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   JINRONG.JR@ALIBABA-INC.COM
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
$\widetilde{O}\left(\quad + \sqrt{\quad_*} \right)$

where $\mathsf{H} = \{\ :\mathsf{X}\ \ \mathbb{R}\}$ is a hypothesis class, $(\mathbf{x}\ )\ \ \mathsf{X} \times \mathbb{R}$ is an instance-label pair sampled from a distribution $\mathbb{D}$, and $(\cdot\ \cdot) : \mathbb{R} \times \mathbb{R}\ \ \mathbb{R}$ is certain loss. In this paper, we mainly focus on the convex version of (1), namely stochastic convex optimization (SCO), where both the domain W and the expected function $(\cdot)$ 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; Nemirovski et al., 2009; Moulines and Bach, 2011; Hazan and Kale, 2011; Rakhlin et al., 2012; Agarwal et al., 2012; Bach and Moulines, 2013; Zhang et al., 2013b; Mahdavi et al., 2015), most theoretical guarantees of ERM are restricted to supervised learning 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 functions $_1 \ldots\ _n$ sampled independently from $\mathbb{P}$, and aim to minimize an empirical objective function:

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

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

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

State-of-the-art risk bounds of ERM include: an $\widetilde{\ }(\sqrt{\ \ })$ bound when the random function $(\cdot)$ is Lipschitz continuous,[1] where is the dimensionality of $\mathbf{w}$; an $(1\ )$ bound when $(\cdot)$ is -strongly convex (Shalev-Shwartz et al., 2009); and an $\widetilde{\ }(\ \ )$ bound when $(\cdot)$ is -exponentially concave (-exp-concave) (Mehta, 2014). From existing studies of ERM for supervised learning (Srebro et al., 2010

F() is

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.

		$\ell(\cdot)$	$\widehat{\phi}(\cdot)$	$\phi(\cdot)$	Risk Bounds
Shalev-Shwartz et al. (2009)		Lip	-	-	$\widetilde{O}(\sqrt{\frac{d}{n}})$
		Lip & sc	-	-	$(\frac{1}{\lambda n})^*$
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	$\widetilde{O}(\frac{d}{n}+\frac{\kappa F_*}{n})$ $(\frac{1}{\lambda n^2}+\frac{\kappa F_*}{n})$ when $\lambda=\widetilde{\Omega}(\epsilon)$
	Theorem 5	nn & sm	c	sc	$\widetilde{O}(\frac{\kappa d}{n}+\frac{\kappa F_*}{n})=\widetilde{O}(\frac{\kappa d}{n})$ $(\frac{1}{\lambda n^2}+\frac{\kappa F_*}{n})$ when $\lambda=\widetilde{\Omega}(\epsilon^2)$
	Theorem 7	nn & sm & gl	c	sc	$(\frac{\kappa}{n}+\frac{\kappa F_*}{n})=O(\frac{\kappa}{n})$ $(\frac{1}{\lambda n^2}+\frac{\kappa F_*}{n})$ when $\lambda=\Omega(\epsilon^2)$

- When $\ell(\cdot)$ is both convex and smooth and $\phi(\cdot)$ is Lipschitz continuous, we establish an $\widetilde{O}(\_ + \sqrt{\_*})$ risk bound (c.f. Theorem 1). In the optimistic case that $\_*$ is small, i.e., $\_* = O(\_^2)$, we obtain an $\widetilde{O}(\_)$ risk bound, which is analogous to the $\widetilde{O}(1-\_)$ optimistic rate of ERM for supervised learning (Srebro et al., 2010).
- If $\phi(\cdot)$ is also λ-strongly convex, we prove an $\widetilde{O}(\_ + \_*)$ risk bound, and improve it to $(1-[\_^2]+\_*)$ when $\lambda = \widetilde{\Omega}(\epsilon)$ (c.f. Theorem 3). Thus, if λ is large and $\_*$ is small, i.e., $\_* = O(1-\_)$, we get an $O(\_^2)$ risk bound, which to the best of our knowledge, is the first $(1-\epsilon^2)$-type of risk bound of ERM.
- When convexity is not present in $\ell(\cdot)$, as long as $\ell(\cdot)$ is smooth, $\widehat{\phi}(\cdot)$ is convex and $\phi(\cdot)$ is strongly convex, we still obtain an improved risk bound of $(1-[\_^2]+\_*)$ when $\lambda = \widetilde{\Omega}(\epsilon^2)$, which will further imply an $O(\_^2)$ risk bound if $\_* = O(1-\_)$ (c.f. Theorem 5).
- Finally, we extend the $(1-[\_^2]+\_*)$ risk bound to supervised learning with a generalized linear form. Our analysis shows that in this case, the lower bound of λ can be replaced with $\Omega(\epsilon^2)$, which is dimensionality-independent (c.f. Theorem 7). 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 ERM.

2.1. ERM for Stochastic Optimization

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

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 $\widehat{\quad}(\cdot)$ to $\quad(\cdot)$ over the hypothesis class $\mathbf{H}$ (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 $\mathbf{H}$, including data-independent measures such as the VC-dimension and data-dependent measures such as the Rademacher complexity.

Generally speaking, when $\mathbf{H}$ has finite VC-dimension, the excess risk can be upper bounded by $\quad(\sqrt{\mathrm{VC}(\mathbf{H})\quad})$, where $\mathrm{VC}(\mathbf{H})$ is the VC-dimension of $\mathbf{H}$. If the loss $\quad(\cdot\ \cdot)$ is Lipschitz continuous with respect to its first argument, we have a risk bound of $\quad(1\quad^- + \mathbf{R}_n(\mathbf{H}))$, where $\mathbf{R}_n(\mathbf{H})$ is the Rademacher complexity of $\mathbf{H}$. The Rademacher complexity typically scales as $\mathbf{R}_n(\mathbf{H}) = \quad(1\quad^-)$, e.g., $\mathbf{H}$ contains linear functions with low-norm, implying an $\quad(1\quad^-)$ risk bound (Bartlett and Mendelson, 2002). There have been intensive efforts to derive rates faster than $\quad(1\quad^-)$ 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 $\quad(\cdot)$ is nonnegative and smooth, Srebro et al. (2010) have established a risk bound of $\widetilde{\quad}(\mathbf{R}_n^2(\quad) + \mathbf{R}_n(\quad)\ \overline{\quad}_*)$, reducing to an $\widetilde{\quad}(1\quad)$ bound if $\mathbf{R}_n(\mathbf{H}) = \quad(1\quad^-)$ and $\quad_* = \quad(1\quad)$. A generalized linear form of (2) is studied by Sridharan et al. (2009), and a risk bound of $\quad(1\quad)$ is proved if the expected function $\quad(\cdot)$ is -strongly convex.

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.

---

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

3.1. Assumptions

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

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

$$\|\mathbf{w}\| \qquad \mathbf{w} \quad \mathbf{W}. \tag{4}$$

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

$$\left\| (\mathbf{w}) - (\mathbf{w}') \right\| \qquad \|\mathbf{w} - \mathbf{w}'\| \quad \mathbf{w}\ \mathbf{w}'\ \mathbf{W}, \qquad \mathbb{P}. \tag{5}$$

Assumption 3 *The expected function* $(\cdot)$ *is -Lipschitz continuous over* **W**, *that is,*

$$| (\mathbf{w}) - (\mathbf{w}')| \qquad \|\mathbf{w} - \mathbf{w}'\| \quad \mathbf{w}\ \mathbf{w}'\ \mathbf{W}. \tag{6}$$

Assumption 4 *We use different combinations of the following assumptions on convexity.*
  (a) *The expected function* $(\cdot)$ *is convex over* **W**.
  (b) *The expected function* $(\cdot)$ *is -strongly convex over* **W**, *that is,*

$$(\mathbf{w}) + \quad (\mathbf{w})\ \mathbf{w}' - \mathbf{w}\ + \frac{}{2}\ \|\mathbf{w}' - \mathbf{w}\|^2 \qquad (\mathbf{w}') \quad \mathbf{w}\ \mathbf{w}'\ \mathbf{W}. \tag{7}$$

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

Assumption 5 *Let* $\mathbf{w}_* \quad \operatorname{argmin}_{\mathbf{w}\in\mathcal{W}}\ (\mathbf{w})$ *be an optimal solution to* ([1](#)). *We assume the gradient of the random function at* $\mathbf{w}_*$ *is upper bounded by* , *that is,*

$$\| (\mathbf{w}_*)\| \qquad \qquad \mathbb{P}. \tag{8}$$

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 $(\cdot)$ implies the expected function $(\cdot)$ is -smooth. By Jensen's inequality, we have

$$\left\| (\mathbf{w}) - (\mathbf{w}') \right\| \quad E_{f\sim\mathbb{P}}\left\| (\mathbf{w}) - (\mathbf{w}') \right\| \qquad \|\mathbf{w} - \mathbf{w}'\| \quad \mathbf{w}\ \mathbf{w}'\ \mathbf{W}.$$

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

3.2. Risk Bounds for SCO

We first present an excess risk bound under the smoothness condition.

Theorem 1 *For any* $0 \quad 1 2, \quad 0$, *define*

$$( ) = 2\left(\log\frac{2}{} + \log\frac{6}{}\right). \tag{9}$$

Under **Assumptions 1**, **2**, **3**, **4(d)**, *and* **5**, *with probability at least* $1 - 2$, *we have*

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

$$\frac{16 \quad ^2 \quad (\quad)}{ } + \frac{8 \quad \log(2 \quad)}{ } + 8$$

Remark 3 The first part of Corollary 4 shows that ERM enjoys an $\tilde{\mathcal{O}}(\quad+\quad_*)$ risk bound for stochastic optimization of strongly convex and smooth functions. In the literature, the most comparable result is the $\mathcal{O}(1\quad)$ 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

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 is increased by a factor of . 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., 2013a) 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 factor in the sample complexity of stochastic gradients.

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 $(1 [^2] + _*)$ risk bound for supervised learning. However, a major limitation of these theorems is that the lower bound of depends on the dimensionality , 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:

$$\min_{\mathbf{w} \in \mathcal{W}} (\mathbf{w}) = \mathrm{E}_{(\mathbf{x},y) \sim \mathbb{D}} [(\mathbf{w} \mathbf{x})] + (\mathbf{w}) \tag{17}$$

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

$$\min_{\mathbf{w} \in \mathcal{W}} \widehat{\ }(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n (\mathbf{w} \mathbf{x}_i \ _i) + (\mathbf{w})$$

We define

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

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 6 *The domain* **W** *is a convex subset of a Hilbert space* **H**, *and is bounded by*

Assumption 10 *Let* $\mathbf{w}_* \in \arg\min_{\mathbf{w}\in\mathcal{W}}\ (\mathbf{w})$ *be an optimal solution to ([17]). We assume the gradient of the random function at* $\mathbf{w}_*$ *is upper bounded by* $\ $, *that is,*

$$\ (\ \mathbf{w}_*^\top \mathbf{x}\ )\ \ \ (\mathbf{x}\ )\ \ \mathbb{D} \tag{22}$$

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

$$\left\|\ \ (\ \mathbf{w}\ \mathbf{x}\ )-\ \ (\ \mathbf{w}'\ \mathbf{x}\ )\right\| = \left\|\ '(\ \mathbf{w}\ \mathbf{x}\ )\mathbf{x} -\ '(\ \mathbf{w}'\ \mathbf{x}\ )\mathbf{x}\right\|$$
$$\overset{(19)}{=}\ \left|\ '(\ \mathbf{w}\ \mathbf{x}\ )-\ '(\ \mathbf{w}'\ \mathbf{x}\ )\right| \overset{(20)}{\ }\ \left|\ \mathbf{w}\ \mathbf{x}\ -\ \mathbf{w}'\ \mathbf{x}\ \right| \overset{(19)}{\ }\ \ ^2\ \mathbf{w}-\mathbf{w}'\ $$

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

We have the following excess risk bound of ERM for supervised learning.

Theorem 7 *For any* $0\ \ \ \ 1\ 2$, *define*

$$\ =4\left(8+\sqrt{2\log\frac{2\log_2(\ )+\log_2(2\ )}{\ }}\right) \tag{23}$$

$$\ _* =\ (\mathbf{w}_*) =\ (\mathbf{w}_*)-\ (\mathbf{w}_*) \tag{24}$$

Under **Assumptions 4(b)**, **4(c)**, **6**, **7**, **8**, **9**, *and* **10** *with probability at least* $1-2\ $, *we have*

$$\ (\widehat{\mathbf{w}})-\ (\mathbf{w}_*)\ \ \max\left(\frac{\ +\ }{\ ^2}+\frac{\ }{\ ^2\ ^4}\ \frac{4\ \ ^2\ ^2\ ^2}{\ }+\frac{4\ \ \log(2\ )}{\ }+\frac{8\ \ _*\log(2\ )}{\ }\right) \tag{25}$$

Furthermore, if

$$\frac{16\ \ ^2\ ^2}{\ ^2}=16\ \ ^2\ ^2 \tag{26}$$

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

$$\ (\widehat{\mathbf{w}})-\ (\mathbf{w}_*)\ \ \max\left(\frac{\ +\ }{\ ^2}+\frac{\ }{\ ^2\ ^4}\ \frac{8\ \ ^2\log^2(2\ )}{\ ^2}+\frac{16\ \ _*\log(2\ )}{\ }\right) \tag{27}$$

Remark 8 The first part of Theorem 7 presents an $(\ \ )$ risk bound,[3] similar to the $(1-\ )$ risk bound of Sridharan et al. (2009). The second part is an $(1-[\ \ ^2]+\ \ _*)$ risk bound, and in this case, the lower bound of $\ $ is $\Omega(\ ^2)$, which is dimensionality-independent. Thus, Theorem 7 can be applied even when the dimensionality is infinite. Generally speaking, the regularizer $\ (\cdot)$ is nonnegative, and thus $\ _*\ \ \ _*$. So, the second bound is even better than those in Theorems 3 and 5. Finally, we note that Theorem 7 should be treated as a counterpart of Theorem 5 for supervised learning, because both of them do not rely on the individual convexity, 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 $\ $ by a factor of $\ $. We will investigate this question as a future work.

---

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

4. Analysis

We here present the key idea of our analysis and the proof of Theorem 1. The omitted ones can be found in appendices.

4.1. The Key Idea

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

Lemma 1 *Under* **Assumptions 2** *and* **4(d)**, *with probability at least* $1 - $

where the last step is due to

$$\widehat{\mathbf{w}} - \mathbf{w}_* \quad \sqrt{\frac{,\ (\cdot\ )(\ (\widehat{\mathbf{w}}) -\ (\mathbf{w}_*))}{}} \quad ,\ (\cdot\ )\ \widehat{\mathbf{w}} - \mathbf{w}_*\ ^2 + \frac{(\widehat{\mathbf{w}}) -\ (\mathbf{w}_*)}{2} \tag{35}$$

$$\widehat{\mathbf{w}} - \mathbf{w}_* \quad \sqrt{\frac{,\ (\cdot\ )\ ,}{}} \quad ,\ (\cdot\ )\ \frac{\widehat{\mathbf{w}} - \mathbf{w}_*\ ^2}{2} + \frac{\prime}{2}, \tag{36}$$

From (34), we get

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

$$\frac{2\ ,\ (\cdot\ )\ \widehat{\mathbf{w}} - \mathbf{w}_*\ ^2}{} + \frac{2\ \log(2\ )\ \widehat{\mathbf{w}} - \mathbf{w}_*}{} + \ \widehat{\mathbf{w}} - \mathbf{w}_* \ \sqrt{\frac{8\ \ _*\log(2\ )}{}}$$

$$+ 2\ ,\ \widehat{\mathbf{w}} - \mathbf{w}_* + \frac{\prime}{2} + \frac{,\ (\cdot\ ),\ \widehat{\mathbf{w}} - \mathbf{w}_*}{}$$

$$^{(4)}\frac{8\ ^2\ ,\ (\cdot\ )}{} + \frac{4\ \log(2\ )}{} + 4\ \sqrt{\frac{2\ \ _*\log(2\ )}{}} + \left(4\ + \frac{}{2} + \frac{2\ ,\ (\cdot\ )}{}\right),$$

which implies (10).

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 $(1\ )$-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 7). A more exciting result is that when is large enough, ERM has $(1\ ^2)$-type of risk bounds under the smoothness, strong convexity, and small minimal risk conditions (i.e., the second part of Theorems 3, 5, and 7).

 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 $_2$-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., $\widetilde{\ }(\ )$ or $\widetilde{\ }(1\ )$, that holds with high probability is still missing.
3. As discussed in **Remark 8**, it is unclear whether the convexity of the loss can be exploited to improve the lower bound of in the second part of Theorem 7. Ideally, we expect that $= \Omega(\ )$ is sufficient to deliver an $(1\ [\ ^2] +\ _*\ )$ risk bound.
4. The $(1\ ^2)$-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 (Hazan et al., 2007).

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

Acknowledgments

(1   )- AND   (1

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, 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, 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, 2013a.

Lijun Zhang, Tianbao Yang, Rong Jin, and Xiaofei He. $(\log)$ projections for stochastic optimization of smooth and strongly convex functions. In *Proceedings of the 30th International Conference on Machine Learning*, 2013b.

Appendix A. Proof of Lemma 1

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

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

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

We first consider a fixed $\mathbf{w} \in \mathbf{N}(\mathbf{W}, r)$. Since $\ell_i(\cdot)$ is γ-smooth, we have

$$\left\| \nabla \ell_i(\mathbf{w}) - \nabla \ell_i(\mathbf{w}_*) \right\| \overset{(5)}{\le} \gamma \|\mathbf{w} - \mathbf{w}_*\|. \tag{37}$$

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

$$\left\| \nabla \ell_i(\mathbf{w}) - \nabla \ell_i(\mathbf{w}_*) \right\|^2 \le \gamma \left(\ell_i(\mathbf{w}) - \ell_i(\mathbf{w}_*) - \langle \nabla \ell_i(\mathbf{w}_*), \mathbf{w} - \mathbf{w}_* \rangle \right).$$

Taking expectation over both sides, we have

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

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

$$\langle \nabla L(\mathbf{w}_*), \mathbf{w} - \mathbf{w}_* \rangle \ge 0, \quad \forall \mathbf{w} \in \mathbf{W}.$$

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

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

$$= \left\| \mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_*) - \frac{1}{n}\sum_{i=1}^{n}[\ell_i(\mathbf{w}) - \ell_i(\mathbf{w}_*)] \right\|$$

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

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

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

$$\log |\mathbf{N}(\mathbf{B}, \epsilon)| \le \log \frac{3}{\epsilon}$$

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

$$\log |\mathbf{N}(\mathbf{W}, \epsilon)| \le \log \left| \mathbf{N}\left(\mathbf{B}(r), \frac{\epsilon}{2}\right) \right| \le \log \frac{6r}{\epsilon}$$

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

Appendix B. Proof of Lemma 2

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

$$\|\nabla \ell_i(\mathbf{w}_*)\|^2 \le 4\beta \ell_i(\mathbf{w}_*)$$

and thus

$$\mathrm{E}\left[\|\nabla \ell_i(\mathbf{w}_*)\|^2\right] \le 4\beta \mathrm{E}[\ell_i(\mathbf{w}_*)] = 4\beta \mathcal{L}_*$$

From **Assumption 5**, we have $\|\nabla \ell_i(\mathbf{w}_*)\| \le G$. Then, according to Lemma 3, with probability at least $1 - \delta$, we have

$$\left\| \nabla \mathcal{L}(\mathbf{w}_*) - \nabla \widehat{\mathcal{L}}(\mathbf{w}_*) \right\| = \left\| \nabla \mathcal{L}(\mathbf{w}_*) - \frac{1}{n}\sum_{i=1}^{n}\nabla \ell_i(\mathbf{w}_*) \right\| \le \frac{2G \log(2/\delta)}{n} + \sqrt{\frac{8\beta \mathcal{L}_* \log(2/\delta)}{n}}$$

Appendix C. Proof of Theorem 3

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

$$\mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_*) + \frac{\mu}{2}\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

$$\le \left(\underbrace{\left\| \nabla \mathcal{L}(\widehat{\mathbf{w}}) - \nabla \mathcal{L}(\mathbf{w}_*) - [\nabla \widehat{\mathcal{L}}(\widehat{\mathbf{w}}) - \nabla \widehat{\mathcal{L}}(\mathbf{w}_*)] \right\|}_{:=A_1} + \underbrace{\left\| \nabla \mathcal{L}(\mathbf{w}_*) - \nabla \widehat{\mathcal{L}}(\mathbf{w}_*) \right\|}_{:=A_2} \right) \|\widehat{\mathbf{w}} - \mathbf{w}_*\|$$

$$\tag{38}$$

16

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

$$
\begin{aligned}
&(\widehat{\mathbf{w}}) - \quad(\mathbf{w}_*) + \frac{}{2}\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \\
&\qquad \frac{\,(\,)\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{} + \|\widehat{\mathbf{w}} - \mathbf{w}_*\|\,\sqrt{\frac{\,(\,)(\,(\widehat{\mathbf{w}}) - \quad(\mathbf{w}_*))}{}} \\
&+ \frac{2\quad\log(2\,)\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{} + \|\widehat{\mathbf{w}} - \mathbf{w}_*\|\,\sqrt{\frac{8\quad_*\log(2\,)}{}} \\
&+ 2\quad\|\widehat{\mathbf{w}} - \mathbf{w}_*\| + \|\widehat{\mathbf{w}} - \mathbf{w}_*\|\,\sqrt{\frac{\,(\,)\,}{}} + \frac{\,(\,)\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{}
\end{aligned}
\tag{39}
$$

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

$$
\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\,\sqrt{\frac{8\quad_*\log(2\,)}{}}\quad\frac{4\quad_*\log(2\,)}{} + \frac{}{2}\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2
$$

into (39), and then obtain

$$
\begin{aligned}
&\frac{1}{2}(\quad(\widehat{\mathbf{w}}) - \quad(\mathbf{w}_*)) \\
&\qquad \frac{2\,(\,)\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{} + \frac{2\quad\log(2\,)\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{} + \frac{4\quad_*\log(2\,)}{} \\
&+ 2\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\| + \frac{}{2} + \frac{\,(\,)\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{} \\
&\stackrel{(4)}{} \frac{8\quad^2\,(\,)}{} + \frac{4\quad\log(2\,)}{} + \frac{4\quad_*\log(2\,)}{} + \left(4\quad + \frac{}{2} + \frac{2\,(\,)}{}\right)\,
\end{aligned}
$$

which implies (11).

To prove (13), we substitute

$$
\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\,\sqrt{\frac{\,(\,)(\,(\widehat{\mathbf{w}}) - \quad(\mathbf{w}_*))}{}}\quad\frac{2\,(\,)(\,(\widehat{\mathbf{w}}) - \quad(\mathbf{w}_*))}{} + \frac{}{8}\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2
$$

$$
\frac{2\quad\log(2\,)\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{}\quad\frac{16\quad^2\log^2(2\,)}{2} + \frac{}{16}\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2
$$

$$
\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\,\sqrt{\frac{8\quad_*\log(2\,)}{}}\quad\frac{64\quad_*\log(2\,)}{} + \frac{}{32}\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2
$$

$$
2\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\quad\frac{64\quad^2\,^2}{} + \frac{}{64}\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2
$$

$$
\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\,\sqrt{\frac{\,(\,)\,}{}}\quad\frac{32\,(\,)\,}{} + \frac{}{128}\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2
$$

$$
\frac{\,(\,)\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{}\quad\frac{32\quad^2\,^2(\,)\,^2}{2} + \frac{}{128}\,\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2
$$

17

into (39), and then obtain

$$
\begin{aligned}
&\mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_*) + \frac{}{4} \left\| \widehat{\mathbf{w}} - \mathbf{w}_* \right\|^2 \\
&\frac{\gamma(\cdot,\cdot)\left\| \widehat{\mathbf{w}} - \mathbf{w}_* \right\|^2}{} + \frac{2\gamma(\cdot,\cdot)(\mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_*))}{} + \frac{16^2 \log^2(2\cdot)}{^2} + \frac{64\,_* \log(2\cdot)}{} \\
&+ \frac{64^{2,2}}{} + \frac{32\,(\cdot,\cdot)\,}{} + \frac{32^{2,2}(\cdot)^2}{^2} \\
&\overset{(12)}{} \frac{}{4}\left\| \widehat{\mathbf{w}} - \mathbf{w}_* \right\|^2 + \frac{1}{2}(\mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_*)) + \frac{16^2 \log^2(2\cdot)}{^2} + \frac{64\,_* \log(2\cdot)}{} \\
&+ \frac{64^{2,2}}{} + 8\, + 2\,^2
\end{aligned}
$$

which implies (13).

Appendix D. 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 - $, *for any* $\mathbf{w} \quad \mathbf{N}(\mathbf{W}\,)$, *we have*

$$
\left\| \mathcal{L}(\mathbf{w}) - \mathcal{L}(\mathbf{w}_*) - [\widehat{\mathcal{L}}(\mathbf{w}) - \widehat{\mathcal{L}}(\mathbf{w}_*)] \right\| \quad \frac{\gamma(\cdot,\cdot)\left\| \mathbf{w} - \mathbf{w}_* \right\|}{} + \left\| \mathbf{w} - \mathbf{w}_* \right\| \sqrt{\frac{(\cdot,\cdot)}{}}
$$

where $(\cdot,\cdot)$ *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 - $, we have

$$
\begin{aligned}
&\left\| \mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_*) - [\widehat{\mathcal{L}}(\widehat{\mathbf{w}}) - \widehat{\mathcal{L}}(\mathbf{w}_*)] \right\| \\
&\frac{\gamma(\cdot,\cdot)\left\| \widetilde{\mathbf{w}} - \mathbf{w}_* \right\|}{} + \left\| \widetilde{\mathbf{w}} - \mathbf{w}_* \right\| \sqrt{\frac{(\cdot,\cdot)}{}} + 2\, \\
&\frac{\gamma(\cdot,\cdot)\left\| \widehat{\mathbf{w}} - \mathbf{w}_* \right\|}{} + \left\| \widehat{\mathbf{w}} - \mathbf{w}_* \right\| \sqrt{\frac{(\cdot,\cdot)}{}} + \frac{\gamma(\cdot,\cdot)}{} + \,\sqrt{\frac{(\cdot,\cdot)}{}} + 2\,
\end{aligned}
\tag{40}
$$

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

$$
\begin{aligned}
&\mathcal{L}(\widehat{\mathbf{w}}) - \mathcal{L}(\mathbf{w}_*) + \frac{}{2}\left\| \widehat{\mathbf{w}} - \mathbf{w}_* \right\|^2 \\
&\frac{\gamma(\cdot,\cdot)\left\| \widehat{\mathbf{w}} - \mathbf{w}_* \right\|^2}{} + \left\| \widehat{\mathbf{w}} - \mathbf{w}_* \right\|^2 \sqrt{\frac{(\cdot,\cdot)}{}} \\
&+ \frac{2\log(2\cdot)\left\| \widehat{\mathbf{w}} - \mathbf{w}_* \right\|}{} + \left\| \widehat{\mathbf{w}} - \mathbf{w}_* \right\| \sqrt{\frac{8\,_* \log(2\cdot)}{}} \\
&+ 2\,\left\| \widehat{\mathbf{w}} - \mathbf{w}_* \right\| + \,\left\| \widehat{\mathbf{w}} - \mathbf{w}_* \right\| \sqrt{\frac{(\cdot,\cdot)}{}} + \frac{\gamma(\cdot,\cdot)\left\| \widehat{\mathbf{w}} - \mathbf{w}_* \right\|}{}
\end{aligned}
\tag{41}
$$

To get (14), we substitute

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \sqrt{\frac{\ (\ ,\ )}{\ }} \quad \frac{\ ^2\ (\ ,\ )\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\ } + \frac{\ }{4}\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\ \sqrt{\frac{8\ _*\log(2\ )}{\ }}\quad \frac{8\ _*\log(2\ )}{\ } + \frac{\ }{4}\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

into (41), and then obtain

$$\ (\widehat{\mathbf{w}}) - \ (\mathbf{w}_*)$$
$$\frac{\ ,\ (\ ,\ )\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\ } + \frac{\ ^2\ (\ ,\ )\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\ } + \frac{2\ \ \log(2\ )\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{\ } + \frac{8\ _*\log(2\ )}{\ }$$

$$+ 2\ ,\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\| + \ ,\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|\ \sqrt{\frac{\ (\ ,\ )}{\ }} + \frac{\ ,\ (\ ,\ ),\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{\ }$$

$$\stackrel{(4)}{\ }\frac{4\ ^2\ ,\ (\ ,\ )}{\ } + \frac{4\ ^2\ ^2\ ,\ (\ ,\ )}{\ } + \frac{4\ \ \log(2\ )}{\ } + \frac{8\ _*\log(2\ )}{\ }$$

$$+ \left(4\ \ + 2\ \ \sqrt{\frac{\ (\ ,\ )}{\ }} + \frac{2\ \ ,\ (\ ,\ )}{\ } \right),$$

which proves (14).

To get (16), we substitute

$$\frac{2\ \ \log(2\ )\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{\ }\quad \frac{8\ ^2\log^2(2\ )}{\ ^2} + \frac{\ }{8}\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\|\ \sqrt{\frac{8\ _*\log(2\ )}{\ }}\quad \frac{32\ _*\log(2\ )}{\ } + \frac{\ }{16}\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

$$2\ ,\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|\quad \frac{32\ ^2,^2}{\ } + \frac{\ }{32}\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

$$,\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|\ \sqrt{\frac{\ (\ ,\ )}{\ }}\quad \frac{16\ ^2,\ (\ ,\ ),\ ^2}{\ } + \frac{\ }{64}\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

$$\frac{\ ,\ (\ ,\ ),\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{\ }\quad \frac{16\ ^2,^2(\ ),\ ^2}{\ ^2} + \frac{\ }{64}\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

into (41), and then obtain

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

$$\frac{\ ,\ (\ ,\ )\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\ } + \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 \sqrt{\frac{\ (\ ,\ )}{\ }} + \frac{8\ ^2\log^2(2\ )}{\ ^2} + \frac{32\ _*\log(2\ )}{\ }$$

$$+ \left(\frac{32\ ^2}{\ } + \frac{16\ ^2,\ (\ ,\ )}{\ } + \frac{16\ ^2,^2(\ )}{\ 2} \right),\ ^2$$

$$\stackrel{(15)}{\ }\frac{\ ^2\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{25} + \frac{\ }{5}\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 + \frac{8\ ^2\log^2(2\ )}{\ 2} + \frac{32\ _*\log(2\ )}{\ }$$

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

$$\stackrel{\lambda/L \le 1}{\ }\frac{6}{25}\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 + \frac{8\ ^2\log^2(2\ )}{\ 2} + \frac{32\ _*\log(2\ )}{\ } + \left(\frac{32\ ^2}{\ } + \frac{416}{625} \right),\ ^2$$

19

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

Appendix E. Proof of Theorem 7

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

$$\|\widehat{\mathbf{w}} - \mathbf{w}_*\| \le \frac{1}{2}$$

Since $(\cdot)$ is -smooth and $(\cdot)$ is -Lipschitz continuous, we have

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

$$\langle \widehat{\mathbf{w}} - \mathbf{w}_*, (\mathbf{w}_*) \rangle + \frac{}{2} \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 + \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \tag{42}$$

$$\langle \widehat{\mathbf{w}} - \mathbf{w}_*, (\mathbf{w}_*) \rangle + \frac{}{2} \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 + \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \quad \frac{+}{2} + \frac{}{2 \cdot 4}$$

where the last step utilizes Jensen's inequality

$$(\mathbf{w}_*) = \left\| E_{(\mathbf{x},y)\sim\mathbb{D}} [(\mathbf{w}_*^\top \mathbf{x})] \right\| \quad E_{(\mathbf{x},y)\sim\mathbb{D}} [(\mathbf{w}_*^\top \mathbf{x})]^{(22)}$$

Next, we study the case

$$\frac{1}{2} \quad \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^{(18)} \quad 2$$

From (29), we have

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

$$(\widehat{\mathbf{w}}) - (\mathbf{w}_*) - [\widehat{}(\widehat{\mathbf{w}}) - \widehat{}(\mathbf{w}_*)] \|\widehat{\mathbf{w}} - \mathbf{w}_*\| + (\mathbf{w}_*) - \widehat{}(\mathbf{w}_*) \|\widehat{\mathbf{w}} - \mathbf{w}_*\|$$

$$= (\widehat{\mathbf{w}}) - (\mathbf{w}_*) - [\widehat{}(\widehat{\mathbf{w}}) - \widehat{}(\mathbf{w}_*)] \|\widehat{\mathbf{w}} - \mathbf{w}_*\| + (\mathbf{w}_*) - \widehat{}(\mathbf{w}_*) \|\widehat{\mathbf{w}} - \mathbf{w}_*\|$$

$$\underbrace{\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\|\widehat{\mathbf{w}}-\mathbf{w}_*\|} \left\langle (\mathbf{w}) - (\mathbf{w}_*) - [\widehat{}(\mathbf{w}) - \widehat{}(\mathbf{w}_*)] \mathbf{w} - \mathbf{w}_* \right\rangle}_{:=B_1}$$

$$+ \underbrace{\left\| (\mathbf{w}_*) - \widehat{}(\mathbf{w}_*) \right\|}_{:=B_2} \|\widehat{\mathbf{w}} - \mathbf{w}_*\|$$

$$\tag{43}$$

We first bound $_1$. To utilize the fact the random variable $\|\widehat{\mathbf{w}} - \mathbf{w}_*\|$ lies in the range $(1 \quad ^2 2 \quad]$, we develop the following lemma.

Lemma 5 *Under* **Assumptions 7** *and* **8***, with probability at least* $1 - $ *, for all*

$$\frac{1}{2} \quad 2$$

the following bound holds:

$$\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma} \left\langle (\mathbf{w}) - (\mathbf{w}_*) - [\widehat{}(\mathbf{w}) - \widehat{}(\mathbf{w}_*)] \mathbf{w} - \mathbf{w}_* \right\rangle \quad \frac{4}{} \left(8 + \sqrt{2 \log -} \right)$$

where $= 2 \log_2( ) + \log_2(2)$.

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

$$_1 \ \frac{4 \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\ } \left(8 + \sqrt{2\log \frac{\ }{\ }}\right) = \ \frac{\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\ } \qquad (44)$$

where $\ $ is defined in (23).

We then proceed to handle $_2$, which can be upper bounded in the same way as A_2. In particular, we have the following lemma.

Lemma 6 *Under* **Assumptions 7, 8,** *and* **10,** *with probability at least* $1 - \ ,$ *we have*

$$\left\| \ (\mathbf{w}_*) - \ \widehat{\ }(\mathbf{w}_*)\right\| \ \frac{2 \ \log(2 \ )}{\ } + \sqrt{\frac{8 \ _* \log(2 \ )}{\ }} \qquad (45)$$

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

$$(\widehat{\mathbf{w}}) - \ (\mathbf{w}_*) + \frac{\ }{2} \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$
$$\ \frac{\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\ } + \frac{2 \ \log(2\ ) \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{\ } + \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \ \sqrt{\frac{8 \ _* \log(2\ )}{\ }} \qquad (46)$$

We substitute

$$\frac{\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\ } \ \frac{^2 \ ^2 \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\ } + \frac{\ }{4} \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$
$$\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \ \sqrt{\frac{8 \ _* \log(2\ )}{\ }} \ \frac{8 \ _* \log(2\ )}{\ } + \frac{\ }{4} \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

into (46), and then have

$$(\widehat{\mathbf{w}}) - \ (\mathbf{w}_*) \ \frac{^2 \ ^2 \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\ } + \frac{2 \ \log(2\ ) \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{\ } + \frac{8 \ _* \log(2\ )}{\ }$$
$$\overset{(18)}{\ } \frac{4 \ ^2 \ ^2 \ ^2}{\ } + \frac{4 \ \log(2\ )}{\ } + \frac{8 \ _* \log(2\ )}{\ }$$

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

To prove (27), we substitute

$$\frac{2 \ \log(2\ ) \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|}{\ } \ \frac{8 \ ^2 \log^2(2\ )}{\ ^2} + \frac{\ }{8} \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$
$$\ \|\widehat{\mathbf{w}} - \mathbf{w}_*\| \ \sqrt{\frac{8 \ _* \log(2\ )}{\ }} \ \frac{16 \ _* \log(2\ )}{\ } + \frac{\ }{8} \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$

into (46), and then have

$$(\widehat{\mathbf{w}}) - \ (\mathbf{w}_*) + \frac{\ }{4} \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2$$
$$\ \frac{\|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2}{\ } + \frac{8 \ ^2 \log^2(2\ )}{\ ^2} + \frac{16 \ _* \log(2\ )}{\ }$$
$$\overset{(26)}{\ } \frac{\ }{4} \ \|\widehat{\mathbf{w}} - \mathbf{w}_*\|^2 + \frac{8 \ ^2 \log^2(2\ )}{\ ^2} + \frac{16 \ _* \log(2\ )}{\ }$$

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

21

Appendix F. Proof of Lemma 5

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

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

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

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

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 7** and **8**, with probability at least $1 - \delta$, we have*

$$
\begin{aligned}
&\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+} \left\langle \nabla(\mathbf{w}) - \nabla(\mathbf{w}_*) - [\nabla\widehat{}(\mathbf{w}) - \nabla\widehat{}(\mathbf{w}_*)], \mathbf{w} - \mathbf{w}_* \right\rangle \\
&\le \frac{(\gamma_k^+)^2}{\sqrt{n}} \left(8 + \sqrt{2\log\frac{1}{\delta}} \right).
\end{aligned}
\tag{48}
$$

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

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

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

$$
\begin{aligned}
&\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma} \left\langle \nabla(\mathbf{w}) - \nabla(\mathbf{w}_*) - [\nabla\widehat{}(\mathbf{w}) - \nabla\widehat{}(\mathbf{w}_*)], \mathbf{w} - \mathbf{w}_* \right\rangle \\
&\overset{(47),(48),(49)}{\le} \frac{4\gamma^2}{\sqrt{n}} \left(8 + \sqrt{2\log\frac{1}{\delta}} \right).
\end{aligned}
$$

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

Appendix G. Proof of Lemma 7

To simplify the notation, we define

$$
\ell_i(\mathbf{w}) = (\mathbf{w}^\top \mathbf{x}_i, y_i), \quad i = 1, \ldots,
$$

$$
(\xi_1, \ldots, \xi_n) = \sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\le\gamma_k^+} \left\langle \nabla(\mathbf{w}) - \nabla(\mathbf{w}_*) - \frac{1}{n}\sum_{i=1}^{n}[\nabla\ell_i(\mathbf{w}) - \nabla\ell_i(\mathbf{w}_*)], \mathbf{w} - \mathbf{w}_* \right\rangle.
$$

To upper bound $(\xi_1, \ldots, \xi_n)$, we utilize the McDiarmid's inequality (McDiarmid, 1989).

Theorem 8 *Let* $h_1, \ldots, h_n$ *be independent random variables taking values in a set A, and assume that* $\Phi : A^n \to \mathbb{R}$ *satisfies*

$$\sup_{x_1,\ldots,x_n,x_i' \in A} \left| \Phi(h_1 \cdots h_n) - \Phi(h_1 \cdots h_{i-1} h_i' h_{i+1} \cdots h_n) \right| \leq c_i$$

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

$$\mathbb{P}\left\{ \Phi(h_1 \cdots h_n) - \mathrm{E}[\Phi(h_1 \cdots h_n)] \geq \epsilon \right\} \leq \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^n c_i^2} \right).$$

As pointed out in **Remark 7**, **Assumptions 7** and **8** imply the random function $h_i(\cdot)$ is μ-smooth, and thus

$$\left| h_i(\mathbf{w}) - h_i(\mathbf{w}_*) \mathbf{w} - \mathbf{w}_* \right| \leq \mu \|\mathbf{w} - \mathbf{w}_*\|^2 \leq \mu (\gamma_k^+)^2.$$

As a result, when a random function h_i changes, the random variable $\Phi(h_1 \cdots h_n)$ can change by no more than $2\mu(\gamma_k^+)^2$. To see this, we have

$$\Phi(h_1 \cdots h_n) - \Phi(h_1 \cdots h_{i-1} h_i' h_{i+1} \cdots h_n)$$
$$\leq \frac{1}{n} \sup_{\mathbf{w}: \|\mathbf{w}-\mathbf{w}_*\| \leq \gamma_k^+} \left\langle h_i'(\mathbf{w}) - h_i'(\mathbf{w}_*) - [h_i(\mathbf{w}) - h_i(\mathbf{w}_*)] \mathbf{w} - \mathbf{w}_* \right\rangle \leq \frac{2\mu}{n} (\gamma_k^+)^2.$$

McDiarmid's inequality implies that with probability at least $1 - \delta$,

$$\Phi(h_1 \cdots h_n) \leq \mathrm{E}[\Phi(h_1 \cdots h_n)] + \mu(\gamma_k^+)^2 \sqrt{\frac{2}{n} \log \frac{1}{\delta}} \tag{50}$$

Let $(h_1' \cdots h_n')$ be an independent copy of $(h_1 \cdots h_n)$, and $\sigma_1 \cdots \sigma_n$ be i.i.d. Rademacher variables with equal probability of being ± 1. Using techniques of Rademacher complexities ([Bartlett and Mendelson, 2002]), we bound $\mathrm{E}[\Phi(h_1 \cdots h_n)]$ as follows:

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

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

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

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

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

23

$$= \frac{1}{-} \mathrm{E}_{h_1,\ldots,h_n,h'_1,\ldots,h'_n,\epsilon_1,\ldots,\epsilon_n} \left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \right.$$

$$\sum_i^n$$

Note that ϕ^2 is 2μ-Lipschitz over $[-\mu, \mu]$, and $\ell_i(\mathbf{w}) + \hat{\ell}_i(\mathbf{w}) \in [-2\gamma_k^+, 2\gamma_k^+]$. Then, from the comparison theorem of Rademacher complexities (Ledoux and Talagrand, 1991), in particular Lemma 5 of Meir and Zhang (2003), we have

$$
\mathbb{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \sum_{i=1}^{n} \sigma_i\left(\ell_i(\mathbf{w}) + \hat{\ell}_i(\mathbf{w})\right)^2\right]
$$
$$
\leq 4\gamma_k^+ \sqrt{\cdot}\, \mathbb{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \sum_{i=1}^{n} \sigma_i\left(\ell_i(\mathbf{w}) + \hat{\ell}_i(\mathbf{w})\right)\right]
$$
$$
\leq 4\gamma_k^+ \sqrt{\cdot}\left(\mathbb{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \sum_{i=1}^{n} \sigma_i \ell_i(\mathbf{w})\right] + \mathbb{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \sum_{i=1}^{n} \sigma_i \hat{\ell}_i(\mathbf{w})\right]\right). \tag{53}
$$

Similarly, we have

$$
\mathbb{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \sum_{i=1}^{n} \sigma_i\left(\ell_i(\mathbf{w}) - \hat{\ell}_i(\mathbf{w})\right)^2\right]
$$
$$
\leq 4\gamma_k^+ \sqrt{\cdot}\left(\mathbb{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \sum_{i=1}^{n} \sigma_i \ell_i(\mathbf{w})\right] + \mathbb{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \sum_{i=1}^{n} \sigma_i \hat{\ell}_i(\mathbf{w})\right]\right). \tag{54}
$$

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

$$
\mathbb{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \sum_{i=1}^{n} \sigma_i\left(\ell_i(\mathbf{w}) - \ell_i(\mathbf{w}_*)\right)\|\mathbf{w} - \mathbf{w}_*\|\right]
$$
$$
\leq 2\gamma_k^+ \sqrt{\cdot}\left(\underbrace{\mathbb{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \sum_{i=1}^{n} \sigma_i \ell_i(\mathbf{w})\right]}_{:=C_1} + \underbrace{\mathbb{E}\left[\sup_{\mathbf{w}:\|\mathbf{w}-\mathbf{w}_*\|\leq\gamma_k^+} \sum_{i=1}^{n} \sigma_i \ell_i(\mathbf{x})\right]}_{:=C_2}\right). \tag{55}
$$

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

$$
\left|\ell_i(\mathbf{w}) - \ell_i(\mathbf{w}')\right| = \frac{1}{\mu}\left|\phi'(\langle \mathbf{w}, \mathbf{x}_i\rangle, y_i) - \phi'(\langle \mathbf{w}', \mathbf{x}_i\rangle, y_i)\right|
$$
$$
\leq \sqrt{\cdot}\left|\langle \mathbf{w}, \mathbf{x}_i\rangle - \langle \mathbf{w}', \mathbf{x}_i\rangle\right| = \sqrt{\cdot}\left|\langle \mathbf{x}_i, \mathbf{w} - \mathbf{w}_*\rangle - \langle \mathbf{x}_i, \mathbf{w}' - \mathbf{w}_*\rangle\right|.
$$

Applying the comparison theorem of Rademacher complexities again, we have

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

ZHANG YANG