Fast Excess Risk Rates via Offset Rademacher Complexity

Chenguang Duan [1 2 3] **Yuling Jiao** [1 2 3] **Lican Kang** [1 4] **Xiliang Lu** [1 2 3] **Jerry Zhijian Yang** [1 2 3]

Abstract

Based on the offset Rademacher complexity, this work outlines a systematical framework for deriving sharp excess risk bounds in statistical learning without Bernstein condition. In addition to recovering fast rates in a unified way for some parametric and nonparametric supervised learning models with minimum identifiability assumptions, we also obtain new and improved results for LAD (sparse) linear regression and deep logistic regression with deep ReLU neural networks, respectively.

1. Introduction

Let $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ denote the predictor and response random variables pair distributed from an unknown probability distribution μ, where $\mathcal{X} \subseteq \mathbb{R}^d$ and $\mathcal{Y} \subseteq \mathbb{R}$. In statistics learning problems, we usually observe an independently and identically distributed (i.i.d.) sample $\mathbb{D} := \{(X_i, Y_i)\}_{i=1}^n$ drawn from μ. With this sample in mind, we can establish a statistical procedure, a measurable function mapping $\mathcal{X}$ to $\mathcal{Y}$. Specifically, let $\mathcal{F}$ be a class of measurable functions and introduce a loss function $\ell(\cdot, \cdot) : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$, then we denote the empirical risk minimizer (ERM) as

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}} R_n(f) := \frac{1}{n} \sum_{i=1}^n \ell(Y_i, f(X_i)). \tag{1}$$

To characterize the finite-sample performance of ERM $\hat{f}_n$ in (1), we can consider the excess risk, denoted by

$$R(\hat{f}_n) - \inf_{f \in \mathcal{J}} R(f) \tag{2}$$

where $R(f) := \mathbb{E}\ell(Y, f(X))$ refers to the population risk of $f \in \mathcal{F}$, and $\mathcal{J} \supseteq \mathcal{F}$ is a reference function class which may be different from the admissible set $\mathcal{F}$. We call the learning model well-specified when $\mathcal{J} = \mathcal{F}$, otherwise it is called misspecified. Obviously, it incurs an additional approximation error denoted by $\mathcal{E}_{\text{app}} = \inf_{f \in \mathcal{F}} R(f) - \inf_{f \in \mathcal{J}} R(f)$ in misspecified settings.

Deriving the convergence rate of (2) i.e., providing a generalization analysis for the ERM $\hat{f}_n$ in (1) stands a central place in machine learning and statistics since it quantifies whether the ERM learning procedure obtained from the training data generalize well on unseen data. With tools of symmetrization and Lipschitz contraction, the classical method for bounding (2) in the well-specified setting is transforming into bounding the global Rademacher average, denoted as

$$\mathbb{E}\Big[\sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n \tau_i f(X_i)\Big], \tag{3}$$

where $\{\tau_i\}_{i=1}^n$ are i.i.d. Rademacher random variables, that is, $P(\tau_i = 1) = P(\tau_i = -1) = 1/2$. Thereafter, global Rademacher average (3) can be bounded in terms of the complexity of function class $\mathcal{F}$, say $\text{Complex}(\mathcal{F})$, such as covering number or VC-dimension, see Van Der Vaart & Wellner (1996); Van de Geer & van de Geer (2000); Van der Vaart (2000); Giné & Nickl (2021) for detailed analysis from the viewpoint of the empirical process. But, global Rademacher average (3) may lead to a suboptimal error bound taking the order of $\mathcal{O}(\sqrt{\frac{\text{Complex}(\mathcal{F})}{n}})$. Subsequently, Bartlett et al. (2005); Koltchinskii (2006) proposed local Rademacher average

$$\mathbb{E}\left[\sup_{f \in \mathcal{F}: \mathbb{E}[f(X)^2] \le r} \frac{1}{n} \sum_{i=1}^n \tau_i f(X_i)\right], \tag{4}$$

where $r > 0$. Obviously, local Rademacher average (4) utilizes the local structure of function class $\mathcal{F}$, and it can reach the sharp error bound with the order of $\mathcal{O}(\frac{\text{Complex}(\mathcal{F})}{n})$ in some scenarios. However, this improvement needs the so-called Bernstein condition which reads for each $f \in \mathcal{F}$,

$$\mathbb{E}[|f(X) - f^*(X)|^2] \le B\mathbb{E}[\ell(Y, f(X)) - \ell(Y, f^*(X))],$$

for some $B > 0$, where f^* minimizes in $\mathcal{F}$ the functional $f \mapsto \mathbb{E}[\ell(Y, f(X))]$. In this case, the application of local Rademacher complexity is restricted to some strongly

[1]School of Mathematics and Statistics, Wuhan University, Wuhan, 430072, P.R. China. [2]Hubei National Center for Applied Mathematics, Wuhan University, Wuhan, 430072, P.R. China. [3]Hubei Key Laboratory of Computational Science, Wuhan University, Wuhan, 430072, P.R. China. [4]Center for Quantitative Medicine, Duke-NUS Medical School, Singapore. Correspondence to: Jerry Zhijian Yang <zjyang.math@whu.edu.cn>.

Proceedings of the 40^{th} International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023.

convex and bounded loss functions. To circumvent this Bernstein condition, Liang et al. (2015) introduced the offset Rademacher complexity as an intuitive alternative to the local Rademacher average to obtain a sharp error bound for the ERM with the least squared loss. Mendelson (2018) studied the behavior of ERM with convex and differentiable loss functions including the log and Huber loss functions by developing the small-ball method. Later, Xu & Zeevi (2021) provided a refined localized procedure to obtain tighter error bounds for the ERM (or estimator obtained via certain optimization algorithms) with convex, differentiable and bounded losses. More recently, Kanade et al. (2022) introduced the offset condition as a replacement of Bernstein condition in the setting of model aggregation to derive an exponential-tail excess risk bound. In model aggregation, the test function class $\mathcal{J}$ in (2) is a subset of admissible class $\mathcal{F}$ in (2), which makes the techniques developed in Kanade et al. (2022) may not be used for general supervised learning setting where $\mathcal{F} \subseteq \mathcal{J}$ is our interest.

Although there is a large body of literature focusing on the error analysis of the excess risk, see Bartlett et al. (2005); Koltchinskii (2006); Liang et al. (2015); Mendelson (2018); Xu & Zeevi (2021); Kanade et al. (2022) and references therein, there still remains some challenging issues to be solved. Especially, in the absence of additional assumptions like the Bernstein condition a sharp error bound $\mathcal{O}(\frac{\text{Complex}(\mathcal{F})}{n})$ for the ERM is not available when the loss function in (1) is only Lipschitz continuous but neither differentiable nor bounded such as the least absolute derivation (LAD) loss. In this work, we make efforts to address these issues. With the tools of offset Rademacher complexity, we obtain a sharp error bound with the order of $\mathcal{O}(\frac{\text{Complex}(\mathcal{F})}{n})$ for the expected excess risk with general Lipschitz continuous loss functions including the LAD. To this end, we use symmetrization and decompose the expected excess risk of ERM into a summation of an approximation error $\mathcal{E}_{\text{app}} = \inf_{f\in\mathcal{F}} R(f) - \inf_{f\in\mathcal{J}} R(f)$ vanishing in well-specified settings and an offset Rademacher complexity term. Instead of using the chaining method in Liang et al. (2015), we apply Hoeffding's inequality to proceed, which leads to an upper bound of the offset Rademacher complexity with the order of $\mathcal{O}(\frac{\log \mathbb{E}_{\mathbb{X}} N_\infty(\delta,\mathcal{F},\mathbb{X})}{n})$, see Theorem 2.3 for details. We highlight that we need neither the Bernstein condition (Bartlett et al., 2005; Koltchinskii, 2006) nor the offset condition (Kanade et al., 2022).

We summarize the contributions of this paper as follows.

(i) We provide sharp error bounds for the expected excess risk of ERM by using the tools of offset Rademacher complexity, wherein Bernstein condition (Bartlett et al., 2005; Koltchinskii, 2006) or offset condition (Kanade et al., 2022) is not required. Thus, the obtained error bound is applicable for general loss functions which may be nonsmooth and unbounded.

(ii) Under conditions that make the models considered are identifiable, we recover fast rates for parametric models including ordinary least squares (ℓ_0 constrained least squares), logistic regression (ℓ_0 constrained logistic regression) and deep nonparametric regression with least square loss in a unified way.

(iii) We obtain a new result for sparse liner regression with LAD loss and derive an improved rate for deep logistic regression with ReLU neural networks under very mild conditions.

1.1. Outlines

The rest of this paper is organized as follows. In Section 2, we sketch out the framework deriving an upper bound of the expected excess risk based on offset Rademacher complexity. Extensions to the error analysis of parametric models incorporating with sparsity and nonparametric models are given in Sections 3 and 4, respectively. Concluding is shown in Section 5. The proofs of all lemmas, corollaries and theorems are deferred in Appendices A to C.

1.2. Notations

In this section, we introduce the notations used throughout this paper. The set of positive integers is denoted by $\mathbb{N} := \{1, 2, \ldots\}$. We also denote $\mathbb{N}_0 := \mathbb{N} \cup \{0\}$ for convenience. We write $\|x\|_q := (\sum_{i=1}^d |x_i|^q)^{\frac{1}{q}}$ as the q-norm ($q \in [1, \infty]$) of a vector $x = (x_1, \ldots, x_d)^T \in \mathbb{R}^d$, and $\|x\|_0$ represents the number of nonzero elements of x. For probability measure ν and measurable function $f : \mathbb{R}^d \to \mathbb{R}$, let $\|f\|^q_{L^q(\nu)} := \mathbb{E}_{X\sim\nu}|f(X)|^q$. For any $a, b \in \mathbb{R}$, $\lceil a \rceil$ denotes the smallest integer no less than a, $a \vee b := \max\{a, b\}$.

2. Main Results

In this section, we construct a unified framework by relating the expected excess risk of ERM to the offset Rademacher complexity. Therefore, we first recall the definition of offset Rademacher complexity (Liang et al., 2015).

2.1. Offset Rademacher complexity

The offset Rademacher complexity is introduced in Liang et al. (2015) to provide sharp bounds for a two-step Star estimator. Denote by $\mathcal{F}$ a class of measurable functions from $\mathcal{X}$ to $\mathbb{R}$, let μ_X be the probability distribution of the predictor X and set $\mathbb{X} := \{X_i\}_{i=1}^n$ as i.i.d. random variables distributed according to μ_X. Let $\{\tau_i\}_{i=1}^n$ be i.i.d. Rademacher random variables. Then, the empirical (or conditional) offset Rademacher complexity (Liang et al., 2015) of $\mathcal{F}$ is defined as the following penalized version of empirical Rademacher

complexity which localizes $\mathcal{F}$ adaptively according to the magnitude of f^2, i.e.,

$$\mathcal{R}_n^{\text{off}}(\mathcal{F}, \beta|\mathbb{X}) := \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \tau_i f(X_i) - \beta f(X_i)^2 \Big| \mathbb{X}\Big],$$

for some $\beta > 0$, and the offset Rademacher complexity of $\mathcal{F}$ is denoted by

$$\begin{aligned}\mathcal{R}_n^{\text{off}}(\mathcal{F}, \beta) &:= \mathbb{E}_{\mathbb{X}}\mathcal{R}_n^{\text{off}}(\mathcal{F}, \beta|\mathbb{X}) \\ &= \mathbb{E}\Big[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \tau_i f(X_i) - \beta f(X_i)^2\Big]. \quad (5)\end{aligned}$$

As shown in Sections 3 and 4, for most of the common parametric and nonparametric supervised learning problems the expected excess risk induced by the Lipschitz continuous loss function can be bounded by the offset Rademacher complexity and approximation errors. Hence, we give the main results of this work on controlling the offset Rademacher complexity with properly chosen classes.

Theorem 2.1. *Let $\mathbb{D} := \{(X_i, Y_i)\}_{i=1}^n$ be i.i.d. copies of $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ distributed from μ, and let $\mathcal{F}$ be a class of measurable functions mapping $\mathcal{X}$ to $[-B, B]$ with $B \geq 1$. Denote the composite function class $\mathcal{G} := \{x \mapsto g(x; f) : x \in \mathcal{X}, f \in \mathcal{F}\}$ given a measurable function g mapping $\mathcal{X} \times \mathcal{F}$ to $\mathbb{R}$. Suppose that $|g(x; f) - g(x; f')| \leq \kappa|f(x) - f'(x)|$ with $\kappa > 0$ and $0 \leq g(x; f) \leq 2\kappa B$, for all $x \in \mathcal{X}$ and $f, f' \in \mathcal{F}$. Then, for any $\omega > 0$ and $f \in \mathcal{F}$ (may depend upon $\mathbb{D}$), it holds*

$$\begin{aligned}&\mathbb{E}_{\mathbb{D}}\Big[\mathbb{E}_X g(X; f)\Big] - (\omega+1)\mathbb{E}_{\mathbb{D}}\Big[\frac{1}{n}\sum_{i=1}^n g(X_i; f)\Big] \quad (6)\\ &\leq (\omega+2)\cdot\mathcal{R}_n^{\text{off}}\left(\mathcal{G}, \frac{\omega}{2B\kappa(\omega+2)}\right),\end{aligned}$$

where

$$\begin{aligned}&\mathcal{R}_n^{\text{off}}(\mathcal{G}, \beta)\\ &:= \mathbb{E}\sup_{f\in\mathcal{F}}\Big(\frac{1}{n}\sum_{i=1}^n \tau_i g(X_i; f) - \frac{\beta}{n}\sum_{i=1}^n g^2(X_i; f)\Big),\end{aligned}$$

with $\beta > 0$, refers to the offset Rademacher complexity of $\mathcal{G}$.

Setting $\omega = 2$, the term in (6) serves as an upper bound for several common well-specified parametric learning problems, see Section 3 for details. Next, it remains to derive an upper bound of the offset Rademacher complexity of function class $\mathcal{G}$ in Theorem 2.1. Hence, we introduce the definition of covering number (Anthony et al., 1999) used to characterize this upper bound in the following Theorem 2.3.

Definition 2.2 (Empirical covering number)**.** Given a sample $\mathbb{X} = \{X_i\}_{i=1}^n$, define the empirical L^p $(1 \leq p \leq \infty)$ metric $\|\cdot\|_{\mathbb{X},p}$ based on the sample $\mathbb{X}$ as

$$\|f\|_{\mathbb{X},p} := \Big(\frac{1}{n}\sum_{i=1}^n |f(X_i)|^p\Big)^{1/p}.$$

A set $\mathcal{F}_\delta$ is called a $\|\cdot\|_{\mathbb{X},p}$ δ-cover for $\mathcal{F}$ if for every $f \in \mathcal{F}$, there exits $f_\delta \in \mathcal{F}_\delta$ such that $\|f - f_\delta\|_{\mathbb{X},p} \leq \delta$. Furthermore,

$$N_p(\delta, \mathcal{F}, \mathbb{X}) := \inf\big\{|\mathcal{F}_\delta| : \mathcal{F}_\delta \text{ is a } \|\cdot\|_{\mathbb{X},p} \text{ } \delta\text{-cover of } \mathcal{F}\big\}$$

is called the L^p δ-covering number of $\mathcal{F}$ conditionally on $\mathbb{X}$.

Theorem 2.3 (An upper bound of offset Rademacher complexity)**.** *Assume the conditions of Theorem 2.1 hold. Then, for any $\delta > 0$, the offset Rademacher complexity of $\mathcal{G}$ satisfies*

$$\begin{aligned}&\mathcal{R}_n^{\text{off}}(\mathcal{G}, \beta)\\ &\leq \frac{1 + \log \mathbb{E}_{\mathbb{X}}\big[N_\infty(\delta, \mathcal{F}, \mathbb{X})\big]}{2\beta n} + (1 + 4B\beta\kappa)\kappa\delta,\end{aligned}$$

where $\beta > 0$, and κ, B are defined in Theorem 2.1.

Remark 2.4. Given $f \in \mathcal{F}, Z = (X, Y) \sim \mu$. Let $\ell(f, Z) = \ell(Y, f(X))$, $\tilde{g}_f(Z) = \ell(f, Z) - \ell(f^*, Z)$, $g(X; f) = \mathbb{E}_{Y|X}[\tilde{g}_f(Z)|X]$ where $f^* := \arg\min_{f\in\mathcal{F}} \mathbb{E}[\ell(Y, f(X))]$ in the well-specified setting and $f^* := \arg\min_{f\in\mathcal{J}} \mathbb{E}[\ell(Y, f(X))]$ in the misspecified setting, respectively. Assuming that the loss function ℓ is Lispchitz continuous in its second element and the function class $\mathcal{F}$ is bounded, the requirement on $\mathcal{G}$ in Theorems 2.1 and 2.3, will meet for most of the common parametric and nonparametric supervised learning problems under mild conditions needed for model identifiability, see Sections 3 and 4 for details. Therefore, we draw a unified blueprint for constructing the error bound of the expected excess risk of ERM in terms of a class of Lipschitz continuous loss functions which may be nonsmooth and unbounded.

Remark 2.5. The fast offset Rademacher complexity rate $\mathcal{O}(1/n)$ obtained in Theorem 2.3 improves the rate $\mathcal{O}(1/\sqrt{n})$ derived in Liang et al. (2015) via chaining method. Note that the Bernstein condition (Bartlett et al., 2005; Koltchinskii, 2006) and offset condition (Kanade et al., 2022) are not imposed. Compared to Liang et al. (2015), we generalize their results since we have a wider range of applicability, rather than just being limited to the analysis of least squares estimations. In additions, we have that the offset Rademacher complexity is closely associated with the covering number, which provides sufficient intuitions enabling us to establish non-asymptotic error bounds of interests in the sequel.

Remark 2.6. There is some "tension" in that there is a bound parameter (here ω) that taken small tightens up the empirical

risk part of the upper bound (Theorem 2.1), but causes typical bounds on the offset Rademacher complexity to grow (Theorem 2.3).

3. Applications in Parametric Models

In this section, we consider excess risk bounds and estimation errors in some well-specified parametric regression and classification models, and there ℓ_0 constraint version under mild assumptions to guarantee that the learning problems are identifiable.

3.1. Linear regression model

Example 3.1 (Well-specified linear regression model). Let $\Theta \subseteq \mathbb{R}^d$ be a parameter space and $\varphi : \mathcal{X} \to \mathbb{R}^d$ be a feature map. The linear regression model is

$$Y = \varphi(X)^T\theta^* + \varepsilon, \tag{7}$$

with $(X, Y) \sim \mu$, $\theta^* \in \Theta$ and ε being a random error term.

Given an i.i.d. sample $\mathbb{D} := \{X_i, Y_i\}_{i=1}^m$ from the distribution μ in model (7), we aim to derive an estimator of θ^* and construct the related non-asymptotic error bounds. To that end, we define the ERM

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \ell(Y_i, \varphi(X_i)^T\theta), \tag{8}$$

where $\ell : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ denotes the loss function.

To bound the estimation error of the ERM $\hat{\theta}_n$ in (8), we impose the following assumption on the loss function ℓ.

Assumption 3.2. Assume that the loss function $\ell : \mathcal{Y}\times\mathbb{R} \to [0, \infty)$ admits the Bayes predictor $\varphi(\cdot)^T\theta^*$, that is,

$$\varphi(x)^T\theta^* = \arg\min_{t\in\mathbb{R}} \mathbb{E}\big[\ell(Y, t) - \ell(Y, \varphi(X)^T\theta^*)\big|X = x\big],$$

for all $x \in \mathcal{X}$.

Assumption 3.2 is a minimum requirement to guarantee that the learning problem is identifiable. For example, $\mathbb{E}[\varepsilon|X] = 0$ and $\mathrm{Med}(\varepsilon|X) = 0$ (conditional medium of ε given X) will imply Assumption 3.2 for least square regression and LAD regression, respectively. Additionally, Assumption 3.2 yields the excess risk

$$\mathcal{E}(\varphi(\cdot)^T\theta; \ell) := \mathbb{E}\big[\ell(Y, \varphi(X)^T\theta) - \ell(Y, \varphi(X)^T\theta^*)\big],$$

$\theta \in \Theta$, and the inner excess risk can be denoted by

$$\begin{aligned} &g(x; \varphi(\cdot)^T\theta) \\ &:= \mathbb{E}\big[\ell(Y, \varphi(X)^T\theta) - \ell(Y, \varphi(X)^T\theta^*)\big|X = x\big], \end{aligned} \tag{9}$$

for all $x \in \mathcal{X}$. It can be easily deduced that $\mathcal{E}(\varphi(\cdot)^T\theta; \ell) = \mathbb{E}_X g(X; \varphi(\cdot)^T\theta)$.

Theorem 3.3 (Excess risk bound of linear regression model (7))**.** *Suppose that the feature map φ is bounded by $B_\varphi > 0$, the parametric space Θ is bounded by $B_\theta > 0$, Assumption 3.2 holds, and the inner excess risk in (9) satisfies $|g(x; \varphi(\cdot)^T\theta) - g(x; \varphi(\cdot)^T\theta')| \le \kappa|\varphi(x)^T\theta - \varphi(x)^T\theta'|$ with $\kappa > 0$ for any $x \in \mathcal{X}$. Then ERM $\hat{\theta}_n$ in (8) satisfies*

$$\mathbb{E}_{\mathbb{D}}\big[\mathcal{E}(\varphi(\cdot)^T\hat{\theta}_n; \ell)\big] \le 4\mathfrak{R}_n^{\mathrm{off}}\Big(\mathcal{G}, \frac{1}{4B_\varphi B_\theta\kappa}\Big),$$

where $\mathcal{G} := \{x \mapsto g(x; \varphi(\cdot)^T\theta) : x \in \mathcal{X}, \theta \in \Theta\}$.

Theorem 3.3 shows that for a class of loss functions satisfying Assumption 3.2, the expected excess risk of ERM $\hat{\theta}_n$ defined in (8) can be bounded by the offset Rademacher complexity. More generally, under some mild conditions we can derive a straightforward error bound between $\hat{\theta}_n$ and θ^* when the loss function ℓ taking a special form such that Assumption 3.2 holds.

Remark 3.4. For simplify of presentation, we assume the feature map is bounded as in Theorem 3.3. Indeed, our method can extend to unbounded data with some distributional assumptions. For instance, we may assume $\varphi(X)$ is sub-Gaussian for $X \sim \mu_X$. Then we can obtain a similar bound via the technique of truncation.

Roughly speaking, we consider the least square and LAD estimation problems, which correspond to the least square loss ($\ell(a, b) = (a - b)^2$) and LAD loss ($\ell(a, b) = |a - b|$).

Corollary 3.5 (Excess risk of least squares regression)**.** *Assume that the conditions of Theorem 3.3 hold, and set the loss function ℓ as a least square loss and $\mathbb{E}[\varepsilon|X] = 0$. Then we have*

$$\begin{aligned} &\mathbb{E}_{\mathbb{D}}\big[\|\hat{\theta}_n - \theta^*\|_\Sigma^2\big] \\ &\qquad \le \frac{8B_\varphi^2B_\theta^2}{n}\Big(1 + d\log\frac{3B_\varphi B_\theta}{\delta}\Big) + 8B_\varphi B_\theta\delta, \end{aligned}$$

for any $\delta \in (0, 1)$, where $\Sigma := \mathbb{E}_X[\varphi(X)\varphi(X)^T]$.

Corollary 3.6 (Excess risk of least absolute deviations regression)**.** *Assume that the conditions of Theorem 3.3 hold, and set the loss function ℓ as a LAD loss and $\mathrm{Med}(\varepsilon|X) = 0$. Then we have*

$$\begin{aligned} &\mathbb{E}_{\mathbb{D}}\big[\mathcal{E}(\varphi(\cdot)^T\hat{\theta}_n; \ell_{\mathrm{lad}})\big] \\ &\qquad \le \frac{2B_\varphi B_\theta}{n}\Big(1 + d\log\frac{3B_\varphi B_\theta}{\delta}\Big) + 2\delta, \end{aligned}$$

for any $\delta \in (0, 1)$.

As far as we know, the convergence rate $\mathcal{O}(1/n)$ established in Corollary 3.6 (Theorem 3.14) for the LAD loss is new. It can not been derived by the localized Rademacher method (Bartlett et al., 2005; Koltchinskii, 2006) or refined localized convergence procedure (Xu & Zeevi, 2021) since neither the

Bernstein condition needed (Bartlett et al., 2005; Koltchinskii, 2006) nor the statistical noise of smooth population risk used in Xu & Zeevi (2021) will hold for the LAD loss. With additional stability assumption as Assumption 3.7 below, we can establish the estimation error bound for $\|\hat{\theta}_n - \theta^*\|$.

Assumption 3.7. There exist some constants $\xi > 0$ and $\tau > 0$ such that for any $|a| \le \xi$,

$$\left|F_{Y|X}\left(\varphi(x)^T\theta^* + a\right) - F_{Y|X}\left(\varphi(x)^T\theta^*\right)\right| \ge \tau|a|, \text{ a.s.},$$

where $F_{Y|X}(\cdot)$ is the cumulative distribution function of Y given X.

Moreover, there exists a constant $C > 0$ such that for any $\theta \in \Theta$,

$$|\varphi(x)^T\theta - \varphi(x)^T\theta^*| \ge C\|\theta - \theta^*\|_2, \text{ a.s.}$$

Corollary 3.8. *Assume that the conditions of Theorem 3.3 and Assumption 3.7 hold for the LAD loss with* $\mathrm{Med}(\varepsilon|X) = 0$. *We have*

$$\min\left\{\mathbb{E}_{\mathbb{D}}\left[\|\hat{\theta}_n^k - \theta^*\|_2^2\right], 2\mathbb{E}_{\mathbb{D}}\left[\|\hat{\theta}_n^k - \theta^*\|_2\right]\right\} \le \frac{2C_{\tau,\xi}B_\varphi B_\theta}{nC}\left(1 + d\log\frac{3B_\varphi B_\theta}{\delta}\right) + \frac{2C_{\tau,\xi}\delta}{C},$$

where $C_{\tau,\xi} := \max\{\frac{4}{\tau}, \frac{8}{\tau\xi}\}$ *and* C *is an absolute constant.*

Remark 3.9. In Corollary 3.5 and Corollary 3.6, the least square and LAD losses satisfy Assumption 3.2 under the assumptions that $\mathbb{E}[\varepsilon|X] = 0$ or $\mathrm{Med}(\varepsilon|X) = 0$. Assumption 3.7 is a mild condition on the conditional distribution such that there exists a neighborhood around $\varphi(x)^T\theta^*$ in which the fluctuation of the conditional probability distribution function $F_{Y|X}(\cdot)$ is larger than that of $\varphi(x)^T\theta^*$. Besides, Padilla & Chatterjee (2020); Shen et al. (2021) also introduced some similar conditions to this condition. Assumption 3.7 is weaker than Condition 2 in He & Shi (1994) where the density function of response is supposed to be lower bounded by some positive constant and is weaker than condition D.1 in Belloni & Chernozhukov (2011) which assumes that the conditional density of Y given $X = x$ is continuously differentiable and bounded away from zero uniformly for all quantile level and all x and is also weaker than Assumption 2 in Hernan Madrid Padilla et al. (2020) where the conditional density of Y given $X = x$ is supposed to have a positive lower bound determined by some positive constant.

3.2. Logistic regression model

Example 3.10 (Well-specified logistic regression model). Let $\Theta \subseteq \mathbb{R}^d$ be a parameter space and $\varphi : \mathcal{X} \to \mathbb{R}^d$ be a feature map. The logistic regression model reads

$$Y \sim \text{Bernoulli}(\eta(X)), \text{ with } \eta(X) = \sigma(\varphi(X)^T\theta^*), \quad (10)$$

where $(X, Y) \sim \mu$, $\theta^* \in \Theta$, and $\sigma(a) = \frac{1}{1+\exp(-a)}$ denotes the sigmoid function.

With Logistic regression model (10) and an i.i.d. sample $\mathbb{D} := \{X_i, Y_i\}_{i=1}^n$, the ERM can be defined as

$$\hat{\theta}_n \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \ell(Y_i, \varphi(X_i)^T\theta)), \quad (11)$$

where

$$\ell(a, b) := \log(1 + \exp(-ab)), a = \pm 1, b \in \mathbb{R}, \quad (12)$$

denotes the logistic loss.

Theorem 3.11 (Excess risk of logistic regression model (10))**.** *Suppose that the feature map* φ *is bounded by* $B_\varphi > 0$, *the parametric space* Θ *is bounded by* $B_\theta > 0$. *Then, for any* $\delta \in (0, 1)$, *the ERM* $\hat{\theta}_n$ *in* (11) *satisfies*

$$\mathbb{E}_{\mathbb{D}}\left[\|\hat{\theta}_n - \theta^*\|_\Sigma^2\right] \le C_B\left\{\frac{2B_\varphi B_\theta}{n}\left(1 + d\log\frac{3B_\varphi B_\theta}{\delta}\right) + 2\delta\right\},$$

where $\Sigma = \mathbb{E}_X[\varphi(X)\varphi(X)^T]$ *and* $C_B = 2(1 + \exp(B))^2\exp(-B)$.

3.3. Sparsity models

In this section, we further consider linear and logistic regression models with sparsity issues (Assumption 3.12). Over the past two decades, to investigate the model selection and variable estimation in high-dimensional and sparse models is a significant research topic in statistics, and there are numerous related studies, one can refer to Tibshirani (1996); Fan & Li (2001); Zou (2006); Zhang & Huang (2008); Zou & Zhang (2009); Ye & Zhang (2010); Zhang & Zhang (2012); Wang et al. (2013); Loh & Wainwright (2015); Huang et al. (2018); Wainwright (2019), etc.

Assumption 3.12 (Sparsity)**.** Assume θ^* defined in (7) or (10) satisfies $\|\theta^*\|_0 \le k$, where k is a positive integer quantifying the sparse level of the underlying regression coefficients θ^*.

With sparsity Assumption 3.12, we can define the ℓ_0-constrained estimation

$$\hat{\theta}_n^k \in \arg\min_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^n \ell(Y_i, \varphi(X_i)^T\theta), \quad \text{s.t. } \|\theta\|_0 \le k. \quad (13)$$

Similar to the analysis in Sections 3.1 and 3.2, we can also derive three types of non-asymptotic error bounds of the ℓ_0-constrained estimators in terms of different loss functions, shown in the following Theorems 3.13 to 3.15.

Theorem 3.13 (Excess risk of ℓ_0-constrained least squares regression)**.** *Set the loss function ℓ as a least square loss in (13). Suppose that the feature map φ is bounded by $B_\varphi > 0$, the parametric space Θ is bounded by $B_\theta > 0$ and and $\mathbb{E}[\varepsilon|X] = 0$ in model (7). Let θ^* be defined in (7) satisfying Assumption 3.12. Then, the ℓ_0-constrained least square estimator $\hat{\theta}_n^k$ satisfies*

$$\begin{aligned}&\mathbb{E}_{\mathbb{D}}\big[\|\hat{\theta}_n^k - \theta^*\|_\Sigma^2\big]\\ &\leq \frac{8B_\varphi^2 B_\theta^2}{n}\Big(1 + k\Big\{1 + \log\frac{d}{k} + \log\Big(\frac{3B_\varphi B_\theta}{\delta}\Big)\Big\}\Big)\\ &\quad + 8B_\varphi B_\theta \delta,\end{aligned}$$

for any $\delta \in (0,1)$, where $\Sigma = \mathbb{E}_X[\varphi(X)\varphi(X)^T]$.

Theorem 3.14 (Excess risk of ℓ_0-constrained least absolute deviations regression)**.** *Set the loss function ℓ as the LAD loss in (13). Suppose that the feature map φ is bounded by $B_\varphi > 0$, the parametric space Θ is bounded by $B_\theta > 0$ and $\mathrm{Med}(\varepsilon|X) = 0$ in model (7). Let θ^* be defined in (7) satisfying Assumption 3.12. Then, for any $\delta \in (0,1)$, the ℓ_0-constrained least absolute deviations estimator $\hat{\theta}_n^k$ satisfies*

$$\begin{aligned}&\mathbb{E}_{\mathbb{D}}\Big[\mathcal{E}(\varphi(\cdot)^T\hat{\theta}_n^k; \ell_{\mathrm{lad}})\Big]\\ &\leq \frac{2B_\varphi B_\theta}{n}\Big(1 + k\Big\{1 + \log\frac{d}{k} + \log\Big(\frac{3B_\varphi B_\theta}{\delta}\Big)\Big\}\Big)\\ &\quad + 2\delta.\end{aligned}$$

Moreover, if Assumption 3.7 holds, we have

$$\begin{aligned}&\min\Big\{\mathbb{E}_{\mathbb{D}}\big[\|\hat{\theta}_n^k - \theta^*\|_2^2\big], 2\mathbb{E}_{\mathbb{D}}\big[\|\hat{\theta}_n^k - \theta^*\|_2\big]\Big\}\\ &\leq \frac{2C_{\tau,\xi}B_\varphi B_\theta}{nC}\Big[1 + k\Big(1 + \log\frac{d}{k} + \log\Big(\frac{3B_\varphi B_\theta}{\delta}\Big)\Big)\Big]\\ &\quad + \frac{2\delta C_{\tau,\xi}}{C},\end{aligned}$$

where $C_{\tau,\xi} = \max\{\frac{4}{\tau}, \frac{8}{\tau\xi}\}$ and C is an absolute constant.

Theorem 3.15 (Excess risk of ℓ_0-constrained logistic regression)**.** *Set the loss function ℓ as the logistic loss (12) in (13). Suppose that the feature map φ is bounded by $B_\varphi > 0$, the parametric space Θ is bounded by $B_\theta > 0$ in model (10). Let θ^* be defined as (10) satisfying Assumption 3.12. Then, for any $\delta \in (0,1)$, the ℓ_0-constrained logistic regression estimator $\hat{\theta}_n^k$ satisfies*

$$\begin{aligned}&\mathbb{E}_{\mathbb{D}}\Big[\|\hat{\theta}_n^k - \theta^*\|_\Sigma^2\Big]\\ &\leq \frac{2C_B B_\varphi B_\theta}{n}\Big[1 + k\Big(1 + \log\frac{d}{k} + \log\Big(\frac{3B_\varphi B_\theta}{\delta}\Big)\Big)\Big]\\ &\quad + 2C_B\delta,\end{aligned}$$

where $\Sigma = \mathbb{E}_X[\varphi(X)\varphi(X)^T]$ and $C_B = 2(1 + \exp(B))^2\exp(-B)$.

Remark 3.16. Setting $\varphi(\cdot)$ in model (7) as the identical operator, then it corresponds to the sparse linear model

$$Y = X^T\theta^* + \varepsilon, \quad \|\theta^*\|_0 \leq k. \tag{14}$$

Note that this linear model (14) is identifiable if and only if, for any k-sparse θ^* and β^*,

$$\boldsymbol{X}\theta^* = \boldsymbol{X}\beta^* \text{ implies } \theta^* = \beta^*, \tag{15}$$

where $\boldsymbol{X} := [X_1, X_2, \ldots, X_n]^T \in \mathbb{R}^{n\times d}$ denotes the design matrix. See Arias-Castro & Lounici (2014) for more details. Besides, this identifiable condition (15) is equivalent to that, for any subset $J \subset \{1, \ldots, d\}$ with $|J| \leq 2k$ ($|J|$ represents the cardinality of J), the submatrix $\boldsymbol{X}_J$ is full-rank (Arias-Castro & Lounici, 2014), and it is also equivalent to $\sigma_{2k} > 0$ denoted by

$$\begin{aligned}\sigma_{2k} := \min\Big\{&\sigma_{\min}(\boldsymbol{X}_{2k}^T\boldsymbol{X}_{2k}/n) :\\ &\boldsymbol{X}_{2k} \in \mathbb{R}^{n\times 2k} \text{consists of } 2k \text{ columns of } \boldsymbol{X}\Big\},\end{aligned}$$

where $\sigma_{\min}(\boldsymbol{X}_{2k}^T\boldsymbol{X}_{2k}/n)$ is the minimum eigenvalue of $\boldsymbol{X}_{2T}^T\boldsymbol{X}_{2T}/n$. We highlight that Theorems 3.13 and 3.14 give the sharp bounds of ℓ_0-constrained least squared and LAD estimators, among which it only requires the identifiability of the model (7) without additional assumptions on the design matrix. However, the existing theoretical studies of the regularized methods require stronger assumptions than the fundamental identifiable condition (15). Van de Geer & Bühlmann (2009) demonstrated that the compatibility condition (CC) is the weakest condition in the existing literature on deriving oracle results for the Lasso (Tibshirani, 1996), while the model identifiability condition $\sigma_{2k} > 0$ is weaker than CC. Zhang & Zhang (2012) showed that obtaining the oracle result of ℓ_0-regularized least square estimator depends on the condition $\sigma_{3k} > 0$. Zhang et al. (2017) illustrated that one can not derive the result of Zhang & Zhang (2012) in other regularized methods including Lasso (Tibshirani, 1996), SCAD (Fan & Li, 2001) and MCP (Zhang, 2010) without additional conditions. Obviously, $\sigma_{3k} > 0$ is a stronger requirement than the condition $\sigma_{2k} > 0$. Furthermore, Theorem 3.15 gives the oracle result of ℓ_0-regularized logistic estimator without imposing the restricted strong convexity condition required in Loh & Wainwright (2015).

4. Applications in Deep Nonparametric Models

In this section, we consider nonparametric regression and classification models in misspecified settings with deep neural networks.

4.1. Nonparametric regression models

We have a nonparametric regression model

$$Y = f_0(X) + \varepsilon, \tag{16}$$

where (X, Y) is the predictor and response variables pair taking values in $\mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$, $f_0 : \mathbb{R}^d \to \mathbb{R}$ is the underlying regression function and $\varepsilon \in \mathbb{R}$ is the random noise. Let μ be the joint distribution of (X, Y) on $\mathcal{X} \times \mathcal{Y}$, and μ_X corresponds to the marginal distribution of X. Next, from the perspective of population and sample levels, we provide a procedure to obtain the estimator for f_0 in (16) using an i.i.d. sample $\mathbb{D} := \{X_i, Y_i\}_{i=1}^n$ drawn from μ.

At population level, given a function $f : \mathcal{X} \to \mathbb{R}$ and a supervised loss $\ell : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$, we can define the population ℓ-risk of f as

$$R(f; \ell) := \mathbb{E}_{(X,Y)}\big[\ell(Y, f(X))\big].$$

Then, the Bayes predictor f^* can be denoted by

$$f^*(x) := \arg\min_{t \in \mathbb{R}} \mathbb{E}\big[\ell(Y, t)|X = x\big], \tag{17}$$

for all $x \in \mathcal{X}$, which is the minimizer of the population ℓ-risk, i.e., by setting $\mathcal{J}$ to be the set of measurable functions,

$$f^* = \arg\min\{R(f; \ell) : f \text{ measurable}\}.$$

In nonparametric regression models, the Bayes ℓ-risk $R(f^*; \ell)$ achieves the optimal performance, then we can define the excess ℓ-risk as the deviation with respect to the optimal risk, that is,

$$\begin{aligned}\mathcal{E}(f; \ell) &:= R(f; \ell) - R(f^*; \ell) \\ &= \mathbb{E}_{(X,Y)}\big[\ell(Y, f(X)) - \ell(Y, f^*(X))\big].\end{aligned}$$

We also define the inner excess risk

$$g(x; f) := \mathbb{E}\big[\ell(Y, f(X)) - \ell(Y, f^*(X))\big|X = x\big], \tag{18}$$

for $x \in \mathcal{X}$.

In practice, we only have access to an i.i.d. sample $\mathbb{D}$. Then, at sample level we can denote the empirical ℓ-risk of f as

$$R_n(f; \ell) := \frac{1}{n}\sum_{i=1}^n \ell(Y_i, f(X_i)).$$

Accordingly, the ERM is defined as

$$\hat{f}_n \in \arg\min_{f \in \mathcal{F}} R_n(f; \ell), \tag{19}$$

where $\mathcal{F}$ is a subclass of $\mathcal{J}$.

As defined in (19), the ERM $\hat{f}_n$ depends on the sample $\mathbb{D}$ and the choice of loss function ℓ. Therefore, we firstly focus on the general Lipschitz continuous inner excess risk (18), then generalize to the least square and robust regression problems by considering different robust loss functions. To evaluate the quality of the ERM $\hat{f}_n$ in (19), we still explore the expected excess risk $\mathbb{E}_{\mathbb{D}}[\mathcal{E}(\hat{f}_n)]$. To that end, we introduce the following boundness assumption.

Assumption 4.1. Assume $\{f^*\} \cup \mathcal{F} \subseteq \{f : \|f\|_{L^\infty(\mu_X)} \leq B\}$ with $B \geq 1$.

Now, we give the non-asymptotic error bound of the excess risk of ERM $\hat{f}_n$ under mainly assuming the Lipschitz continuity of the inner excess risk in (18), shown in the following theorem.

Theorem 4.2 (Excess risk bound of nonparametric models)**.** *Let f^* be Bayes predictor defined in (17) and $\hat{f}_n$ be the ERM defined in (19). Suppose that the inner excess risk defined in (18) satisfies $|g(x; f) - g(x; f')| \leq \kappa|f(x) - f'(x)|$ with $\kappa > 0$, $f, f' \in \mathcal{F}$, for any $x \in \mathcal{X}$, and Assumption 4.1 holds. Then it follows that*

$$\mathbb{E}_{\mathbb{D}}\big[\mathcal{E}(\hat{f}_n; \ell)\big] \leq 4\mathcal{R}_n^{\text{off}}\Big(\mathcal{G}, \frac{1}{4B\kappa}\Big) + 3\inf_{f \in \mathcal{F}} \mathcal{E}(f; \ell), \tag{20}$$

where $\mathcal{G} := \{x \mapsto g(x; f) : x \in \mathcal{X}, f \in \mathcal{F}\}$.

Remark 4.3. On the right hand of (20), it includes two terms, i.e, statistical error $\mathcal{R}_n^{\text{off}}\left(\mathcal{G}, \frac{1}{4B\kappa}\right)$ and approximation error $\inf_{f \in \mathcal{F}} \mathcal{E}(f; \ell)$. Note that this approximation error is a derivative of model misspecified settings, which can be controlled by the approximation theory. Additionally, the Lispchitz continuity of the inner excess risk can be naturally met when taking certain forms of loss functions such as the least square and robust losses.

Next, we consider the least square loss ($\ell_{\text{ls}}(a, b) = (a-b)^2$). In this case, we can conclude that the minimizer f^* of ℓ_{ls}-risk exactly coincides with the underlying regression function $f_0(x) = \mathbb{E}[Y|X = x], x \in \mathcal{X}$, under the assumption that $\mathbb{E}[\varepsilon|X] = 0$.

Lemma 4.4 (Excess risk bound of least squares regression)**.** *Set the loss in (17) and (19) as the least square loss. Suppose that Assumption 4.1 holds and $\mathbb{E}[\varepsilon|X] = 0$. Then, for any $\delta \in (0, 1)$, the least square estimation $\hat{f}_n$ satisfies*

$$\begin{aligned}&\mathbb{E}_{\mathbb{D}}\big[\|\hat{f}_n - f_0\|_{L^2(\mu_X)}^2\big] \\ &\leq \frac{32B^2}{n}\big(1 + \log \mathbb{E}_{\mathbb{X}}\big[N_\infty(\delta, \mathcal{F}, \mathbb{X})\big]\big) + 32B\delta \\ &\quad + 12B\inf_{f \in \mathcal{F}} \|f - f_0\|_{L^2(\mu_X)}^2.\end{aligned}$$

In Lemma 4.4, the statistical error takes the order of $\mathcal{O}(\frac{\log \mathbb{E}_{\mathbb{X}} N_\infty(\delta, \mathcal{F}, \mathbb{X})}{n})$ ignoring other terms and the approximation error degenerates to $\inf_{f \in \mathcal{F}} \|f - f_0\|_{L^2(\mu_X)}^2$, and both of them can be determined by the trade-off of bias and variance. More recently, the approximation capability of neural networks to smooth functions has been established,

see Yarotsky (2018); Yarotsky & Zhevnerchuk (2020); Shen et al. (2019); Shen (2020); Lu et al. (2021); Petersen & Voigtlaender (2018) and references therein for more details. It sparks interests of many researchers in studying nonparametric regression using deep neural networks (Bauer & Kohler, 2019; Kohler & Langer, 2021; Schmidt-Hieber, 2020; Nakada & Imaizumi, 2020; Farrell et al., 2021; Jiao et al., 2021; Suzuki & Nitanda, 2021). Thence, we take the function space $\mathcal{F}$ as neural networks with rectified linear unit (ReLU) activation defined in Definition 4.5 and set the target function f_0 as a Hölder continuous function in Definition 4.6 to give the statistical guarantee of deep nonparametric regression, see Corollary 4.7 below. To this end, we firstly recall the definition of ReLU neural networks, then introduce the concept of Hölder class.

Definition 4.5 (ReLU neural networks)**.** A class of feedforward neural networks $f_\theta : \mathbb{R}^d \to \mathbb{R}$ with parameter θ, depth $\mathcal{D}$, width $\mathcal{W}$ and size $\mathcal{S}$ can be defined as

$$f_\theta(x) = v_{\mathcal{D}} \circ \rho \circ v_{\mathcal{D}-1} \circ \rho \circ \cdots \circ \rho \circ v_1 \circ \rho \circ v_0(x),$$

for $x \in \mathbb{R}^d$, where $\rho(x) := \max(0, x)$ is the ReLU activation function and operates pointwisely on x and

$$v_i(x) = A_i x + b_i, \quad i = 0, 1, \ldots, \mathcal{D},$$

$A_i \in \mathbb{R}^{d_{i+1} \times d_i}$ is the weight matrix, $b_i \in \mathbb{R}^{d_{i+1}}$ is the bias vector, and d_i is the width of the i-th layer. The feedforward neural network f_θ has $\mathcal{D}$ hidden layers and $(\mathcal{D}+1)$ layers in total. We use a $(\mathcal{D}+1)$-vector $(d_0, d_1, \ldots, d_{\mathcal{D}})^T$ to describe the width of each layer; in particular, $d_0 = d$ is the dimension of the input X and $d_{\mathcal{D}} = 1$ is the dimension of the output Y. The width $\mathcal{W}$ is defined as the maximum width of hidden layers, i.e., $\mathcal{W} = \max\{d_1, \ldots, d_{\mathcal{D}}\}$. The size $\mathcal{S}$ is defined as the total number of parameters in the neural network f_θ.

Definition 4.6 (Hölder class)**.** Let $\tilde{\alpha}$ be an n-tuple of non-negative integers $\tilde{\alpha}_j$. A partial derivative of f of order $\|\tilde{\alpha}\|_1 = \tilde{\alpha}_1 + \ldots + \tilde{\alpha}_n$ is defined by

$$\partial^{\tilde{\alpha}} f = \frac{\partial^{|\tilde{\alpha}|} f}{\partial x_1^{\tilde{\alpha}_1} \cdots \partial x_n^{\tilde{\alpha}_n}}.$$

For $\varsigma > 0$ with $\varsigma = s + r$, where $s \in \mathbb{N}_0$ and $r \in (0, 1]$ and $d \in \mathbb{N}$, we denote the Hölder class $\mathcal{H}^\varsigma\left([0,1]^d, B\right)$ as

$$\mathcal{H}^\varsigma\left([0,1]^d, B\right) = \Big\{ f : [0,1]^d \to \mathbb{R}, \max_{\|\tilde{\alpha}\|_1 \le s} \left\| \partial^{\tilde{\alpha}} f \right\|_\infty \le B, \max_{\|\tilde{\alpha}\|_1 = s} \sup_{x \neq y} \frac{\left|\partial^{\tilde{\alpha}} f(x) - \partial^{\tilde{\alpha}} f(y)\right|}{\|x - y\|_2^r} \le B \Big\}.$$

For any subset $\mathcal{X} \subseteq [0,1]^d$, we denote $\mathcal{H}^\varsigma(\mathcal{X}, B) := \left\{ f : \mathcal{X} \to \mathbb{R}, f \in \mathcal{H}^\varsigma\left([0,1]^d, B\right) \right\}$.

Based on the above discussion, we can construct a non-asymptotic error bound of deep nonparametric estimator with the square loss by incorporating statistical and approximation errors. This obtained error bound reaches the minimax optimal rate in nonparametric regression (Stone, 1982; Gyorfi et al., 2002; Tsybakov, 2009), shown in the following corollary.

Corollary 4.7. *Let $\mathcal{F}$ be deep ReLU neural networks and f_0 be a Hölder continuous function in Definition 4.6 with $\varsigma = s + r > 0$, $s \in \mathbb{N}_0$ and $r \in (0, 1]$. Suppose the assumptions of Lemma 4.4 hold and μ_X is absolutely continuous with respect to Lebesgue measure. Setting the depth $\mathcal{D} = \mathcal{O}((s+1)^2 n^{\frac{d}{8\varsigma+4d}} \log(n))$ and width $\mathcal{W} = \mathcal{O}((s+1)^2 d^{s+1} n^{\frac{d}{8\varsigma+4d}} \log(n))$, then we have*

$$\mathbb{E}_{\mathbb{D}}\left[\|\hat{f}_n - f_0\|_{L^2(\mu_X)}^2\right] \le B^2 (s+1)^4 d^{2s+(\varsigma \vee 1)} \mathcal{O}\left(n^{-\frac{2\varsigma}{2\varsigma+d}}\right).$$

More generally, we can further consider the robust nonparametric regression including the following robust loss functions (i)-(iv) denoted by ℓ_{robust} uniformly. Table 1 shows that these robust losses considered here are Lipschitz continuous, and we let κ be the Lipschitz constant for notation simplification.

(i) LAD loss: $\ell(a, b) = |a - b|$, $(a, b) \in \mathbb{R}^2$.

(ii) Quantile loss: $\ell(a, b) = \rho_\tau(a - b)$, $(a, b) \in \mathbb{R}^2$, where $\rho_\tau(x) = \tau|x|$ if $x \ge 0$ and $\rho_\tau(x) = (1 - \tau)|x|$ if $x < 0$ for some $\tau \in (0, 1)$.

(iii) Huber loss: $\ell(a, b) = h_\gamma(a - b)$, $(a, b) \in \mathbb{R}^2$, for some $\gamma > 0$, where $h_\gamma(x) = x^2/2$ if $|x| \le \gamma$ and $h_\gamma(x) = \gamma|x| - \gamma^2/2$ otherwise.

(iv) Cauchy loss: $\ell(a, b) = \log\{1 + \zeta^2(a - b)^2\}$, $(a, b) \in \mathbb{R}^2$, for some $\zeta > 0$.

Table 1. Summary of different Lipschitz loss functions.

	Hyper parameter	Lipschitz constant κ
LAD	N/A	1
Quantile	$\tau \in (0, 1)$	$\max\{\tau, 1 - \tau\}$
Huber	$\gamma \in (0, \infty)$	γ
Cauchy	$\zeta \in (0, \infty)$	ζ

Lemma 4.8 (Excess risk bound of robust nonparametric regressions)**.** *Set the loss in (17) and (19) as the robust loss in Table 1. Let f^* be the Bayes predictor of ℓ_{robust}-risk defined in (17) and $\hat{f}_n$ be the empirical ℓ_{robust}-risk minimizer defined in (19). Suppose Assumption 4.1 holds.*

Then, for any $\delta \in (0,1)$, we have

$$\mathbb{E}_{\mathbb{D}}\big[\mathcal{E}(\hat{f}_n;\ell_{\text{robust}})\big] \le \frac{8B\kappa}{n}\big(1+\log \mathbb{E}_{\mathbb{X}}\big[N_\infty(\delta,\mathcal{F},\mathbb{X})\big]\big) + 8\kappa\delta + 3\kappa \inf_{f\in\mathcal{F}} \|f-f^*\|_{L^1(\mu_X)}.$$

4.2. Nonparametric classification models

Now we turn to consider the nonparametric binary classification problems. In this case, the response variable Y takes values in $\{-1,1\}$, then we apply the logistic and hinge losses defined in Definition 4.9, Definition 4.13. Recall that the logistic loss (12) can be rewritten as follows.

Definition 4.9 (Logistic loss)**.** $\ell_{\text{logist}}(a,b) = \log(1+\exp(-ab))$ for $a \in \{-1,+1\}$ and $b \in \mathbb{R}$.

Lemma 4.10 (Excess risk bound of logistic loss)**.** *Set the loss in (17) and (19) as the logistic loss in Definition 4.9. Let f^* be Bayes predictor of ℓ_{logist}-risk defined in (17) and $\hat{f}_n$ be the empirical ℓ_{logist}-risk minimizer defined in (19). Suppose Assumption 4.1 holds. Then, for any $\delta \in (0,1)$, we have*

$$\mathbb{E}_{\mathbb{D}}\big[\mathcal{E}(\hat{f}_n;\ell_{\text{logist}})\big] \le \frac{8B}{n}\big(1+\log \mathbb{E}_{\mathbb{X}}\big[N_\infty(\delta,\mathcal{F},\mathbb{X})\big]\big) + 8\delta + \frac{3}{8}\inf_{f\in\mathcal{F}} \|f-f^*\|^2_{L^2(\mu_X)}.$$

Moreover, it holds

$$\mathbb{E}_{\mathbb{D}}\big[\|\hat{f}_n - f^*\|^2_{L^2(\mu_X)}\big] \le C_B\Big\{\frac{8B}{n}\big(1+\log \mathbb{E}_{\mathbb{X}}\big[N_\infty(\delta,\mathcal{F},\mathbb{X})\big]\big) + 8\delta + \frac{3}{8}\inf_{f\in\mathcal{F}} \|f-f^*\|^2_{L^2(\mu_X)}\Big\},$$

where $C_B = 2(1+\exp(B))^2\exp(-B)$.

Corollary 4.11. *Under the same assumptions in Lemma 4.10 and assume that f^* is a Hölder continuous function in Definition 4.6 with $\varsigma = s + r > 0$, $s \in \mathbb{N}_0$ and $r \in (0,1]$, and μ_X is absolutely continuous with respect to Lebesgue measure. Let $\mathcal{F}$ be deep ReLU neural networks. Setting the depth $\mathcal{D} = \mathcal{O}((s+1)^2 n^{\frac{d}{8\varsigma+4d}}\log(n))$ and width $\mathcal{W} = \mathcal{O}((s+1)^2 d^{s+1} n^{\frac{d}{8\varsigma+4d}}\log(n))$, then we have*

$$\mathbb{E}_{\mathbb{D}}\big[\|\hat{f}_n - f^*\|^2_{L^2(\mu_X)}\big] \le C_B B(s+1)^4 d^{2s+(\varsigma\vee 1)}\cdot n^{-\frac{2\varsigma}{2\varsigma+d}},$$

where $C_B = 2(1+\exp(B))^2\exp(-B)$.

Remark 4.12. The excess risk bound for deep logistic regression derived in Corollary 4.11 is minimax optimal and it improves the recent sub-optimal rate $\mathcal{O}(n^{-\frac{\varsigma}{2\varsigma+d}})$ in Shen et al. (2022).

Definition 4.13 (Hinge loss)**.** $\ell_{\text{hinge}}(a,b) = \max\{0, 1-ab\}$ for $a \in \{-1,+1\}$ and $b \in \mathbb{R}$.

Lemma 4.14 (Excess risk bound of hinge loss)**.** *Set the loss in (17) and (19) as the hinge loss in Definition 4.13. Let f^* be Bayes predictor of ℓ_{hinge}-risk defined in (17) and $\hat{f}_n$ be the empirical ℓ_{hinge}-risk minimizer defined in (19). Suppose Assumption 4.1 holds. Then, for any $\delta \in (0,1)$, we have*

$$\mathbb{E}_{\mathbb{D}}\big[\mathcal{E}(\hat{f}_n;\ell_{\text{hinge}})\big] \le \frac{8B}{n}\big(1+\log \mathbb{E}_{\mathbb{X}}\big[N_\infty(\delta,\mathcal{F},\mathbb{X})\big]\big) + 8\delta + 3\inf_{f\in\mathcal{F}} \|f-f^*\|_{L^1(\mu_X)}.$$

5. Conclusion and Further Work

In this work, we introduce a unified framework establishing a sharp bound for the expected excess risk of ERM in terms of general Lipschtiz loss functions which may be nonsmooth and unbounded via offset Rademacher complexity. We not only recover some known sharp results in several parametric and nonparametric supervised learning scenarios but also make nontrivial improvement in analysing LAD estimations, sparse linear regression models, and deep logistic regression with ReLU neural networks.

In future research, we intend to extend the approach in this paper to nonparametric kernel methods. In addition, extending this approach to other algorithms including regularization is indeed a rewarding problem. Furthermore, performing simulations to validate the theoretical results presented in this paper is also an important issue.

Acknowledgements

We appreciate all the anonymous reviewers for their insightful and constructive comments. This work is supported by the National Key Research and Development Program of China (No. 2020YFA0714200), by the National Science Foundation of China (No. 12125103, No. 12071362, No. 11971468), by the research fund of KLATASDSMOE of China.

References

Anthony, M., Bartlett, P. L., Bartlett, P. L., et al. *Neural network learning: Theoretical foundations*, volume 9. cambridge university press Cambridge, 1999.

Arias-Castro, E. and Lounici, K. Estimation and variable selection with exponential weights. *Electronic Journal of Statistics*, 8(1):328–354, 2014.

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

Bartlett, P. L., Harvey, N., Liaw, C., and Mehrabian, A. Nearly-tight vc-dimension and pseudodimension bounds for piecewise linear neural networks. *The Journal of Machine Learning Research*, 20(1):2285–2301, 2019.

Bauer, B. and Kohler, M. On deep learning as a remedy for the curse of dimensionality in nonparametric regression. *The Annals of Statistics*, 47(4):2261–2285, 2019.

Belloni, A. and Chernozhukov, V. ℓ_1-penalized quantile regression in high-dimensional sparse models. *The Annals of Statistics*, 39(1):82–130, 2011.

Fan, J. and Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360, 2001.

Farrell, M. H., Liang, T., and Misra, S. Deep neural networks for estimation and inference. *Econometrica*, 89(1): 181–213, 2021.

Giné, E. and Nickl, R. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

Gyorfi, L., Kohler, M., Krzyzak, A., and Walk, H. *A distribution-free theory of nonparametric regression*, volume 1. Springer, 2002.

He, X. and Shi, P. Convergence rate of b-spline estimators of nonparametric conditional quantile functions. *Journaltitle of Nonparametric Statistics*, 3(3-4):299–308, 1994.

Hernan Madrid Padilla, O., Tansey, W., and Chen, Y. Quantile regression with deep relu networks: Estimators and minimax rates. *arXiv e-prints*, pp. arXiv–2010, 2020.

Huang, J., Jiao, Y., Liu, Y., and Lu, X. A constructive approach to l0 penalized regression. *The Journal of Machine Learning Research*, 19(1):403–439, 2018.

Jiao, Y., Shen, G., Lin, Y., and Huang, J. Deep nonparametric regression on approximately low-dimensional manifolds. *arXiv preprint arXiv:2104.06708*, 2021.

Kanade, V., Rebeschini, P., and Vaskevicius, T. Exponential tail local rademacher complexity risk bounds without the bernstein condition. *arXiv preprint arXiv:2202.11461*, 2022.

Kohler, M. and Langer, S. On the rate of convergence of fully connected deep neural network regression estimates. *The Annals of Statistics*, 49(4):2231–2249, 2021.

Koltchinskii, V. Local rademacher complexities and oracle inequalities in risk minimization. *The Annals of Statistics*, 34(6):2593–2656, 2006.

Liang, T., Rakhlin, A., and Sridharan, K. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pp. 1260–1285. PMLR, 2015.

Loh, P.-L. and Wainwright, M. J. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 16: 559–616, 2015.

Lu, J., Shen, Z., Yang, H., and Zhang, S. Deep network approximation for smooth functions. *SIAM Journal on Mathematical Analysis*, 53(5):5465–5506, 2021.

Mendelson, S. Learning without concentration for general loss functions. *Probability Theory and Related Fields*, 171(1):459–502, 2018.

Nakada, R. and Imaizumi, M. Adaptive approximation and generalization of deep neural network with intrinsic dimensionality. *J. Mach. Learn. Res.*, 21:174–1, 2020.

Padilla, O. H. M. and Chatterjee, S. Risk bounds for quantile trend filtering. *arXiv preprint arXiv:2007.07472*, 2020.

Petersen, P. and Voigtlaender, F. Optimal approximation of piecewise smooth functions using deep relu neural networks. *Neural Networks*, 108:296–330, 2018.

Schmidt-Hieber, J. Nonparametric regression using deep neural networks with relu activation function. *The Annals of Statistics*, 48(4):1875–1897, 2020.

Shen, G., Jiao, Y., Lin, Y., Horowitz, J. L., and Huang, J. Deep quantile regression: Mitigating the curse of dimensionality through composition. *arXiv preprint arXiv:2107.04907*, 2021.

Shen, G., Jiao, Y., Lin, Y., and Huang, J. Approximation with cnns in sobolev space: with applications to classification. In *NeurIPS*, 2022.

Shen, Z. Deep network approximation characterized by number of neurons. *Communications in Computational Physics*, 28(5):1768–1811, 2020.

Shen, Z., Yang, H., and Zhang, S. Nonlinear approximation via compositions. *Neural Networks*, 119:74–84, 2019.

Stone, C. J. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pp. 1040–1053, 1982.

Suzuki, T. and Nitanda, A. Deep learning is adaptive to intrinsic dimensionality of model smoothness in anisotropic besov space. *Advances in Neural Information Processing Systems*, 34, 2021.

Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

Tsybakov, A. B. *Introduction to Nonparametric Estimation*. Springer, 2009.

Van de Geer, S. and Bühlmann, P. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Van de Geer, S. A. and van de Geer, S. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

Van der Vaart, A. W. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Van Der Vaart, A. W. and Wellner, J. A. Weak convergence and empirical processes: With applications to statistics. In *Weak convergence and empirical processes*, pp. 16–28. Springer, 1996.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*, volume 47 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge university press, 2018.

Wainwright, M. J. *High-dimensional Statistics: A Non-asymptotic Viewpoint*, volume 48. Cambridge University Press, 2019.

Wang, L., Kim, Y., and Li, R. Calibrating non-convex penalized regression in ultra-high dimension. *Annals of Statistics*, 41(5):2505, 2013.

Xu, Y. and Zeevi, A. Towards optimal problem dependent generalization error bounds in statistical learning theory. In *NeurIPS*, 2021.

Yarotsky, D. Optimal approximation of continuous functions by very deep relu networks. In *Conference on Learning Theory*, pp. 639–649. PMLR, 2018.

Yarotsky, D. and Zhevnerchuk, A. The phase diagram of approximation rates for deep neural networks. *Advances in neural information processing systems*, 33:13005–13015, 2020.

Ye, F. and Zhang, C.-H. Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *Journal of Machine Learning Research*, 11(Dec):3519–3540, 2010.

Zhang, C.-H. Nearly unbiased variable selection under minimax concave penalty. *The Annals of statistics*, 38(2): 894–942, 2010.

Zhang, C.-H. and Huang, J. The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics*, 36(4):1567–1594, 2008.

Zhang, C.-H. and Zhang, T. A general theory of concave regularization for high-dimensional sparse estimation problems. *Statistical Science*, 27(4):576–593, 2012.

Zhang, Y., Wainwright, M. J., and Jordan, M. I. Optimal prediction for sparse linear models? lower bounds for coordinate-separable m-estimators. *Electronic Journal of Statistics*, 11(1):752–799, 2017.

Zou, H. The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429, 2006.

Zou, H. and Zhang, H. H. On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics*, 37 (4):1733, 2009.

A. Proofs in Section 2

A.1. Proof of Theorem 2.1

Proof. For any $\omega > 0$, we have

$$\begin{aligned}
&\mathbb{E}_{\mathbb{D}}\mathbb{E}_X g(X; f_n) \\
&= \mathbb{E}_{\mathbb{D}}\Big[\mathbb{E}_X g(X; f_n) - \frac{\omega+1}{n}\sum_{i=1}^n g(X_i; f_n)\Big] + (\omega+1)\mathbb{E}_{\mathbb{D}}\Big[\frac{1}{n}\sum_{i=1}^n g(X_i; f_n)\Big] \\
&\le \mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big(\mathbb{E}_X g(X; f) - \frac{\omega+1}{n}\sum_{i=1}^n g(X_i; f)\Big) + (\omega+1)\mathbb{E}_{\mathbb{D}}\Big[\frac{1}{n}\sum_{i=1}^n g(X_i; f_n)\Big]. \qquad \text{(A.1)}
\end{aligned}$$

Then, it suffices to bound the first term on the right hand of (A.1). Since $\|f\|_{L^\infty(\mu_X)} \le B$ for $f \in \mathcal{F}$, $g(x;\cdot)$ is κ-Lipschitz and $0 \le g(x;\cdot) \le 2B\kappa$ for all $x \in \mathcal{X}$, then for $\omega > 0$, some algebra calculation yields

$$\begin{aligned}
&\mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big(\mathbb{E}_X g(X; f) - \frac{\omega+1}{n}\sum_{i=1}^n g(X_i; f)\Big) \\
&\le \mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big(\frac{\omega+2}{2}\mathbb{E}_X g(X; f) - \frac{\omega}{2}\mathbb{E}_X g(X; f) - \frac{\omega+2}{2n}\sum_{i=1}^n g(X_i; f) - \frac{\omega}{2n}\sum_{i=1}^n g(X_i; f)\Big) \\
&\le \mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big(\frac{\omega+2}{2}\mathbb{E}_X g(X; f) - \frac{\omega}{4B\kappa}\mathbb{E}_X g^2(X; f) - \frac{\omega+2}{2n}\sum_{i=1}^n g(X_i; f) - \frac{\omega}{4B\kappa n}\sum_{i=1}^n g^2(X_i; f)\Big),
\end{aligned}$$

where the last inequality holds due to $g^2(x;\cdot) \le 2B\kappa g(x;\cdot), x \in \mathcal{X}$. Further, we introduce a ghost sample $\mathbb{X}' = \{X_i'\}_{i=1}^n$ drawn i.i.d. from μ_X independent of $\mathbb{X}$, and let $\tau = \{\tau_i\}_{i=1}^n$ be a sequence of i.i.d. Rademacher variables independent of $\mathbb{X}$ and $\mathbb{X}'$. By means of the technique of symmetrization, we can replace the expectation inside the supremum by an empirical mean based on a ghost sample. Because of the independent and identical distribution of $\{X_1, \ldots, X_n, X_1', \ldots, X_n'\}$, the joint distribution of $\mathbb{X}$ and $\mathbb{X}'$ is not affected if one randomly interchanges the corresponding components of $\mathbb{X}$ and $\mathbb{X}'$. Then by the convexity of supremum and Jensen's inequality, we obtain

$$\begin{aligned}
&\mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big(\frac{\omega+2}{2}\mathbb{E}_X g(X; f) - \frac{\omega}{4B\kappa}\mathbb{E}_X g^2(X; f) - \frac{\omega+2}{2n}\sum_{i=1}^n g(X_i; f) - \frac{\omega}{4B\kappa n}\sum_{i=1}^n g^2(X_i; f)\Big) \\
&= \mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big(\mathbb{E}_{\mathbb{X}'}\Big[\frac{\omega+2}{2n}\sum_{i=1}^n g(X_i'; f) - \frac{\omega}{4B\kappa n}\sum_{i=1}^n g^2(X_i'; f)\Big] - \frac{\omega+2}{2n}\sum_{i=1}^n g(X_i; f) - \frac{\omega}{4B\kappa n}\sum_{i=1}^n g^2(X_i; f)\Big) \\
&\le \mathbb{E}_{\mathbb{X}}\mathbb{E}_{\mathbb{X}'} \sup_{f\in\mathcal{F}}\Big(\frac{\omega+2}{2n}\sum_{i=1}^n (g(X_i'; f) - g(X_i; f)) - \frac{\omega}{4B\kappa n}\sum_{i=1}^n (g^2(X_i'; f) + g^2(X_i; f))\Big) \\
&= \mathbb{E}_{\mathbb{X}}\mathbb{E}_{\mathbb{X}'}\mathbb{E}_\tau \sup_{f\in\mathcal{F}}\Big(\frac{\omega+2}{2n}\sum_{i=1}^n \tau_i(g(X_i'; f) - g(X_i; f)) - \frac{\omega}{4B\kappa n}\sum_{i=1}^n (g^2(X_i'; f) + g^2(X_i; f))\Big) \\
&\le \mathbb{E}_{\mathbb{X}'}\mathbb{E}_\tau \sup_{f\in\mathcal{F}}\Big(\frac{\omega+2}{2n}\sum_{i=1}^n \tau_i g(X_i'; f) - \frac{\omega}{4B\kappa n}\sum_{i=1}^n g^2(X_i'; f)\Big) \\
&\quad + \mathbb{E}_{\mathbb{X}}\mathbb{E}_\tau \sup_{f\in\mathcal{F}}\Big(\frac{\omega+2}{2n}\sum_{i=1}^n (-\tau_i) g(X_i; f) - \frac{\omega}{4B\kappa n}\sum_{i=1}^n g^2(X_i; f)\Big) \\
&= \mathbb{E}_{\mathbb{X}}\mathbb{E}_\tau \sup_{f\in\mathcal{F}}\Big(\frac{\omega+2}{n}\sum_{i=1}^n \tau_i g(X_i; f) - \frac{\omega}{2B\kappa n}\sum_{i=1}^n g^2(X_i; f)\Big),
\end{aligned}$$

where the last inequality follows from the fact that $(-\tau_i)g(X_i; f)$ and $\tau_i g(X_i'; f)$ have the same distribution as $\tau_i g(X_i; f)$. Therefore, setting $\beta = \frac{\omega}{2B\kappa(\omega+2)}$ in (5) yields the desired result. □

A.2. Proof of Theorem 2.3

Proof. As $\tau = \{\tau_i\}_{i=1}^n$ is a sequence of i.i.d. Rademacher variables independent of $\mathbb{X} = \{X_i\}_{i=1}^n$, then conditionally on $\mathbb{X}$ we have

$$\begin{aligned}&\mathbb{E}_{\tau|\mathbb{X}}\Big[\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\tau_i g(X_i;f)-\frac{\beta}{n}\sum_{i=1}^n g^2(X_i;f)\Big|\mathbb{X}=(X_i)_{i=1}^n\Big]\\&=\mathbb{E}_\tau\sup_{f\in\mathcal{F}}\Big(\frac{1}{n}\sum_{i=1}^n\tau_i g(X_i;f)-\frac{\beta}{n}\sum_{i=1}^n g^2(X_i;f)\Big).\end{aligned}$$

Let $\delta>0$ and let $\mathcal{F}_\delta$ be an L^∞ δ-cover of $\mathcal{F}$ on $\mathbb{X}=(X_1,\ldots,X_n)$. For any fixed $f\in\mathcal{F}$, there exists a $f_\delta\in\mathcal{F}_\delta$ such that $\|f-f_\delta\|_{\mathbb{X},\infty}\le\delta$. Thus

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n\tau_i g(X_i;f)&\le\frac{1}{n}\sum_{i=1}^n\tau_i g(X_i;f_\delta)+\frac{1}{n}\sum_{i=1}^n|\tau_i||g(X_i;f)-g(X_i;f_\delta)|\\&\le\frac{1}{n}\sum_{i=1}^n\tau_i g(X_i;f_\delta)+\kappa\delta,\end{aligned}\tag{A.2}$$

where we use the Lipschitz continuity of $g(x;\cdot)$, $x\in\mathcal{X}$. Using $|g(x;f)|\le 2B\kappa$ and $|g(x;f_\delta)|\le 2B\kappa$ and the Lipschitz continuity of $g(x;\cdot)$, one obtains

$$\begin{aligned}-\frac{1}{n}\sum_{i=1}^n g^2(X_i;f)&=-\frac{1}{n}\sum_{i=1}^n g^2(X_i;f_\delta)+\frac{1}{n}\sum_{i=1}^n(g(X_i;f_\delta)+g(X_i;f))(g(X_i;f_\delta)-g(X_i;f))\\&\le-\frac{1}{n}\sum_{i=1}^n g^2(X_i;f_\delta)+4B\kappa\frac{1}{n}\sum_{i=1}^n|g(X_i;f_\delta)-g(X_i;f)|\\&\le-\frac{1}{n}\sum_{i=1}^n g^2(X_i;f_\delta)+4B\kappa^2\delta.\end{aligned}\tag{A.3}$$

Hence, it follows from (A.2)-(A.3) that

$$\begin{aligned}&\mathbb{E}_\tau\sup_{f\in\mathcal{F}}\Big(\frac{1}{n}\sum_{i=1}^n\tau_i g(X_i;f)-\frac{\beta}{n}\sum_{i=1}^n g^2(f,X_i)\Big)\\&\le\mathbb{E}_\tau\max_{f_\delta\in\mathcal{F}_\delta}\Big(\frac{1}{n}\sum_{i=1}^n\tau_i g(X_i;f_\delta)-\frac{\beta}{n}\sum_{i=1}^n g^2(X_i;f_\delta)\Big)+(1+4B\beta\kappa)\kappa\delta.\end{aligned}$$

Since $\{\tau_i g(X_i;f_\delta)\}_{i=1}^n$ are independent random variables conditioning on $\mathbb{X}=(X_i)_{i=1}^n$, then

$$\mathbb{E}_\tau\big[\tau_i g(X_i;f_\delta)\big]=0,\quad\text{and}\quad-g(X_i;f_\delta)\le\tau_i g(X_i;f_\delta)\le g(X_i;f_\delta),\quad i=1,\ldots,n.$$

By Hoeffding's inequality, it yields that for any $f_\delta\in\mathcal{F}_\delta$ and $\xi>0$,

$$\begin{aligned}\mathbb{P}_\tau\Big\{\frac{1}{n}\sum_{i=1}^n\tau_i g(X_i;f_\delta)>\xi+\frac{\beta}{n}\sum_{i=1}^n g^2(X_i;f_\delta)\Big\}&\le\exp\Big(-\frac{(n\xi+\beta\sum_{i=1}^n g^2(X_i;f_\delta))^2}{2\sum_{i=1}^n g^2(X_i;f_\delta)}\Big)\\&\le\exp(-2\beta n\xi),\end{aligned}$$

where the last inequality holds due to the numeric inequality

$$\frac{(a+y)^2}{y}\ge\frac{(a+a)^2}{a}=4a,\quad y\in\mathbb{R}_+.$$

Therefore, we obtain an estimator for the tail probability

$$\begin{aligned}
&\mathbb{P}_\tau\left\{\max_{f_\delta\in\mathcal{F}_\delta}\left(\frac{1}{n}\sum_{i=1}^n\tau_i g(X_i;f_\delta)-\frac{\beta}{n}\sum_{i=1}^n g^2(X_i;f_\delta)\right)>\xi\right\}\\
&\le N_\infty(\delta,\mathcal{F},\mathbb{X})\max_{f_\delta\in\mathcal{F}_\delta}\mathbb{P}_\tau\Big\{\frac{1}{n}\sum_{i=1}^n\tau_i g(X_i;f_\delta)>\xi+\frac{\beta}{n}\sum_{i=1}^n g^2(X_i;f_\delta)\Big\}\\
&\le N_\infty(\delta,\mathcal{F},\mathbb{X})\exp(-2\beta n\xi).
\end{aligned}$$

Consequently, for any $A>0$,

$$\begin{aligned}
&\mathbb{E}_\tau\max_{f_\delta\in\mathcal{F}_\delta}\Big(\frac{1}{n}\sum_{i=1}^n\tau_i g(X_i;f_\delta)-\frac{\beta}{n}\sum_{i=1}^n g^2(X_i;f_\delta)\Big)\\
&\le\int_0^\infty\mathbb{P}_\tau\Big\{\max_{f_\delta\in\mathcal{F}_\delta}\left(\frac{1}{n}\sum_{i=1}^n\tau_i g(X_i;f_\delta)-\frac{\beta}{n}\sum_{i=1}^n g^2(X_i;f_\delta)>\xi\right)\Big\}d\xi\\
&\le A+\int_A^\infty N_\infty(\delta,\mathcal{F},\mathbb{X})\exp(-2\beta n\xi)d\xi\\
&\le A+\frac{N_\infty(\delta,\mathcal{F},\mathbb{X})}{2\beta n}\exp(-2\beta nA).
\end{aligned}$$

Setting $A=\frac{\log N_\infty(\delta,\mathcal{F},\mathbb{X})}{2\beta n}$ leads to

$$\mathbb{E}_\tau\max_{f_\delta\in\mathcal{F}_\delta}\Big(\frac{1}{n}\sum_{i=1}^n\tau_i g(f_\delta,X_i)-\frac{\beta}{n}\sum_{i=1}^n g^2(f_\delta,X_i)\Big)\le\frac{1+\log N_\infty(\delta,\mathcal{F},\mathbb{X})}{2\beta n}.$$

As a consequence, we have

$$\mathbb{E}\sup_{f\in\mathcal{F}}\Big(\frac{1}{n}\sum_{i=1}^n\tau_i g(X_i;f)-\frac{\beta}{n}\sum_{i=1}^n g^2(X_i;f)\Big)\le\frac{1+\log\mathbb{E}_{\mathbb{X}}\big[N_\infty(\delta,\mathcal{F},\mathbb{X})\big]}{2\beta n}+(1+4B\beta\kappa)\kappa\delta.$$

This completes the proof. □

B. Proofs in Section 3

B.1. Proof of Theorem 3.3

Proof. To begin with, for any $\theta\in\Theta$, let us define

$$\tilde{g}(x,y;\varphi(\cdot)^T\theta):=\ell(y,\varphi(x)^T\theta)-\ell(y,\varphi(x)^T\theta^*),\ (x,y)\in\mathcal{X}\times\mathcal{Y}.$$

Then, we can obtain the following decomposition by the fact that $\hat{\theta}_n$ is the ERM, that is,

$$\begin{aligned}
&\mathcal{E}(\varphi(\cdot)^T\hat{\theta}_n;\ell)=\mathbb{E}_{(X,Y)}\tilde{g}(X,Y;\varphi(\cdot)^T\hat{\theta}_n)\\
&=\mathbb{E}_{(X,Y)}\tilde{g}(X,Y;\varphi(\cdot)^T\hat{\theta}_n)-\frac{3}{n}\sum_{i=1}^n\tilde{g}(X_i,Y_i;\varphi(\cdot)^T\hat{\theta}_n)+\frac{3}{n}\sum_{i=1}^n\tilde{g}(X_i,Y_i;\varphi(\cdot)^T\hat{\theta}_n)\\
&\le\mathbb{E}_{(X,Y)}\tilde{g}(X,Y;\varphi(\cdot)^T\hat{\theta}_n)-\frac{3}{n}\sum_{i=1}^n\tilde{g}(X_i,Y_i;\varphi(\cdot)^T\hat{\theta}_n)+\frac{3}{n}\sum_{i=1}^n\tilde{g}(X_i,Y_i;\varphi(\cdot)^T\theta),
\end{aligned}$$

for any $\theta \in \Theta$. Taking expectation with respect to $\mathbb{D} = \{(X_i, Y_i)\}_{i=1}^n$ on both sides of the above inequality gives

$$\begin{aligned}
\mathbb{E}_{\mathbb{D}}\big[\mathcal{E}(\varphi(\cdot)^T\hat{\theta}_n;\ell)\big] &\le \sup_{f\in\mathcal{F}} \mathbb{E}_{\mathbb{D}}\Big[\mathbb{E}_{(X,Y)}\tilde{g}(X,Y;f) - \frac{3}{n}\sum_{i=1}^n \tilde{g}(X_i,Y_i;f)\Big] + 3\mathbb{E}_{\mathbb{D}}\Big[\frac{1}{n}\sum_{i=1}^n \tilde{g}(X_i,Y_i;\varphi(\cdot)^T\theta)\Big] \\
&= \sup_{f\in\mathcal{F}} \mathbb{E}_{\mathbb{X}}\Big[\mathbb{E}_X g(X;f) - \frac{3}{n}\sum_{i=1}^n g(X_i;f)\Big] + 3\mathcal{E}(\varphi(\cdot)^T\theta;\ell) \\
&\le \mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big[\mathbb{E}_X g(X;f) - \frac{3}{n}\sum_{i=1}^n g(X_i;f)\Big] + 3\mathcal{E}(\varphi(\cdot)^T\theta;\ell),
\end{aligned}$$

for any $\theta \in \Theta$, where $\mathcal{F} := \{x \mapsto \varphi(x)^T\theta : \theta \in \Theta\}$ and the last inequality follows form the convexity of supremum and Jensen's inequality. In Example 3.1, we see that $\mathcal{E}(\varphi(\cdot)^T\theta^*;\ell) = 0$ with $\theta^* \in \Theta$. Hence

$$\mathbb{E}_{\mathbb{D}}\big[\mathcal{E}(\varphi(\cdot)^T\hat{\theta}_n;\ell)\big] \le \mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big[\mathbb{E}_X g(X;f) - \frac{3}{n}\sum_{i=1}^n g(X_i;f)\Big].$$

Observe that $0 \le g(x;\cdot) \le 2B_\varphi B_\theta \kappa$ for all $x \in \mathcal{X}$, then it implies $g^2(x;\cdot) \le 2B_\varphi B_\theta \kappa g(x;\cdot)$ by Assumption 3.2. Consequently,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big[\mathbb{E}_X g(X;f) - \frac{3}{n}\sum_{i=1}^n g(X_i;f)\Big] \\
&= \mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big[2\mathbb{E}_X g(X;f) - \mathbb{E}_X g(X;f) - \frac{2}{n}\sum_{i=1}^n g(X_i;f) - \frac{1}{n}\sum_{i=1}^n g(X_i;f)\Big] \\
&\le \mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big(2\mathbb{E}_X g(X;f) - \frac{1}{2B_\varphi B_\theta\kappa}\mathbb{E}_X g^2(X;f) - \frac{2}{n}\sum_{i=1}^n g(X_i;f) - \frac{1}{2B_\varphi B_\theta \kappa n}\sum_{i=1}^n g^2(X_i;f)\Big).
\end{aligned}$$

We introduce an i.i.d. ghost sample $\mathbb{X}' := \{X_i'\}_{i=1}^n$ drawn from μ_X and independent of $\mathbb{X}$, and let $\tau = \{\tau_i\}_{i=1}^n$ be a sequence of i.i.d. Rademacher variables independent of $\mathbb{X}$ and $\mathbb{X}'$. By means of the technique of symmetrization, we can replace the expectation inside the supremum by an empirical mean based on the ghost sample. Because of the independent and identical distribution of $\{X_1,\ldots,X_n,X_1',\ldots,X_n'\}$, the joint distribution of $\mathbb{X}$ and $\mathbb{X}'$ is not affected if one randomly interchanges the corresponding components of $\mathbb{X}$ and $\mathbb{X}'$. Then by the convexity of supremum and Jensen's inequality, we obtain

$$\begin{aligned}
&\mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big(2\mathbb{E}_X g(X;f) - \frac{1}{2B_\varphi B_\theta\kappa}\mathbb{E}_X g^2(X;f) - \frac{2}{n}\sum_{i=1}^n g(X_i;f) - \frac{1}{2B_\varphi B_\theta \kappa n}\sum_{i=1}^n g^2(X_i;f)\Big) \\
&= \mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big(\mathbb{E}_{\mathbb{X}'}\Big[\frac{2}{n}\sum_{i=1}^n g(X_i';f) - \frac{1}{2B_\varphi B_\theta\kappa n}\sum_{i=1}^n g^2(X_i';f)\Big] - \frac{2}{n}\sum_{i=1}^n g(X_i;f) - \frac{1}{2B_\varphi B_\theta \kappa n}\sum_{i=1}^n g^2(X_i;f)\Big) \\
&\le \mathbb{E}_{\mathbb{X}}\mathbb{E}_{\mathbb{X}'} \sup_{f\in\mathcal{F}}\Big(\frac{2}{n}\sum_{i=1}^n (g(X_i';f) - g(X_i;f)) - \frac{1}{2B_\varphi B_\theta \kappa n}\sum_{i=1}^n (g^2(X_i';f) + g^2(X_i;f))\Big) \\
&= \mathbb{E}_{\mathbb{X}}\mathbb{E}_{\mathbb{X}'}\mathbb{E}_\tau \sup_{f\in\mathcal{F}}\Big(\frac{2}{n}\sum_{i=1}^n \tau_i(g(X_i';f) - g(X_i;f)) - \frac{1}{2B_\varphi B_\theta \kappa n}\sum_{i=1}^n (g^2(X_i';f) + g^2(X_i;f))\Big) \\
&= 2\mathbb{E}_{\mathbb{X}'}\mathbb{E}_\tau \sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\Big(\tau_i g(X_i';f) - \frac{1}{4B_\varphi B_\theta\kappa} g^2(X_i';f)\Big) \\
&\quad + 2\mathbb{E}_{\mathbb{X}}\mathbb{E}_\tau \sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\Big(-\tau_i g(X_i;f) - \frac{1}{4B_\varphi B_\theta\kappa} g^2(X_i;f)\Big) \\
&= 2\mathbb{E}_{\mathbb{X}'}\mathcal{R}_n^{\mathrm{off}}\Big(\mathcal{G}, \frac{1}{4B_\varphi B_\theta\kappa}\Big|\mathbb{X}'\Big) + 2\mathbb{E}_{\mathbb{X}}\mathcal{R}_n^{\mathrm{off}}\Big(\mathcal{G}, \frac{1}{4B_\varphi B_\theta\kappa}\Big|\mathbb{X}\Big) \\
&= 4\mathcal{R}_n^{\mathrm{off}}\Big(\mathcal{G}, \frac{1}{4B_\varphi B_\theta\kappa}\Big).
\end{aligned}$$

We complete the proof. □

B.2. Proofs of Corollaries 3.5 and 3.6 and Theorem 3.11

We first introduce and prove the following Lemmas B.1 and B.3 preparing for the proofs of Corollaries 3.5 and 3.6 and Theorem 3.11.

Lemma B.1. *Let $\varphi : \mathcal{X} \to \mathbb{R}^d$ be a feature map satisfying $\|\varphi(x)\|_2 \le B_\varphi$ for each $x \in \mathcal{X}$. Denote $\mathcal{F} := \{x \mapsto \varphi(x)^T\theta : \theta \in \mathbb{R}^d, \|\theta\|_2 \le B_\theta\}$. Then*

$$\log N_\infty(\delta, \mathcal{F}) \le d \log\Big(\frac{3B_\varphi B_\theta}{\delta}\Big).$$

Proof. Denote $\Theta := \{\theta \in \mathbb{R}^d : \|\theta\|_2 \le B_\theta\}$. Let Θ_ε be a $\|\cdot\|_2$ ε-net of Θ with $|\Theta_\varepsilon| = N(\varepsilon, \Theta, \|\cdot\|_2)$. Then, for any $\theta \in \Theta$, there exists a $\theta_\varepsilon \in \Theta_\varepsilon$, such that $\|\theta - \theta_\varepsilon\|_2 \le \varepsilon$. By Cauchy-Schwardz inequality, we have

$$|\varphi(x)^T\theta - \varphi(x)^T\theta_\varepsilon| \le \|\varphi(x)\|_2\|\theta - \theta_\varepsilon\|_2 \le B_\varphi\varepsilon,$$

and consequently, $N_\infty(B_\varphi\varepsilon, \mathcal{F}) \le N(\varepsilon, \Theta, \|\cdot\|_2)$. Using Corollary 4.2.13 in Vershynin (2018) gives

$$N(\varepsilon, \Theta, \|\cdot\|_2) \le \Big(\frac{3B_\theta}{\varepsilon}\Big)^d,$$

for all $0 \le \varepsilon \le 1$. By setting $\delta = B_\varphi\varepsilon$, we conclude the result. □

Lemma B.2. *Let $\mathcal{Y} = \{+1, -1\}$. Let $\ell(a, b) : \mathcal{Y} \times \mathbb{R} \to [0, \infty)$ be a margin-based loss functions, namely there exists a representing function $\psi : \mathbb{R} \to [0, \infty)$ such that $\ell(a, b) = \psi(ab)$ for $a \in \mathcal{Y}$ and $b \in \mathbb{R}$. If*

(i) $\psi(\cdot)$ and $f(x; \cdot)$ are convex for all $x \in \mathcal{X}$, and

(ii) $\eta(x)\psi'(f(x; \theta^*)) = (1 - \eta(x))\psi'(-f(x; \theta^*))$ for all $x \in \mathcal{X}$,

then for any $x \in \mathcal{X}$, it holds $\min_{\theta\in\mathbb{R}^p} \mathbb{E}\big[\ell(Y, f(X; \theta))\big|X\big] = \mathbb{E}\big[\ell(Y, f(X; \theta^*))\big|X\big]$.

Proof. Observe that

$$\mathbb{E}\big[\ell(Y, f(X; \theta)) - \ell(Y, f(X; \theta^*))\big|X\big] = \mathbb{E}\big[\psi(Yf(X; \theta))\big|X\big] - \mathbb{E}\big[\psi(Yf(X; \theta^*))\big|X\big],$$

and we define

$$\begin{aligned} h(X; \theta) &= \mathbb{E}\big[\psi(Yf(X; \theta))\big|X\big] \\ &= \eta(X)\psi(f(X; \theta)) + (1 - \eta(X))\psi(-f(X; \theta)). \end{aligned}$$

Note that $h(x; \cdot)$ is convex provided $\psi(\cdot)$ and $f(x; \cdot)$ are convex for any $x \in \mathcal{X}$ and

$$0 \in \partial_\theta h(x; \theta^*) = \big\{\eta(x)\psi'(f(x; \theta^*)) - (1 - \eta(x))\psi'(-f(x; \theta^*))\big\}\partial_\theta f(x; \theta^*),$$

then we obtain the result. □

Lemma B.3. *Let f^* be a Bayes predictor of ℓ_{logist}-risk such that $\|f^*\|_{L^\infty(\mu_X)} \le B$ with $B \ge 1$, then it follows that*

$$\frac{\exp(B)}{2(1 + \exp(B))^2}\|f - f^*\|^2_{L^2(\mu_X)} \le \mathcal{E}(f; \ell_{\text{logist}}) \le \frac{1}{8}\|f - f^*\|^2_{L^2(\mu_X)},$$

for any f satisfying $\|f\|_{L^\infty(\mu_X)} \le B$.

Proof. We define the conditional loss at $X = x$ as

$$h_x(t) = \mathbb{E}\big[\log(1 + \exp(-Yt))\big|X = x\big].$$

Since f^* is a Bayes predictor of $R(f;\ell_{\mathrm{logist}})$, we have $t^* = f^*(x) = \arg\min_{t\in\mathbb{R}} h_x(t)$, which means that $h_x'(t^*) = 0$. Using Taylor expansion implies

$$h_x(t) = h_x(t^*) + h_x'(t^*)(t-t^*) + \frac{h_x''(\xi)}{2}(t-t^*)^2 = h_x(t^*) + \frac{h_x''(\xi)}{2}(t-t^*)^2.$$

Note that

$$h_x''(\xi) = \mathbb{E}\Big[\frac{Y^2\exp(Y\xi)}{(1+\exp(Y\xi))^2}\Big|X=x\Big] = \mathbb{E}\Big[\frac{1}{\exp(-Y\xi)+\exp(Y\xi)+2}\Big|X=x\Big]$$

with $|\xi| \le B$, then we can deduce that

$$\frac{\exp(B)}{(1+\exp(B))^2} \le h_x''(\xi) \le \frac{1}{4}.$$

Therefore,

$$\frac{\exp(B)}{2(1+\exp(B))^2}(f(x)-f^*(x))^2 \le h_x(t) - h_x(t^*) \le \frac{1}{8}(f(x)-f^*(x))^2,$$

and integrating with respect to x over μ_X completes the proof. □

Now, we prove Corollary 3.5.

Proof. Define

$$g(x;\varphi(\cdot)^T\theta) := \mathbb{E}\big[(\varphi(X)^T\theta - Y)^2 - (\varphi(X)^T\theta^* - Y)^2\big|X=x\big], \quad x\in\mathcal{X}.$$

Then, we see

$$\begin{aligned} g(X;\varphi(\cdot)^T\theta) &= \mathbb{E}\big[(\varphi(X)^T\theta - \varphi(X)^T\theta^* + \varphi(X)^T\theta^* - Y)^2 - (\varphi(X)^T\theta^* - Y)^2\big|X\big] \\ &= \mathbb{E}\big[(\varphi(X)^T\theta - \varphi(X)^T\theta^*)^2 + 2(\varphi(X)^T\theta - \varphi(X)^T\theta^*)(\varphi(X)^T\theta^* - Y)\big|X\big] \\ &= (\varphi(X)^T\theta - \varphi(X)^T\theta^*)^2. \end{aligned}$$

Thus $\mathcal{E}(f;\ell_{\mathrm{ls}}) = \|\theta-\theta^*\|_\Sigma^2$ for any $f\in\mathcal{F}$ and Assumption 3.2 holds. Observe that $|g(x;\cdot)|$ is $4B_\varphi B_\theta$-Lipschitz continuous for any $x\in\mathcal{X}$. Using Theorems 2.3 and 3.3 and Lemma B.1 yields the result. □

Then we prove Corollary 3.6.

Proof. Denote

$$g(x;\varphi(\cdot)^T\theta) = \mathbb{E}\big[|\varphi(X)^T\theta - Y| - |\varphi(X)^T\theta^* - Y|\big|X=x\big], \quad x\in\mathcal{X}.$$

We deduce

$$\begin{aligned} g(X;\varphi(\cdot)^T\theta) &= \mathbb{E}\big[|\varphi(X)^T\theta - Y| - |\varphi(\cdot)^T\theta^* - Y|\big|X\big] \\ &= \mathbb{E}\big[|\varphi(X)^T\theta - \varphi(X)^T\theta^* - \varepsilon|\big|X\big] - \mathbb{E}\big[|\varepsilon|\big|X\big]. \end{aligned}$$

Define $h(a) := \mathbb{E}[|a-\varepsilon||X]$, and note that $0\in\partial_a h(0)$ and $h(a)$ is convex, thus $h(a)\ge h(0) = \mathbb{E}[|\varepsilon||X]$ for all $a\in\mathbb{R}$. Hence Assumption 3.2 holds. Using Theorems 2.3 and 3.3 and Lemma B.1, one obtains the first result. Combining Lemma 5 in Shen et al. (2021) with Assumption 3.7 yields the second desired result. □

Last, we prove Theorem 3.11.

Proof. By Lemma B.2, we see that $g(f(\cdot;\theta),X)\ge 0$ for all $\theta\in\Theta$ and $X\in\mathcal{X}$. In fact, for logistic loss, the representation function is given by $\psi(z) = \log(1+\exp(-z))$. Then we see $\psi'(z) = \sigma(z) - 1$ and $\psi'(-z) = -\sigma(z)$ where $\sigma(\cdot)$ is the sigmoid function, and thus Assumption 3.2 holds by Lemma B.2. Note that $g(\cdot,X)$ is 1-Lipschitz continuous. By applying Theorem 3.3, Lemma B.3 and Lemma B.1, we arrive at the final estimate. □

B.3. Proofs of Theorems 3.13 and 3.15

Lemma B.4. *Let $k \in \mathbb{N}$, $k \le d$ and $\varphi : \mathcal{X} \to \mathbb{R}^d$ be a feature map satisfying $\|\varphi(x)\|_2 \le B_\varphi$ for each $x \in \mathcal{X}$. Denote $\mathcal{F}_k := \{x \mapsto \varphi(x)^T\theta : \theta \in \mathbb{R}^d, \|\theta\|_2 \le B_\theta, \|\theta\|_0 \le k\}$. Then*

$$\log N_\infty(\delta, \mathcal{F}_k) \le k\Big\{1 + \log\frac{d}{k} + \log\Big(\frac{3B_\varphi B_\theta}{\delta}\Big)\Big\}.$$

Proof. By Lemma B.1, it is easy to check that

$$\log N_\infty(\delta, \mathcal{F}_k) \le k\Big(1 + \log\frac{d}{k}\Big) + k\log\Big(\frac{3B_\varphi B_\theta}{\delta}\Big),$$

where we used the numerical inequality $\log\binom{d}{k} \le k(1 + \log\frac{d}{k})$. This completes the proof. □

Thence, Theorems 3.13 and 3.15 can be deduced by Lemma B.4, Corollaries 3.5 and 3.6, and Theorem 3.11.

C. Proofs in Section 4

C.1. Proof of Theorem 4.2

Proof. Denote

$$\tilde{g}(x, y; f) := \ell(y, f(x)) - \ell(y, f^*(x)),$$

for any $(x, y) \in \mathcal{X} \times \mathcal{Y}$ and $f \in \mathcal{F}$. Then we have the following decomposition by the fact that $\hat{f}_n$ is the empirical risk minimizer, i.e.,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{D}}\big[\mathcal{E}(\hat{f}_n; \ell)\big] = \mathbb{E}_{\mathbb{D}}\big[\mathbb{E}_{(X,Y)}\tilde{g}(X, Y; \hat{f}_n)\big] \\
&= \mathbb{E}_{\mathbb{D}}\Big[\mathbb{E}_{(X,Y)}\tilde{g}(X, Y; \hat{f}_n) - \frac{3}{n}\sum_{i=1}^n \tilde{g}(X_i, Y_i; \hat{f}_n)\Big] + 3\mathbb{E}_{\mathbb{D}}\Big[\frac{1}{n}\sum_{i=1}^n \tilde{g}(X_i, Y_i; \hat{f}_n)\Big] \\
&\le \sup_{f\in\mathcal{F}} \mathbb{E}_{\mathbb{D}}\Big[\mathbb{E}_{(X,Y)}\tilde{g}(X, Y; f) - \frac{3}{n}\sum_{i=1}^n \tilde{g}(X_i, Y_i; f)\Big] + 3\mathbb{E}_{\mathbb{D}}\Big[\frac{1}{n}\sum_{i=1}^n \tilde{g}(X_i, Y_i; f)\Big] \\
&= \sup_{f\in\mathcal{F}} \mathbb{E}_{\mathbb{X}}\Big[\mathbb{E}_X g(X; f) - \frac{3}{n}\sum_{i=1}^n g(X_i; f)\Big] + 3\mathcal{E}(f; \ell) \\
&\le \mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big[\mathbb{E}_X g(X; f) - \frac{3}{n}\sum_{i=1}^n g(X_i; f)\Big] + 3\mathcal{E}(f; \ell),
\end{aligned}$$

for any $f \in \mathcal{F}$. Here we used the convexity of supremum and Jensen's inequality. Hence

$$\mathbb{E}_{\mathbb{D}}\big[\mathcal{E}(\hat{f}_n; \ell)\big] \le \mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big[\mathbb{E}_X g(X; f) - \frac{3}{n}\sum_{i=1}^n g(X_i; f)\Big] + 3\inf_{f\in\mathcal{F}} \mathcal{E}(f; \ell).$$

Observe that $0 \le g(x; \cdot) \le 2B\kappa$ for all $x \in \mathcal{X}$ implies $g^2(x; \cdot) \le 2B\kappa g(x; \cdot)$, and consequently,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big[\mathbb{E}_X g(X; f) - \frac{3}{n}\sum_{i=1}^n g(X_i; f)\Big] \\
&= \mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big[2\mathbb{E}_X g(X; f) - \mathbb{E}_X g(X; f) - \frac{2}{n}\sum_{i=1}^n g(X_i; f) - \frac{1}{n}\sum_{i=1}^n g(X_i; f)\Big] \\
&\le \mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}}\Big(2\mathbb{E}_X g(X; f) - \frac{1}{2B\kappa}\mathbb{E}_X g^2(X; f) - \frac{2}{n}\sum_{i=1}^n g(X_i; f) - \frac{1}{2B\kappa n}\sum_{i=1}^n g^2(X_i; f)\Big).
\end{aligned}$$

We introduce a independent copy of $\mathbb{X}$, that is, $\mathbb{X}' := \{X_i'\}_{i=1}^n$, and let $\tau = \{\tau_i\}_{i=1}^n$ be a sequence of i.i.d. Rademacher variables independent of $\mathbb{X}$ and $\mathbb{X}'$. By the technique of symmetrization and the convexity of supremum and Jensen's inequality, we obtain

$$
\begin{aligned}
&\mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}} \Big(2\mathbb{E}_X g(X;f) - \frac{1}{2B\kappa}\mathbb{E}_X g^2(X;f) - \frac{2}{n}\sum_{i=1}^n g(X_i;f) - \frac{1}{2B\kappa n}\sum_{i=1}^n g^2(X_i;f)\Big) \\
&= \mathbb{E}_{\mathbb{X}} \sup_{f\in\mathcal{F}} \Big(\mathbb{E}_{\mathbb{X}'}\Big[\frac{2}{n}\sum_{i=1}^n g(X_i';f) - \frac{1}{2B\kappa n}\sum_{i=1}^n g^2(X_i';f)\Big] - \frac{2}{n}\sum_{i=1}^n g(X_i;f) - \frac{1}{2B\kappa n}\sum_{i=1}^n g^2(X_i;f)\Big) \\
&\le \mathbb{E}_{\mathbb{X}}\mathbb{E}_{\mathbb{X}'} \sup_{f\in\mathcal{F}} \Big(\frac{2}{n}\sum_{i=1}^n (g(X_i';f) - g(X_i;f)) - \frac{1}{2B\kappa n}\sum_{i=1}^n (g^2(X_i';f) + g^2(X_i;f))\Big) \\
&= \mathbb{E}_{\mathbb{X}}\mathbb{E}_{\mathbb{X}'}\mathbb{E}_{\tau} \sup_{f\in\mathcal{F}} \Big(\frac{2}{n}\sum_{i=1}^n \tau_i(g(X_i';f) - g(X_i;f)) - \frac{1}{2B\kappa n}\sum_{i=1}^n (g^2(X_i';f) + g^2(X_i;f))\Big) \\
&= 2\mathbb{E}_{\mathbb{X}'}\mathbb{E}_{\tau} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \Big(\tau_i g(X_i';f) - \frac{1}{4B\kappa} g^2(X_i';f)\Big) + 2\mathbb{E}_{\mathbb{X}}\mathbb{E}_{\tau} \sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \Big(-\tau_i g(X_i;f) - \frac{1}{4B\kappa} g^2(X_i;f)\Big) \\
&= 2\mathbb{E}_{\mathbb{X}'}\mathcal{R}_n^{\text{off}}\Big(\mathcal{G}, \frac{1}{4B\kappa}\Big|\mathbb{X}'\Big) + 2\mathbb{E}_{\mathbb{X}}\mathcal{R}_n^{\text{off}}\Big(\mathcal{G}, \frac{1}{4B\kappa}\Big|\mathbb{X}\Big) \\
&= 4\mathcal{R}_n^{\text{off}}\Big(\mathcal{G}, \frac{1}{4B\kappa}\Big).
\end{aligned}
$$

We complete the proof. □

C.2. Proof of Corollary 4.7

Proof. By the relationship between the covering number and the VC-dimension of the ReLU neural networks $\mathcal{F}$ (Anthony et al., 1999), we have

$$
\mathbb{E}_{\mathbb{X}}\big[N_\infty(\delta, \mathcal{F}, \mathbb{X})\big] \le \Big(\frac{eBn}{\delta \mathrm{VC}_{\mathcal{F}}}\Big)^{\mathrm{VC}_{\mathcal{F}}}, \tag{C.1}
$$

where $\mathrm{VC}_{\mathcal{F}}$ denotes the VC-dimension of $\mathcal{F}$. Moreover, the VC-dimension for the ReLU neural network $\mathcal{F}$ satisfies

$$
c_1 \cdot \mathcal{D}\mathcal{S}\log(\mathcal{S}/\mathcal{D}) \le \mathrm{VC}_{\mathcal{F}} \le c_2 \cdot \mathcal{D}\mathcal{S}\log\mathcal{S} \tag{C.2}
$$

with universal constant c_1 and c_2, see Bartlett et al. (2019).

By Theorem 3.3 of Jiao et al. (2021), for any $W, L \in \mathbb{N}$, there exists a function ϕ belonging to the ReLU neural networks $\mathcal{F}$ with width $\mathcal{W} = 38(s+1)^2 d^{s+1} W \lceil \log_2(8W) \rceil$ and depth $\mathcal{D} = 21(s+1)^2 L \lceil \log_2(8L) \rceil$ such that

$$
|f_0(x) - \phi(x)| \le 18B(s+1)^2 d^{s+(\varsigma\vee 1)/2} (WL)^{-2\varsigma/d}, \tag{C.3}
$$

for any $x \in \cup_\theta Q_\theta$, where

$$
Q_\theta = \Big\{x = (x_1, x_2, \ldots, x_d) : x_i \in \Big[\tfrac{\theta_i}{K}, \tfrac{\theta_i+1}{K} - \widetilde{\delta}\cdot 1_{\{\theta_i < K-1\}}\Big], i = 1, 2, \ldots, d\Big\}
$$

with $\theta = (\theta_1, \theta_2, \ldots, \theta_d) \in \{0, 1, \ldots, K-1\}^d$ and $0 < \widetilde{\delta} \le \frac{1}{3K}$. Then the Lebesgue measure of $[0,1]^d \backslash Q_\theta$ is no more than $dK\widetilde{\delta}$ which can be arbitrarily small if $\widetilde{\delta}$ is arbitrarily small. Since μ_X is absolutely continuous with respect to the Lebesgue measure, we have

$$
\inf_{f\in\mathcal{F}} \|f - f_0\|_{L^2(\mu_X)}^2 \le 324B^2(s+1)^4 d^{2s+(\varsigma\vee 1)} (WL)^{-4\varsigma/d}.
$$

By (C.1) to (C.3), setting $W = \mathcal{O}(n^{\frac{d}{8\varsigma+4d}})$ and $L = \mathcal{O}(n^{\frac{d}{8\varsigma+4d}})$ yields the desired result. □

C.3. Proofs of Lemmas 4.4, 4.8, 4.10 and 4.14 and Corollary 4.11

We first prove Lemma 4.4.

Proof. Define

$$g(x;f) = \mathbb{E}\big[(f(X)-Y)^2 - (f^*(X)-Y)^2 \big| X = x\big], x \in \mathcal{X}.$$

Then, we have

$$\begin{aligned} g(X;f) &= \mathbb{E}\big[(f(X)-f^*(X)+f^*(X)-Y)^2 - (f^*(X)-Y)^2 \big| X\big] \\ &= \mathbb{E}\big[(f(X)-f^*(X))^2 + 2(f(X)-f^*(X))(f^*(X)-Y) \big| X\big] \\ &= (f(X)-f^*(X))^2. \end{aligned}$$

Thus $\mathcal{E}(f;\ell_{\text{ls}}) = \|f-f^*\|^2_{L^2(\mu_X)}$ for any $f \in \mathcal{F}$. Observe that $|g(x;\cdot)|$ is $4B$-Lipschitz continuous for any $x \in \mathcal{X}$, then using Theorems 2.3 and 4.2 yields the desired result. □

Lemmas 4.8 and 4.14 are directly followed from Theorems 2.3 and 4.2. In addition, combing Theorems 2.3 and 4.2 and Lemma B.3, we can obtain Lemma 4.10. Similar to the proof of Corollary 4.7, we can deduce Corollary 4.11.