\mathbf{w}, \mathbf{u} \in \mathcal{W}$,

$$
0 \leq f(\mathbf{w}, \mathbf{z}) - f(\mathbf{u}, \mathbf{z}) - \nabla f(\mathbf{u}, \mathbf{z})^\top (\mathbf{w} - \mathbf{u}) \leq \frac{L}{2} \|\mathbf{w} - \mathbf{u}\|_2^2, \quad \forall \mathbf{z} \in \mathcal{Z}.
$$

We first prove the following theorem. Theorem 2 is a corollary of the following theorem by setting $\varepsilon = 1/n$.

Theorem 4. *Let $\varepsilon > 0$ be any constant and $C(\varepsilon) = 2(\log(2/\delta) + d\log(6R/\varepsilon)$. Under* **Assumptions 1, 2, 3***, and that $r(\mathbf{w})$ is convex and G'-Lipschitz continuous over $\mathcal{W}$, with probability at least $1 - 2\delta$, we have*

$$
P(\widehat{\mathbf{w}}) - P_* \leq \frac{4(6LR^2 + \bar{G}R)C(\varepsilon)}{n} + 2\left(1 \vee \alpha^{1/\theta}\right)\left(\frac{4LC(\epsilon)P_*}{n}\right)^{\frac{1}{2-\theta}} + 2\left(12RL + \frac{\bar{G}}{4} + \frac{4LRC(\varepsilon)}{n}\right)\varepsilon,
$$

where $\bar{G} = G + G'$. Furthermore, if $n \geq \left(256LC(\varepsilon)\alpha^{\frac{1}{\theta}}\right)^{(2-\theta)}$, we also have

$$
\begin{aligned}
P(\widehat{\mathbf{w}}) - P_* \leq\ & 34LC(\varepsilon)\left(\frac{1}{n}\right)^{\frac{2}{2-\theta}} + 2\left(1 \vee 4\alpha^{1/\theta}\right)\left(\frac{\bar{G}C(\varepsilon)}{n}\right)^{\frac{2}{2-\theta}} + 2\left(1 \vee 64\alpha^{1/\theta}\right)\left(\frac{4LC(\varepsilon)P_*}{n}\right)^{\frac{1}{2-\theta}} \\
&+ 4LC(\varepsilon)\left(1 \vee 64\alpha^{1/\theta}\right)\left(\frac{\varepsilon}{n}\right)^{\frac{2}{2-\theta}} + 12L\left(1 \vee 64\alpha^{1/\theta}\right)\varepsilon^{\frac{2}{2-\theta}} + 2\left(1 \vee 64\alpha^{1/\theta}\right)\left(\frac{4L\bar{G}C(\varepsilon)\varepsilon}{n}\right)^{\frac{1}{2-\theta}}.
\end{aligned}
$$

To prove the theorem, we need the following lemmas.

Lemma 3. *Under* **Assumptions 1***, with probability at least $1 - \delta$, for any $\mathbf{w} \in \mathcal{W}$, we have*

$$
\begin{aligned}
&\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}^*) - [\nabla F_n(\mathbf{w}) - \nabla F_n(\mathbf{w}^*)]\|_2 \\
&\leq \frac{LC(\varepsilon)\|\mathbf{w} - \mathbf{w}^*\|_2}{n} + \frac{2LC(\varepsilon)\varepsilon}{n} + \sqrt{\frac{LC(\varepsilon)(P(\mathbf{w}) - P_*)}{n}} + 2\sqrt{\frac{L\bar{G}C(\varepsilon)\varepsilon}{n}} + 4L\varepsilon.
\end{aligned}
$$

where $\mathbf{w}^*$ is the closest optimal solution to $\mathbf{w}$ and $C(\varepsilon)$ is define in Theorem 4.

Lemma 4. *Under* **Assumption 1***, with probability at least $1 - \delta$, for any $\mathbf{w}_* \in \mathcal{W}_*$, we have*

$$
\|\nabla F(\mathbf{w}_*) - \nabla F_n(\mathbf{w}_*)\|_2 \leq \frac{GC(\varepsilon)}{n} + \sqrt{\frac{4LC(\varepsilon)P_*}{n}} + 2L\varepsilon. \tag{16}
$$

Lemma 5. *Let A be a nonnegative number. Under the EBC(θ, α) condition with $\theta \in (0, 1]$ and $0 < \alpha < \infty$, for any $\epsilon > 0$ and $\mathbf{w} \in \mathcal{W}$, we have*

$$
\|\mathbf{w} - \mathbf{w}^*\|_2 \sqrt{A} \leq \left(1 \vee \frac{\alpha^{1/\theta}}{4\epsilon}\right) A^{\frac{1}{2-\theta}} + \epsilon(P(\mathbf{w}) - P_*)
$$

C.1 Proof of Theorem 4

Proof. Using the Lemma 3 and Lemma 4 to proceed bounding the inequality in Lemma 2, with probability at least $1 - 2\delta$, we have

$$
\begin{aligned}
P(\widehat{\mathbf{w}}) - P_* \leq\ & \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2^2}{n} + \frac{\bar{G}C(\varepsilon)\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2}{n} + \frac{2LC(\varepsilon)\varepsilon\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2}{n} + 6L\varepsilon\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2 \\
&+ \|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2 \sqrt{\frac{LC(\varepsilon)(P(\widehat{\mathbf{w}}) - P_*)}{n}} + \|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2 \sqrt{\frac{4LC(\varepsilon)P_*}{n}} + \|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2 \sqrt{\frac{4L\bar{G}C(\varepsilon)\varepsilon}{n}}.
\end{aligned} \tag{17}
$$

Next, we will bound the three terms that have a $1/\sqrt{n}$ factor.

$$\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2 \sqrt{\frac{LC(\varepsilon)(P(\widehat{\mathbf{w}}) - P_*)}{n}} \leq \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2^2}{n} + \frac{P(\widehat{\mathbf{w}}) - P_*}{4}, \tag{18}$$

$$\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2 \sqrt{\frac{4L\bar{G}C(\varepsilon)\varepsilon}{n}} \leq \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2^2}{n} + \bar{G}\varepsilon \tag{19}$$

$$\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2 \sqrt{\frac{4LC(\varepsilon)P_*}{n}} \leq \left(1 \vee \alpha^{1/\theta}\right)\left(\frac{4LC(\varepsilon)P_*}{n}\right)^{\frac{1}{2-\theta}} + \frac{P(\widehat{\mathbf{w}}) - P_*}{4} \tag{20}$$

where the last inequality follows Lemma 5. Combining the inequalities in (17), (18), (19), and (20), with probability $1-\delta$ we have

$$\begin{aligned}
\frac{P(\widehat{\mathbf{w}}) - P_*}{2} &\leq \frac{3LC(\varepsilon)\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2^2}{n} + \frac{\bar{G}C(\varepsilon)\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2}{n} + \frac{2LC(\varepsilon)\varepsilon\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2}{n} + 6L\varepsilon\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2 \\
&\quad + \bar{G}\varepsilon + \left(1 \vee \alpha^{1/\theta}\right)\left(\frac{4LC(\varepsilon)P_*}{n}\right)^{\frac{1}{2-\theta}} \\
&\leq \frac{(12LR^2 + 2\bar{G}R)C(\varepsilon)}{n} + \left(1 \vee \alpha^{1/\theta}\right)\left(\frac{4LC(\varepsilon)P_*}{n}\right)^{\frac{1}{2-\theta}} + \left(12RL + \bar{G} + \frac{4LRC(\varepsilon)}{n}\right)\varepsilon,
\end{aligned}$$

which finishes the first part of the theorem.

To prove the second part, we need more refined analysis. The following inequalities will be proved later.

$$\frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2^2}{n} \leq \max\left(LC(\varepsilon)\left(\frac{1}{n}\right)^{\frac{2}{2-\theta}}, \epsilon(P(\widehat{\mathbf{w}}) - P_*)\right), \text{ for } n \geq \left(LC(\varepsilon)\alpha^{\frac{1}{\theta}}/\epsilon\right)^{(2-\theta)} \tag{21}$$

$$\begin{aligned}
&\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2 \sqrt{\frac{LC(\varepsilon)(P(\widehat{\mathbf{w}}) - P_*)}{n}} \leq \epsilon(P(\widehat{\mathbf{w}}) - P_*) + \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2^2}{4\epsilon n} \\
&\leq \epsilon(P(\widehat{\mathbf{w}}) - P_*) + \max\left(\frac{LC(\varepsilon)}{\epsilon}\left(\frac{1}{n}\right)^{\frac{2}{2-\theta}}, \epsilon(P(\widehat{\mathbf{w}}) - P_*)\right), \text{ for } n \geq \left(LC(\varepsilon)\alpha^{\frac{1}{\theta}}/\epsilon^2\right)^{(2-\theta)}
\end{aligned} \tag{22}$$

$$\frac{GC(\varepsilon)\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2}{n} \leq \left\{\left(1 \vee \frac{\alpha^{1/\theta}}{4\epsilon}\right)\left(\frac{GC(\varepsilon)}{n}\right)^{\frac{2}{2-\theta}} + \epsilon(P(\widehat{\mathbf{w}}) - P_*)\right\} \tag{23}$$

$$\frac{2LC(\varepsilon)\varepsilon\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2}{n} \leq 2LC(\varepsilon)\left\{\left(1 \vee \frac{\alpha^{1/\theta}}{4\epsilon}\right)\left(\frac{\varepsilon}{n}\right)^{\frac{2}{2-\theta}} + \epsilon(P(\widehat{\mathbf{w}}) - P_*)\right\} \tag{24}$$

$$6L\varepsilon\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2 \leq 6L\left\{\left(1 \vee \frac{\alpha^{1/\theta}}{4\epsilon}\right)\varepsilon^{\frac{2}{2-\theta}} + \epsilon(P(\widehat{\mathbf{w}}) - P_*)\right\} \tag{25}$$

$$\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2 \sqrt{\frac{4LC(\varepsilon)P_*}{n}} \leq \left(1 \vee \frac{\alpha^{1/\theta}}{4\epsilon}\right)\left(\frac{4LC(\varepsilon)P_*}{n}\right)^{\frac{1}{2-\theta}} + \epsilon(P(\widehat{\mathbf{w}}) - P_*) \tag{26}$$

$$\|\widehat{\mathbf{w}} - \widehat{\mathbf{w}}^*\|_2 \sqrt{\frac{4LGC(\varepsilon)\varepsilon}{n}} \leq \left(1 \vee \frac{\alpha^{1/\theta}}{4\epsilon}\right)\left(\frac{4LGC(\varepsilon)\varepsilon}{n}\right)^{\frac{1}{2-\theta}} + \epsilon(P(\widehat{\mathbf{w}}) - P_*) \tag{27}$$

Plugging appropriate small constant values of ϵ in each inequality, we have

$$\begin{aligned}
\frac{P(\widehat{\mathbf{w}}) - P_*}{2} &\leq 17LC(\varepsilon)\left(\frac{1}{n}\right)^{\frac{2}{2-\theta}} + \left(1 \vee 4\alpha^{1/\theta}\right)\left(\frac{GC(\varepsilon)}{n}\right)^{\frac{2}{2-\theta}} + 2LC(\varepsilon)\left(1 \vee 64\alpha^{1/\theta}\right)\left(\frac{\varepsilon}{n}\right)^{\frac{2}{2-\theta}} \\
&\quad + 6L\left(1 \vee 64\alpha^{1/\theta}\right)\varepsilon^{\frac{2}{2-\theta}} + \left(1 \vee 64\alpha^{1/\theta}\right)\left(\frac{4LGC(\varepsilon)\varepsilon}{n}\right)^{\frac{1}{2-\theta}} + \left(1 \vee \frac{\alpha^{1/\theta}}{4\epsilon}\right)\left(\frac{4LC(\varepsilon)P_*}{n}\right)^{\frac{1}{2-\theta}}.
\end{aligned}$$

□

C.2 Proof of Inequality (21)

Proof. If $\|\widehat{\mathbf{w}}-\widehat{\mathbf{w}}^*\|_2^2 \leq (\frac{1}{n})^{\frac{\theta}{2-\theta}}$, then $\frac{LC(\varepsilon)\|\widehat{\mathbf{w}}-\widehat{\mathbf{w}}^*\|_2^2}{n} \leq LC(\varepsilon)(\frac{1}{n})^{\frac{2}{2-\theta}}$. If $\|\widehat{\mathbf{w}}-\widehat{\mathbf{w}}^*\|_2^2 \geq (\frac{1}{n})^{\frac{\theta}{2-\theta}}$, then

$$\frac{1}{\|\widehat{\mathbf{w}}-\widehat{\mathbf{w}}^*\|_2^{\frac{2}{\theta}-2}} \leq n^{\frac{1-\theta}{2-\theta}}, \tag{28}$$

so when $n \geq \left(LC(\varepsilon)\alpha^{\frac{1}{\theta}}/\epsilon\right)^{(2-\theta)}$, we have

$$\frac{LC(\varepsilon)\|\widehat{\mathbf{w}}-\widehat{\mathbf{w}}^*\|_2^2}{n} = \frac{LC(\varepsilon)\|\widehat{\mathbf{w}}-\widehat{\mathbf{w}}^*\|_2^{\frac{2}{\theta}}\|\widehat{\mathbf{w}}-\widehat{\mathbf{w}}^*\|_2^{2-\frac{2}{\theta}}}{n} \leq \frac{LC(\varepsilon)\alpha^{\frac{1}{\theta}}(P(\widehat{\mathbf{w}})-P_*)}{n^{\frac{1}{2-\theta}}} \leq \epsilon(P(\widehat{\mathbf{w}})-P_*),$$

where the first inequality holds by employing the EBC and the inequality (28), and the second inequality holds due to the fact that $n \geq \left(LC(\varepsilon)\alpha^{\frac{1}{\theta}}/\epsilon\right)^{(2-\theta)}$. Combining two cases together, we complete the proof. □

C.3 Proof of Inequality (22)

Proof. The first inequality in the inequality (22) obviously holds, and now we prove the second inequality.

- If $\|\widehat{\mathbf{w}}-\widehat{\mathbf{w}}^*\|_2^2 \leq 4(\frac{1}{n})^{\frac{\theta}{2-\theta}}$, then

$$\frac{LC(\varepsilon)\|\widehat{\mathbf{w}}-\widehat{\mathbf{w}}^*\|_2^2}{4\epsilon n} \leq \frac{LC(\varepsilon)}{\epsilon}(\frac{1}{n})^{\frac{2}{2-\theta}}.$$

- If $\|\widehat{\mathbf{w}}-\widehat{\mathbf{w}}^*\|_2^2 \geq 4(\frac{1}{n})^{\frac{\theta}{2-\theta}}$, then

$$\frac{1}{\|\widehat{\mathbf{w}}-\widehat{\mathbf{w}}^*\|_2^{2-\frac{2}{\theta}}} \geq \frac{1}{2^{2-\frac{2}{\theta}}}n^{\frac{\theta-1}{2-\theta}} \geq \frac{1}{4}n^{\frac{\theta-1}{2-\theta}}, \tag{29}$$

  so when $n \geq \left(LC(\varepsilon)\alpha^{\frac{1}{\theta}}/\epsilon^2\right)^{(2-\theta)}$, we have

$$\frac{LC(\varepsilon)\|\widehat{\mathbf{w}}-\widehat{\mathbf{w}}^*\|_2^2}{4\epsilon n} = \frac{LC(\varepsilon)\|\widehat{\mathbf{w}}-\widehat{\mathbf{w}}^*\|_2^{\frac{2}{\theta}}\|\widehat{\mathbf{w}}-\widehat{\mathbf{w}}_*\|_2^{2-\frac{2}{\theta}}}{4\epsilon n} \leq \frac{LC(\varepsilon)\alpha^{\frac{1}{\theta}}(P(\widehat{\mathbf{w}})-P_*)4n^{\frac{1-\theta}{2-\theta}}}{4\epsilon n}$$
$$\leq \epsilon(P(\widehat{\mathbf{w}})-P_*),$$

  where the first inequality holds by employing the EBC and the inequality (29), and the second inequality holds due to the fact that $n \geq \left(LC(\varepsilon)\alpha^{\frac{1}{\theta}}/\epsilon^2\right)^{(2-\theta)}$.

Combining two cases together, we complete the proof. □

C.4 Proof of Inequalities (23)–(27)

Proof. In Lemma 5, taking A to be

$$\left(\frac{GC(\varepsilon)}{n}\right)^2, \left(\frac{\varepsilon}{n}\right)^2, \varepsilon^2, \frac{4LC(\varepsilon)P_*}{n}, \frac{4LGC(\varepsilon)\varepsilon}{n}$$

yields inequalities (23)–(27) respectively. □

D Proof of Lemma 3

Lemma 6. *[40]. Let $\mathcal{H}$ be a Hilbert space and let ξ be a random variable with values in $\mathcal{H}$. Assume $\|\xi\| \leq G < \infty$ almost surely. Denote $\sigma^2(\xi) = \mathbb{E}\left[\|\xi\|^2\right]$. Let $\{\xi_i\}_{i=1}^m$ be m ($m < \infty$) independent drawers of ξ. For any $0 < \delta < 1$, with confidence $1 - \delta$,*

$$\left\| \frac{1}{m} \sum_{i=1}^{m} [\xi_i - \mathbb{E}[\xi_i]] \right\| \leq \frac{2G \log(2/\delta)}{m} + \sqrt{\frac{2\sigma^2(\xi) \log(2/\delta)}{m}}.$$

Proof of Lemma 3. In order to prove the high probability bounds for all $\mathbf{w} \in \mathcal{W}$, we first consider the points in the ε-net of $\mathcal{W}$ with minimal cardinality. To this end, let $\mathcal{N}(\mathcal{W}, \varepsilon)$ denote the ε-net of $\mathcal{W}$ with minimal cardinality. Since $\mathcal{W} \subseteq \mathcal{B}^d(R)$, where $\mathcal{B}^d(R)$ denotes a d-dimentional bounded ball with radius R. Following the standard results of covering numbers, we have

$$\log |\mathcal{N}(\mathcal{W}, \varepsilon)| \leq \log |\mathcal{N}(\mathcal{B}^d(R), \varepsilon/2)| \leq d \log \frac{6R}{\epsilon}.$$

We first consider a fixed $\mathbf{w} \in \mathcal{N}(\mathcal{W}, \varepsilon)$. Denote by $\mathbf{w}^*$ the closest optimal solution to $\mathbf{w}$. Let $f_i(\mathbf{w}) = f(\mathbf{w}, \mathbf{z}_i)$. Since $f_i(\cdot)$ is L-smooth, we have

$$\|\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}^*)\|_2 \leq L \|\mathbf{w} - \mathbf{w}^*\|_2. \tag{30}$$

Because $f_i(\cdot)$ is both convex and L-smooth, by (2.1.7) of [30], we have

$$\|\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}^*)\|_2^2 \leq L \left(f_i(\mathbf{w}) - f_i(\mathbf{w}^*) - \langle \nabla f_i(\mathbf{w}_*), \mathbf{w} - \mathbf{w}^* \rangle \right).$$

Taking expectation over both sides, we have

$$\mathbb{E}\left[\|\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}^*)\|_2^2 \right] \leq L \left(F(\mathbf{w}) - F(\mathbf{w}^*) - \langle \nabla F(\mathbf{w}^*), \mathbf{w} - \mathbf{w}^* \rangle \right) \leq L \left(P(\mathbf{w}) - P(\mathbf{w}^*) \right),$$

where the last inequality follows from the optimality condition of $\mathbf{w}^*$, i.e., there exists $\mathbf{v}_* \in \partial R(\mathbf{w}^*)$

$$\langle \nabla F(\mathbf{w}^*) + \mathbf{v}_*, \mathbf{w} - \mathbf{w}_* \rangle \geq 0, \ \forall \mathbf{w} \in \mathcal{W}.$$

and the convexity of $R(\mathbf{w})$ and $F(\mathbf{w})$, i.e., $\langle \nabla F(\mathbf{w}^*), \mathbf{w} - \mathbf{w}^* \rangle \leq F(\mathbf{w}) - F(\mathbf{w}^*)$ and $\langle \mathbf{v}_*, \mathbf{w} - \mathbf{w}^* \rangle \leq R(\mathbf{w}) - R(\mathbf{w}^*)$.

Following Lemma 6, with probability at least $1 - \delta$, we have

$$\left\| \nabla F(\mathbf{w}) - \nabla F(\mathbf{w}^*) - \frac{1}{n} \sum_{i=1}^{n} [\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}^*)] \right\|_2 \leq \frac{2L \|\mathbf{w} - \mathbf{w}^*\|_2 \log(2/\delta)}{n} + \sqrt{\frac{2L(P(\mathbf{w}) - P(\mathbf{w}^*)) \log(2/\delta)}{n}}.$$

By taking the union bound over $\mathcal{N}(\mathcal{W}, \varepsilon)$, we have for any $\mathbf{w} \in \mathcal{N}(\mathcal{W}, \varepsilon)$, with probability $1 - \delta$,

$$\|\nabla P(\mathbf{w}) - \nabla P(\mathbf{w}^*) - [\nabla P_n(\mathbf{w}) - \nabla P_n(\mathbf{w}^*)]\|_2 = \left\| \nabla F(\mathbf{w}) - \nabla F(\mathbf{w}^*) - \frac{1}{n} \sum_{i=1}^{n} [\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}^*)] \right\|_2$$
$$\leq \frac{2L \|\mathbf{w} - \mathbf{w}_*\|_2 (\log(2/\delta) + d \log(6R/\varepsilon))}{n} + \sqrt{\frac{2L(P(\mathbf{w}) - P(\mathbf{w}^*))(\log(2/\delta) + d \log(6R/\varepsilon))}{n}}.$$

To finish the proof of Lemma 3, for any $\mathbf{w} \in \mathcal{W}$. There exists $\widetilde{\mathbf{w}} \in \mathcal{N}(\mathcal{W}, \varepsilon)$ such that $\|\mathbf{w} - \widetilde{\mathbf{w}}\| \leq \varepsilon$. Let $\widetilde{\mathbf{w}}^*$ denote the closest optimal solution to $\widetilde{\mathbf{w}}$. Then by non-expansiveness of projection onto a convex set we have $\|\mathbf{w}^* - \widetilde{\mathbf{w}}^*\|_2 \leq \|\mathbf{w} - \widetilde{\mathbf{w}}\|_2 \leq \varepsilon$. In addition, we have

$$\|\widetilde{\mathbf{w}} - \widetilde{\mathbf{w}}^*\|_2 \leq \|\widetilde{\mathbf{w}} - \mathbf{w}\|_2 + \|\mathbf{w} - \mathbf{w}^*\|_2 + \|\mathbf{w}^* - \widetilde{\mathbf{w}}^*\|_2 \leq 2\varepsilon + \|\mathbf{w} - \mathbf{w}^*\|_2 \tag{31}$$

$$\begin{aligned} P(\widetilde{\mathbf{w}}) - P(\widetilde{\mathbf{w}}^*) &\leq P(\widetilde{\mathbf{w}}) - P(\mathbf{w}) + P(\mathbf{w}) - P(\mathbf{w}^*) + P(\mathbf{w}^*) - P(\widetilde{\mathbf{w}}^*) \\ &\leq \bar{G} \|\widetilde{\mathbf{w}} - \mathbf{w}\|_2 + P(\mathbf{w}) - P(\mathbf{w}^*) + \bar{G} \|\mathbf{w}^* - \widetilde{\mathbf{w}}^*\|_2 \\ &\leq 2\bar{G}\varepsilon + P(\mathbf{w}) - P(\mathbf{w}^*) \end{aligned} \tag{32}$$

Then with probability $1-\delta$, we have

$$\begin{aligned}
&\|\nabla P(\mathbf{w}) - \nabla P(\mathbf{w}^*) - [\nabla P_n(\mathbf{w}) - \nabla P_n(\mathbf{w}^*)]\|_2 \\
&\leq \|\nabla P(\widetilde{\mathbf{w}}) - \nabla P(\widetilde{\mathbf{w}}^*) - [\nabla P_n(\widetilde{\mathbf{w}}) - \nabla P_n(\widetilde{\mathbf{w}}^*)]\|_2 + 2L\|\mathbf{w} - \widetilde{\mathbf{w}}\|_2 + 2L\|\mathbf{w}^* - \widetilde{\mathbf{w}}^*\|_2 \\
&\leq \frac{2L\|\widetilde{\mathbf{w}} - \widetilde{\mathbf{w}}^*\|_2(\log(2/\delta) + 2d\log(6R/\varepsilon))}{n} + \sqrt{\frac{2L(P(\widetilde{\mathbf{w}}) - P(\widetilde{\mathbf{w}}^*))(\log(2/\delta) + 2d\log(6R/\varepsilon))}{n}} + 4L\varepsilon \\
&\leq \frac{2L(\|\mathbf{w} - \mathbf{w}^*\|_2 + 2\varepsilon)(\log(2/\delta) + 2d\log(6R/\varepsilon))}{n} + \sqrt{\frac{2L(2\bar{G}\varepsilon + (P(\mathbf{w}) - P(\mathbf{w}^*)))(\log(2/\delta) + 2d\log(6R/\varepsilon))}{n}} + \\
&4L\varepsilon \\
&\leq \frac{LC(\varepsilon)\|\mathbf{w} - \mathbf{w}^*\|_2}{n} + \frac{2LC(\varepsilon)\varepsilon}{n} + \sqrt{\frac{LC(\varepsilon)(P(\mathbf{w}) - P_*)}{n}} + 2\sqrt{\frac{L\bar{G}C(\varepsilon)\varepsilon}{n}} + 4L\varepsilon.
\end{aligned}$$

□

E Proof of Lemma 4

Proof. We first consider a fixed $\mathbf{w}_* \in \mathcal{N}(\mathcal{W}_*, \varepsilon) \subseteq \mathcal{W}_*$. To apply Lemma 6, we need an upper bound of $\mathbb{E}\left[\|\nabla f_i(\mathbf{w}_*)\|_2^2\right]$. Since $f_i(\cdot)$ is L-smooth and nonnegative, from Lemma 4.1 of [41], we have

$$\|\nabla f_i(\mathbf{w}_*)\|_2^2 \leq 4Lf_i(\mathbf{w}_*)$$

and thus

$$\mathbb{E}\left[\|\nabla f_i(\mathbf{w}_*)\|_2^2\right] \leq 4L\mathbb{E}\left[f_i(\mathbf{w}_*)\right] = 4LF_*.$$

By **Assumption 1**, we have $\|\nabla f_i(\mathbf{w}_*)\|_2 \leq G$. Then, according to Lemma 6, with probability at least $1-\delta$, we have

$$\|\nabla F(\mathbf{w}_*) - \nabla F_n(\mathbf{w}_*)\|_2 = \left\|\nabla F(\mathbf{w}_*) - \frac{1}{n}\sum_{i=1}^n \nabla f_i(\mathbf{w}_*)\right\|_2 \leq \frac{2G\log(2/\delta)}{n} + \sqrt{\frac{8LF_*\log(2/\delta)}{n}}.$$

By taking the union bound over $\mathcal{N}(\mathcal{W}_*, \varepsilon)$, for any $\mathbf{w}_* \in \mathcal{N}(\mathcal{W}_*, \varepsilon)$, with probability $1-\delta$ we have

$$\|\nabla F(\mathbf{w}_*) - \nabla F_n(\mathbf{w}_*)\|_2 \leq \frac{GC(\varepsilon)}{n} + \sqrt{\frac{4LF_*C(\varepsilon)}{n}}.$$

For any $\mathbf{w}^* \in \mathcal{W}_*$, there exists $\widetilde{\mathbf{w}}^* \in \mathcal{N}(\mathcal{W}_*, \varepsilon)$ such that $\|\mathbf{w}^* - \widetilde{\mathbf{w}}^*\| \leq \varepsilon$. Then

$$\begin{aligned}
\|\nabla F(\mathbf{w}^*) - \nabla F_n(\mathbf{w}^*)\|_2 &\leq \|\nabla F(\widetilde{\mathbf{w}}^*) - \nabla F_n(\widetilde{\mathbf{w}}^*)\|_2 + \|\nabla F(\mathbf{w}^*) - \nabla F(\widetilde{\mathbf{w}}^*)\|_2 \\
&\quad + \|\nabla F_n(\mathbf{w}^*) - \nabla F_n(\widetilde{\mathbf{w}}^*)\|_2 \\
&\leq \frac{GC(\varepsilon)}{n} + \sqrt{\frac{4LF_*C(\varepsilon)}{n}} + 2L\varepsilon.
\end{aligned}$$

□

F Proof of Lemma 5

Proof. We consider two cases. First, $\|\mathbf{w} - \mathbf{w}^*\|_2 \leq A^{\frac{\theta}{4-2\theta}}$, under which the inequality follows trivially. Next, we consider $\|\mathbf{w} - \widehat{\mathbf{w}}^*\|_2 \geq A^{\frac{\theta}{4-2\theta}}$. Then

$$\begin{aligned}
&\|\mathbf{w} - \mathbf{w}^*\|_2\sqrt{A} = \frac{\|\mathbf{w} - \mathbf{w}^*\|_2^{1/\theta}}{\|\mathbf{w} - \mathbf{w}^*\|_2^{1/\theta - 1}}\sqrt{A} \leq \|\mathbf{w} - \mathbf{w}^*\|_2^{1/\theta} A^{\frac{1}{2(2-\theta)}} \leq \frac{\epsilon\|\mathbf{w} - \mathbf{w}^*\|_2^{2/\theta}}{\alpha^{1/\theta}} + \frac{\alpha^{1/\theta}}{4\epsilon}A^{\frac{1}{2-\theta}} \\
&\leq \epsilon(P(\mathbf{w}) - P_*) + \frac{\alpha^{1/\theta}}{4\epsilon}A^{\frac{1}{2-\theta}}
\end{aligned}$$

where the last inequality follows the EBC. □

G Proof of Theorem 3

Before proceeding to the proof, we first present a standard result for SSG, which is the Lemma 10 of [14].

Proposition 1. *Suppose* **Assumptions 1 and 2** *hold. Let* $0 < \delta < 1$, $\mathbf{w}^* \in \mathcal{W}_*$ *be the closest optimal solution to* $\mathbf{w}_1$*, and* R_0 *be an upper bound on* $\|\mathbf{w}_1 - \mathbf{w}^*\|_2$*. Apply T iterations of the update* $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}\cap\mathcal{B}(\mathbf{w}_1,R_0)}(\mathbf{w}_t - \gamma g_t)$*, where* g_t *is a stochastic subgradient of* $P(\mathbf{w})$ *at* $\mathbf{w}_t$*. With probability at least* $1-\delta$*, we have*

$$P(\widehat{\mathbf{w}}_T) - P_* \leq \frac{\gamma G^2}{2} + \frac{\|\mathbf{w}_1 - \mathbf{w}^*\|_2^2}{2\gamma(T+1)} + \frac{4GR_0\sqrt{2\log(2/\delta)}}{\sqrt{T+1}}.$$

where $\widehat{\mathbf{w}}_T = \frac{1}{T+1}\sum_{t=1}^{T+1}\mathbf{w}_t$*. Moreover, choose* $\gamma = \frac{R_0}{G\sqrt{T+1}}$*, and then with probability at least* $1-\delta$*,*

$$P(\widehat{\mathbf{w}}_T) - P_* \leq R_0 G\left(\frac{1}{\sqrt{T+1}} + \frac{4\sqrt{2\log(2/\delta)}}{\sqrt{T+1}}\right).$$

It is easy to derive a similar lemma as Proposition 1, which is stated in Lemma 7.

Lemma 7. *Suppose* **Assumptions 1 and 2** *hold. Let* $0 < \delta < 1$*,* R_0 *be any nonnegative real number. Apply T iterations of the update* $\mathbf{w}_{t+1} = \Pi_{\mathcal{W}\cap\mathcal{B}(\mathbf{w}_1,R_0)}(\mathbf{w}_t - \gamma g_t)$*, where* g_t *is a stochastic subgradient of* $P(\mathbf{w})$ *at* $\mathbf{w}_t$*. With probablity at least* $1-\delta$*, we have*

$$P(\widehat{\mathbf{w}}_T) - P(\mathbf{w}_1) \leq \frac{\gamma G^2}{2} + \frac{4GR_0\sqrt{2\log(2/\delta)}}{\sqrt{T+1}},$$

where $\widehat{\mathbf{w}}_T = \frac{1}{T+1}\sum_{t=1}^{T+1}\mathbf{w}_t$*. Moreover, choose* $\gamma = \frac{R_0}{G\sqrt{T+1}}$*, and then with probability at least* $1-\delta$*,*

$$P(\widehat{\mathbf{w}}_T) - P(\mathbf{w}_1) \leq R_0 G\left(\frac{1}{\sqrt{T+1}} + \frac{4\sqrt{2\log(2/\delta)}}{\sqrt{T+1}}\right).$$

Proof. Denote $\mathbb{E}_{t-1}(X)$ by the expectation conditioned on the randomness until round $t-1$, then we have $\mathbb{E}_{t-1}(\hat{g}_t) = g_t$, and $X_t = g_t(\mathbf{w}_t - \mathbf{w}_1) - \hat{g}_t(\mathbf{w}_t - \mathbf{w}_1)$ is a martingale difference sequence. Note that $\|g_t\|_2 = \|\mathbb{E}_{t-1}(\hat{g}_t)\|_2 \leq \mathbb{E}_{t-1}(\|\hat{g}_t\|_2) \leq G$, so we have

$$|X_t| \leq \|g_t\|_2\|\mathbf{w}_t - \mathbf{w}_1\|_2 + \|\hat{g}_t\|_2\|\mathbf{w}_t - \mathbf{w}_1\|_2 \leq 4GR_0,$$

since the update needs to project the gradient update onto the intersection of $\mathcal{W}$ and a ball with radius R_0.

By Azuma-Hoeffding's inequality, we have with probability at least $1-\delta$,

$$\frac{1}{T+1}\sum_{t=1}^{T+1} g_t(\mathbf{w}_t - \mathbf{w}_1) - \frac{1}{T+1}\sum_{t=1}^{T}\hat{g}_t(\mathbf{w}_t - \mathbf{w}_1) \leq \frac{4GR_0\sqrt{2\log(1/\delta)}}{\sqrt{T+1}}. \tag{33}$$

By the convexity of P, we have $P(\mathbf{w}_t) - P(\mathbf{w}_1) \leq g_t(\mathbf{w}_t - \mathbf{w}_1)$, then using a standard result in online gradient descent [56], we have

$$\frac{1}{T+1}\sum_{t=1}^{T}\hat{g}_t(\mathbf{w}_t - \mathbf{w}_1) \leq \frac{\gamma G^2}{2} + \frac{\|\mathbf{w}_1 - \mathbf{w}_1\|_2^2}{2\gamma(T+1)} = \frac{\gamma G^2}{2}. \tag{34}$$

Combining inequality (33) and (34) suffices to derive the conclusion. □

With the above proposition and lemma, the proof of Theorem 3 proceeds similarly as that of Theorem 5.3 in [15]. The difference is that our analysis only relies on the EBC instead of the uniform convexity.

Proof. Define $\bar{\delta} = \frac{2\delta}{\log_2 n}$, and

$$a(n,\bar{\delta}) = G\left(\frac{1}{\sqrt{n+1}} + \frac{4\sqrt{2\log(2/\bar{\delta})}}{\sqrt{n+1}}\right).$$

We set $\mu_0 = 2R_0^{1-\frac{2}{\theta}} a(n_0, \bar{\delta})$, $\mu_k = 2^{(\frac{2}{\theta}-1)k}\mu_0$ and $R_k = R_0/2^k$, where $k = 1, \ldots, m$. Then we have $\mu_k R_k^{\frac{2}{\theta}} = 2^{-k}\mu_0 R_0^{\frac{2}{\theta}}$. We can also assume that α is large enough such that $\alpha \geq R_0^{2-\theta}/G^\theta$, i.e., $\alpha^{-\frac{1}{\theta}} \leq GR_0^{1-\frac{2}{\theta}}$, otherwise we can set $\alpha = R_0^{2-\theta}/G^\theta$, which makes the EBC still hold.

By definition of m, when $n \geq 100$,

$$0 < \frac{1}{2}\log_2 \frac{2n}{\log_2 n} - 2 \leq m \leq \frac{1}{2}\log_2 \frac{2n}{\log_2 n} - 1 \leq \frac{1}{2}\log_2 n, \tag{35}$$

so we have

$$2^m \geq \frac{1}{4}\sqrt{\frac{2n}{\log_2 n}}. \tag{36}$$

When $n \geq 100$, we have

$$\begin{aligned}
\mu_m &= 2^{(\frac{2}{\theta}-1)m}\mu_0 \geq 2^m \mu_0 \\
&\geq \frac{1}{4}\sqrt{\frac{2n}{\log_2 n}} 4GR_0^{1-\frac{2}{\theta}} \left(\frac{1}{2\sqrt{n_0+1}} + \frac{2\sqrt{2\log(\log_2 n)}}{\sqrt{n_0+1}}\right) \\
&\geq GR_0^{1-\frac{2}{\theta}}\sqrt{\frac{2n}{\log_2 n}} \left(\frac{1}{2\sqrt{\frac{n}{m}+1}} + \frac{2\sqrt{2\log(\log_2 n)}}{\sqrt{\frac{n}{m}+1}}\right) \\
&\geq GR_0^{1-\frac{2}{\theta}}\sqrt{\frac{2n}{\log_2 n}} \left(\frac{1}{2\sqrt{\frac{2n}{\log_2 2n - \log_2\log_2 n - 4}+1}} + \frac{2\sqrt{2\log(\log_2 n)}}{\sqrt{\frac{2n}{\log_2 2n - \log_2\log_2 n - 4}+1}}\right) \\
&\geq GR_0^{1-\frac{2}{\theta}}\sqrt{\frac{2n}{\log_2 n}} \frac{2\sqrt{\sqrt{2\log(\log_2 n)}}}{\sqrt{\frac{2n}{\log_2 2n - \log_2\log_2 n - 4}+1}} \\
&= GR_0^{1-\frac{2}{\theta}} \frac{2\sqrt{\sqrt{2\log(\log_2 n)}}}{\sqrt{\frac{1}{1-\frac{\log_2\log_2 n+3}{\log_2 n}} + \frac{\log_2 n}{2n}}} \geq GR_0^{1-\frac{2}{\theta}},
\end{aligned}$$

where the first inequality holds because $\theta \in (0, 1]$, the second inequality comes from (36) and the fact that $0 < \delta < 1$, the third and fourth inequalities hold because of the definition of n_0 and inequality (35), the fifth inequality holds by utilizing $a+b \geq 2\sqrt{ab}$, and the sixth inequality holds since $n \geq 100$ and the function is monotonically increasing with respect to n. So $\alpha^{-\frac{1}{\theta}} \leq \mu_m$.

Below, given $\widehat{\mathbf{w}}_k$ we denote by $\widehat{\mathbf{w}}_k^*$ the closest optimal solution to $\widehat{\mathbf{w}}_k$. Next, we consider two cases.

Case 1. If $\alpha^{-\frac{1}{\theta}} \geq \mu_0$, then $\mu_0 \leq \alpha^{-\frac{1}{\theta}} \leq \mu_m$. We have the following lemma.

Lemma 8. *Let k^* satisfy $\mu_{k^*} \leq \alpha^{-\frac{1}{\theta}} \leq 2^{\frac{2}{\theta}-1}\mu_{k^*}$. Then for any $1 \leq k \leq k^*$, there exists a Borel set $\mathcal{A}_k \subset \Omega$ of probability at least $1 - k\bar{\delta}$, such that for $\omega \in \mathcal{A}_k$, the points $\{\widehat{\mathbf{w}}_k\}_{k=1}^m$ generated by the Algorithm 2 satisfy*

$$\|\widehat{\mathbf{w}}_{k-1} - \widehat{\mathbf{w}}_{k-1}^*\|_2 \leq R_{k-1} = 2^{-k+1}R_0, \tag{37}$$

$$P(\widehat{\mathbf{w}}_k) - P_* \leq \mu_k R_k^{\frac{2}{\theta}} = 2^{-k}\mu_0 R_0^{\frac{2}{\theta}}. \tag{38}$$

Moreover, for $k > k^*$ there is a Borel set $\mathcal{C}_k \subset \Omega$ of probability at least $1 - (k-k^*)\bar{\delta}$ such that on $\mathcal{C}_k$, we have

$$P(\widehat{\mathbf{w}}_k) - P(\widehat{\mathbf{w}}_{k^*}) \leq \mu_{k^*} R_{k^*}^{\frac{2}{\theta}}. \tag{39}$$

Proof. (Proof of Lemma 8) We prove (37) and (38) by induction. Note that (37) holds for $k = 1$. Assume it is true for some $k > 1$ on $\mathcal{A}_{k-1}$. According to the Proposition 1, there exists a Borel set

$\mathcal{B}_k$ with $\Pr(\mathcal{B}_k) \geq 1 - \bar{\delta}$ such that

$$P(\widehat{\mathbf{w}}_k) - P_* \leq R_{k-1} G \left(\frac{1}{\sqrt{n_0+1}} + \frac{4\sqrt{2\log(2/\bar{\delta})}}{\sqrt{n_0+1}} \right) = R_{k-1} a(n_0, \bar{\delta})$$
$$= \frac{1}{2} \mu_k 2^{(1-\frac{2}{\theta})k} R_0^{\frac{2}{\theta}-1} R_{k-1} = \mu_k R_k^{\frac{2}{\theta}},$$

which is (38). By the inductive hypothesis, $\|\widehat{\mathbf{w}}_{k-1} - \mathbf{w}_{k-1}^*\|_2 \leq R_{k-1}$ on the set $\mathcal{A}_{k-1}$. Define $\mathcal{A}_k = \mathcal{A}_{k-1} \cap \mathcal{B}_k$. Note that

$$\Pr(\mathcal{A}_k) \geq \Pr(\mathcal{A}_{k-1}) + \Pr(\mathcal{B}_k) - 1 \geq 1 - k\bar{\delta},$$

and on $\mathcal{A}_k$, by the EBC and the definition of k^*, we have

$$\|\widehat{\mathbf{w}}_k - \widehat{\mathbf{w}}_k^*\|_2^{\frac{2}{\theta}} \leq \alpha^{\frac{1}{\theta}} (P(\widehat{\mathbf{w}}_k) - P_*) \leq \frac{P(\widehat{\mathbf{w}}_k) - P_*}{\mu_{k^*}} \leq \frac{\mu_k R_k^{\frac{2}{\theta}}}{\mu_{k^*}} \leq R_k^{\frac{2}{\theta}},$$

which is (37) for $k+1$.

Now we prove (39). For $k > k^*$, by Lemma 7, there exists a Borel set $\mathcal{B}_k$ with $\Pr(\mathcal{B}_k) \geq 1 - \bar{\delta}$ such that

$$P(\widehat{\mathbf{w}}_k) - P(\widehat{\mathbf{w}}_{k-1}) \leq \frac{\gamma_k G^2}{2} + \frac{4G R_{k-1}\sqrt{2\log(2/\delta)}}{\sqrt{n_0+1}} \leq R_{k-1} a(n_0, \bar{\delta}) = 2^{k^*-k} R_{k^*-1} a(n_0, \bar{\delta})$$
$$= 2^{k^*-k} \mu_{k^*} R_{k^*}^{\frac{2}{\theta}} = \mu_k R_k^{\frac{2}{\theta}},$$

which implies that on $\mathcal{C}_k = \cap_{j=k^*+1}^k \mathcal{B}_j$, we have

$$P(\widehat{\mathbf{w}}_k) - P(\widehat{\mathbf{w}}_{k^*}) = \sum_{j=k^*+1}^{k} (P(\widehat{\mathbf{w}}_j) - P(\widehat{\mathbf{w}}_{j-1})) \leq \sum_{j=k^*+1}^{k} 2^{k^*-j} \mu_{k^*} R_{k^*}^{\frac{2}{\theta}} \leq \mu_{k^*} R_{k^*}^{\frac{2}{\theta}}.$$

By union bound, we have $\Pr(\cap_{j=k^*+1}^k \mathcal{B}_j) \geq 1 - (k-k^*)\bar{\delta}$. Here completes the proof.

□

Now we proceed the proof as follows. Note that $\mu_0 \leq \alpha^{-\frac{1}{\theta}} \leq \mu_m$. At the end of k^*-th stage, on the Borel set $\mathcal{A}_{k^*}$ of probability at least $1 - k^*\bar{\delta}$, we have

$$P(\widehat{\mathbf{w}}_{k^*}) - P_* \leq \mu_{k^*} R_{k^*}^{\frac{2}{\theta}}.$$

Then on the Borel set $\mathcal{D}_m = \mathcal{C}_m \cap \mathcal{A}_{k^*} = (\cap_{j=k^*+1}^m \mathcal{B}_j) \cap A_{k^*}$ with $\Pr(\mathcal{D}_m) \geq 1 - m\bar{\delta}$, we have

$$P(\widehat{\mathbf{w}}_m) - P_* = P(\widehat{\mathbf{w}}_m) - P(\widehat{\mathbf{w}}_{k^*}) + (P(\widehat{\mathbf{w}}_{k^*}) - P_*) \leq 2\mu_{k^*} R_{k^*}^{\frac{2}{\theta}} \leq 4(\frac{\mu_{k^*}}{\alpha^{-\frac{1}{\theta}}})^{\frac{1}{\frac{2}{\theta}-1}} \mu_{k^*} R_{k^*}^{\frac{2}{\theta}}$$
$$= 4\left(\frac{2^{(\frac{2}{\theta}-1)k^*} \mu_0}{\alpha^{-\frac{1}{\theta}}} \right)^{\frac{1}{\frac{2}{\theta}-1}} \mu_{k^*} R_{k^*}^{\frac{2}{\theta}} = 4(2^{k^*} \mu_{k^*} R_{k^*}^{\frac{2}{\theta}} \mu_0^{\frac{\theta}{2-\theta}} \alpha^{\frac{1}{2-\theta}})$$
$$= 4(\mu_0 R_0^{\frac{2}{\theta}} \mu_0^{\frac{\theta}{2-\theta}} \alpha^{\frac{1}{2-\theta}}) = 4[(2R_0^{1-\frac{2}{\theta}} a(n_0, \bar{\delta}))^{\frac{2}{2-\theta}} R_0^{\frac{2}{\theta}} \alpha^{\frac{1}{2-\theta}}] = 4(2\sqrt{\alpha} \cdot a(n_0, \bar{\delta}))^{\frac{2}{2-\theta}}$$
$$= (2^{2-\theta} 2\sqrt{\alpha} \cdot a(n_0, \bar{\delta}))^{\frac{2}{2-\theta}}.$$

By the definition of m and $\bar{\delta}$, and the fact that $m \leq \frac{1}{2}\log_2 n$, we have $m\bar{\delta} \leq \delta$. So $\Pr(\mathcal{D}_m) \geq 1 - \delta$.

Case 2. If $\alpha^{-\frac{1}{\theta}} < \mu_0$, then on $\mathcal{A}_1 = \mathcal{B}_1$,

$$P(\widehat{\mathbf{w}}_1) - P_* \leq R_0 \cdot a(n_0, \bar{\delta}) = \frac{R_0}{a(n_0, \bar{\delta})^{\frac{\theta}{2-\theta}}} \cdot a(n_0, \bar{\delta})^{\frac{2}{2-\theta}}$$
$$= \frac{2^{\frac{\theta}{2-\theta}}}{\mu_0^{\frac{\theta}{2-\theta}}} a(n_0, \bar{\delta})^{\frac{2}{2-\theta}} \leq 2^{\frac{\theta}{2-\theta}} \left(\sqrt{\alpha} \cdot a(n_0, \bar{\delta}) \right)^{\frac{2}{2-\theta}}.$$

Hence on $\mathcal{A}_1 \cap \mathcal{C}_m$, by a similar argument as in case 1, we have

$$P(\widehat{\mathbf{w}}_m) - P_* = P(\widehat{\mathbf{w}}_m) - P(\widehat{\mathbf{w}}_1) + P(\widehat{\mathbf{w}}_1) - P_* \leq 2R_0 \cdot a(n_0, \bar{\delta}) \leq (2\sqrt{\alpha} \cdot a(n_0, \bar{\delta}))^{\frac{2}{2-\theta}},$$

where $\Pr(\mathcal{A}_1 \cap \mathcal{C}_m) \geq 1 - \delta$.

Combining the two cases, we have with probability at least $1 - \delta$,

$$\begin{aligned}
&P(\widehat{\mathbf{w}}_m) - P_* \\
&\leq (8\sqrt{\alpha} \vee 2\sqrt{\alpha})^{\frac{2}{2-\theta}} \left(G\left(\frac{1}{\sqrt{n_0+1}} + \frac{4\sqrt{2\log(2/\bar{\delta})}}{\sqrt{n_0+1}} \right) \right)^{\frac{2}{2-\theta}} \\
&\leq (64\alpha)^{\frac{1}{2-\theta}} \left(\frac{G\left(1 + 4\sqrt{2\log(\frac{\log_2 n}{\delta})}\right)}{\sqrt{\frac{n}{\frac{1}{2}\log_2 n}}} \right)^{\frac{2}{2-\theta}} = \left(\frac{128\alpha G^2 \log_2 n \left(1 + 4\sqrt{2\log(\frac{\log_2 n}{\delta})}\right)^2}{n} \right)^{\frac{1}{2-\theta}},
\end{aligned}$$

where the second inequality stems from the fact that $n_0 + 1 \geq \frac{n}{m} \geq \frac{n}{\frac{1}{2}\log_2 n}$. □

H Detailed Analysis of Examples Satisfying EBC

Risk Minimization Problems over an ℓ_2 ball.

Lemma 9. *Consider the following problem*

$$\min_{\|\mathbf{w}\|_2 \leq B} P(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{z}}[f(\mathbf{w}, \mathbf{z})] \tag{40}$$

If $\min_{\mathbf{w}\in\mathbb{R}^d} P(\mathbf{w}) < \min_{\|\mathbf{w}\|_2 \leq B} P(\mathbf{w})$*, then the above problem satisfies EBC*$(\theta = 1, \alpha)$.

Proof. The proof is similar to that of Theorem 3.5 of [24]. Denote $\mathbf{w}_*$ by an optimal solution of Example 4. Let $\Omega = \{\mathbf{w} \in \mathbb{R}^d \mid \|\mathbf{w}\|_2 \leq B\}$, and $F(\mathbf{w}) = P(\mathbf{w}) + I_\Omega(\mathbf{w})$, where $I_\Omega(\mathbf{w}) = 0$ if $\mathbf{w} \in \Omega$, and otherwise $I_\Omega(\mathbf{w}) = +\infty$. Then we have $\arg\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w}) = \arg\min_{\|\mathbf{w}\|_2 \leq B} P(\mathbf{w})$. Let $\mathbf{w}_* \in \arg\min_{\mathbf{w}\in\mathbb{R}^d} F(\mathbf{w})$ denote an optimal solution.

Since $B > 0$, so the optimization problem is strictly feasible, then by the Lagrangian theory, there exists some $\lambda \geq 0$, such that

$$\begin{aligned}
F(\mathbf{w}_*) &= \min_{\|\mathbf{w}\|_2 \leq B} P(\mathbf{w}) = \min_{\mathbf{w}\in\mathbb{R}^d} (P(\mathbf{w}) + \lambda(\|\mathbf{w}\|_2^2 - B^2)) \\
&= P(\mathbf{w}_*) + \lambda(\|\mathbf{w}_*\|_2^2 - B^2).
\end{aligned}$$

Note that $\min_{\mathbf{w}\in\mathbb{R}^d} P(\mathbf{w}) < \min_{\|\mathbf{w}\|_2 \leq B} P(\mathbf{w})$, as a result $\lambda > 0$. Then by complementary slackness, we know that $\|\mathbf{w}_*\|_2 = B$. Denote by $P_\lambda(\mathbf{w}) = P(\mathbf{w}) + \lambda(\|\mathbf{w}\|_2^2 - B^2)$. Then according to Theorem 28.1 [34], we have

$$\mathbf{w}_* \in \arg\min F = \{\mathbf{w} \mid \|\mathbf{w}\|_2 = B\} \cap \arg\min_{\mathbf{w}\in\mathbb{R}^d} P_\lambda(\mathbf{w}). \tag{41}$$

Since $P_\lambda(\mathbf{w})$ is strongly convex due to $\lambda > 0$, its optimal solution is unique. As a result,

$$\mathbf{w}_* = \arg\min F = \arg\min_{\mathbf{w}\in\mathbb{R}^d} P_\lambda(\mathbf{w}). \tag{42}$$

In addition, there exists $\mu > 0$ such that (due to the strong convexity of $P_\lambda(\mathbf{w})$),

$$\begin{aligned}
&\|\mathbf{w} - \arg\min P_\lambda(\mathbf{w})\|_2 \leq \mu(P_\lambda(\mathbf{w}) - \min_{\mathbf{w}} P_\lambda(\mathbf{w}))^{1/2} \\
&= \mu(P(\mathbf{w}) + \lambda(\|\mathbf{w}\|_2^2 - B^2) - P(\mathbf{w}_*))^{1/2} \\
&\leq \mu(P(\mathbf{w}) - P(\mathbf{w}_*))^{1/2}.
\end{aligned}$$

Then according to (42), we know that

$$\|\mathbf{w} - \mathbf{w}_*\|_2^2 \leq \mu^2(P(\mathbf{w}) - P(\mathbf{w}_*)),$$

which is EBC$(\theta = 1, \mu^2)$. □

Quadratic Problems.

Lemma 10. *Consider the following problem*

$$\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w}) \triangleq \mathbf{w}^\top \mathbb{E}_{\mathbf{z}}[A(\mathbf{z})]\mathbf{w} + \mathbf{w}^\top \mathbb{E}_{\mathbf{z}'}[\mathbf{b}(\mathbf{z}')] + c \tag{43}$$

If $\mathbb{E}_{\mathbf{z}}[A(\mathbf{z})]$ is PSD and $\mathcal{W}$ is a bounded polyhedron, then the above problem satisfies EBC($\theta = 1, \alpha$).

Proof. Let us consider $\mathbb{E}_{\mathbf{z}}[A(\mathbf{z})] \neq 0$; otherwise it reduces to PLP.

Note that $\mathbb{E}_{\mathbf{z}}[A(\mathbf{z})]$ is PSD, so there exists a nonzero matrix A such that $\mathbb{E}_{\mathbf{z}}[A(\mathbf{z})] = A^\top A$. The original optimization problem is equivalent to

$$\min_{\mathbf{w}\in\mathcal{W}} g(A\mathbf{w}) + \mathbf{w}^\top \mathbb{E}_{\mathbf{z}'}[b(\mathbf{z}')] + c, \tag{44}$$

where $g(\mathbf{u}) = \mathbf{u}^\top \mathbf{u}$ is a strongly convex function of $\mathbf{u}$. Since the constraint is a polyhedral function of $\mathbf{w}$, according to the Lemma 12 of [52], we know that the optimization problem satisfies EBC($\theta = 1, \alpha$).

□

Piecewise Linear Problems (PLP)

Lemma 11. *Consider the problem*

$$\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w}) \triangleq \mathbb{E}[f(\mathbf{w}, \mathbf{z})] \tag{45}$$

where $\mathbb{E}[f(\mathbf{w}, \mathbf{z})]$ is a piecewise linear function and $\mathcal{W}$ is a bounded polyhedron. Then the problem (45) satisfies EBC($\theta = 1, \alpha$).

Proof. According to weak sharp minima condition [5] (e.g., Lemma 8 in [52]), we have

$$\|\mathbf{w} - \mathbf{w}^*\|_2^2 \leq c(P(\mathbf{w}) - P(\mathbf{w}^*))^2,$$

Since $P(\mathbf{w})$ is piecewise linear, then $P(\mathbf{w}) - P(\mathbf{w}_*)$ is bounded on a bounded set. Then there exists $\alpha > 0$ such that

$$\|\mathbf{w} - \mathbf{w}^*\|_2^2 \leq \alpha(P(\mathbf{w}) - P(\mathbf{w}^*)),$$

□

ℓ_1 regularized problems

Lemma 12. *Consider the problem: for ℓ_1 regularized risk minimization:*

$$\min_{\|\mathbf{w}\|_1\leq B} F(\mathbf{w}) \triangleq P(\mathbf{w}) + \lambda\|\mathbf{w}\|_1, \tag{46}$$

If $P(\mathbf{w})$ is convex quadratic or piecewise linear, then the problem (46) satisfies EBC($\theta = 1, \alpha$).

Proof. It is easy to see that $P(\mathbf{w})$ is either piecewise linear or piecewise convex quadratic. According to Lemma 3.3 of [23], we have

- When $P(\mathbf{w})$ is piecewise linear, there exists $\alpha_1, \alpha > 0$, such that

  $$\|\mathbf{w} - \mathbf{w}^*\|_2^2 \leq \alpha_1(P(\mathbf{w}) - P(\mathbf{w}^*))^2 \leq \alpha(P(\mathbf{w}) - P(\mathbf{w}^*)),$$

  where we use the fact $P(\mathbf{w}) - P(\mathbf{w}_*)$ is bounded over a bounded domain due to its Lipschitz continuity.

- When $P(\mathbf{w})$ is piecewise convex quadratic, there exists $\alpha_2 > 0$, such that

  $$\|\mathbf{w} - \mathbf{w}^*\|_2^2 \leq \alpha_2(P(\mathbf{w}) - P(\mathbf{w}^*)).$$

□

Algorithm 3 SSGS($\mathbf{w}_1, \beta, T$)

Input: $\mathbf{w}_1 \in \mathcal{W}$, $\beta > 0$ and T
Output: $\widehat{\mathbf{w}}_T$

1: **for** $t = 1, \ldots, T$ **do**
2: $\quad \mathbf{w}'_{t+1} = (1 - \frac{2}{t})\mathbf{w}_t + \frac{2}{t}\mathbf{w}_1 - \frac{2\beta}{t} g_t$
3: $\quad \mathbf{w}_{t+1} = \Pi_{\mathcal{W}}(\mathbf{w}'_{t+1})$
4: **end for**
5: $\widehat{\mathbf{w}}_T = \frac{1}{T+1} \sum_{t=1}^{T+1} \mathbf{w}_t$
6: return $\widehat{\mathbf{w}}_T$

Algorithm 4 ASA2($\mathbf{w}_1, n, R_0$)

Input: $\mathbf{w}_1 \in \mathcal{W}$, n and $R_0 = 2R$
Output: $\widehat{\mathbf{w}}_m$

1: Set $\widehat{\mathbf{w}}_0 = \mathbf{w}_1$, $m = \lfloor \frac{1}{2} \log_2 \frac{2n}{\log_2 n} \rfloor - 1$, $n_0 = \lfloor n/m \rfloor$
2: **for** $k = 1, \ldots, m$ **do**
3: $\quad$ Set $\beta_k = \frac{R_{k-1}\sqrt{n_0}}{2G}$ and $R_k = R_{k-1}/2$
4: $\quad \widehat{\mathbf{w}}_k = \text{SSGS}(\widehat{\mathbf{w}}_{k-1}, \beta_k, n_0)$
5: **end for**

Lemma 13. *Consider the problem:*

$$\min_{\mathbf{w} \in \mathcal{W}} F(\mathbf{w}) \triangleq P(\mathbf{w}) + \lambda \|\mathbf{w}\|_p^p \tag{47}$$

If $P(\mathbf{w})$ is convex quadratic, and $\mathcal{W}$ is a bounded polyheron, then the above problem satisfies EBC($\theta = 1/p, \alpha$).

Proof. According to Theorem 5.2 [53], the objective function is p-th order convex polynomial function and $\forall \mathbf{w} \in \mathcal{W}$ there exists $\tau > 0$ such that

$$\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \tau(P(\mathbf{w}) - P(\mathbf{w}^*) + (P(\mathbf{w}) - P(\mathbf{w}^*))^{1/p}).$$

There exists $c > 0$ such that $P(\mathbf{w}) - P(\mathbf{w}^*) \leq c$ for any $\mathbf{w} \in \mathcal{W}$. Then

$$\|\mathbf{w} - \mathbf{w}^*\|_2 \leq \tau(c^{1-1/p} + 1)(P(\mathbf{w}) - P(\mathbf{w}^*))^{1/p},$$

i.e.,

$$\|\mathbf{w} - \mathbf{w}^*\|_2^2 \leq \tau^2(c^{1-1/p} + 1)^2(P(\mathbf{w}) - P(\mathbf{w}^*))^{2/p}.$$

□

I A Variant of ASA without projection into intersection

Now we provide a different variant of ASA, which utilizes SSGS (Algorithm 2 in [49]) as a subroutine to avoid the projection onto the intersection of $\mathcal{W}$ and a bounded ball in the vanilla ASA. SSGS is an algorithm which adds a strongly convex regualarizer to the original loss function, i.e.,

$$\min_{\mathbf{w} \in \mathcal{W}} P(\mathbf{w}) + \frac{1}{2\beta}\|\mathbf{w} - \mathbf{w}_1\|_2^2,$$

where $\mathbf{w}_1 \in \mathcal{W}$ is called reference point. For completeness, we describe the SSGS and the corresponding ASA2 algorithms in Algorithm 3 and Algorithm 4 respectively.

We first present a result for analyzing SSGS, which is the Corollary 5 in [49].

Proposition 2. *Suppose* **Assumptions 1 and 2** *hold. Let $0 < \delta < 1/e$, $T \geq 3$, $\mathbf{w}^* \in \mathcal{W}_*$ be the closest optimal solution to $\mathbf{w}_1$, and R_0 be an upper bound on $\|\mathbf{w}_1 - \mathbf{w}^*\|_2$. Apply T iterations of the SSGS (Algorithm 3) and return the average solution, where g_t is a stochastic subgradient of $P(\mathbf{w})$ at $\mathbf{w}_t$. With probability at least $1 - \delta$, we have*

$$P(\widehat{\mathbf{w}}_T) - P_* \leq \frac{1}{2\beta}\|\mathbf{w}_1 - \mathbf{w}^*\|_2^2 + \frac{34\beta G^2(1 + \log T + \log(4 \log T/\delta))}{T}.$$

where $\widehat{\mathbf{w}}_T = \frac{1}{T+1} \sum_{t=1}^{T+1} \mathbf{w}_t$. Moreover, choose $\beta = \frac{R_0\sqrt{T}}{2G}$, and then with probability at least $1 - \delta$,

$$P(\widehat{\mathbf{w}}_T) - P_* \leq R_0 G \left(\frac{1}{\sqrt{T}} + \frac{17\,(1 + \log T + \log\,(4 \log T/\delta))}{\sqrt{T}} \right).$$

Similarly, for any nonnegative R_0, by choosing $\beta = \frac{R_0\sqrt{T}}{2G}$, and then with probability at least $1 - \delta$,

$$P(\widehat{\mathbf{w}}_T) - P(\mathbf{w}_1) \leq R_0 G \left(\frac{1}{\sqrt{T}} + \frac{17\,(1 + \log T + \log\,(4 \log T/\delta))}{\sqrt{T}} \right).$$

Then we provide the high probability analysis of ASA2, which is Theorem 5.

Theorem 5. *Suppose Assumptions 1, and 2 hold. Let $\widehat{\mathbf{w}}_m$ be the returned solution of the Algorithm 4. For $n \geq 100$ and any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$P(\widehat{\mathbf{w}}_m) - P_* \leq O\left(\frac{\alpha G^2 \log(n)(\log n + \log(\frac{\log n}{\sqrt{\delta}}))^2}{n}\right)^{\frac{1}{2-\theta}}.$$

Proof. We use the same notation as that in the proof of Theorem 3 unless specified. Define

$$a(n, \bar{\delta}) = G\left(\frac{1}{\sqrt{n}} + \frac{17(1 + \log n + \log(4\log n/\bar{\delta}))}{\sqrt{n}}\right). \tag{48}$$

First we show that when $n \geq 100$, we have

$$\frac{1}{2}\sqrt{\frac{2n}{\log_2 n}}\left(\frac{1}{\sqrt{n_0}} + \frac{17(1 + \log n_0 + \log(4\log n_0/\bar{\delta}))}{\sqrt{n_0}}\right) \geq 1.$$

Note that

$$\begin{aligned}
\text{LHS} &\geq \sqrt{\frac{2n}{\log_2 n}}\left(\frac{\sqrt{17(1 + \log n_0 + \log(4\log n_0/\bar{\delta}))}}{\sqrt{n_0}}\right) \\
&\geq \sqrt{\frac{34m(1 + \log(\frac{n}{m} - 1) + \log(4\log(\frac{n}{m} - 1)/\bar{\delta}))}{\log_2 n}} \\
&\geq \sqrt{\frac{17(\log_2 n - \log_2\log_2 n - 3)\cdot \mathcal{F}_1}{\log_2 n}} \\
&\geq \sqrt{17(1 - \frac{\log_2\log_2 n + 3}{\log_2 n})} \geq 1 = \text{RHS},
\end{aligned}$$

where $\mathcal{F}_1 = (1 + \log(\frac{n}{m} - 1) + \log(2\log(\frac{n}{m} - 1)\log_2 n/\delta))$. The first inequality holds by utilizing the fact that $a + b \geq 2\sqrt{ab}$, the second inequality holds since $n \geq 100$, and then $3 \leq \frac{n}{m} - 1 \leq n_0 = \lfloor \frac{n}{m} \rfloor \leq \frac{n}{m}$, the third inequality holds because of $m \geq \frac{1}{2}\log_2 \frac{2n}{\log_2 n} - 2 > 0$ and definition of $\bar{\delta}$, the fourth and fifth inequalities hold since $n \geq 100$ and $m \leq \frac{1}{2}\log_2 n$.

We can duplicate the rest of the proof of Theorem 3 other than using the definition of $a(n_0, \bar{\delta})$ according to (48). Finally, we have with probablity at least $1-\delta$,

$$P(\widehat{\mathbf{w}}_m) - P_* \leq (64\alpha)^{\frac{1}{2-\theta}} a(n_0, \bar{\delta})^{\frac{2}{2-\theta}} \leq \left(\frac{64\alpha G^2(1 + 17\mathcal{F}_2)^2}{\frac{2n}{\log_2 n} - 1}\right)^{\frac{1}{2-\theta}},$$

where

$$\mathcal{F}_2 = 1 + \log\left(\frac{n}{\frac{1}{2}\log_2 \frac{2n}{\log_2 n} - 2}\right) + \log\left(2\log\left(\frac{n}{\frac{1}{2}\log_2 \frac{2n}{\log_2 n} - 2}\right)\log_2 n/\delta\right).$$

The second inequality holds since $n_0 = \lfloor \frac{n}{m} \rfloor \geq \frac{n}{m} - 1$, $\frac{1}{2}\log_2 \frac{2n}{\log_2 n} - 2 \leq m \leq \frac{1}{2}\log_2 n$. □

J A variant of ASA with a subroutine using proximal mapping

In this section, we consider the nonsmooth composite optimization problem (2), which is

$$\min_{\mathbf{w}\in\mathcal{W}} P(\mathbf{w}) \triangleq \mathbb{E}_{\mathbf{z}\sim\mathbb{P}}[f(\mathbf{w}, \mathbf{z})] + r(\mathbf{w}).$$

We introduce a variant of ASA, i.e., ASA3 (Algorithm 6), with a theoretical guarantee. ASA3 is a multistage scheme of proximal SGD (Algorithm 5).

Before analysis, we first present a standard result of proximal SGD, which is the Lemma 5 of [49].

Algorithm 5 PSG($\mathbf{w}_1, \gamma, T, \mathcal{W}$)

Input: $\mathbf{w}_1 \in \mathcal{W}$, $\gamma > 0$ and T
Output: $\widehat{\mathbf{w}}_T$
1: **for** $t = 1, \ldots, T$ **do**
2: Compute

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}} \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|_2^2 + \eta g_t^\top \mathbf{w} + \eta r(\mathbf{w}),$$

where g_t is the stochastic subgradient of $\mathbb{E}_{\mathbf{z}\sim\mathbb{P}}[f(\mathbf{w}, \mathbf{z})]$ evaluated at $\mathbf{w}_t$
3: **end for**
4: $\widehat{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$
5: return $\widehat{\mathbf{w}}_T$

Algorithm 6 ASA3($\mathbf{w}_1, n, R_0$)

Input: $\mathbf{w}_1 \in \mathcal{W}$, n and $R_0 = 2R$
Output: $\widehat{\mathbf{w}}_m$
1: Set $\widehat{\mathbf{w}}_0 = \mathbf{w}_1$, $m = \lfloor \frac{1}{2}\log_2 \frac{2n}{\log_2 n} \rfloor - 1$, $n_0 = \lfloor n/m \rfloor$
2: **for** $k = 1, \ldots, m$ **do**
3: Set $\gamma_k = \frac{R_{k-1}}{G\sqrt{n_0}}$ and $R_k = R_{k-1}/2$
4:

$$\widehat{\mathbf{w}}_k = \text{PSG}(\widehat{\mathbf{w}}_{k-1}, \gamma_k, n_0, \mathcal{W} \cap \mathcal{B}(\widehat{\mathbf{w}}_{k-1}, R_{k-1}))$$

5: **end for**
6: return $\widehat{\mathbf{w}}_m$

Proposition 3. *Suppose* **Assumptions 1 and 2** *hold. In addition, we assume the proximal mapping in terms of $r(\mathbf{w})$ has a closed form, and $r(\mathbf{w})$ is ρ-Lipschitz continuous for any $\mathbf{w} \in \mathcal{W}$. Let $\epsilon \geq 0$ and D be the upper bound of $\|\mathbf{w}_1 - \mathbf{w}_{1,\epsilon}^\dagger\|_2$, where $\mathbf{w}_{1,\epsilon}^\dagger$ is the point closed to ϵ-sublevel set of $P(\mathbf{w})$. Denote g_t by the stochastic subgradient of $\mathbb{E}_{\mathbf{z}\sim\mathbb{P}}[f(\mathbf{w}, \mathbf{z})]$ at $\mathbf{w}_t$. Apply T-iterations of the following steps:*

$$\mathbf{w}_{t+1} = \arg\min_{\mathbf{w}\in\mathcal{W}\cap\mathcal{B}(\mathbf{w}_1, D)} \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|_2^2 + \eta g_t^\top \mathbf{w} + \eta r(\mathbf{w}).$$

Given $\mathbf{w}_1$, for any $\delta \in (0, 1)$, we have with probability at least $1 - \delta$,

$$P(\widehat{\mathbf{w}}_T) - P(\mathbf{w}_{1,\epsilon}^\dagger) \leq \frac{\eta G^2}{2} + \frac{\|\mathbf{w}_1 - \mathbf{w}_{1,\epsilon}^\dagger\|_2^2}{2\eta T} + \frac{4GD\sqrt{3\log(1/\delta)}}{\sqrt{T}} + \frac{\rho D}{T},$$

where $\widehat{\mathbf{w}}_T = \frac{1}{T}\sum_{t=1}^T \mathbf{w}_t$.

Theorem 6. *Suppose* **Assumptions 1 and 2** *hold. In addition, we assume the proximal mapping in terms of $r(\mathbf{w})$ has a closed form, and $r(\mathbf{w})$ is ρ-Lipschitz continuous for any $\mathbf{w} \in \mathcal{W}$. $\|\mathbf{w}_1 - \mathbf{w}^*\|_2 \leq R_0$, where $\mathbf{w}^*$ is the closest optimal solution to $\mathbf{w}_1$. For $n \geq 100$, $n_0 \geq \frac{\rho^2}{G^2}$ and any $\delta \in (0, 1)$, with probability at least $1 - \delta$, the Algorithm ASA3 guarantees that*

$$P(\widehat{\mathbf{w}}_m) - P_* \leq O\left(\frac{\bar{\alpha}(\log(n)\log(\log(n)/\delta))}{n}\right)^{\frac{1}{2-\theta}}.$$

where $\bar{\alpha} = \max(\alpha G^2, (R_0 G)^{2-\theta})$.

Proof. At first we derive the parallel version of the Proposition 1 and Lemma 7 in the case of solving problem (2), which is not difficult by utilizing the Proposition 3.

- We first prove the parallel version of the Proposition 1. By taking $\epsilon = 0$, then $\mathbf{w}_{1,\epsilon}^\dagger$ is the projection of $\mathbf{w}_1$ onto the optimal set $\mathcal{W}_*$, and we define it to be $\mathbf{w}^*$. If R_0 is a upper bound

of $\|\mathbf{w}_1 - \mathbf{w}^*\|_2$, by taking $\eta = \frac{R_0}{G\sqrt{T}}$, then applying T iterations of

$$\mathbf{w}_{t+1} = \operatorname*{arg\,min}_{\mathbf{w} \in \mathcal{W} \cap \mathcal{B}(\mathbf{w}_1, R_0)} \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|_2^2 + \eta g_t^\top \mathbf{w} + \eta r(\mathbf{w})$$

has the guarantee that with probability at least $1 - \delta$,

$$P(\widehat{\mathbf{w}}_T) - P_* \leq R_0 G \left(\frac{1}{\sqrt{T}} + \frac{4\sqrt{3\log(1/\delta)}}{\sqrt{T}} \right) + \frac{\rho R_0}{T}.$$

By choosing $T \geq \frac{\rho^2}{G^2}$, i.e., $\frac{\rho R_0}{T} \leq \frac{R_0 G}{\sqrt{T}}$, and we have

$$P(\widehat{\mathbf{w}}_T) - P_* \leq R_0 G \left(\frac{2}{\sqrt{T}} + \frac{4\sqrt{3\log(1/\delta)}}{\sqrt{T}} \right).$$

- We then prove the parallel version of the Lemma 7. We make choose ϵ large enough such that $\mathbf{w}_{1,\epsilon}^\dagger = \mathbf{w}_1$. By utilizing the Proposition 3, we know that for any nonnegative R_0, taking $\eta = \frac{R_0}{G\sqrt{T}}$ and applying T iterations of

$$\mathbf{w}_{t+1} = \operatorname*{arg\,min}_{\mathbf{w} \in \mathcal{W} \cap \mathcal{B}(\mathbf{w}_1, R_0)} \frac{1}{2}\|\mathbf{w} - \mathbf{w}_t\|_2^2 + \eta g_t^\top \mathbf{w} + \eta r(\mathbf{w})$$

have the guarantee that with probability at least $1 - \delta$,

$$P(\widehat{\mathbf{w}}_T) - P(\mathbf{w}_1) \leq R_0 G \left(\frac{1}{\sqrt{T}} + \frac{4\sqrt{3\log(1/\delta)}}{\sqrt{T}} \right) + \frac{\rho R_0}{T}.$$

By choosing $T \geq \frac{\rho^2}{G^2}$, i.e., $\frac{\rho R_0}{T} \leq \frac{R_0 G}{\sqrt{T}}$, and we have

$$P(\widehat{\mathbf{w}}_T) - P_* \leq R_0 G \left(\frac{2}{\sqrt{T}} + \frac{4\sqrt{3\log(1/\delta)}}{\sqrt{T}} \right).$$

The rest of the proof is similar to the proof of Theorem 3. □

Finally, we mention that a stochastic mirror descent algorithm with a non-Euclidean norm prox-function can be used, e.g., the Composite Objective Mirror Descent algorithm with p-norm divergence in [8], Similar analysis based on Theorem 8 in [8] can be derived. When leveraging the error bound, we can use a p-norm version (i.e., changing the Euclidean norm to the p-norm and the corresponding parameter α).