Empirical Risk Minimization for Stochastic Convex Optimization: $O(1/n)$- and $O(1/n^2)$-type of Risk Bounds

Lijun Zhang   ZHANGLJ@LAMDA.NJU.EDU.CN

National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China

Tianbao Yang   TIANBAO-YANG@UIOWA.EDU

Department of Computer Science
the University of Iowa, Iowa City, IA 52242, USA

Rong Jin   RONGJIN@CSE.MSU.EDU

Alibaba Group, Seattle, USA

Abstract

Although there exist plentiful theories of empirical risk minimization (ERM) for supervised learning, current theoretical understandings of ERM for a related problem—stochastic convex optimization (SCO), are limited. In this work, we strengthen the realm of ERM for SCO by exploiting smoothness and strong convexity conditions to improve the risk bounds. First, we establish an $\widetilde{O}(d/n + \sqrt{F_*/n})$ risk bound when the random function is nonnegative, convex and smooth, and the expected function is Lipschitz continuous, where d is the dimensionality of the problem, n is the number of samples, and F_* is the minimal risk. Thus, when F_* is small we obtain an $\widetilde{O}(d/n)$ risk bound, which is analogous to the $\widetilde{O}(1/n)$ optimistic rate of ERM for supervised learning. Second, if the objective function is also λ-strongly convex, we prove an $\widetilde{O}(d/n + \kappa F_*/n)$ risk bound where κ is the condition number, and improve it to $O(1/[\lambda n^2] + \kappa F_*/n)$ when $n = \widetilde{\Omega}(\kappa d)$. As a result, we obtain an $O(\kappa/n^2)$ risk bound under the condition that n is large and F_* is small, which to the best of our knowledge, is the *first* $O(1/n^2)$-type of risk bound of ERM. Third, we stress that the above results are established in a unified framework, which allows us to derive new risk bounds under weaker conditions, e.g., without convexity of the random function and Lipschitz continuity of the expected function. Finally, we demonstrate that to achieve an $O(1/[\lambda n^2] + \kappa F_*/n)$ risk bound for supervised learning, the $\widetilde{\Omega}(\kappa d)$ requirement on n can be replaced with $\Omega(\kappa^2)$, which is dimensionality-independent.

Keywords: Empirical Risk Minimization, Stochastic Convex Optimization, Excess Risk

1. Introduction

Stochastic optimization occurs in almost all areas of science and engineering, such as machine learning, statistics and operations research (Shapiro et al., 2014). In this problem, the goal is to optimize the value of an expected objective function $F(\cdot)$ over some set $\mathcal{W}$, i.e.,

$$\min_{\mathbf{w} \in \mathcal{W}} \ F(\mathbf{w}) = \mathrm{E}_{f \sim \mathbb{P}} \left[f(\mathbf{w}) \right], \tag{1}$$

where $f(\cdot) : \mathcal{W} \mapsto \mathbb{R}$ is a random function sampled from a (possibly unknown) distribution $\mathbb{P}$. A well-known special case is the risk minimization problem in supervised learning (Vapnik,

1998, 2000), which takes the following form

$$\min_{h \in \mathcal{H}} \ F(h) = \mathrm{E}_{(\mathbf{x},y) \sim \mathbb{D}} \left[\ell(h(\mathbf{x}), y) \right], \tag{2}$$

where $\mathcal{H} = \{h : \mathcal{X} \mapsto \mathbb{R}\}$ is a hypothesis class, $(\mathbf{x}, y) \in \mathcal{X} \times \mathbb{R}$ is an instance-label pair sampled from a distribution $\mathbb{D}$, and $\ell(\cdot, \cdot) : \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ is certain loss. In this paper, we mainly focus on the convex version of (1), namely stochastic convex optimization (SCO), where both the domain $\mathcal{W}$ and the expected function $F(\cdot)$ are convex.

Two classical approaches for solving stochastic optimization are stochastic approximation (SA) (Kushner and Yin, 2003) and the sample average approximation (SAA), the latter of which is also referred to as empirical risk minimization (ERM) in the machine learning community (Vapnik, 1998). While both SA and ERM have been extensively studied in recent years (Bartlett and Mendelson, 2002; Bartlett et al., 2005; Koltchinskii, 2011; Nemirovski et al., 2009; Moulines and Bach, 2011), most theoretical guarantees of ERM are restricted to the supervised learning problem in (2). As pointed out in a seminal work of Shalev-Shwartz et al. (2009), the success of ERM for supervised learning cannot be directly extended to stochastic optimization. Actually, Shalev-Shwartz et al. (2009) have constructed an instance of SCO that is learnable by SA but cannot be solved by ERM. Literatures about ERM for stochastic optimization (including SCO) are quite limited, and we still lack a full understanding of the theory.

In ERM, we are given n i.i.d. functions $f_1, \ldots, f_n$ sampled from $\mathbb{P}$, and minimize an empirical objective function:

$$\min_{\mathbf{w} \in \mathcal{W}} \ \widehat{F}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} f_i(\mathbf{w}). \tag{3}$$

Let $\widehat{\mathbf{w}} \in \operatorname{argmin}_{\mathbf{w} \in \mathcal{W}} \widehat{F}(\mathbf{w})$ be the empirical minimizer. The performance of ERM is measured in terms of the excess risk defined as

$$F(\widehat{\mathbf{w}}) - \min_{\mathbf{w} \in \mathcal{W}} \ F(\mathbf{w}).$$

State-of-the-art risk bounds of ERM include: an $\widetilde{O}(\sqrt{d/n})$ bound when the random function $f(\cdot)$ is Lipschitz continuous,[1] where d is the dimensionality of $\mathbf{w}$; an $O(1/\lambda n)$ bound when $f(\cdot)$ is λ-strongly convex (Shalev-Shwartz et al., 2009); and an $\widetilde{O}(d/\eta n)$ bound when $f(\cdot)$ is η-exponentially concave (η-exp-concave) (Mehta, 2016). From existing studies of ERM for supervised learning (Srebro et al., 2010), we know that smoothness can be utilized to boost the risk bound. Thus, it is natural to ask whether smoothness can also be exploited to improve the performance of ERM for SCO. This paper provides an affirmative answer to this question. Indeed, we propose a general approach for analyzing the excess risk bound of ERM, which brings several improved risk bounds and new risk bounds as well.

To state our results, we first introduce some notations. Let $F_* = \min_{\mathbf{w} \in \mathcal{W}} F(\mathbf{w})$ be the minimal risk, λ be the modulus of strong convexity of $F(\cdot)$ and L be the modulus of smoothness of $f(\cdot)$. Denote by $\kappa = L/\lambda$ the condition number of the problem. Our and previous results of ERM for SCO are summarized in Table 1, where we make explicit the

---

1. We use the $\widetilde{O}$ and $\widetilde{\Omega}$ notations to hide constant factors as well as polylogarithmic factors in d and n.

Table 1: Summary of Excess Risk Bounds of ERM for SCO. All bounds hold with high probability except the one marked by *, which holds in expectation. Abbreviations: bounded $\to$ b, convex $\to$ c, generalized linear $\to$ gl, Lipschitz continuous $\to$ Lip, nonnegative $\to$ nn, strongly convex $\to$ sc, smooth $\to$ sm, η-exponentially concave $\to$ η-exp.

		$f(\cdot)$	$\widehat{F}(\cdot)$	$F(\cdot)$	Risk Bounds
Shalev-Shwartz et al. (2009)		Lip	-	-	$\widetilde{O}(\sqrt{\frac{d}{n}})$
		Lip & sc	-	-	$O(\frac{1}{\lambda n})^*$
Mehta (2016)		η-exp & Lip & b	-	-	$\widetilde{O}(\frac{d}{\eta n})$
This work	Theorem 1	nn & c & sm	-	Lip	$\widetilde{O}(\frac{d}{n} + \sqrt{\frac{F_*}{n}})$
	Theorem 3	nn & c & sm	-	Lip & sc	$\widetilde{O}(\frac{d}{n} + \frac{\kappa F_*}{n})$ $O(\frac{1}{\lambda n^2} + \frac{\kappa F_*}{n})$ when $n = \widetilde{\Omega}(\kappa d)$
	Theorem 5	nn & sm	c	sc	$\widetilde{O}(\frac{\kappa d}{n} + \frac{\kappa F_*}{n}) = \widetilde{O}(\frac{\kappa d}{n})$ $O(\frac{1}{\lambda n^2} + \frac{\kappa F_*}{n})$ when $n = \widetilde{\Omega}(\kappa^2 d)$
	Theorem 7	nn & sm	c	c	$\widetilde{O}(\sqrt{\frac{d}{n}} + \sqrt{\frac{F_*}{n}}) = \widetilde{O}(\sqrt{\frac{d}{n}})$
	Theorem 8	nn & sm & gl	c	sc	$O(\frac{\kappa}{n} + \frac{\kappa F_*}{n}) = O(\frac{\kappa}{n})$ $O(\frac{1}{\lambda n^2} + \frac{\kappa F_*}{n})$ when $n = \Omega(\kappa^2)$

assumptions on the random function $f(\cdot)$, the empirical function $\widehat{F}(\mathbf{w})$ and the expected function $F(\cdot)$. For our results of ERM for SCO, we assume the domain is bounded, and the random function is nonnegative. We highlight the significance of this work as follows:

- When $f(\cdot)$ is both convex and smooth and $F(\cdot)$ is Lipschitz continuous, we establish an $\widetilde{O}(d/n + \sqrt{F_*/n})$ risk bound (c.f. Theorem 1). In the optimistic case that F_* is small, i.e., $F_* = O(d^2/n)$, we obtain an $\widetilde{O}(d/n)$ risk bound, which is analogous to the $\widetilde{O}(1/n)$ optimistic rate of ERM for supervised learning (Srebro et al., 2010) and also matches a recent lower bound of ERM for SCO (Feldman, 2016, Theorem 3.10).
- If $F(\cdot)$ is also λ-strongly convex, we prove an $\widetilde{O}(d/n + \kappa F_*/n)$ risk bound, and improve it to $O(1/[\lambda n^2] + \kappa F_*/n)$ when $n = \widetilde{\Omega}(\kappa d)$ (c.f. Theorem 3). Thus, if n is large and F_* is small, i.e., $F_* = O(1/n)$, we get an $O(\kappa/n^2)$ risk bound, which to the best of our knowledge, is the first $O(1/n^2)$-type of risk bound of ERM.
- When neither convexity is present in $f(\cdot)$ nor Lipschitz continuity is present in $F(\cdot)$, as long as $f(\cdot)$ is smooth, $\widehat{F}(\cdot)$ is convex and $F(\cdot)$ is strongly convex, we still obtain an improved risk bound of $O(1/[\lambda n^2] + \kappa F_*/n)$ when $n = \widetilde{\Omega}(\kappa^2 d)$, which will further implies an $O(\kappa/n^2)$ risk bound if $F_* = O(1/n)$ (c.f. Theorem 5).
- If strong convexity is also absent in $F(\cdot)$, assuming $f(\cdot)$ is smooth and both $\widehat{F}(\cdot)$ and $F(\cdot)$ are convex, we obtain an $\widetilde{O}(\sqrt{d/n})$ risk bound (c.f. Theorem 7). This result

3

breaks the barrier of non-learnability of bounded convex functions (Feldman, 2016, Theorem 5.2) by exploiting the smoothness of random functions.

- Finally, we extend the $O(1/[\lambda n^2] + \kappa F_*/n)$ risk bound to supervised learning with a generalized linear form. Our analysis shows that in this case, the lower bound of n can be replaced with $\Omega(\kappa^2)$, which is dimensionality-independent (c.f. Theorem 8). Thus, this result can be applied to infinite dimensional cases, e.g., learning with kernels.

2. Related Work

In this section, we give a brief introduction to previous work on stochastic optimization.

2.1 ERM for Stochastic Optimization

As we mentioned earlier, there are few works devoted to ERM for stochastic optimization. When $\mathcal{W} \subset \mathbb{R}^d$ is bounded and $f(\cdot)$ is Lipschitz continuous, Shalev-Shwartz et al. (2009) demonstrate that $\widehat{F}(\mathbf{w})$ converges to $F(\mathbf{w})$ uniformly over $\mathcal{W}$ with an $\widetilde{O}(\sqrt{d/n})$ error bound that holds with high probability, implying an $\widetilde{O}(\sqrt{d/n})$ risk bound of ERM. They further establish an $O(1/\lambda n)$ risk bound of ERM that holds in expectation when $f(\cdot)$ is λ-strongly convex and Lipschitz continuous. Stochastic optimization with exp-concave functions is studied recently (Koren and Levy, 2015),[2] and Mehta (2016) proves an $\widetilde{O}(d/\eta n)$ bound of ERM that holds with high probability when $f(\cdot)$ is η-exp-concave, Lipschitz continuous, and bounded. Lower bounds of ERM for stochastic optimization is investigated by Feldman (2016), who exhibits (i) a lower bound of $\Omega(d/\epsilon^2)$ sample complexity for uniform convergence that nearly matches the upper bound of Shalev-Shwartz et al. (2009); and (ii) a lower bound of $\Omega(d/\epsilon)$ sample complexity of ERM, which is matched by our $\widetilde{O}(d/n + \sqrt{F_*/n})$ bound when F_* is small.

It is worth mentioning the difference among proof techniques in these works. The uniform convergence result of Shalev-Shwartz et al. (2009) leverages the covering number to bound $|\widehat{F}(\mathbf{w}) - F(\mathbf{w})|$ for any $\mathbf{w} \in \mathcal{W}$. The analysis for strongly convex functions by Shalev-Shwartz et al. (2009) and exp-concave functions by Koren and Levy (2015) utilize the tool of stability, which only produces risk bounds that hold in expectation. A simple way to achieve a high probability bound is to use ERM combined with a generic or specific boosting-the-confidence method (Mehta, 2016; Haussler et al., 1991), but the guarantee is not directly on the empirical minimizer as noted by Shalev-Shwartz et al. (2009). The convergence of ERM given by Mehta (2016) relies on a central condition or "stochastic mixability" of the exp-concave function. In this paper, we present a general approach for analyzing ERM for SCO of smooth functions. In particular, our analysis is based on a uniform convergence of $\nabla \widehat{F}(\mathbf{w}) - \nabla \widehat{F}(\mathbf{w}_*)$ to $\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}_*)$ for any $\mathbf{w} \in \mathcal{W}$, and a concentration inequality of $\|\nabla \widehat{F}(\mathbf{w}_*) - \nabla F(\mathbf{w}_*)\|$, where $\mathbf{w}_*$ is the optimal solution to (1).

2.2 ERM for Supervised Learning

We note that there are extensive studies on ERM for supervised learning, and hence the review here is non-exhaustive. In the context of supervised learning, the performance of ERM is closely related to the uniform convergence of $\widehat{F}(\cdot)$ to $F(\cdot)$ over the hypothesis

---

2. Their excess risk bound is for a regularized empirical risk minimizer.

class $\mathcal{H}$ (Koltchinskii, 2011). In fact, uniform convergence is a sufficient condition for learnability (Shalev-Shwartz and Ben-David, 2014), and in some special cases such as binary classification, it is also a necessary condition (Vapnik, 1998). The accuracy of uniform convergence, as well as the quality of the empirical minimizer, can be upper bounded in terms of the complexity of the hypothesis class $\mathcal{H}$, including data-independent measures such as the VC-dimension and data-dependent measures such as the Rademacher complexity.

Generally speaking, when $\mathcal{H}$ has finite VC-dimension, the excess risk can be upper bounded by $O(\sqrt{\mathrm{VC}(\mathcal{H})/n})$, where $\mathrm{VC}(\mathcal{H})$ is the VC-dimension of $\mathcal{H}$. If the loss $\ell(\cdot, \cdot)$ is Lipschitz continuous with respect to its first argument, we have a risk bound of $O(1/\sqrt{n} + \mathcal{R}_n(\mathcal{H}))$, where $\mathcal{R}_n(\mathcal{H})$ is the Rademacher complexity of $\mathcal{H}$. The Rademacher complexity typically scales as $\mathcal{R}_n(\mathcal{H}) = O(1/\sqrt{n})$, e.g., $\mathcal{H}$ contains linear functions with low-norm, implying an $O(1/\sqrt{n})$ risk bound (Bartlett and Mendelson, 2002). There have been intensive efforts to derive rates faster than $O(1/\sqrt{n})$ under various conditions (Lee et al., 1996; Panchenko, 2002; Bartlett et al., 2005; Gonen and Shalev-Shwartz, 2016), such as low-noise (Tsybakov, 2004), smoothness (Srebro et al., 2010), strong convexity (Sridharan et al., 2009), to name a few amongst many. Specifically, when the random function $f(\cdot)$ is nonnegative and smooth, Srebro et al. (2010) have established a risk bound of $\widetilde{O}(\mathcal{R}_n^2(H) + \mathcal{R}_n(H)\sqrt{F_*})$, reducing to an $\widetilde{O}(1/n)$ bound if $\mathcal{R}_n(\mathcal{H}) = O(1/\sqrt{n})$ and $F_* = O(1/n)$. A generalized linear form of (2) is studied by Sridharan et al. (2009), and a risk bound of $O(1/\lambda n)$ is proved if the expected function $F(\cdot)$ is λ-strongly convex.

2.3 SA for Stochastic Optimization

Stochastic approximation (SA) solves the stochastic optimization problem via noisy observations of the expected function (Kushner and Yin, 2003). For brevity, we only discuss first-order methods for SCO, and in this case, n is the number of stochastic gradients consumed by the algorithm. For Lipschitz continuous convex functions, stochastic gradient descent (SGD) exhibits the optimal $O(1/\sqrt{n})$ risk bound (Nemirovski and Yudin, 1983). When the random function $f(\cdot)$ is nonnegative and smooth, SGD (with a suitable step size) has a risk bound of $O(1/n + \sqrt{F_*/n})$, becoming $O(1/n)$ if $F_* = O(1/n)$ (Srebro et al., 2010, Corollary 4). If $F(\cdot)$ is λ-strongly convex, some variants of SGD (Hazan and Kale, 2011; Rakhlin et al., 2012) achieve an $O(1/\lambda n)$ rate which is known to be minimax optimal (Agarwal et al., 2012). For the square loss and the logistic loss, an $O(1/n)$ rate is attainable without any strong convexity assumptions (Bach and Moulines, 2013). When the random function $f(\cdot)$ is η-exp-concave, the online Newton step (ONS) is equipped with an $\widetilde{O}(d/\eta n)$ risk bound (Hazan et al., 2007; Mahdavi et al., 2015).

3. Faster Rates of ERM

We first introduce all the assumptions used in our analysis, then present theoretical results under different combinations of them, and finally discuss a special case of supervised learning.

3.1 Assumptions

In the following, we use $\| \cdot \|$ to denote the ℓ_2-norm of vectors.

Assumption 1 *The domain $\mathcal{W}$ is a convex subset of $\mathbb{R}^d$, and is bounded by R, that is,*

$$\|\mathbf{w}\| \leq R, \ \forall \mathbf{w} \in \mathcal{W}. \tag{4}$$

Assumption 2 *The random function $f(\cdot)$ is nonnegative, and L-smooth over $\mathcal{W}$, that is,*

$$\left\|\nabla f(\mathbf{w}) - \nabla f(\mathbf{w}')\right\| \leq L\|\mathbf{w} - \mathbf{w}'\|, \ \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}, \ f \sim \mathbb{P}. \tag{5}$$

Assumption 3 *The expected function $F(\cdot)$ is G-Lipschitz continuous over $\mathcal{W}$, that is,*

$$|F(\mathbf{w}) - F(\mathbf{w}')| \leq G\|\mathbf{w} - \mathbf{w}'\|, \ \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}. \tag{6}$$

Assumption 4 *We use different combinations of the following assumptions on convexity.*

(a) *The expected function $F(\cdot)$ is convex over $\mathcal{W}$.*

(b) *The expected function $F(\cdot)$ is λ-strongly convex over $\mathcal{W}$, that is,*

$$F(\mathbf{w}) + \langle \nabla F(\mathbf{w}), \mathbf{w}' - \mathbf{w} \rangle + \frac{\lambda}{2}\|\mathbf{w}' - \mathbf{w}\|^2 \leq F(\mathbf{w}'), \ \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}. \tag{7}$$

(c) *The empirical function $\widehat{F}(\cdot)$ is convex.*

(d) *The random function $f(\cdot)$ is convex.*

Remark 1 First, note that **Assumption 4(a)** is implied by either **Assumption 4(b)** or **Assumption 4(d)**, and **Assumption 4(c)** is implied by **Assumption 4(d)**. Second, the smoothness assumption of $f(\cdot)$ implies the expected function $F(\cdot)$ is L-smooth. By Jensen's inequality, we have

$$\left\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}')\right\| \leq \mathrm{E}_{f\sim\mathbb{P}}\left\|\nabla f(\mathbf{w}) - \nabla f(\mathbf{w}')\right\| \leq L\|\mathbf{w} - \mathbf{w}'\|, \ \forall \mathbf{w}, \mathbf{w}' \in \mathcal{W}.$$

Similarly, the empirical function $\widehat{F}(\cdot)$ is also L-smooth. The *condition number* κ of $F(\cdot)$ is defined as the ratio between L and λ, i.e., $\kappa = L/\lambda \geq 1$.

3.2 Risk Bounds for SCO

Let $\mathbf{w}_* \in \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$ and $\widehat{\mathbf{w}} \in \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w})$ be optimal solutions to (1) and (3), respectively. We first present an excess risk bound under the smoothness condition.

Theorem 1 *For any $0 < \delta < 1$, define*

$$M = \sup_{f\sim\mathbb{P}} \|\nabla f(\mathbf{w}_*)\|, \tag{8}$$

$$C(\varepsilon) = 2\left(\log\frac{2}{\delta} + d\log\frac{6R}{\varepsilon}\right). \tag{9}$$

*Under **Assumptions 1**, **2**, **3**, and **4(d)**, with probability at least $1 - 2\delta$, we have*

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*)$$
$$\leq \frac{16R^2 LC(\varepsilon)}{n} + \frac{8RM\log(2/\delta)}{n} + 8R\sqrt{\frac{2LF_*\log(2/\delta)}{n}} + \left(8RL + G + \frac{4RLC(\varepsilon)}{n}\right)\varepsilon, \tag{10}$$

where $F_* = F(\mathbf{w}_*)$ is the minimal risk.

By choosing ε small enough, the last term in (10) that contains ε becomes non-dominating. To be specific, we have the following corollary.

Corollary 2 *By setting $\varepsilon = 1/n$ in Theorem 1, we have $C(1/n) = 2\left(\log\frac{2}{\delta} + d\log(6nR)\right) = \Theta(d\log n)$, and with high probability*

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{d\log n}{n} + \sqrt{\frac{F_*}{n}}\right) = \widetilde{O}\left(\frac{d}{n} + \sqrt{\frac{F_*}{n}}\right).$$

Remark 2 The above corollary implies that under the smoothness and other common assumptions, ERM achieves an $\widetilde{O}(d/n + \sqrt{F_*/n})$ risk bound for SCO. When the minimal risk is small, i.e., $F_* = O(d^2/n)$, the rate is improved to $\widetilde{O}(d/n)$. Note that even under the smoothness assumption, the linear dependence on d is unavoidable (Feldman, 2016, Theorem 3.7).

We next present excess risk bounds under both the smoothness and strong convexity conditions.

Theorem 3 *Under* **Assumptions 1**, **2**, **3**, **4(b)**, *and* **4(d)**, *with probability at least $1 - 2\delta$, we have*

$$
\begin{aligned}
&F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) \\
&\leq \frac{16R^2 LC(\varepsilon)}{n} + \frac{8RM\log(2/\delta)}{n} + \frac{8LF_*\log(2/\delta)}{\lambda n} + \left(8RL + G + \frac{4RLC(\varepsilon)}{n}\right)\varepsilon.
\end{aligned}
\tag{11}
$$

Furthermore, if

$$n \geq \frac{4LC(\varepsilon)}{\lambda} = 4\kappa C(\varepsilon), \tag{12}$$

we also have

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) \leq \frac{32M^2\log^2(2/\delta)}{\lambda n^2} + \frac{128LF_*\log(2/\delta)}{\lambda n} + \left(\frac{128L^2\varepsilon^2}{\lambda} + 16G\varepsilon + 4\lambda\varepsilon^2\right). \tag{13}$$

The above theorem can be simplified by choosing different values of ε.

Corollary 4 *By setting $\varepsilon = 1/n$ in Theorem 3, we have $C(1/n) = \Theta(d\log n)$, and with high probability*

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{d\log n}{n} + \frac{\kappa F_*}{n}\right) = \widetilde{O}\left(\frac{d}{n} + \frac{\kappa F_*}{n}\right).$$

Setting $\varepsilon = 1/n^2$, we have $C(1/n^2) = 2\left(\log\frac{2}{\delta} + d\log(6n^2R)\right) = \Theta(d\log n)$ and when $n = \Omega(\kappa d\log n) = \widetilde{\Omega}(\kappa d)$, with high probability

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{1}{\lambda n^2} + \frac{\kappa F_*}{n}\right).$$

7

Remark 3 The first part of Corollary 4 shows that ERM enjoys an $\widetilde{O}\left(d/n + \kappa F_*/n\right)$ risk bound for stochastic optimization of strongly convex and smooth functions. In the literature, the most comparable result is the $O(1/\lambda n)$ risk bound proved by Shalev-Shwartz et al. (2009) but with striking differences highlighted in Table 1. Since the risk bound of Shalev-Shwartz et al. (2009) is independent of the dimensionality d, it is natural to ask whether it is possible to prove a dimensionality-independent $\widetilde{O}(\kappa/n)$ bound that holds with high probability. The second part of Corollary 4 indeed provides such a bound, but under an additional condition $n = \widetilde{\Omega}(\kappa d)$.

Remark 4 The second part implies that when n is large enough, i.e., $n = \widetilde{\Omega}(\kappa d)$, the risk bound can be tightened to $O(1/[\lambda n^2] + \kappa F_*/n)$. In particular, when the minimal risk is small, i.e., $F_* = O(1/n)$, we obtain an $O(\kappa/n^2)$ bound. To the best of our knowledge, this is the first $O(1/n^2)$-type of risk bound of ERM, and even in the studies of stochastic approximation, we have not found similar theoretical guarantees. Finally, it is worth to point out the following two features of the second part:

- Although the lower bound of n depends on d, the risk bound is independent of d.
- The domain size R only appears in the lower bound of n, and the dependence is logarithmic.

Our next result shows that the individual convexity assumption, i.e., **Assumption 4(d)**, and the Lipschitz continuity assumption, i.e., **Assumptions 3**, in Theorem 3 can be relaxed. To be specific, **Assumptions 4(d)** and **3** can be replaced with **Assumption 4(c)**.

Theorem 5 *Under* **Assumptions 1**, **2**, **4(b)**, *and* **4(c)**, *with probability at least* $1 - 2\delta$, *we have*

$$
\begin{aligned}
F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) \leq & \frac{4R^2 LC(\varepsilon)}{n} + \frac{4R^2 L^2 C(\varepsilon)}{\lambda n} + \frac{4RM\log(2/\delta)}{n} + \frac{8LF_*\log(2/\delta)}{\lambda n} \\
& + \left(4RL + 2RL\sqrt{\frac{C(\varepsilon)}{n}} + \frac{2RLC(\varepsilon)}{n}\right)\varepsilon.
\end{aligned}
\tag{14}
$$

Furthermore, if

$$
n \geq \frac{25L^2 C(\varepsilon)}{\lambda^2} = 25\kappa^2 C(\varepsilon),
\tag{15}
$$

we also have

$$
F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) \leq \frac{8M^2\log^2(2/\delta)}{\lambda n^2} + \frac{32LF_*\log(2/\delta)}{\lambda n} + \left(\frac{32L^2}{\lambda} + \frac{416\lambda}{625}\right)\varepsilon^2.
\tag{16}
$$

We have the following corollary to simplify the above theorem.

Corollary 6 *By setting* $\varepsilon = 1/n$ *in Theorem 5, we have* $C(1/n) = \Theta(d\log n)$, *and with high probability*

$$
F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{\kappa d\log n}{n} + \frac{\kappa F_*}{n}\right) = \widetilde{O}\left(\frac{\kappa d}{n} + \frac{\kappa F_*}{n}\right) = \widetilde{O}\left(\frac{\kappa d}{n}\right).
$$

Setting $\varepsilon = 1/n^2$, *we have* $C(1/n^2) = \Theta(d\log n)$, *and when* $n = \Omega\left(\kappa^2 d\log n\right) = \widetilde{\Omega}(\kappa^2 d)$, *with high probability*

$$
F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) = O\left(\frac{1}{\lambda n^2} + \frac{\kappa F_*}{n}\right).
$$

8

Remark 5 Comparing the first part of Corollaries 6 and 4, we observe that without the individual convexity and Lipschitz continuity, the risk bound is increased from $\widetilde{O}\left(d/n + \kappa F_*/n\right)$ to $\widetilde{O}(\kappa d/n + \kappa F_*/n)$.

Remark 6 Comparing the second part of Corollaries 6 and 4, we can see that the risk bound is on the same order, but the lower bound of n is increased by a factor of κ. It is interesting to mention that a similar phenomenon also happens in stochastic approximation. Recently, a variance reduction technique named SVRG (Johnson and Zhang, 2013) or EMGD (Zhang et al., 2013) was proposed for stochastic optimization when both full gradients and stochastic gradients are available. In the analysis, SVRG assumes the stochastic function is convex, while EMGD does not. From their theoretical results, we observe that the individual convexity leads to a difference of κ factor in the sample complexity of stochastic gradients.

Finally, we want to mention that even when the strong convexity assumption in Theorem 5 is missing, a risk bound of $\widetilde{O}(\sqrt{d/n})$ is still attainable.

Theorem 7 *Under* **Assumptions 1**, **2**, **4(a)**, *and* **4(c)**, *with probability at least* $1 - 2\delta$, *we have*

$$
\begin{aligned}
F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) \leq & \frac{4R^2 L C(\varepsilon)}{n} + \frac{4RM \log(2/\delta)}{n} + 4R^2 L \sqrt{\frac{C(\varepsilon)}{n}} + 2R \sqrt{\frac{8LF_* \log(2/\delta)}{n}} \\
& + \left(4RL + 2RL \sqrt{\frac{C(\varepsilon)}{n}} + \frac{2RLC(\varepsilon)}{n} \right) \varepsilon.
\end{aligned}
$$

Remark 7 In a recent work, Feldman (2016) shows that SCO without the Lipschitz condition cannot be solved by ERM. Theorem 7 exhibits that as long as the random function is smooth, SCO is learnable by ERM.

3.3 Risk Bounds for Supervised Learning

If the conditions of Theorem 3 or Theorem 5 are satisfied, we can directly use them to establish an $O(1/[\lambda n^2] + \kappa F_*/n)$ risk bound for supervised learning. However, a major limitation of these theorems is that the lower bound of n depends on the dimensionality d, and thus cannot be applied to infinite dimensional cases, e.g., kernel methods (Schölkopf and Smola, 2002). In this section, we exploit the structure of supervised learning to make the theory dimensionality-independent.

We focus on the generalized linear form of supervised learning:

$$
\min_{\mathbf{w} \in \mathcal{W}} \ F(\mathbf{w}) = \mathrm{E}_{(\mathbf{x}, y) \sim \mathbb{D}} \left[\ell(\langle \mathbf{w}, \mathbf{x} \rangle, y) \right] + r(\mathbf{w}), \tag{17}
$$

where $\ell(\langle \mathbf{w}, \mathbf{x} \rangle, y)$ is the loss of predicting $\langle \mathbf{w}, \mathbf{x} \rangle$ when the true target is y, and $r(\cdot)$ is a regularizer. Given n training examples $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$ independently sampled from $\mathbb{D}$, the empirical objective is

$$
\min_{\mathbf{w} \in \mathcal{W}} \ \widehat{F}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^{n} \ell(\langle \mathbf{w}, \mathbf{x}_i \rangle, y_i) + r(\mathbf{w}).
$$

We define

$$H(\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\ell(\langle\mathbf{w},\mathbf{x}\rangle,y)\right] \text{ and } \widehat{H}(\mathbf{w}) = \frac{1}{n}\sum_{i=1}^{n}\ell(\langle\mathbf{w},\mathbf{x}_i\rangle,y_i)$$

to capture the stochastic component.

Besides **4(b)** and **4(c)**, we introduce the following additional assumptions. We abuse the same notation $\|\cdot\|$ to denote the norm induced by the inner product of a Hilbert space.

Assumption 5 *The domain $\mathcal{W}$ is a convex subset of a Hilbert space $\mathcal{H}$, and is bounded by R, that is,*

$$\|\mathbf{w}\| \leq R, \ \forall\mathbf{w}\in\mathcal{W}. \tag{18}$$

Assumption 6 *The norm of the random data $\mathbf{x}\in\mathcal{H}$ is upper bounded by a constant D, that is,*

$$\|\mathbf{x}\| \leq D, \ \forall(\mathbf{x},y)\sim\mathbb{D}. \tag{19}$$

Assumption 7 *For any $(\mathbf{x},y)\sim\mathbb{D}$, $\ell(\cdot,y)$ is nonnegative, and β-smooth over $[-DR, DR]$, that is,*

$$|\ell'(u,y) - \ell'(v,y)| \leq \beta|u-v|, \ \forall u,v\in[-DR,DR]. \tag{20}$$

Assumption 8 *The regularizer $r(\cdot)$ is P-Lipschitz continuous over $\mathcal{W}$, that is,*

$$|r(\mathbf{w}) - r(\mathbf{w}')| \leq P\|\mathbf{w}-\mathbf{w}'\|, \ \forall\mathbf{w},\mathbf{w}'\in\mathcal{W}. \tag{21}$$

Remark 8 The above assumptions allow us to model many popular losses in machine learning, such as (regularized) least squares and (regularized) logistic regression. **Assumptions 6** and **7** imply the random function $\ell(\langle\cdot,\mathbf{x}\rangle,y)$ is βD^2-smooth over $\mathcal{W}$. To see this, for any $\mathbf{w},\mathbf{w}'\in\mathcal{W}$, we have

$$\left\|\nabla\ell(\langle\mathbf{w},\mathbf{x}\rangle,y) - \nabla\ell(\langle\mathbf{w}',\mathbf{x}\rangle,y)\right\| = \left\|\ell'(\langle\mathbf{w},\mathbf{x}\rangle,y)\mathbf{x} - \ell'(\langle\mathbf{w}',\mathbf{x}\rangle,y)\mathbf{x}\right\|$$

$$\overset{(19)}{\leq} D|\ell'(\langle\mathbf{w},\mathbf{x}\rangle,y) - \ell'(\langle\mathbf{w}',\mathbf{x}\rangle,y)| \overset{(20)}{\leq} \beta D|\langle\mathbf{w},\mathbf{x}\rangle - \langle\mathbf{w}',\mathbf{x}\rangle| \overset{(19)}{\leq} \beta D^2\|\mathbf{w}-\mathbf{w}'\|.$$

By Jensen's inequality, $H(\cdot)$ is also βD^2-smooth. Notice that βD^2 is the modulus of smoothness of $H(\cdot)$, and λ is the modulus of strong convexity of $F(\cdot)$. With a slight abuse of notation, we define $L = \beta D^2$, and the condition number κ as the ratio between L and λ, i.e., $\kappa = L/\lambda$. Finally, we note that the regularizer $r(\cdot)$ could be *non-smooth*.

Recall that $\mathbf{w}_* \in \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}} F(\mathbf{w})$ and $\widehat{\mathbf{w}} \in \mathrm{argmin}_{\mathbf{w}\in\mathcal{W}} \widehat{F}(\mathbf{w})$. We have the following excess risk bound of ERM for supervised learning.

Theorem 8 *For any $0 < \delta < 1$, define*

$$M = \sup_{(\mathbf{x},y)\sim\mathbb{D}} \|\nabla\ell(\langle\mathbf{w}_*,\mathbf{x}\rangle,y)\|, \tag{22}$$

$$C = 4\left(8 + \sqrt{2\log\frac{\lceil 2\log_2(n) + \log_2(2R)\rceil}{\delta}}\right), \tag{23}$$

$$H_* = H(\mathbf{w}_*) = F(\mathbf{w}_*) - r(\mathbf{w}_*). \tag{24}$$

10

Under **Assumptions 4**(b), **4**(c), **5**, **6**, **7**, *and* **8**, *with probability at least* $1 - 2\delta$, *we have*

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) \leq \max\left(\frac{M+P}{n^2} + \frac{L}{2n^4}, \frac{4R^2L^2C^2}{\lambda n} + \frac{4RM\log(2/\delta)}{n} + \frac{8LH_*\log(2/\delta)}{\lambda n}\right).$$
(25)

Furthermore, if

$$n \geq \frac{16L^2C^2}{\lambda^2} = 16\kappa^2C^2,$$
(26)

with probability at least $1 - 2\delta$, *we have*

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) \leq \max\left(\frac{M+P}{n^2} + \frac{L}{2n^4}, \frac{8M^2\log^2(2/\delta)}{\lambda n^2} + \frac{16LH_*\log(2/\delta)}{\lambda n}\right).$$
(27)

Remark 9 The first part of Theorem 8 presents an $O(\kappa/n)$ risk bound,[3] similar to the $O(1/\lambda n)$ risk bound of Sridharan et al. (2009). The second part is an $O(1/[\lambda n^2] + \kappa H_*/n)$ risk bound, and in this case, the lower bound of n is $\Omega(\kappa^2)$, which is dimensionality-independent. Thus, Theorem 8 can be applied even when the dimensionality is infinite. Generally speaking, the regularizer $r(\cdot)$ is nonnegative, and thus $H_* \leq F_*$. So, the second bound is even better than those in Theorems 3 and 5. Finally, we note that Theorem 8 should be treated as a counterpart of Theorem 5 for supervised learning, because both of them do not rely on the individual complexity, i.e., **Assumption 4**(d). One may wonder whether it is possible to derive a counterpart of Theorem 3, that is, whether it is possible to utilize the individual convexity to reduce the lower bound of n by a factor of κ. We will investigate this question as a future work.

4. Analysis

We here present the proofs of main theorems. The omitted ones can be found in appendices.

4.1 The Key Idea

By the convexity of $\widehat{F}(\cdot)$ and the optimality condition of $\widehat{\mathbf{w}}$ (Boyd and Vandenberghe, 2004), we have

$$\langle \nabla\widehat{F}(\widehat{\mathbf{w}}), \mathbf{w} - \widehat{\mathbf{w}}\rangle \geq 0, \ \forall \mathbf{w} \in \mathcal{W}.$$
(28)

Our theoretical analysis is built upon the following inequality:

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \leq \langle \nabla F(\widehat{\mathbf{w}}), \widehat{\mathbf{w}} - \mathbf{w}_*\rangle$$
$$= \langle \nabla F(\widehat{\mathbf{w}}) - \nabla F(\mathbf{w}_*), \widehat{\mathbf{w}} - \mathbf{w}_*\rangle + \langle \nabla F(\mathbf{w}_*), \widehat{\mathbf{w}} - \mathbf{w}_*\rangle$$
$$= \langle \nabla F(\widehat{\mathbf{w}}) - \nabla F(\mathbf{w}_*) - [\nabla\widehat{F}(\widehat{\mathbf{w}}) - \nabla\widehat{F}(\mathbf{w}_*)], \widehat{\mathbf{w}} - \mathbf{w}_*\rangle$$
$$+ \langle \nabla\widehat{F}(\widehat{\mathbf{w}}) - \nabla\widehat{F}(\mathbf{w}_*) + \nabla F(\mathbf{w}_*), \widehat{\mathbf{w}} - \mathbf{w}_*\rangle$$
$$\overset{(28)}{\leq} \langle \nabla F(\widehat{\mathbf{w}}) - \nabla F(\mathbf{w}_*) - [\nabla\widehat{F}(\widehat{\mathbf{w}}) - \nabla\widehat{F}(\mathbf{w}_*)], \widehat{\mathbf{w}} - \mathbf{w}_*\rangle + \langle \nabla F(\mathbf{w}_*) - \nabla\widehat{F}(\mathbf{w}_*), \widehat{\mathbf{w}} - \mathbf{w}_*\rangle,$$
(29)

---

3. For brevity, we treat C as a constant because it only has a *double* logarithmic dependence on n.

11

where $\lambda > 0$ is the strong convexity modulus of $F(\cdot)$ if exists otherwise it is zero.

In Theorems 1, 3, 5, and 7, we utilize the covering number to upper bound the first term on the last line of (29), and thus introduce a linear dependence on the dimensionality d. In Theorem 8, we use the Rademacher complexity to upper bound it, leading to a dimensionality-independent bound. The second term on the last line of (29) is upper bounded by the concentration inequality for vectors, which produces a quantity containing F_*.

4.2 Proof of Theorem 1

We set $\lambda = 0$ in (29), and upper bound the last line as

$$
\begin{aligned}
&F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) \\
&\leq \left(\underbrace{\left\| \nabla F(\widehat{\mathbf{w}}) - \nabla F(\mathbf{w}_*) - [\nabla \widehat{F}(\widehat{\mathbf{w}}) - \nabla \widehat{F}(\mathbf{w}_*)] \right\|}_{:=A_1} + \underbrace{\left\| \nabla F(\mathbf{w}_*) - \nabla \widehat{F}(\mathbf{w}_*) \right\|}_{:=A_2} \right) \| \widehat{\mathbf{w}} - \mathbf{w}_* \| .
\end{aligned}
\tag{30}
$$

We first bound A_1. Let $\mathcal{N}(\mathcal{W}, \varepsilon)$ be the ε-net of $\mathcal{W}$ with minimal cardinality, which is referred to as the covering numbers.[4] Based on the concentration inequality of vectors (Smale and Zhou, 2007), we establish a uniform convergence of $\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}_*)$ to $\nabla \widehat{F}(\mathbf{w}) - \nabla \widehat{F}(\mathbf{w}_*)$ over any $\mathbf{w} \in \mathcal{N}(\mathcal{W}, \varepsilon)$.

Lemma 1 *Under* **Assumptions 2** *and* **4(d)***, with probability at least $1 - \delta$, for any $\mathbf{w} \in \mathcal{N}(\mathcal{W}, \varepsilon)$, we have*

$$
\left\| \nabla F(\mathbf{w}) - \nabla F(\mathbf{w}_*) - [\nabla \widehat{F}(\mathbf{w}) - \nabla \widehat{F}(\mathbf{w}_*)] \right\| \leq \frac{LC(\varepsilon)\|\mathbf{w} - \mathbf{w}_*\|}{n} + \sqrt{\frac{LC(\varepsilon)(F(\mathbf{w}) - F(\mathbf{w}_*))}{n}}.
$$

where $C(\varepsilon)$ is define in (9).

Then, we extend the uniform convergence over $\widehat{\mathbf{w}}$. From the property of ε-net, we know that there exists an point $\widetilde{\mathbf{w}} \in \mathcal{N}(\mathcal{W}, \varepsilon)$ such that $\|\widehat{\mathbf{w}} - \widetilde{\mathbf{w}}\| \leq \varepsilon$. From the smoothness of $F(\cdot)$ and $\widehat{F}(\cdot)$, we have

$$
\begin{aligned}
&\left\| \nabla F(\widehat{\mathbf{w}}) - \nabla F(\mathbf{w}_*) - [\nabla \widehat{F}(\widehat{\mathbf{w}}) - \nabla \widehat{F}(\mathbf{w}_*)] \right\| \\
&\leq \left\| \nabla F(\widetilde{\mathbf{w}}) - \nabla F(\mathbf{w}_*) - [\nabla \widehat{F}(\widetilde{\mathbf{w}}) - \nabla \widehat{F}(\mathbf{w}_*)] \right\| + 2L\varepsilon.
\end{aligned}
\tag{31}
$$

---

4. A subset $\mathcal{N} \subseteq \mathcal{K}$ is called an ε-net of $\mathcal{K}$ if for every $\mathbf{w} \in \mathcal{K}$ one can find $\widetilde{\mathbf{w}} \in \mathcal{N}$ so that $\|\mathbf{w} - \widetilde{\mathbf{w}}\| \leq \varepsilon$.

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

$$
\begin{aligned}
&\left\| \nabla F(\widehat{\mathbf{w}}) - \nabla F(\mathbf{w}_*) - [\nabla \widehat{F}(\widehat{\mathbf{w}}) - \nabla \widehat{F}(\mathbf{w}_*)] \right\| \\
\leq & \frac{LC(\varepsilon)\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|}{n} + \sqrt{\frac{LC(\varepsilon)(F(\widetilde{\mathbf{w}}) - F(\mathbf{w}_*))}{n}} + 2L\varepsilon \\
\leq & \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} + \frac{LC(\varepsilon)\varepsilon}{n} + 2L\varepsilon \\
& + \sqrt{\frac{LC(\varepsilon)(F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*))}{n}} + \sqrt{\frac{LC(\varepsilon)(|F(\widehat{\mathbf{w}}) - F(\widetilde{\mathbf{w}})|)}{n}} \\
\overset{(6)}{\leq} & \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} + \sqrt{\frac{LC(\varepsilon)(F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*))}{n}} + \frac{LC(\varepsilon)\epsilon}{n} + \sqrt{\frac{LC(\varepsilon)G\varepsilon}{n}} + 2L\varepsilon.
\end{aligned}
\tag{32}
$$

Next, we proceed to bound A_2 in (30), and develop the following lemma.

Lemma 2 *Under* **Assumption 2***, with probability at least $1 - \delta$, we have*

$$
\left\| \nabla F(\mathbf{w}_*) - \nabla \widehat{F}(\mathbf{w}_*) \right\| \leq \frac{2M \log(2/\delta)}{n} + \sqrt{\frac{8LF_* \log(2/\delta)}{n}}.
\tag{33}
$$

Substituting (32) and (33) into (30), with probability at least $1 - 2\delta$, we have

$$
\begin{aligned}
& F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) \\
\leq & \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{n} + \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \sqrt{\frac{LC(\varepsilon)(F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*))}{n}} \\
& + \frac{2M \log(2/\delta)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} + \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \sqrt{\frac{8LF_* \log(2/\delta)}{n}} \\
& + 2L\varepsilon \|\widehat{\mathbf{w}} - \mathbf{w}_*\| + \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \sqrt{\frac{LC(\varepsilon)G\varepsilon}{n}} + \frac{LC(\varepsilon)\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} \\
\overset{(35),\ (36)}{\leq} & \frac{2LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{n} + \frac{2M \log(2/\delta)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} + \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \sqrt{\frac{8LF_* \log(2/\delta)}{n}} \\
& + \frac{F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*)}{2} + 2L\varepsilon \|\widehat{\mathbf{w}} - \mathbf{w}_*\| + \frac{G\varepsilon}{2} + \frac{LC(\varepsilon)\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n}
\end{aligned}
\tag{34}
$$

where the last step is due to

$$
\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \sqrt{\frac{LC(\varepsilon)(F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*))}{n}} \leq \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{2n} + \frac{F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*)}{2},
\tag{35}
$$

$$
\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \sqrt{\frac{LC(\varepsilon)G\varepsilon}{n}} \leq \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{2n} + \frac{G\varepsilon}{2}.
\tag{36}
$$

13

From (34), we get

$$\frac{1}{2}\left(F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*)\right)$$

$$\leq \frac{2LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{n} + \frac{2M\log(2/\delta)\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} + \|\widehat{\mathbf{w}} - \mathbf{w}_*\|\sqrt{\frac{8LF_*\log(2/\delta)}{n}}$$

$$+ 2L\varepsilon\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\| + \frac{G\varepsilon}{2} + \frac{LC(\varepsilon)\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n}$$

$$\overset{(4)}{\leq} \frac{8R^2LC(\varepsilon)}{n} + \frac{4RM\log(2/\delta)}{n} + 4R\sqrt{\frac{2LF_*\log(2/\delta)}{n}} + \left(4RL + \frac{G}{2} + \frac{2RLC(\varepsilon)}{n}\right)\varepsilon,$$

which implies (10).

4.3 Proof of Lemma 1

We introduce Lemma 2 of Smale and Zhou (2007).

Lemma 3 *Let $\mathcal{H}$ be a Hilbert space and let ξ be a random variable with values in $\mathcal{H}$. Assume $\|\xi\| \leq M < \infty$ almost surely. Denote $\sigma^2(\xi) = \mathrm{E}\left[\|\xi\|^2\right]$. Let $\{\xi_i\}_{i=1}^m$ be m $(m < \infty)$ independent drawers of ξ. For any $0 < \delta < 1$, with confidence $1 - \delta$,*

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

We first consider a fixed $\mathbf{w} \in \mathcal{N}(\mathcal{W}, \varepsilon)$. Since $f_i(\cdot)$ is L-smooth, we have

$$\|\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}_*)\| \overset{(5)}{\leq} L\|\mathbf{w} - \mathbf{w}_*\|. \tag{37}$$

Because $f_i(\cdot)$ is both convex and L-smooth, by (2.1.7) of Nesterov (2004), we have

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

Taking expectation over both sides, we have

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

where the last inequality follows from the optimality condition of $\mathbf{w}_*$, i.e.,

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

Following Lemma 3, with probability at least $1 - \delta$, we have

$$\left\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}_*) - [\nabla \widehat{F}(\mathbf{w}) - \nabla \widehat{F}(\mathbf{w}_*)]\right\|$$

$$= \left\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}_*) - \frac{1}{n}\sum_{i=1}^n [\nabla f_i(\mathbf{w}) - \nabla f_i(\mathbf{w}_*)]\right\|$$

$$\leq \frac{2L\|\mathbf{w} - \mathbf{w}_*\|\log(2/\delta)}{n} + \sqrt{\frac{2L(F(\mathbf{w}) - F(\mathbf{w}_*))\log(2/\delta)}{n}}.$$

14

We obtain Lemma 1 by taking the union bound over all $\mathbf{w} \in \mathcal{N}(\mathcal{W}, \varepsilon)$. To this end, we need an upper bound of the covering number $|\mathcal{N}(\mathcal{W}, \varepsilon)|$.

Let $\mathcal{B}$ be an unit ball of d dimension, and $\mathcal{N}(\mathcal{B}, \varepsilon)$ be its ε-net with minimal cardinality. According to a standard volume comparison argument (Pisier, 1989), we have

$$\log |\mathcal{N}(\mathcal{B}, \varepsilon)| \leq d \log \frac{3}{\varepsilon}.$$

Let $\mathcal{B}(R)$ be a ball centered at origin with radius R. Since we assume $\mathcal{W} \subseteq \mathcal{B}(R)$, it follows that

$$\log |\mathcal{N}(\mathcal{W}, \varepsilon)| \leq \log \left| \mathcal{N}\left(\mathcal{B}(R), \frac{\varepsilon}{2}\right) \right| \leq d \log \frac{6R}{\varepsilon}$$

where the first inequality is because the covering numbers are (almost) increasing by inclusion (Plan and Vershynin, 2013, (3.2)).

4.4 Proof of Lemma 2

To apply Lemma 3, we need an upper bound of $\mathrm{E}\left[\|\nabla f_i(\mathbf{w}_*)\|^2\right]$. Since $f_i(\cdot)$ is L-smooth and nonnegative, from Lemma 4.1 of Srebro et al. (2010), we have

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

and thus

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

From the definition in (8), we have $\|\nabla f_i(\mathbf{w}_*)\| \leq M$. Then, according to Lemma 3, with probability at least $1 - \delta$, we have

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

4.5 Proof of Theorem 3

The proof follows the same logic as that of Theorem 1. Under **Assumption 4(b)**, (30) becomes

$$
\begin{aligned}
&F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \\
&\leq \left(\underbrace{\left\|\nabla F(\widehat{\mathbf{w}}) - \nabla F(\mathbf{w}_*) - [\nabla \widehat{F}(\widehat{\mathbf{w}}) - \nabla \widehat{F}(\mathbf{w}_*)]\right\|}_{:=A_1} + \underbrace{\left\|\nabla F(\mathbf{w}_*) - \nabla \widehat{F}(\mathbf{w}_*)\right\|}_{:=A_2} \right) \|\widehat{\mathbf{w}} - \mathbf{w}_*\|.
\end{aligned}
$$

(38)

Substituting (32) and (33) into (38), with probability at least $1 - 2\delta$, we have

$$
\begin{aligned}
&F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \\
&\leq \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{n} + \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \sqrt{\frac{LC(\varepsilon)(F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*))}{n}} \\
&+ \frac{2M\log(2/\delta)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} + \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \sqrt{\frac{8LF_*\log(2/\delta)}{n}} \\
&+ 2L\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\| + \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \sqrt{\frac{LC(\varepsilon)G\varepsilon}{n}} + \frac{LC(\varepsilon)\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n}
\end{aligned}
\tag{39}
$$

To prove (11), we substitute (35), (36), and

$$
\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \sqrt{\frac{8LF_*\log(2/\delta)}{n}} \leq \frac{4LF_*\log(2/\delta)}{\lambda n} + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2
$$

into (39), and then obtain

$$
\begin{aligned}
&\frac{1}{2}\left(F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*)\right) \\
&\leq \frac{2LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{n} + \frac{2M\log(2/\delta)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} + \frac{4LF_*\log(2/\delta)}{\lambda n} \\
&+ 2L\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\| + \frac{G\varepsilon}{2} + \frac{LC(\varepsilon)\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} \\
&\overset{(4)}{\leq} \frac{8R^2LC(\varepsilon)}{n} + \frac{4RM\log(2/\delta)}{n} + \frac{4LF_*\log(2/\delta)}{\lambda n} + \left(4RL + \frac{G}{2} + \frac{2RLC(\varepsilon)}{n}\right)\varepsilon.
\end{aligned}
$$

which implies (11).

To prove (13), we substitute

$$
\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \sqrt{\frac{LC(\varepsilon)(F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*))}{n}} \leq \frac{2LC(\varepsilon)(F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*))}{\lambda n} + \frac{\lambda}{8}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2,
$$

$$
\frac{2M\log(2/\delta)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} \leq \frac{16M^2\log^2(2/\delta)}{\lambda n^2} + \frac{\lambda}{16}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2,
$$

$$
\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \sqrt{\frac{8LF_*\log(2/\delta)}{n}} \leq \frac{64LF_*\log(2/\delta)}{\lambda n} + \frac{\lambda}{32}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2,
$$

$$
2L\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \leq \frac{64L^2\varepsilon^2}{\lambda} + \frac{\lambda}{64}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2,
$$

$$
\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \sqrt{\frac{LC(\varepsilon)G\varepsilon}{n}} \leq \frac{32LC(\varepsilon)G\varepsilon}{\lambda n} + \frac{\lambda}{128}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2,
$$

$$
\frac{LC(\varepsilon)\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} \leq \frac{32L^2C^2(\varepsilon)\varepsilon^2}{\lambda n^2} + \frac{\lambda}{128}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2
$$

16

into (39), and then obtain

$$
\begin{aligned}
&F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{4}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \\
\leq& \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{n} + \frac{2LC(\varepsilon)(F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*))}{\lambda n} + \frac{16M^2\log^2(2/\delta)}{\lambda n^2} + \frac{64LF_*\log(2/\delta)}{\lambda n} \\
&+ \frac{64L^2\varepsilon^2}{\lambda} + \frac{32LC(\varepsilon)G\varepsilon}{\lambda n} + \frac{32L^2C^2(\varepsilon)\varepsilon^2}{\lambda n^2} \\
\overset{(12)}{\leq}& \frac{\lambda}{4}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 + \frac{1}{2}\left(F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*)\right) + \frac{16M^2\log^2(2/\delta)}{\lambda n^2} + \frac{64LF_*\log(2/\delta)}{\lambda n} \\
&+ \frac{64L^2\varepsilon^2}{\lambda} + 8G\varepsilon + 2\lambda\varepsilon^2
\end{aligned}
$$

which implies (13).

4.6 Proof of Theorem 5

Without **Assumption 4(d)**, Lemma 1 which is used in the proofs of Theorems 1 and 3 does not hold anymore. Instead, we will use the following version that only relies on the smoothness condition.

Lemma 4 *Under **Assumption 2**, with probability at least $1 - \delta$, for any $\mathbf{w} \in \mathcal{N}(\mathcal{W}, \varepsilon)$, we have*

$$
\left\|\nabla F(\mathbf{w}) - \nabla F(\mathbf{w}_*) - [\nabla\widehat{F}(\mathbf{w}) - \nabla\widehat{F}(\mathbf{w}_*)]\right\| \leq \frac{LC(\varepsilon)\|\mathbf{w} - \mathbf{w}_*\|}{n} + L\|\mathbf{w} - \mathbf{w}_*\|\sqrt{\frac{C(\varepsilon)}{n}}
$$

where $C(\varepsilon)$ is define in (9).

The above lemma is a direct consequence of (37), Lemma 3 and the union bound.

The rest of the proof is similar to those of Theorems 1 and 3. We first derive a counterpart of (32) under Lemma 4. Combining (31) with Lemma 4, with probability at least $1 - \delta$, we have

$$
\begin{aligned}
&\left\|\nabla F(\widehat{\mathbf{w}}) - \nabla F(\mathbf{w}_*) - [\nabla\widehat{F}(\widehat{\mathbf{w}}) - \nabla\widehat{F}(\mathbf{w}_*)]\right\| \\
\leq& \frac{LC(\varepsilon)\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|}{n} + L\|\widetilde{\mathbf{w}} - \mathbf{w}_*\|\sqrt{\frac{C(\varepsilon)}{n}} + 2L\varepsilon \\
\leq& \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} + L\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\sqrt{\frac{C(\varepsilon)}{n}} + \frac{LC(\varepsilon)\varepsilon}{n} + L\varepsilon\sqrt{\frac{C(\varepsilon)}{n}} + 2L\varepsilon.
\end{aligned} \tag{40}
$$

Substituting (40) and (33) into (38), with probability at least $1 - 2\delta$, we have

$$
\begin{aligned}
&F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \\
\leq& \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{n} + L\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2\sqrt{\frac{C(\varepsilon)}{n}} \\
&+ \frac{2M\log(2/\delta)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} + \|\widehat{\mathbf{w}} - \mathbf{w}_*\|\sqrt{\frac{8LF_*\log(2/\delta)}{n}} \\
&+ 2L\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\| + L\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\sqrt{\frac{C(\varepsilon)}{n}} + \frac{LC(\varepsilon)\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n}.
\end{aligned} \tag{41}
$$

17

To get (14), we substitute

$$L\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \sqrt{\frac{C(\varepsilon)}{n}} \leq \frac{L^2 C(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\lambda n} + \frac{\lambda}{4}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2,$$

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\sqrt{\frac{8LF_* \log(2/\delta)}{n}} \leq \frac{8LF_* \log(2/\delta)}{\lambda n} + \frac{\lambda}{4}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

into (41), and then obtain

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*)$$

$$\leq \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{n} + \frac{L^2 C(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\lambda n} + \frac{2M \log(2/\delta)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} + \frac{8LF_* \log(2/\delta)}{\lambda n}$$

$$+ 2L\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\| + L\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\sqrt{\frac{C(\varepsilon)}{n}} + \frac{LC(\varepsilon)\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n}$$

$$\overset{(4)}{\leq} \frac{4R^2 LC(\varepsilon)}{n} + \frac{4R^2 L^2 C(\varepsilon)}{\lambda n} + \frac{4RM \log(2/\delta)}{n} + \frac{8LF_* \log(2/\delta)}{\lambda n}$$

$$+ \left(4RL + 2RL\sqrt{\frac{C(\varepsilon)}{n}} + \frac{2RLC(\varepsilon)}{n}\right)\varepsilon$$

which proves (14).

To get (16), we substitute

$$\frac{2M \log(2/\delta)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} \leq \frac{8M^2 \log^2(2/\delta)}{\lambda n^2} + \frac{\lambda}{8}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2,$$

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\sqrt{\frac{8LF_* \log(2/\delta)}{n}} \leq \frac{32LF_* \log(2/\delta)}{\lambda n} + \frac{\lambda}{16}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2,$$

$$2L\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \leq \frac{32L^2\varepsilon^2}{\lambda} + \frac{\lambda}{32}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2,$$

$$L\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\sqrt{\frac{C(\varepsilon)}{n}} \leq \frac{16L^2 C(\varepsilon)\varepsilon^2}{\lambda n} + \frac{\lambda}{64}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2,$$

$$\frac{LC(\varepsilon)\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} \leq \frac{16L^2 C^2(\varepsilon)\varepsilon^2}{\lambda n^2} + \frac{\lambda}{64}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

into (41), and then obtain

$$F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{4}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

$$\leq \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{n} + L\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2\sqrt{\frac{C(\varepsilon)}{n}} + \frac{8M^2 \log^2(2/\delta)}{\lambda n^2} + \frac{32LF_* \log(2/\delta)}{\lambda n}$$

$$+ \left(\frac{32L^2}{\lambda} + \frac{16L^2 C(\varepsilon)}{\lambda n} + \frac{16L^2 C^2(\varepsilon)}{\lambda n^2}\right)\varepsilon^2$$

$$\overset{(15)}{\leq} \frac{\lambda^2\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{25L} + \frac{\lambda}{5}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 + \frac{8M^2 \log^2(2/\delta)}{\lambda n^2} + \frac{32LF_* \log(2/\delta)}{\lambda n}$$

$$+ \left(\frac{32L^2}{\lambda} + \frac{16\lambda}{25} + \frac{16\lambda^3}{625L^2}\right)\varepsilon^2$$

$$\overset{\lambda/L \leq 1}{\leq} \frac{6\lambda}{25}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 + \frac{8M^2 \log^2(2/\delta)}{\lambda n^2} + \frac{32LF_* \log(2/\delta)}{\lambda n} + \left(\frac{32L^2}{\lambda} + \frac{416\lambda}{625}\right)\varepsilon^2.$$

18

By subtracting $\lambda \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2/4$ from both sides we complete the proof of (16).

4.7 Proof of Theorem 8

We consider two cases. In the first case, we assume that

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \leq \frac{1}{n^2}.$$

Since $H(\cdot)$ is L-smooth and $r(\cdot)$ is P-Lipschitz continuous, we have

$$
\begin{aligned}
F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) &= H(\widehat{\mathbf{w}}) + r(\widehat{\mathbf{w}}) - H(\mathbf{w}_*) - r(\mathbf{w}_*) \\
&\leq \langle \widehat{\mathbf{w}} - \mathbf{w}_*, \nabla H(\mathbf{w}_*) \rangle + \frac{L}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 + P\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \\
&\leq \|\widehat{\mathbf{w}} - \mathbf{w}_*\|\|\nabla H(\mathbf{w}_*)\| + \frac{L}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 + P\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \leq \frac{M+P}{n^2} + \frac{L}{2n^4}
\end{aligned}
\tag{42}
$$

where the last step utilizes Jensen's inequality

$$\|\nabla H(\mathbf{w}_*)\| = \left\|\mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\nabla\ell(\langle\mathbf{w}_*,\mathbf{x}\rangle,y)\right]\right\| \leq \mathrm{E}_{(\mathbf{x},y)\sim\mathbb{D}}\left[\|\nabla\ell(\langle\mathbf{w}_*,\mathbf{x}\rangle,y)\|\right] \overset{(22)}{\leq} M.$$

Next, we study the case

$$\frac{1}{n^2} < \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \overset{(18)}{\leq} 2R.$$

From (29), we have

$$
\begin{aligned}
&F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \\
\leq &\langle \nabla F(\widehat{\mathbf{w}}) - \nabla F(\mathbf{w}_*) - [\nabla\widehat{F}(\widehat{\mathbf{w}}) - \nabla\widehat{F}(\mathbf{w}_*)], \widehat{\mathbf{w}} - \mathbf{w}_* \rangle + \langle \nabla F(\mathbf{w}_*) - \nabla\widehat{F}(\mathbf{w}_*), \widehat{\mathbf{w}} - \mathbf{w}_* \rangle \\
= &\langle \nabla H(\widehat{\mathbf{w}}) - \nabla H(\mathbf{w}_*) - [\nabla\widehat{H}(\widehat{\mathbf{w}}) - \nabla\widehat{H}(\mathbf{w}_*)], \widehat{\mathbf{w}} - \mathbf{w}_* \rangle + \langle \nabla H(\mathbf{w}_*) - \nabla\widehat{H}(\mathbf{w}_*), \widehat{\mathbf{w}} - \mathbf{w}_* \rangle \\
\leq &\underbrace{\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\|\widehat{\mathbf{w}}-\mathbf{w}_*\|} \left\langle \nabla H(\mathbf{w}) - \nabla H(\mathbf{w}_*) - [\nabla\widehat{H}(\mathbf{w}) - \nabla\widehat{H}(\mathbf{w}_*)], \mathbf{w} - \mathbf{w}_* \right\rangle}_{:=B_1} \\
&+ \underbrace{\left\|\nabla H(\mathbf{w}_*) - \nabla\widehat{H}(\mathbf{w}_*)\right\|}_{:=B_2} \|\widehat{\mathbf{w}} - \mathbf{w}_*\|.
\end{aligned}
\tag{43}
$$

We first bound B_1. To utilize the fact the random variable $\|\widehat{\mathbf{w}} - \mathbf{w}_*\|$ lies in the range $(1/n^2, 2R]$, we develop the following lemma.

Lemma 5 *Under* **Assumptions 6** *and* **7**, *with probability at least* $1 - \delta$, *for all*

$$\frac{1}{n^2} < \gamma \leq 2R$$

the following bound holds:

$$\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma} \left\langle \nabla H(\mathbf{w}) - \nabla H(\mathbf{w}_*) - [\nabla\widehat{H}(\mathbf{w}) - \nabla\widehat{H}(\mathbf{w}_*)], \mathbf{w} - \mathbf{w}_* \right\rangle \leq \frac{4L\gamma^2}{\sqrt{n}}\left(8 + \sqrt{2\log\frac{s}{\delta}}\right)$$

where $s = \lceil 2\log_2(n) + \log_2(2R)\rceil$.

19

Based on the above lemma, we have with probability at least $1 - \delta$,

$$B_1 \leq \frac{4L\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\sqrt{n}}\left(8 + \sqrt{2\log\frac{s}{\delta}}\right) = \frac{LC\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\sqrt{n}} \tag{44}$$

where C is defined in (23).

We then proceed to handle B_2, which can be upper bounded in the same way as A_2. In particular, we have the following lemma.

Lemma 6 *Under* **Assumptions 6** *and* **7**, *with probability at least $1 - \delta$, we have*

$$\left\|\nabla H(\mathbf{w}_*) - \nabla \widehat{H}(\mathbf{w}_*)\right\| \leq \frac{2M\log(2/\delta)}{n} + \sqrt{\frac{8LH_*\log(2/\delta)}{n}}. \tag{45}$$

Substituting (44) and (45) into (43), with probability at least $1 - 2\delta$, we have

$$
\begin{aligned}
&F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \\
&\leq \frac{LC\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\sqrt{n}} + \frac{2M\log(2/\delta)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} + \|\widehat{\mathbf{w}} - \mathbf{w}_*\|\sqrt{\frac{8LH_*\log(2/\delta)}{n}}.
\end{aligned}
\tag{46}
$$

We substitute

$$\frac{LC\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\sqrt{n}} \leq \frac{L^2C^2\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\lambda n} + \frac{\lambda}{4}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2,$$

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\sqrt{\frac{8LH_*\log(2/\delta)}{n}} \leq \frac{8LH_*\log(2/\delta)}{\lambda n} + \frac{\lambda}{4}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

into (46), and then have

$$
\begin{aligned}
F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) &\leq \frac{L^2C^2\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\lambda n} + \frac{2M\log(2/\delta)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} + \frac{8LH_*\log(2/\delta)}{\lambda n} \\
&\overset{(4)}{\leq} \frac{4R^2L^2C^2}{\lambda n} + \frac{4RM\log(2/\delta)}{n} + \frac{8LH_*\log(2/\delta)}{\lambda n}.
\end{aligned}
$$

Combining the above inequality with (42), we obtain (25).

To prove (27), we substitute

$$\frac{2M\log(2/\delta)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} \leq \frac{8M^2\log^2(2/\delta)}{\lambda n^2} + \frac{\lambda}{8}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2,$$

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\sqrt{\frac{8LH_*\log(2/\delta)}{n}} \leq \frac{16LH_*\log(2/\delta)}{\lambda n} + \frac{\lambda}{8}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

into (46), and then have

$$
\begin{aligned}
&F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) + \frac{\lambda}{4}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \\
&\leq \frac{LC\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\sqrt{n}} + \frac{8M^2\log^2(2/\delta)}{\lambda n^2} + \frac{16LH_*\log(2/\delta)}{\lambda n} \\
&\overset{(26)}{\leq} \frac{\lambda}{4}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 + \frac{8M^2\log^2(2/\delta)}{\lambda n^2} + \frac{16LH_*\log(2/\delta)}{\lambda n}.
\end{aligned}
$$

Combining the above inequality with (42), we obtain (27).

5. Conclusions and Future work

In this paper, we study the excess risk of ERM for SCO. Our theoretical results show that it is possible to achieve $O(1/n)$-type of risk bounds under (i) the smoothness and small minimal risk conditions (i.e., Theorem 1) or (ii) the smoothness and strong convexity conditions (i.e., the first part of Theorems 3, 5, and 8). A more exciting result is that when n is large enough, ERM has $O(1/n^2)$-type of risk bounds under the smoothness, strong convexity, and small minimal risk conditions (i.e., the second part of Theorems 3, 5, and 8).

In the context of SCO, there remain many open problems about ERM.

1. Our current results are restricted to the Hilbert or Euclidean space, because the smoothness and strong convexity are defined in terms of the ℓ_2-norm. We will extend our analysis to other geometries in the future.
2. As mentioned in **Remark 3**, under the strong convexity condition, a dimensionality-independent risk bound, e.g., $\widetilde{O}(\kappa/n)$ or $\widetilde{O}(1/\lambda n)$, that holds with high probability is still missing.
3. As discussed in **Remark 9**, it is unclear whether the convexity of the loss can be exploited to improve the lower bound of n in the second part of Theorem 8. Ideally, we expect that $n = \Omega(\kappa)$ is sufficient to deliver an $O(1/[\lambda n^2] + \kappa H_*/n)$ risk bound.
4. The $O(1/n^2)$-type of risk bounds require both the smoothness and strong convexity conditions. One may investigate whether strong convexity can be relaxed to other weaker conditions, such as exponential concavity.

Finally, as far as we know, there are no $O(1/n^2)$-type of risk bounds for stochastic approximation (SA). We will try to establish such bounds for SA.

References

Alekh Agarwal, Peter L. Bartlett, Pradeep Ravikumar, and Martin J. Wainwright. Information-theoretic lower bounds on the oracle complexity of stochastic convex optimization. *IEEE Transactions on Information Theory*, 58(5):3235–3249, 2012.

Francis Bach and Eric Moulines. Non-strongly-convex smooth stochastic approximation with convergence rate $O(1/n)$. In *Advances in Neural Information Processing Systems 26*, pages 773–781. 2013.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Peter L. Bartlett, Olivier Bousquet, and Shahar Mendelson. Local rademacher complexities. *The Annals of Statistics*, 33(4):1497–1537, 2005.

Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

Giulia Desalvo, Mehryar Mohri, and Umar Syed. Learning with deep cascades. In *Proceedings of the 26th International Conference on Algorithmic Learning Theory*, pages 254–269, 2015.

Vitaly Feldman. Generalization of erm in stochastic convex optimization: The dimension strikes back. *ArXiv e-prints*, arXiv:1608.04414, 2016.

Alon Gonen and Shai Shalev-Shwartz. Average stability is invariant to data preconditioning. implications to exp-concave empirical risk minimization. *ArXiv e-prints*, arXiv:1601.04011, 2016.

David Haussler, Michael Kearns, Nick Littlestone, and Manfred K. Warmuth. Equivalence of models for polynomial learnability. *Information and Computation*, 95(2):129–161, 1991.

Elad Hazan and Satyen Kale. Beyond the regret minimization barrier: an optimal algorithm for stochastic strongly-convex optimization. In *Proceedings of the 24th Annual Conference on Learning Theory*, pages 421–436, 2011.

Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems 26*, pages 315–323, 2013.

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. Springer, 2011.

Tomer Koren and Kfir Levy. Fast rates for exp-concave empirical risk minimization. In *Advances in Neural Information Processing Systems 28*, pages 1477–1485. 2015.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, second edition, 2003.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer, 1991.

Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. The importance of convexity in learning with squared loss. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pages 140–146, 1996.

Mehrdad Mahdavi, Lijun Zhang, and Rong Jin. Lower and upper bounds on the generalization of stochastic exponentially concave optimization. In *Proceedings of the 28th Conference on Learning Theory*, 2015.

Colin McDiarmid. On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188, 1989.

Nishant A. Mehta. Fast rates with high probability in exp-concave statistical learning. *ArXiv e-prints*, arXiv:1605.01288, 2016.

Ron Meir and Tong Zhang. Generalization error bounds for bayesian mixture algorithms. *Journal of Machine Learning Research*, 4:839–860, 2003.

Eric Moulines and Francis R. Bach. Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems 24*, pages 451–459. 2011.

A. Nemirovski and D. B. Yudin. *Problem complexity and method efficiency in optimization.* John Wiley & Sons Ltd, 1983.

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

Yurii Nesterov. *Introductory lectures on convex optimization: a basic course*, volume 87 of *Applied optimization.* Kluwer Academic Publishers, 2004.

Dmitriy Panchenko. Some extensions of an inequality of vapnik and chervonenkis. *Electronic Communications in Probability*, 7:55–65, 2002.

Gilles Pisier. *The volume of convex bodies and Banach space geometry.* Cambridge Tracts in Mathematics (No. 94). Cambridge University Press, 1989.

Yaniv Plan and Roman Vershynin. One-bit compressed sensing by linear programming. *Communications on Pure and Applied Mathematics*, 66(8):1275–1297, 2013.

Alexander Rakhlin, Ohad Shamir, and Karthik Sridharan. Making gradient descent optimal for strongly convex stochastic optimization. In *Proceedings of the 29th International Conference on Machine Learning*, pages 449–456, 2012.

Bernhard Schölkopf and Alexander J. Smola. *Learning with kernels : support vector machines, regularization, optimization, and beyond.* MIT Press, 2002.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms.* Cambridge University Press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on Stochastic Programming: Modeling and Theory.* SIAM, second edition, 2014.

Steve Smale and Ding-Xuan Zhou. Learning theory estimates via integral operators and their approximations. *Constructive Approximation*, 26(2):153–172, 2007.

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

Karthik Sridharan, Shai Shalev-shwartz, and Nathan Srebro. Fast rates for regularized objectives. In *Advances in Neural Information Processing Systems 21*, pages 1545–1552, 2009.

Alexandre B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *The Annals of Statistics*, 32:135–166, 2004.

Vladimir Vapnik. *The Nature of Statistical Learning Theory*. Springer, second edition edition, 2000.

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

Lijun Zhang, Mehrdad Mahdavi, and Rong Jin. Linear convergence with condition number independent access of full gradients. In *Advance in Neural Information Processing Systems 26*, pages 980–988, 2013.

Appendix A. Proof of Theorem 7

This result is actually a byproduct of Theorem 5. Since strong convexity is absent, we set $\lambda = 0$ in (41) and obtain

$$
\begin{aligned}
&F(\widehat{\mathbf{w}}) - F(\mathbf{w}_*) \\
&\leq \frac{LC(\varepsilon)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{n} + L\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \sqrt{\frac{C(\varepsilon)}{n}} \\
&\quad + \frac{2M\log(2/\delta)\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} + \|\widehat{\mathbf{w}} - \mathbf{w}_*\|\sqrt{\frac{8LF_*\log(2/\delta)}{n}} \\
&\quad + 2L\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\| + L\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\sqrt{\frac{C(\varepsilon)}{n}} + \frac{LC(\varepsilon)\varepsilon\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{n} \\
&\overset{(4)}{\leq} \frac{4R^2LC(\varepsilon)}{n} + \frac{4RM\log(2/\delta)}{n} + 4R^2L\sqrt{\frac{C(\varepsilon)}{n}} + 2R\sqrt{\frac{8LF_*\log(2/\delta)}{n}} \\
&\quad + \left(4RL + 2RL\sqrt{\frac{C(\varepsilon)}{n}} + \frac{2RLC(\varepsilon)}{n}\right)\varepsilon.
\end{aligned}
$$

Appendix B. Proof of Lemma 5

First, we partition the range $(1/n^2, 2R]$ into $s = \lceil 2\log_2(n) + \log_2(2R)\rceil$ consecutive segments $\Delta_1, \Delta_2, \ldots, \Delta_s$ such that

$$
\Delta_k = \left(\underbrace{\frac{2^{k-1}}{n^2}}_{:=\gamma_k^-}, \underbrace{\frac{2^k}{n^2}}_{:=\gamma_k^+}\right], \quad k = 1, \ldots, s.
$$

Then, we consider the case $\gamma \in \Delta_k$ for a fixed value of k. We have

$$
\begin{aligned}
&\sup_{\mathbf{w}:\|\mathbf{w} - \mathbf{w}_*\| \leq \gamma} \left\langle \nabla H(\mathbf{w}) - \nabla H(\mathbf{w}_*) - [\nabla\widehat{H}(\mathbf{w}) - \nabla\widehat{H}(\mathbf{w}_*)], \mathbf{w} - \mathbf{w}_* \right\rangle \\
&\leq \sup_{\mathbf{w}:\|\mathbf{w} - \mathbf{w}_*\| \leq \gamma_k^+} \left\langle \nabla H(\mathbf{w}) - \nabla H(\mathbf{w}_*) - [\nabla\widehat{H}(\mathbf{w}) - \nabla\widehat{H}(\mathbf{w}_*)], \mathbf{w} - \mathbf{w}_* \right\rangle.
\end{aligned} \tag{47}
$$

Based on the McDiarmid's inequality (McDiarmid, 1989) and the Rademacher complexity (Bartlett and Mendelson, 2002), we have the following lemma to upper bound the last term.

Lemma 7 *Under* **Assumptions 6** *and* **7**, *with probability at least* $1 - \delta$, *we have*

$$\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \left\langle \nabla H(\mathbf{w}) - \nabla H(\mathbf{w}_*) - [\nabla\widehat{H}(\mathbf{w}) - \nabla\widehat{H}(\mathbf{w}_*)], \mathbf{w} - \mathbf{w}_* \right\rangle$$

$$\leq \frac{L\left(\gamma_k^+\right)^2}{\sqrt{n}}\left(8 + \sqrt{2\log\frac{1}{\delta}}\right). \tag{48}$$

Since $\gamma \in \Delta_k$, we have

$$\gamma_k^+ = 2\gamma_k^- \leq 2\gamma. \tag{49}$$

Thus, with probability at least $1 - \delta$, we have

$$\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma} \left\langle \nabla H(\mathbf{w}) - \nabla H(\mathbf{w}_*) - [\nabla\widehat{H}(\mathbf{w}) - \nabla\widehat{H}(\mathbf{w}_*)], \mathbf{w} - \mathbf{w}_* \right\rangle$$

$$\overset{(47),(48),(49)}{\leq} \frac{4L\gamma^2}{\sqrt{n}}\left(8 + \sqrt{2\log\frac{1}{\delta}}\right).$$

We complete the proof by taking the union bound over s segments.

Appendix C. Proof of Lemma 7

To simplify the notation, we define

$$h_i(\mathbf{w}) = \ell(\langle\mathbf{w}, \mathbf{x}_i\rangle, y_i), \ \ i = 1, \ldots, n,$$

$$l(h_1, \ldots, h_n) = \sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \left\langle \nabla H(\mathbf{w}) - \nabla H(\mathbf{w}_*) - \frac{1}{n}\sum_{i=1}^n[\nabla h_i(\mathbf{w}) - \nabla h_i(\mathbf{w}_*)], \mathbf{w} - \mathbf{w}_* \right\rangle.$$

To upper bound $l(h_1, \ldots, h_n)$, we utilize the McDiarmid's inequality (McDiarmid, 1989).

Theorem 9 *Let* $X_1, \ldots, X_n$ *be independent random variables taking values in a set* A, *and assume that* $f : A^n \mapsto \mathbb{R}$ *satisfies*

$$\sup_{x_1,\ldots,x_n,x_i'\in A} \left|H(x_1, \ldots, x_n) - H(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n)\right| \leq c_i$$

for every $1 \leq i \leq n$. *Then, for every* $t > 0$,

$$P\{H(X_1, \ldots, X_n) - \mathrm{E}\left[H(X_1, \ldots, X_n)\right] \geq t\} \leq \exp\left(-\frac{2t^2}{\sum_{i=1}^n c_i^2}\right).$$

As pointed out in Remark 7, **Assumptions 6** and **7** imply the random function $h_i(\cdot)$ is L-smooth, and thus

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

25

As a result, when a random function h_i changes, the random variable $l(h_1, \ldots, h_n)$ can change by no more than $2L\left(\gamma_k^+\right)^2 / n$. McDiarmid's inequality implies that with probability at least $1 - \delta$

$$l(h_1, \ldots, h_n) \leq \mathrm{E}\left[l(h_1, \ldots, h_n)\right] + L\left(\gamma_k^+\right)^2 \sqrt{\frac{2}{n} \log \frac{1}{\delta}}. \tag{50}$$

Let $(h_1', \ldots, h_n')$ be an independent copy of $(h_1, \ldots, h_n)$, and $\epsilon_1, \ldots, \epsilon_n$ be n i.i.d. Rademacher variables with equal probability of being ± 1. Using techniques of Rademacher complexities (Bartlett and Mendelson, 2002), we bound $\mathrm{E}\left[l(h_1, \ldots, h_n)\right]$ as follows:

$$\mathrm{E}_{h_1, \ldots, h_n}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\| \leq \gamma_k^+} \left\langle \nabla H(\mathbf{w}) - \nabla H(\mathbf{w}_*) - \frac{1}{n}\sum_{i=1}^n [\nabla h_i(\mathbf{w}) - \nabla h_i(\mathbf{w}_*)], \mathbf{w} - \mathbf{w}_* \right\rangle\right]$$

$$= \frac{1}{n}\mathrm{E}_{h_1, \ldots, h_n}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\| \leq \gamma_k^+}\right.$$

$$\left.\mathrm{E}_{h_1', \ldots, h_n'}\left[\sum_{i=1}^n \left\langle \nabla h_i'(\mathbf{w}) - \nabla h_i'(\mathbf{w}_*), \mathbf{w} - \mathbf{w}_* \right\rangle\right] - \sum_{i=1}^n \left\langle \nabla h_i(\mathbf{w}) - \nabla h_i(\mathbf{w}_*), \mathbf{w} - \mathbf{w}_* \right\rangle\right]$$

$$\leq \frac{1}{n}\mathrm{E}_{h_1, \ldots, h_n, h_1', \ldots, h_n'}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\| \leq \gamma_k^+}\right.$$

$$\left.\sum_{i=1}^n \left\langle \nabla h_i'(\mathbf{w}) - \nabla h_i'(\mathbf{w}_*), \mathbf{w} - \mathbf{w}_* \right\rangle - \sum_{i=1}^n \left\langle \nabla h_i(\mathbf{w}) - \nabla h_i(\mathbf{w}_*), \mathbf{w} - \mathbf{w}_* \right\rangle\right]$$

$$= \frac{1}{n}\mathrm{E}_{h_1, \ldots, h_n, h_1', \ldots, h_n', \epsilon_1, \ldots, \epsilon_n}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\| \leq \gamma_k^+}\right.$$

$$\left.\sum_{i=1}^n \epsilon_i \left(\left\langle \nabla h_i'(\mathbf{w}) - \nabla h_i'(\mathbf{w}_*), \mathbf{w} - \mathbf{w}_* \right\rangle - \left\langle \nabla h_i(\mathbf{w}) - \nabla h_i(\mathbf{w}_*), \mathbf{w} - \mathbf{w}_* \right\rangle\right)\right]$$

$$\leq \frac{2}{n}\mathrm{E}_{h_1, \ldots, h_n, \epsilon_1, \ldots, \epsilon_n}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\| \leq \gamma_k^+} \sum_{i=1}^n \epsilon_i \left\langle \nabla h_i(\mathbf{w}) - \nabla h_i(\mathbf{w}_*), \mathbf{w} - \mathbf{w}_* \right\rangle\right].$$

Substituting the above inequality into (50), we obtain

$$l(h_1, \ldots, h_n)$$
$$\leq L\left(\gamma_k^+\right)^2 \sqrt{\frac{2}{n} \log \frac{1}{\delta}} + \frac{2}{n}\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\| \leq \gamma_k^+} \sum_{i=1}^n \epsilon_i \left\langle \nabla h_i(\mathbf{w}) - \nabla h_i(\mathbf{w}_*), \mathbf{w} - \mathbf{w}_* \right\rangle\right]. \tag{51}$$

To upper bound the last term of (51), we use the Rademacher complexity of the product of two functions (Desalvo et al., 2015), and develop the following lemma.

Lemma 8

$$\mathrm{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\| \leq \gamma_k^+} \sum_{i=1}^n \epsilon_i \left\langle \nabla h_i(\mathbf{w}) - \nabla h_i(\mathbf{w}_*), \mathbf{w} - \mathbf{w}_* \right\rangle\right] \leq 4L\left(\gamma_k^+\right)^2 \sqrt{n}.$$

We complete the proof by substituting the above inequality into (51).

Appendix D. Proof of Lemma 8

Define

$$p_i(\mathbf{w}) = \frac{1}{\sqrt{\beta}} \left(\ell'(\langle \mathbf{w}, \mathbf{x}_i \rangle, y_i) - \ell'(\langle \mathbf{w}_*, \mathbf{x}_i \rangle, y_i) \right) \in [-\gamma_k^+ D\sqrt{\beta}, \gamma_k^+ D\sqrt{\beta}],$$

$$q_i(\mathbf{w}) = \sqrt{\beta} \langle \mathbf{x}_i, \mathbf{w} - \mathbf{w}_* \rangle \in [-\gamma_k^+ D\sqrt{\beta}, \gamma_k^+ D\sqrt{\beta}]$$

such that

$$\langle \nabla h_i(\mathbf{w}) - \nabla h_i(\mathbf{w}_*), \mathbf{w} - \mathbf{w}_* \rangle = \langle \nabla \ell(\langle \mathbf{w}, \mathbf{x}_i \rangle, y_i)) - \nabla \ell(\langle \mathbf{w}_*, \mathbf{x}_i \rangle, y_i), \mathbf{w} - \mathbf{w}_* \rangle$$
$$= \left(\ell'(\langle \mathbf{w}, \mathbf{x}_i \rangle, y_i) - \ell'(\langle \mathbf{w}_*, \mathbf{x}_i \rangle, y_i) \right) \langle \mathbf{x}_i, \mathbf{w} - \mathbf{w}_* \rangle = p_i(\mathbf{w}) q_i(\mathbf{w}).$$

From the equality $ab = \frac{1}{4} \left((a+b)^2 - (a-b)^2 \right)$, we have

$$\mathrm{E} \left[\sup_{\mathbf{w}: \|\mathbf{w} - \mathbf{w}_*\| \le \gamma_k^+} \sum_{i=1}^n \epsilon_i \langle \nabla h_i(\mathbf{w}) - \nabla h_i(\mathbf{w}_*), \mathbf{w} - \mathbf{w}_* \rangle \right]$$
$$\le \frac{1}{4} \mathrm{E} \left[\sup_{\mathbf{w}: \|\mathbf{w} - \mathbf{w}_*\| \le \gamma_k^+} \sum_{i=1}^n \epsilon_i \left(p_i(\mathbf{w}) + q_i(\mathbf{w}) \right)^2 \right] + \frac{1}{4} \mathrm{E} \left[\sup_{\mathbf{w}: \|\mathbf{w} - \mathbf{w}_*\| \le \gamma_k^+} \sum_{i=1}^n \epsilon_i \left(p_i(\mathbf{w}) - q_i(\mathbf{w}) \right)^2 \right]. \tag{52}$$

Note that the function x^2 is $2a$-Lipschitz over $[-a, a]$, and $p_i(\mathbf{w}) + q_i(\mathbf{w}) \in [-2\gamma_k^+ D\sqrt{\beta}, 2\gamma_k^+ D\sqrt{\beta}]$. Then, from the comparison theorem of Rademacher complexities (Ledoux and Talagrand, 1991), in particular Lemma 5 of Meir and Zhang (2003), we have

$$\mathrm{E} \left[\sup_{\mathbf{w}: \|\mathbf{w} - \mathbf{w}_*\| \le \gamma_k^+} \sum_{i=1}^n \epsilon_i \left(p_i(\mathbf{w}) + q_i(\mathbf{w}) \right)^2 \right]$$
$$\le 4\gamma_k^+ D\sqrt{\beta} \, \mathrm{E} \left[\sup_{\mathbf{w}: \|\mathbf{w} - \mathbf{w}_*\| \le \gamma_k^+} \sum_{i=1}^n \epsilon_i \left(p_i(\mathbf{w}) + q_i(\mathbf{w}) \right) \right] \tag{53}$$
$$\le 4\gamma_k^+ D\sqrt{\beta} \left(\mathrm{E} \left[\sup_{\mathbf{w}: \|\mathbf{w} - \mathbf{w}_*\| \le \gamma_k^+} \sum_{i=1}^n \epsilon_i p_i(\mathbf{w}) \right] + \mathrm{E} \left[\sup_{\mathbf{w}: \|\mathbf{w} - \mathbf{w}_*\| \le \gamma_k^+} \sum_{i=1}^n \epsilon_i q_i(\mathbf{w}) \right] \right).$$

Similarly, we have

$$\mathrm{E} \left[\sup_{\mathbf{w}: \|\mathbf{w} - \mathbf{w}_*\| \le \gamma_k^+} \sum_{i=1}^n \epsilon_i \left(p_i(\mathbf{w}) - q_i(\mathbf{w}) \right)^2 \right]$$
$$\le 4\gamma_k^+ D\sqrt{\beta} \left(\mathrm{E} \left[\sup_{\mathbf{w}: \|\mathbf{w} - \mathbf{w}_*\| \le \gamma_k^+} \sum_{i=1}^n \epsilon_i p_i(\mathbf{w}) \right] + \mathrm{E} \left[\sup_{\mathbf{w}: \|\mathbf{w} - \mathbf{w}_*\| \le \gamma_k^+} \sum_{i=1}^n \epsilon_i q_i(\mathbf{w}) \right] \right). \tag{54}$$

Combining (52), (53), and (54), we arrive at

$$E\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+}\sum_{i=1}^n \epsilon_i\left\langle\nabla h_i(\mathbf{w})-\nabla h_i(\mathbf{w}_*),\mathbf{w}-\mathbf{w}_*\right\rangle\right]$$

$$\leq 2\gamma_k^+ D\sqrt{\beta}\left(\underbrace{E\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+}\sum_{i=1}^n \epsilon_i p_i(\mathbf{w})\right]}_{:=C_1}+\underbrace{E\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+}\sum_{i=1}^n \epsilon_i q_i(\mathbf{x})\right]}_{:=C_2}\right). \quad (55)$$

We proceed to upper bound C_1 in (55). From our definition of $p_i(\mathbf{w})$, we have

$$\left|p_i(\mathbf{w})-p_i(\mathbf{w}')\right| = \frac{1}{\sqrt{\beta}}\left|\ell'(\langle\mathbf{w},\mathbf{x}_i\rangle,y_i)-\ell'(\langle\mathbf{w}',\mathbf{x}_i\rangle,y_i)\right|$$

$$\leq\sqrt{\beta}\left|\langle\mathbf{w},\mathbf{x}_i\rangle-\langle\mathbf{w}',\mathbf{x}_i\rangle\right| = \sqrt{\beta}\left|\langle\mathbf{x}_i,\mathbf{w}-\mathbf{w}_*\rangle-\langle\mathbf{x}_i,\mathbf{w}'-\mathbf{w}_*\rangle\right|.$$

Applying the comparison theorem of Rademacher complexities again, we have

$$C_1 \leq \sqrt{\beta}E\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+}\sum_{i=1}^n \epsilon_i\langle\mathbf{x}_i,\mathbf{w}-\mathbf{w}_*\rangle\right] = C_2. \quad (56)$$

Next, we upper bound C_2 as follows:

$$\sqrt{\beta}E\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+}\sum_{i=1}^n \epsilon_i\langle\mathbf{x}_i,\mathbf{w}-\mathbf{w}_*\rangle\right] \leq \sqrt{\beta}E\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+}\left\|\sum_{i=1}^n \epsilon_i\mathbf{x}_i\right\|\|\mathbf{w}-\mathbf{w}_*\|\right]$$

$$\leq\gamma_k^+\sqrt{\beta}E\left[\left\|\sum_{i=1}^n \epsilon_i\mathbf{x}_i\right\|\right] \leq \gamma_k^+\sqrt{E\left[\|\mathbf{x}_i\|^2+\sum_{u\neq v}\epsilon_u\epsilon_v\mathbf{x}_u^\top\mathbf{x}_v\right]} \leq \gamma_k^+ D\sqrt{\beta n}. \quad (57)$$

We complete the proof by combining (55), (56) and (57).