Toward fast rates: A review of localization analysis for statistical learning

Ilsang Ohn[1] and Yongdai Kim[2*]

[1]Department of Statistics, Inha University, 100 Inha-ro, Michuhol-gu, Incheon, 22212, Republic of Korea.
[2*]Department of Statistics, Seoul National University, 1 Gwanak-ro, Gwanak-gu, Seoul, 08826, Republic of Korea.

*Corresponding author(s). E-mail(s): ydkim0903@gmail.com;
Contributing authors: ilsang.ohn@inha.ac.kr;

Abstract

Deriving convergence rates of the excess risks of machine learning methods is an important task for theoretical statistics. A classical approach utilizes the uniform convergence of an empirical process which is established under the control of the global complexity of the corresponding function class. However, the global complexity may not be a sharp tool to describe the behavior of the empirical process, and, as a result, this classical approach only allows slower rates than $n^{-1/2}$ where n denotes the sample size. To get faster and probably optimal rates, we need a refined theoretical technique called localization analysis. Despite its importance and usefulness, an elementary and user-friendly introduction to the localization analysis is limited. In this paper, we attempt to give such an introduction together with several applications.

Keywords: Convergence rate, Empirical risk minimizer, Local Rademacher complexity, Metric entropy, Penalized estimator, Talagrand inequality

1 Introduction

Providing a theoretical understanding of machine learning procedures is useful for several important tasks, for example, anticipating their generalization behaviors, ways to improve them, and finding the best one among them, etc. In a decision-theoretic framework, we evaluate the *risk* of an estimator, which is the expected loss incurred by using the estimator, and then compare it with that of the *Bayes predictor* that is the function minimizing the risk. The

difference between the two risk values is referred to as the *excess risk*. Using several probabilistic tools, we aim to find a high-probability upper bound of the excess risk, in terms of the sample size. This sequence indexed by the sample size is called a *convergence rate*. By obtaining the convergence rates of some estimators, we can compare them and choose the best one from a theoretical perspective.

The first approach goes back to a seminal work by Vapnik and Chervonenkis [1]. They established a worst-case upper bound of the excess risk of an empirical risk minimization estimator by leveraging the uniform convergence theory of empirical processes due to Glivenko, Cantelli, and Donsker. The uniform convergence relies on the complexity of its function class and Vapnik and Chervonenkis [1] developed a notion of the VC dimension, which measures the complexity of a function class suitably. Another widely used complexity measure is the Rademacher complexity, which was used to analyze machine learning for the first time in Koltchinskii [2], Koltchinskii and Panchenko [3], Bartlett et al. [4], Bartlett and Mendelson [5]. The Rademacher complexity can be estimated in terms of the metric entropy via Dudely's entropy integral [6].

However, as we will explain, these complexity measures are not appropriate in deriving the rates of convergence. This is mainly due to the fact that they provide *global* complexity, that is, they measure the complexity of the whole function class. In this sense, the algorithmic power of estimation procedures, which guarantees the estimator lands down to a good subset of the function class with high probability, cannot be fully reflected in the global complexity measures. As a result, the convergence rate derived by the global complexity measures may be too slow and turns out to be no faster than $n^{-1/2}$, where n denotes the sample size. This "slow" $n^{-1/2}$ rate is suboptimal for some statistical problems.

A localization analysis was proposed to overcome the drawback of the global complexity analysis and enables us to attain faster rate than $n^{-1/2}$. A main intuition is that when our estimator is expected to be close to the Bayes predictor, the complexity of a small neighborhood of the Bayes predictor may be sufficient to establish a upper bound of the excess risk. In general, the excess risk can be bounded by a *fixed point* of the *locally measured* complexity of the function class. A localized entropy integral method was introduced by Geer [7] and the use of bracketing entropy was studied in Shen and Wong [8]. The local Rademacher complexity, which can be estimated in a data-dependent fashion and so used in model selection, was introduced in Bartlett et al. [9]. A unified and general overview of the localization analysis can be found in Koltchinskii [10]. See also Massart and Nédélec [11]. In this paper, we provide a user-friendly tutorial for the localization analysis.

1.1 Setup

We introduce our setup for statistical inference. A reader might refer to Section 1.2 to get the meaning of the notations used in this paper. Let $(\mathcal{Z}, \mathfrak{A}, \mathsf{P})$ be a probability space and let Ξ Let $Z_{1:n} := (Z_1, \ldots, Z_n)$ be a sample of i.i.d. random variables taking the values in $\mathcal{Z}$ with common distribution P. Let P_n denote the corresponding empirical distribution. We denote by $\mathbb{P}$ the law of the sample $Z_{1:n}$ and $\mathbb{E}$ be the corresponding expectation operator. Let Ξ be a certain class of measurable functions from $\mathcal{Z}$ to a measurable space $\mathcal{S}$. Let $\ell : \mathcal{Z} \times \mathcal{S} \mapsto [0, \infty)$ be a *loss* function such that we evaluate the performance of a function f at a point $z \in \mathcal{Z}$ by $\ell(z, f(z))$. For notational simplicity, let $\ell \circ f : \mathcal{Z} \mapsto \mathbb{R}$ be a function such that $\ell \circ f(z) = \ell(z, f(z))$ for any $z \in \mathcal{Z}$. We define the *risk* of f as $\mathsf{P}[\ell \circ f]$. Let $f_\star$ be the Bayes predictor that is a minimizer of

the risk, i.e.,

$$f_\star = \operatorname*{argmin}_{f\in\Xi} \mathsf{P}[\ell\circ f].$$

The *excess risk* (with respect to the best $f_\star$) of a function f is defined as

$$\mathcal{E}(f) := \mathsf{P}[\ell\circ f] - \mathsf{P}[\ell\circ f_\star] = \mathsf{P}[\ell\circ f] - \min_{\tilde{f}\in\Xi} \mathsf{P}[\ell\circ \tilde{f}].$$

The aim is to find an estimator that has a small excess risk. A popular one is the empirical risk minimization (ERM) estimator, which is obtained by minimizing the empirical risk $\mathsf{P}_n[\ell\circ f]$ over a certain function $\mathcal{F}\in\Xi$ chosen by a user for estimation, that is,

$$\hat{f}_n := \widehat{f}_n(\mathcal{F}) = \operatorname*{argmin}_{f\in\mathcal{F}} \mathsf{P}_n[\ell\circ f] \tag{1.1}$$

Another type of an estimator is the *penalized* ERM estimator which minimizes the sum of the empirical risk and a certain deterministic penalty $\Gamma : \mathcal{F}\mapsto[0,\infty)$, that is,

$$\widehat{f}_{n,\lambda} := \widehat{f}_{n,\lambda}(\mathcal{F},\Gamma) = \operatorname*{argmin}_{f\in\mathcal{F}} \left\{\mathsf{P}_n[\ell\circ f] + \lambda\Gamma(f)\right\}, \tag{1.2}$$

where $\lambda\geq 0$ denotes a tuning parameter that controls the degree of penalization. In this paper, we study the above two types of an estimator.

1.2 Notation

For a probability measure Q, we write $\mathsf{Q}[g] = \int g\mathrm{d}\mathsf{Q}$ and $\mathsf{Var}_\mathsf{Q}(g) := \mathsf{Q}[g^2] - (\mathsf{Q}[g])^2$ to denote the expectation and variance of a function g with respect to Q. For a random variable X, we denote by $\mathbb{E}_X$ the expectation operator with respect to the law of the random variable X, for instance, for $X\sim\mathsf{Q}$, $\mathbb{E}_X[g(X)] = \int g(x)\mathrm{d}\mathsf{Q}(x)$. We write $\|g\|_\infty := \sup_{z\in\mathcal{Z}}|g(z)|$ and $\|g\|_{\mathfrak{L}^q(\mathsf{Q})} := (\int |g(z)|^q\mathrm{d}\mathsf{Q}(z))^{1/q}$ for $q>0$. For a vector valued function $g=(g_1,\ldots,g_d)^\top$, we write $\|g\|_{\mathfrak{L}^q(\mathsf{Q})} := (\sum_{j=1}^d \|g_j\|^q_{\mathfrak{L}^q(\mathsf{Q})})^{1/q}$. For a natural number $N\in\mathbb{N}$, we denote $[N] := \{1,2,\ldots,N\}$. For two real numbers a and b, we write $a\vee b := \max\{a,b\}$ and $a\wedge b := \min\{a,b\}$. For two positive sequences $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, we write $a_n\lesssim b_n$ or $b_n\gtrsim a_n$, if there exists a positive constant $C>0$ such that $a_n\leq Cb_n$ for any $n\in\mathbb{N}$. We write $a_n\asymp b_n$ if both $a_n\lesssim b_n$ and $a_n\gtrsim b_n$ hold. We denote by $\mathbb{1}(\cdot)$ the indicator function. Absolute constants $C_1,C_2,\ldots,$ may vary from place to place.

1.3 Outline

The rest of the paper is organized as follows. In Section 2, we provide the intuition as well as the detailed technical arguments of the localization analysis for the ERM estimator. In Section 3, we introduce several local complexity measures and some examples. The localization analysis for the penalized ERM estimator is given in Section 4. In Section 5, we study several concrete statistical problems with the localization analysis developed in the previous sections.

2 Localization analysis for fast rates

In this section, we give an outline for deriving fast rates with localization analysis for the ERM estimator. For technical simplicity, we impose the assumption that the loss function is bounded. Although satisfied in many applications including classification and quantile regression, this does not cover many important problems such as regression with unbounded outputs. But this limitation can be handled without much difficulty, as illustrated in some examples in Section 5.

Assumption 1 (Bounded loss and functions)**.** There exists an absolute constant $B > 0$ such that $\|\ell \circ f - \ell \circ f_\star\|_\infty \le B$.

2.1 Basic inequality and global complexity analysis

Basic inequality. The first step is to translate the randomness of the ERM estimator $\widehat{f}_n$ to the one easier to analyze. This is attained by using the optimization optimality of $\widehat{f}_n$ such that

$$\mathsf{P}_n[\ell \circ \widehat{f}_n] \le \mathsf{P}_n[\ell \circ f] \text{ for any } f \in \mathcal{F}, \tag{2.1}$$

which is called the *basic inequality*. From the basic inequality, when $f_\star \in \mathcal{F}$, we have

$$\begin{aligned} \mathcal{E}(\widehat{f}_n) &= \mathsf{P}[\ell \circ \widehat{f}_n] - \mathsf{P}[\ell \circ f_\star] \\ &= (\mathsf{P} - \mathsf{P}_n)[\ell \circ \widehat{f}_n - \ell \circ f_\star] + \mathsf{P}_n[\ell \circ \widehat{f}_n - \ell \circ f_\star] \\ &\le (\mathsf{P} - \mathsf{P}_n)[\ell \circ \widehat{f}_n - \ell \circ f_\star], \end{aligned} \tag{2.2}$$

where the last line will be further bounded by analyzing the behavior of the empirical process $f \mapsto (\mathsf{P} - \mathsf{P}_n)[\ell \circ f - \ell \circ f_\star]$. When our function class does not include $f_\star$, we need to find a good approximation of $\bar{f} \in \mathcal{F}$ of $f_\star$. We will discuss this issue in Section 2.4.

Sub-optimality of global complexity analysis. A sub-optimal approach for finding a upper bound of the empirical process $(\mathsf{P} - \mathsf{P}_n)[\ell \circ \widehat{f}_n - \ell \circ f_\star]$ is to employ the *global* bound such as

$$(\mathsf{P} - \mathsf{P}_n)[\ell \circ \widehat{f}_n - \ell \circ f_\star] \le \sup_{f \in \mathcal{F}} (\mathsf{P} - \mathsf{P}_n)[\ell \circ f - \ell \circ f_\star].$$

Then we give a high probability upper bound of the right-hand side of the preceding display, by the *Talagrand inequality*, or the functional Bernstein inequality, stated as the next lemma.

Lemma 2.1 (Talagrand)**.** *Let $\mathcal{G}$ be a class of functions on $\mathcal{Z}$ such that $\sup_{g \in \mathcal{G}} \|g\|_\infty \le B$ for some $B > 0$. Then*

$$\mathbb{P}\left(\sup_{g \in \mathcal{G}} (\mathsf{P} - \mathsf{P}_n)[g] \ge 2\mathbb{E}\big[\sup_{g \in \mathcal{G}} (\mathsf{P} - \mathsf{P}_n)[g]\big] + \sqrt{\frac{2t}{n} \sigma_{\mathsf{P}}^2(\mathcal{G})} + \frac{4Bt}{3n} \right) \le \mathrm{e}^{-t}$$

where we denote $\sigma_{\mathsf{P}}^2(\mathcal{G}) := \sup_{g \in \mathcal{G}} \mathsf{Var}_{\mathsf{P}}(g)$.

There are many resources providing the proof. For example, see the proof of Theorem 3.3.9 of Giné and Nickl [12]. Here, we give a more convenient version using the AM-GM

inequality $\sqrt{(2\mathbb{E}[W])(2Bt/n)} \leq \mathbb{E}[W] + Bt/n$ compared with the standard form

$$\mathbb{P}\Big(W \geq \mathbb{E}[W] + \sqrt{\frac{2t}{n}\left\{\sigma_{\mathsf{P}}^2(\mathcal{G}) + 2B\mathbb{E}[W]\right\}} + \frac{Bt}{3n}\Big) \leq \mathrm{e}^{-t}$$

due to Bousquet [13], where we denote $W := \sup_{g\in\mathcal{G}}(\mathsf{P}-\mathsf{P}_n)[g]$.

We apply the Talagrand inequality in Lemma 2.1 to the class

$$\ell(\mathcal{F}_{-f_\star}) := \{\ell\circ f - \ell\circ f_\star : f \in \mathcal{F}\}.$$

Then since $\mathsf{Var}_{\mathsf{P}}(\ell\circ f - \ell\circ f_\star) \leq B^2$ for any $f\in\mathcal{F}$, we have

$$\sup_{f\in\mathcal{F}}(\mathsf{P}-\mathsf{P}_n)[\ell\circ f - \ell\circ f_\star] \leq 2\mathbb{E}\Big[\sup_{f\in\mathcal{F}}(\mathsf{P}-\mathsf{P}_n)[\ell\circ f - \ell\circ f_\star]\Big] + \sqrt{\frac{2B^2t}{n}} + \frac{4Bt}{3n}$$

with probability at least $1-\mathrm{e}^{-t}$. The second term of this upper bound is of order $n^{-1/2}$. Moreover, a usual upper bound of the expected global supremum of the empirical process is also of order $n^{-1/2}$. This is the case even for a finite function class.

Example 1 (Finite function class)**.** Let $\mathcal{F} = \{f_1,\dots,f_N\}$ be a class of a finite number of functions such that Assumption 1 holds. We let $g_j := \ell\circ f_j - \ell\circ f_\star$ for simplicity. Then we have that for any $t>0$

$$\begin{aligned}
\exp\Big(t\mathbb{E}\Big[\sup_{f\in\mathcal{F}}(\mathsf{P}-\mathsf{P}_n)[\ell\circ f - \ell\circ f_\star]\Big]\Big) &= \exp\Big(t\mathbb{E}\Big[\max_{j\in[N]}(\mathsf{P}-\mathsf{P}_n)[g_j]\Big]\Big) \\
&\leq \mathbb{E}\Big[\exp\Big(t\max_{j\in[N]}(\mathsf{P}-\mathsf{P}_n)[g_j]\Big)\Big] \\
&\leq \sum_{j=1}^{N}\mathbb{E}\Big[\exp\big(t(\mathsf{P}-\mathsf{P}_n)[g_j]\big)\Big] \\
&= \sum_{j=1}^{N}\prod_{i=1}^{n}\mathbb{E}_{Z_i}\Big[\exp\big(\frac{t}{n}(\mathsf{P}[g_j] - g_j(Z_i))\big)\Big] \\
&\leq N\exp(t^2B^2/(2n)),
\end{aligned}$$

where the first inequality follows from Jensen's inequality and the last inequality from Hoeffiding's lemma. Taking logarithms on both sides and setting $t = t^* := \sqrt{2n\log N}/B$, we have

$$\mathbb{E}\Big[\sup_{f\in\mathcal{F}}(\mathsf{P}-\mathsf{P}_n)[\ell\circ f - \ell\circ f_\star]\Big] \leq \frac{1}{t^*}\log N + \frac{B^2}{2n}t^* = \sqrt{\frac{2B^2\log N}{n}}.$$

2.2 Localization analysis

We have shown that the global complexity analysis only yields slow rates, and so we need new theoretical tools. In this section, we introduce the general idea of *localization analysis* which allows us to have fast rates.

Heuristic of localization analysis. To give readers intuition, we provide a heuristic explanation of the localization technique. We define a *localized* subset of $\mathcal{F}$ as

$$\mathcal{F}(\delta) := \{f \in \mathcal{F} : \mathcal{E}(f) \le \delta\}$$

for a positive number $\delta > 0$. We denote the bounding term in the Talagrand inequality applied to the localized class $\mathcal{F}(\delta)$ by

$$U_n(\delta, t) := 2\mathbb{E}\Big[\sup_{f\in\mathcal{F}(\delta)} (\mathsf{P}-\mathsf{P}_n)(\ell\circ f - \ell\circ f_\star)\Big] + \sqrt{\frac{2t}{n}\sup_{f\in\mathcal{F}(\delta)} \mathsf{Var}_{\mathsf{P}}(\ell\circ f - \ell\circ f_\star)} + \frac{4Bt}{3n}.$$

Suppose that Assumption 1 holds. We first take $\delta_1 := B$ so that $\mathcal{F}(\delta_1) = \mathcal{F}$. Then by the Talagrand inequality, for $t_1 > 0$, the event defined as

$$\mathfrak{A}_1 := \left\{ Z_{1:n} \in \mathcal{Z}^n : \sup_{f\in\mathcal{F}(\delta_1)} (\mathsf{P}-\mathsf{P}_n)[\ell\circ f - \ell\circ f_\star] \le \delta_2 := U_n(\delta_1, t_1) \right\} \tag{2.3}$$

occurs with probability at least $1 - \mathrm{e}^{-t_1}$. Then on $\mathfrak{A}_1$, it follows that $\mathcal{E}(\widehat{f}_n) \le \delta_2$, that is, $\widehat{f}_n \in \mathcal{F}(\delta_2)$. On the other hand, for $t_2 > 0$, the event defined as

$$\mathfrak{A}_2 := \left\{ Z_{1:n} \in \mathcal{Z}^n : \sup_{f\in\mathcal{F}(\delta_2)} (\mathsf{P}-\mathsf{P}_n)[\ell\circ f - \ell\circ f_\star] \le \delta_3 := U_n(\delta_2, t_2) \right\} \tag{2.4}$$

occurs with probability at least $1 - \mathrm{e}^{-t_2}$ by the Talagrand inequality. Thus, we have that

$$\begin{aligned}\mathbb{P}(\mathcal{E}(\widehat{f}_n) > \delta_3) &\le \mathbb{P}\big(\{\mathcal{E}(\widehat{f}_n) > \delta_3\} \cap \mathfrak{A}_1\big) + \mathbb{P}(\mathfrak{A}_1^{\complement}) \\ &\le \mathbb{P}(\delta_2 \ge \mathcal{E}(\widehat{f}_n) > \delta_3) + \mathrm{e}^{-t_1} \le \mathbb{P}(\mathfrak{A}_2^{\complement}) + \mathrm{e}^{-t_1} \le \mathrm{e}^{-t_1} + \mathrm{e}^{-t_2}.\end{aligned}$$

That is, we have $\mathcal{E}(\widehat{f}_n) \le \delta_3$ with probability at least $1 - \mathrm{e}^{-t_1} - \mathrm{e}^{-t_2}$. In other words, we first show that the ERM estimaor $\widehat{f}_n$ belongs to a localized function class $\mathcal{F}(\delta_2)$, and then, working on the localized function class, we get a high probability bound δ_3 of the excess risk smaller than δ_2. By proceeding in this manner, we can obtain a sharper bound. We define the sequence $(\delta_j)_{j\in\mathbb{N}}$ recursively as $\delta_{j+1} := U_n(\delta_j, t_j)$, then we have $\mathcal{E}(\widehat{f}_n) \le \delta_N$ with probability at least $1 - \sum_{j=1}^{N} \mathrm{e}^{-t_j}$. The upper bound δ_N may be substantially smaller than the upper bound δ_2 obtained by the global complexity analysis. Indeed, when $U_n(\delta, t)$ is a concave function of δ, this iterative argument can be beneficial, as illustrated in Figure 1.

Fixed point method. We expressed the aforementioned iterative argument in a formal manner. We will see that the resulting bound after the iteration can be expressed by a *fixed point* of a certain function we will call a *local complexity function*. By doing so, it is easier to describe the excess risk bound by the localization analysis. Also, we have the freedom to choose the sequence $(\delta_j)_{j\in\mathbb{N}}$ other than the iteratively defined one $\delta_{j+1} := U_n(\delta_j, t_j)$.

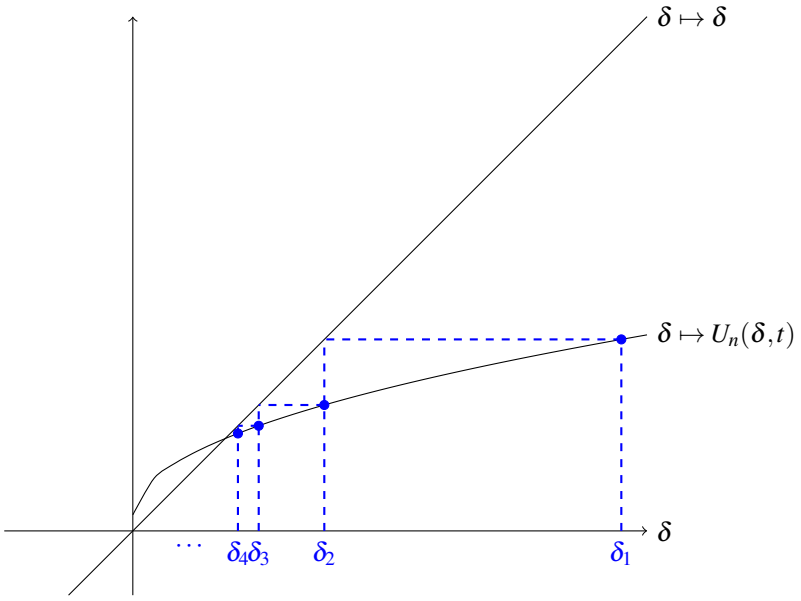

Fig. 1: Illustration of the iterative arguement

Definition 1. Let $(\delta_j)_{j\in\mathbb{N}}$ be a decreasing sequence of positive numbers and $(t_j)_{j\in\mathbb{N}}$ be a sequence of positive numbers. Then we say that a monotonically increasing function $\Psi_n : [0,\infty) \mapsto [0,\infty)$ a *local complexity function* of $\mathcal{F}$, if it satisfies

$$\Psi_n(\delta_{j+1}) \geq U_n(\delta_j, t_j) \text{ for any } j \in \mathbb{N}.$$

The *fixed point* $\delta_n^\dagger$ of Ψ_n is defined as

$$\delta_n^\dagger := \sup\{\delta \in [0,\infty) : \delta \leq \Psi_n(\delta)\}. \tag{2.5}$$

The next proposition states that the excess risk of $\widehat{f}_n$ is bounded above by the fixed point with high probability.

Proposition 2.2. *Suppose that Assumption 1 holds and that $f_\star \in \mathcal{F}$. Consider the setup in Definition 1 and choose $\delta_1 := B$. Then for any $\delta > \delta_n^\dagger$,*

$$\mathbb{P}\big(\mathcal{E}(\widehat{f}_n) > \delta\big) \leq \sum_{j\in\mathbb{N}:\delta_j>\delta} \exp(-t_j). \tag{2.6}$$

As a particular consequence of the above, we have

$$\mathbb{P}\big(\mathcal{E}(\widehat{f}_n) > \delta\big) \leq \log_2\left(\frac{2B}{\delta}\right)\exp(-t) \tag{2.7}$$

for any $t > 0$.

Proof. Let $J := J(\delta) := \sup\{j \in \mathbb{N} : \delta_j > \delta\}$. Then since $\mathcal{E}(\widehat{f}_n) \le \delta_1 := B$ by assumption, from the union bound,

$$\mathbb{P}\big(\mathcal{E}(\widehat{f}_n) > \delta\big) \le \mathbb{P}\big(\mathcal{E}(\widehat{f}_n) > \delta_J\big) \le \sum_{j=1}^{J-1} \mathbb{P}\Big(\delta_{j+1} < \mathcal{E}(\widehat{f}_n) \le \delta_j\Big).$$

Since it follows that $\mathcal{E}(\widehat{f}_n) \le \sup_{f \in \mathcal{F}(\delta_j)} (\mathsf{P} - \mathsf{P}_n)[\ell \circ f - \ell \circ f_\star]$ on the event $\{\widehat{f}_n \in \mathcal{F}(\delta_j)\}$, we have

$$\begin{aligned}\mathbb{P}\big(\delta_{j+1} < \mathcal{E}(\widehat{f}_n) \le \delta_j\big) &= \mathbb{P}\big(\delta_{j+1} < \mathcal{E}(\widehat{f}_n) \text{ and } f \in \mathcal{F}(\delta_j)\big) \\ &\le \mathbb{P}\Big(\sup_{f \in \mathcal{F}(\delta_j)} (\mathsf{P} - \mathsf{P}_n)[\ell \circ f - \ell \circ f_\star] > \delta_{j+1}\Big).\end{aligned}$$

Then by definition of $\delta_n^\dagger$, we have $\delta_{j+1} > \Psi_n(\delta_{j+1})$ for any $j \in [J-1]$. Moreover, by the definition of Ψ_n, we have $\Psi_n(\delta_{j+1}) \ge U_n(\delta_j, t_j)$. Therefore, by the Talagrand inequality, we get

$$\mathbb{P}\big(\delta_{j+1} < \mathcal{E}(\widehat{f}_n) \le \delta_j\big) \le \mathbb{P}\Big(\sup_{f \in \mathcal{F}(\delta_j)} (\mathsf{P} - \mathsf{P}_n)[\ell \circ f - \ell \circ f_\star] > U_n(\delta_j, t_j)\Big) \le \mathrm{e}^{-t_j},$$

which completes the proof of the result (2.6)

For the second assertion (2.7), we take $\delta_j = 2^{-j}(2B)$ and $t_j = t$ for any $j \in \mathbb{N}$. Then since the inequality $\delta_j = 2^{-j}(2B) > \delta$ is equivalent to $j < \log_2(2B/\delta)$, we get the desired result. □

The next hypothetical example illustrates a situation where the localization analysis enables us to obtain a fast rate n^{-1}.

Example 2. Suppose that a local complexity function is given by

$$\Psi_n(\delta) = c_1 \Big\{ \sqrt{\frac{\delta}{n}} + \frac{1}{n} \Big\}$$

for some constant $c_1 \in (0,1)$. Then the global complexity with $\delta = B$ leads to the bound of order $n^{-1/2}$. But using the AM-GM inequality, we have $\sqrt{\delta/n} \le \delta/2 + 1/(2n)$, and thus we can see that the fixed point $\delta_n^\dagger$ of Ψ_n satisfies

$$\delta_n^\dagger \le \frac{3c_1}{2 - c_1} \frac{1}{n}.$$

and so the localization analysis gives a fate rate n^{-1}.

2.3 Bernstein condition: Toward sub-root local complexity

The fixed point method reveals that obtaining fast rates is closely related to a "shape" of the local complexity function Ψ_n. Intuitively, the function Ψ_n should be concave on $[0, \delta)$. We introduce the Bernstein condition to get a concave Ψ_n.

Definition 2. Let $\kappa \in (0,1]$ and $R > 0$. We say that $\mathcal{G}$ is a (κ, R)-*Bernstein class* with respect to a probability measure P, if

$$\mathsf{Var}_{\mathsf{P}}[g] \leq R(\mathsf{P}[g])^{\kappa}$$

for every $g \in \mathcal{G}$. We call κ the *Bernstein exponent*.

Assumption 2 (Bernstein)**.** The class $\ell(\mathcal{F}_{-f_\star}) := \{\ell \circ f - \ell \circ f_\star : f \in \mathcal{F}\}$ is a (κ, R)-Bernstein class with respect to a probability measure P. That is, $\mathsf{Var}_{\mathsf{P}}(\ell \circ f - \ell \circ f_\star) \leq R[\mathsf{P}[\ell \circ f - \ell \circ f_\star])^{\kappa} = R(\mathcal{E}(f))^{\kappa}$ for any $f \in \mathcal{F}$.

Example 3 (Bonded regression)**.** Consider a bounded regression problem where each sample point is a pair of input and output $Z_i = (X_i, Y_i)$ with $X_i \in \mathcal{X}$ and $Y_i \in [-B, B]$ for some $B > 0$, where $\mathcal{X}$ denotes the input space. Assume that $\|f\|_\infty \leq B$ for any $f \in \mathcal{F}$. Consider a square loss function ℓ_{sq} such that $\ell_{\mathrm{sq}} \circ f(Z) = (Y - f(X))^2$. Then it is clear that $\|f_\star\|_\infty \leq B$. Thus, we have

$$\begin{aligned}\left\{(Y - f(X))^2 - (Y - f_\star(X))^2\right\}^2 &= \left\{(2Y - f(X) - f_\star(X))(f(X) - f_\star(X))\right\}^2 \\ &\leq 16B^2 \left\{f(X) - f_\star(X)\right\}^2\end{aligned}$$

for any $f \in \mathcal{F}$. This implies that

$$\begin{aligned}\mathsf{Var}_{\mathsf{P}}(\ell \circ f - \ell \circ f_\star) &\leq \mathsf{P}[(\ell \circ f - \ell \circ f_\star)^2] \\ &\leq 16B^2 \mathsf{P}[(f - f_\star)^2] = 16B^2 \mathcal{E}(f).\end{aligned}$$

Hence $\ell_{\mathrm{sq}}(\mathcal{F}_{-f_\star})$ is a $(1, 16B^2)$-Bernstein class.

Next, we see the results under the Bernstein assumption.

Bernstein for the variance. Under the Bernstein assumption, we have

$$\sup_{f \in \mathcal{F}(\delta)} \mathsf{Var}_{\mathsf{P}}(\ell \circ f - \ell \circ f_\star) \leq R \sup_{f \in \mathcal{F}(\delta)} \mathcal{E}(f)^{\kappa} \leq R\delta^{\kappa}$$

Then by Lemma A.1 with $a = (2(16/\kappa)^{\kappa} BRt_j/n)^{1/2}$, $b = ((\kappa\delta/16)^{\kappa})^{1/2}$, $p = 2/(2-\kappa)$ and $q = 2/\kappa$, the variance term in $U_n(\delta, t)$ can be bounded as

$$\sqrt{\frac{2t}{n} \sup_{f \in \mathcal{F}(\delta)} \mathsf{Var}_{\mathsf{P}}(\ell \circ f - \ell \circ f_\star)} \leq \sqrt{\frac{2Rt\delta^{\kappa}}{n}} \leq \frac{1}{16}\delta + C_\kappa \left(\frac{Rt}{n}\right)^{1/(2-\kappa)} \tag{2.8}$$

where we let $C_\kappa := (2-\kappa)\{2(16/\kappa)^{\kappa}\}^{1/(2-\kappa)}/2$. Thus we can see that the $n^{-1/2}$ order can be improved to $n^{-1/(2-\kappa)}$, which is much faster for $\kappa \in (0,1]$.

Bernstein for the expected supremum of the empirical process. The Bernstein assumption also plays an important role in analyzing the expected supremum in $U_n(\delta, t)$. We see the detailed relationship in the next section, and now we introduce an assumption to deal with this term. This is a simpler but easier-to-use version of the assumptions in the related literature [9–11].

Assumption 3 (Sub-root complexity)**.** There exists a constant $\rho \in (0,1]$ and sequences of positive numbers $(\phi_{0,n})_{n\in\mathbb{N}}$ and $(\phi_{1,n})_{n\in\mathbb{N}}$ such that

$$\varphi_n(\delta,\mathcal{F}) := \mathbb{E}\Big[\sup_{f\in\mathcal{F}(\delta)} (\mathsf{P}-\mathsf{P}_n)(\ell\circ f-\ell\circ f_\star)\Big] \le \phi_{1,n}\delta^{\rho/2}+\phi_{0,n} \tag{2.9}$$

for any $\delta > 0$.

Usually, the sequence $(\phi_{1,n})_{n\in\mathbb{N}}$ is not faster than $n^{-1/2}$. But the sub-root dependence of $\varphi_n(\delta)$ on δ can lead to a faster rate as a similar argument used for deriving (2.8) as

$$\phi_{1,n}\delta^{\rho/2}+\phi_{0,n} \le \frac{1}{16}\delta + C_\rho\,(\phi_{1,n})^{2/(2-\rho)}+\phi_{0,n} \tag{2.10}$$

with $C_\rho := (2-\rho)(16/\rho)^{\rho/(2-\rho)}/2$. Thus, since $2/(2-\rho) > 2$ for $\rho \in (0,1]$, the rate $(\phi_{1,n})^{2/(2-\rho)}$ can be faster than $n^{-1/2}$. Motivated by (2.10), we define the quantity $\bar{\phi}_n$, which represents the "effective" complexity of $\mathcal{F}$, as

$$\bar{\phi}_n(\mathcal{F}) := (\phi_{1,n})^{2/(2-\rho)} \vee \phi_{0,n} \tag{2.11}$$

which we call the *estimation error* of the function class $\mathcal{F}$.

Combining (2.8) and (2.10), we know that the excess risk can be upper bounded by $\max\{\bar{\phi}_n(\mathcal{F}), (Rt/n)^{1/(2-\kappa)}, Bt/n\}$ up to a constant. For details, see the proof of Theorem 2.4 given later.

2.4 Incorporating approximation error

When $f_\star \notin \mathcal{F}$, it is no longer guaranteed that $\mathsf{P}_n[\ell\circ\widehat{f}_n] \le \mathsf{P}_n[\ell\circ f_\star]$. In this case, we use the basic inequality in a slightly different way, namely,

$$\begin{aligned}\mathcal{E}(\widehat{f}_n) &\le (\mathsf{P}-\mathsf{P}_n)[\ell\circ\widehat{f}_n-\ell\circ f_\star]+\mathsf{P}_n[\ell\circ\widehat{f}_n-\ell\circ f_\star]\\ &\le (\mathsf{P}-\mathsf{P}_n)[\ell\circ\widehat{f}_n-\ell\circ f_\star]+\mathsf{P}_n[\ell\circ\bar{f}-\ell\circ f_\star]\end{aligned}$$

for an arbitrary function $\bar{f}\in\mathcal{F}$. If the function $\bar{f}$ appearing in the preceding display is close to the best function $f_\star$, we expect that the second term is sufficiently small. The following proposition illustrates this.

Proposition 2.3. *Let $\bar{f}\in\mathcal{F}$. Under Assumptions 1 and 2,*

$$\mathbb{P}\left(\mathsf{P}_n[\ell\circ\bar{f}-\ell\circ f_\star] \ge (2+\kappa)\mathcal{E}(\bar{f})+\frac{2-\kappa}{2}\left(\frac{2Rt}{n}\right)^{1/(2-\kappa)}+\frac{4Bt}{3n}\right) \le \mathrm{e}^{-t}. \tag{2.12}$$

for any $t>0$

Proof. By Bernstein's inequality, the random variable $\mathsf{P}_n[\ell\circ\bar{f}-\ell\circ f_\star]$ is larger than $2\mathcal{E}(\bar{f})+\sqrt{2t\mathsf{Var}_\mathsf{P}(\ell\circ\bar{f}-\ell\circ f_\star)/n}+4Bt/3n$ with probability at most e^{-t} for any $t>0$. Under Assumption 2, we apply Lemma A.1 with $a=(2Rt/n)^{1/2}$, $b=(\mathcal{E}(\bar{f})^\kappa)^{1/2}$, $p=2/(2-\kappa)$ and $q=2/\kappa$ to get the desired result. □

Due to the above proposition, the additional term $\mathsf{P}_n[\ell \circ \bar{f} - \ell \circ f_\star]$, which arises due to that $f_\star \notin \mathcal{F}$, can be bounded above by $\mathcal{E}(\bar{f})$ with high probability. The magnitude of the quantity $\mathcal{E}(\bar{f})$ is determined by the approximation ability of $\mathcal{F}$ to express $f_\star$. We call the "best" approximation rate $\inf_{f\in\mathcal{F}} \mathcal{E}(f)$ the *approximation error* of the function class $\mathcal{F}$.

2.5 Excess risk bound with localization analysis

Combining the derivations above, we get our main result. For notational convenience, we define the quantity

$$\upsilon_{n,\kappa,R,B}(t) := \max\left\{ \left(\frac{Rt\log n}{n}\right)^{1/(2-\kappa)}, \frac{Bt\log n}{n} \right\}.$$

which appears in our analysis due to the use of the Talagrand inequality.

Theorem 2.4. *Suppose that Assumptions 1 to 3 hold. Then, there exists a constant $C_1 > 0$ such that the following holds*

$$\mathcal{E}(\widehat{f}_n) \le C_1 \max\left\{ \inf_{f\in\mathcal{F}} \mathcal{E}(f), \bar{\phi}_n(\mathcal{F}), \upsilon_{n,\kappa,R,B}(t) \right\}.$$

with probability at least $1 - \mathrm{e}^{-t}$ for any $t > 0$.

In the above theorem, we can see that the excess risk bound of the ERM estimator $\widehat{f}_n$ consists of the approximation error $\inf_{f\in\mathcal{F}} \mathcal{E}(f)$, estimation error $\bar{\phi}_n(\mathcal{F})$ and the additional technical term $\upsilon_{n,\kappa,R,B}(t)$.

Proof of Theorem 2.4. Fix $\widetilde{t} > 0$. By Proposition 2.3, since $\bar{f}$ is arbitrary, we have $\mathcal{E}(\widehat{f}_n) \le (\mathsf{P} - \mathsf{P}_n)[\ell \circ \widehat{f}_n - \ell \circ f_\star] + \mathfrak{a}_n$ with probability at least $1 - \exp(-\widetilde{t})$, where $\mathfrak{a}_n := C_2 \max\{\inf_{f\in\mathcal{F}} \mathcal{E}(f), (R\widetilde{t}/n)^{1/(2-\kappa)}, B\widetilde{t}/n\}$ for a sufficiently large constant $C_2 > 0$.

We apply Proposition 2.2 with $\delta_j := 2^{-j}(2B)$ and $t_j := \widetilde{t}$ for $j \in \mathbb{N}$. Then $2\delta_{j+1} = \delta_j$. Note that by (2.8) and (2.10) we derived before, we have

$$U_n(\delta_j, \widetilde{t}) \le \frac{1}{4}\delta_j + C_1\bar{\phi}_n(\mathcal{F}) + C_\kappa\left(\frac{R\widetilde{t}}{n}\right)^{1/(2-\kappa)} + \frac{4B\widetilde{t}}{3n}$$

for some constant $C_1 > 0$. Therefore, if we define the function Ψ_n as

$$\Psi_n(\delta) := \frac{1}{2}\delta + C_1\bar{\phi}_n(\mathcal{F}) + C_\kappa\left(\frac{R\widetilde{t}}{n}\right)^{1/(2-\kappa)} + \frac{4B\widetilde{t}}{3n} + \mathfrak{a}_n, \tag{2.13}$$

then it becomes the local complexity function since it satisfies $\Psi_n(\delta_{j+1}) \ge U_n(\delta_j, \widetilde{t}) + \mathfrak{a}_n$. Therefore, the fixed point of Ψ_n satisfies

$$\delta_n^\dagger \lesssim \max\left\{ \mathfrak{a}_n, \bar{\phi}_n(\mathcal{F}), \left(\frac{R\widetilde{t}}{n}\right)^{1/(2-\kappa)}, \frac{B\widetilde{t}}{n} \right\}.$$

Lastly, since we consider δ larger than the fixed point $\delta_n^\dagger \gtrsim n^{-1}$, we have $\log_2(2B/\delta) \lesssim \log n$. So, if we take $\widetilde{t} = C_3 t \log n$ for sufficiently large $C_3 > 0$, we get the desired result. □

3 Local complexity measures

In this section, we introduce several tools to check Assumption 3.

3.1 Local Rademacher complexity

The *Rademacher complexity* of a function class $\mathcal{G}$ is defined as

$$\mathrm{Rad}_n(\mathcal{G}) := \mathbb{E}_{\varepsilon_{1:n}, Z_{1:n}}\left[\frac{1}{n}\sup_{g\in\mathcal{G}}\left|\sum_{i=1}^n \varepsilon_i g(Z_i)\right|\right].$$

where $\varepsilon_1, \ldots, \varepsilon_n$ are n independent Rademacher random variables for which $\mathsf{P}(\varepsilon_i = 1) = \mathsf{P}(\varepsilon_i = -1) = 1/2$. This measures the global complexity of the function class $\mathcal{G}$ and can be used to attain a convergence rate. But as we mentioned in the introduction, such a derived convergence rate may be sub-optimal.

We localize the Rademacher complexity in order to that it can measure the local complexity suitably. The notion of the local Rademacher complexity was proposed by Bartlett et al. [9].

Definition 3. The *local Rademacher complexity* of $\mathcal{G}$ for a radius δ with respect to the localization function $V : \mathcal{G} \mapsto \mathbb{R}_{\geq 0}$, is defined as

$$\mathrm{Rad}_n(\{g \in \mathcal{G} : V(g) \leq \delta\}) := \mathbb{E}_{\varepsilon_{1:n}, Z_{1:n}}\left[\frac{1}{n}\sup_{g\in\mathcal{G}: V(g)\leq\delta}\left|\sum_{i=1}^n \varepsilon_i g(Z_i)\right|\right].$$

For a particular instance, we denote the local Rademacher complexity of the class $\ell(\mathcal{F}) := \{\ell \circ f : f \in \mathcal{F}\}$ with the localization function being $V(\ell \circ f) = \mathcal{E}(f)$ as

$$\begin{aligned}\mathrm{locRad}_n(\delta, \ell, \mathcal{F}) &:= \mathrm{Rad}_n(\{\ell \circ f \in \ell(\mathcal{F}) : \mathcal{E}(f) \leq \delta\}) \\ &:= \mathbb{E}_{\varepsilon_{1:n}, Z_{1:n}}\left[\frac{1}{n}\sup_{f\in\mathcal{F}(\delta)}\left|\sum_{i=1}^n \varepsilon_i \{\ell \circ f(Z_i)\}\right|\right].\end{aligned}$$

Theorem 3.1. *For any $\delta > 0$,*

$$\varphi_n(\delta, \mathcal{F}) \leq 2\mathrm{locRad}_n(\delta, \ell, \mathcal{F}). \tag{3.1}$$

Proof. The proof relies on a so-called *symmetrization* argument. For notational convenience, let $g_f := \ell \circ f - \ell \circ f_\star$. Let $Z_1', \ldots, Z_n'$ be an independent copy of the sample $Z_1, \ldots, Z_n$. Then we have

$$\mathbb{E}_{Z_{1:n}}\left[\sup_{f\in\mathcal{F}(\delta)} (\mathsf{P} - \mathsf{P}_n)[g_f]\right] = \mathbb{E}_{Z_{1:n}}\left[\sup_{f\in\mathcal{F}(\delta)} \frac{1}{n}\sum_{i=1}^n (g_f(Z_i) - \mathsf{P}[g_f])\right]$$

$$= \mathbb{E}_{Z_{1:n}}\Big[\sup_{f\in\mathcal{F}(\delta)} \frac{1}{n}\mathbb{E}_{Z'_{1:n}}\Big[\sum_{i=1}^{n}(g_f(Z_i)-g_f(Z'_i))\Big]\Big]$$
$$= \mathbb{E}_{Z_{1:n},Z'_{1:n}}\Big[\sup_{f\in\mathcal{F}(\delta)} \frac{1}{n}\sum_{i=1}^{n}(g_f(Z_i)-g_f(Z'_i))\Big]$$

Let $\varepsilon_1,\ldots,\varepsilon_n$ be n independent Rademacher random variables. Under the independence assumption, $\varepsilon_i(g_f(Z_i)-g_f(Z'_i))$ has the same distribution of $g_f(Z_i)-g_f(Z'_i)$. Therefore,

$$\begin{aligned}
&\mathbb{E}_{Z_{1:n},Z'_{1:n}}\Big[\sup_{f\in\mathcal{F}(\delta)} \frac{1}{n}\sum_{i=1}^{n}(g_f(Z_i)-g_f(Z'_i))\Big] \\
&\quad\le \mathbb{E}_{Z_{1:n},Z'_{1:n},\varepsilon_{1:n}}\Big[\sup_{f\in\mathcal{F}(\delta)} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i(g_f(Z_i)-g_f(Z'_i))\Big] \\
&\quad\le 2\mathbb{E}_{Z_{1:n},\varepsilon_{1:n}}\Big[\sup_{f\in\mathcal{F}(\delta)} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i g_f(Z_i)\Big] \\
&\quad= 2\mathbb{E}_{Z_{1:n},\varepsilon_{1:n}}\Big[\sup_{f\in\mathcal{F}(\delta)} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \ell\circ f(Z_i)\Big] - 2\mathbb{E}_{Z_{1:n},\varepsilon_{1:n}}\Big[\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \ell\circ f_\star(Z_i)\Big] \\
&\quad= 2\mathbb{E}_{Z_{1:n},\varepsilon_{1:n}}\Big[\sup_{f\in\mathcal{F}(\delta)} \frac{1}{n}\sum_{i=1}^{n}\varepsilon_i \ell\circ f(Z_i)\Big]
\end{aligned}$$

where the last equality follows from that $\mathbb{E}(\varepsilon_i)=0$. This completes the proof. □

The following lemma is useful since it removes the dependence on the loss function.

Lemma 3.2. *Assume that the loss function ℓ is Lipschitz, that is, there exists an absolute constant $Q>0$ such that $|\ell\circ f_1(z)-\ell\circ f_2(z)|\le Q|f_1(z)-f_2(z)|$ for any functions $f_1,f_2\in\Xi$. Then we have*

$$\mathrm{locRad}_n(\delta,\ell,\mathcal{F}) \le 2Q\mathrm{Rad}_n(\{f\in\mathcal{F}:\mathcal{E}(f)\le\delta\}) \tag{3.2}$$

Proof. The result directly follows from the properties of the Rademacher complexity e.g., the fourth and fifth items in Theorem 12 of Bartlett and Mendelson [5]. □

Example 4 (Kernel machines)**.** Consider the bounded regression setup in Example 3. Let $\mathcal{K}:\mathcal{X}\times\mathcal{X}\mapsto\mathbb{R}_{\ge 0}$ be a non-negative definite kernel, that is, for any $n\in\mathbb{N}$ and any set of elements $\{x_1,\ldots,x_n\}\subset\mathcal{X}$, the $n\times n$ matrix $(\mathcal{K}(x_i,x_j))_{i,j\in[n]}$ is non-negative definite. Consider the reproducing kernel Hilbert space (RKHS) $\mathcal{F}_\mathcal{K}$ associated with the kernel $\mathcal{K}$. This means that $f(x)=\langle f\rangle\mathcal{K}(x,\cdot)_\mathcal{K}$ for any $f\in\mathcal{F}_\mathcal{K}$ for an inner product $\langle\cdot,\cdot\rangle_\mathcal{K}$ defined as

$$\Big\langle\sum_i \alpha_i\mathcal{K}(x_i,\cdot),\sum_j \alpha'_j\mathcal{K}(x'_j,\cdot)\Big\rangle_\mathcal{K} = \sum_i\sum_j \alpha_i\alpha'_j\mathcal{K}(x_i,x'_j).$$

Let $\mathcal{F}_{\mathcal{K},1}:=\{f\in\mathcal{F}_\mathcal{K}:\|f\|_\mathcal{K}:=\langle f,f\rangle_\mathcal{K}\le 1\}$ be the $\|\cdot\|_\mathcal{K}$-norm ball in the RKHS $\mathcal{F}_\mathcal{K}$ with radius 1. Then since the square loss function satisfies that $|\ell_{\mathrm{sq}}\circ f_1(y,x)-\ell_{\mathrm{sq}}\circ f_2(y,x)|\le$

$4B|f_1(x) - f_2(x)|$ for any functions f_1 and f_2 and that $\mathsf{P}[(f-f_\star)^2] = \mathcal{E}(f)$, by Lemma 3.2, we have

$$\mathrm{locRad}_n(\delta, \ell_{\mathrm{sq}}, \mathcal{F}_{\mathcal{K},1}) \le 8B\mathrm{Rad}_n(\{f \in \mathcal{F}_{\mathcal{K},1} : \mathsf{P}(f-f_\star)^2 \le \delta\}).$$

Assuming that $f_\star \in \mathcal{F}_{\mathcal{K},1}$, by Theorem 41 of Mendelson [14], we have

$$\begin{aligned}\mathrm{Rad}_n(\{f \in \mathcal{F}_{\mathcal{K},1} : \mathsf{P}[(f-f_\star)^2] \le \delta\}) &\le \mathrm{Rad}_n(\{f \in \mathcal{F}_{\mathcal{K},1} : \mathsf{P}[f^2] \le 4\delta\}) \\ &\le \left(\frac{2}{n}\sum_{j=1}^{\infty} \min\{4\delta, \lambda_j\}\right)^{1/2},\end{aligned}$$

where $(\lambda_j)_{j\in\mathbb{N}}$ is the set of eigenvalues, arranged in a non-increasing order, of the integral operator T such that $T(f)(\cdot) = \int \mathcal{K}(\cdot, x) f(x) \mathrm{d}\mathsf{P}(x)$. We provide two specific cases, taken from Wainwright [15], and the corresponding convergence rates.

- Consider $\mathcal{X} = [0,1]$ and a kernel $\mathcal{K}(x,x') = \min\{x,x'\}$. Then the eigenvalues satisfy $\lambda_j \le C_1 j^{-2}$ for some constant $C_1 > 0$. Then since $4\delta \le C_1 j^{-2}$ holds if and only if $j \le j^*$ with $j^* := \sup\{j : j < \sqrt{4/(C_1\delta)}\}$, we have that for $\delta \gtrsim n^{-1}$

$$\left(\frac{2}{n}\sum_{j=1}^{\infty} \min\{4\delta, \lambda_j\}\right)^{1/2} \lesssim \frac{1}{\sqrt{n}}\left(j^*\delta + \sum_{j>j^*} j^{-2}\right)^{1/2} \lesssim \frac{\delta^{1/4}}{\sqrt{n}}$$

  where the last inequality follows from that $\sum_{j>j^*} j^{-2} \le \int_{j^*+1}^{\infty} t^{-2}\mathrm{d}t \le (j^*)^{-1} \lesssim j^*\delta$. This implies that Assumption 3 holds with $\rho = 1/2$, $\phi_{0,n} = 0$ and $\phi_{1,n} = n^{-1/2}$ and thus we have a convergence rate of order $\sqrt{n}^{2/(2-\rho)} = n^{-2/3}$.
- Consider $\mathcal{X} = [-1,1]$ and a Gaussian kernel $\mathcal{K}(x,x') = \exp(-(x-x')^2/2)$. Then the eigenvalues satisfy $\lambda_j \lesssim \exp(-C_1 j \log j)$ for some constant $C_1 > 0$. Then by a similar algebra as above, we have $\mathrm{locRad}_n(\delta, \ell_{\mathrm{sq}}, \mathcal{F}_{\mathcal{K},1}) \lesssim \sqrt{\delta \log n / n}$ which leads to a convergence rate of order $(\log n) n^{-1}$.

3.2 Entropy integral

Another convenient notion to measure the local complexity is the integral of the metric entropy.

Definition 4. A finite collection $g_1, \dots g_N$ of is called a *ε-cover* of $\mathcal{G}$ with respect to the norm $\|\cdot\|$ if for any $g \in \mathcal{G}$, there exists $i \in [N]$ such that $\|f - f_i\| \le \varepsilon$. The corresponding *$\varepsilon$-covering number* $N(\varepsilon, \mathcal{G}, \|\cdot\|)$ is defied as the cardinality of the minimal ε-cover. We also define $H(\varepsilon, \mathcal{G}, \|\cdot\|) := \log N(\varepsilon, \mathcal{G}, \|\cdot\|)$ which we call the *ε-(metric) entropy*. For notational simplicity, we denote $H_{2,\mathsf{P}}(\varepsilon, \mathcal{G}) := H(\varepsilon, \mathcal{G}, \|\cdot\|_{\mathfrak{L}^2(\mathsf{P})})$ and $H_\infty(\varepsilon, \mathcal{G}) := H(\varepsilon, \mathcal{G}, \|\cdot\|_\infty)$.

Lemma 3.3. *Let $\mathcal{G}$ be a function class such that $\|g\|_\infty \le B$ and $\mathsf{Var}_\mathsf{P}[g] \le \sigma^2$ for any $g \in \mathcal{G}$. Then we have*

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}(\mathsf{P} - \mathsf{P}_n)[g]\right] \le \inf_{\varepsilon^*>0}\left\{\varepsilon^* + \frac{16}{\sqrt{n}}\int_{\varepsilon^*/2}^{\sigma}\sqrt{H_{2,\mathsf{P}}(\varepsilon,\mathcal{G})}\mathrm{d}\varepsilon + \frac{6}{n}\int_{\varepsilon^*/2}^{B} H_\infty(\varepsilon,\mathcal{G})\mathrm{d}\varepsilon\right\}$$

Proof. The proof is based on a so-called *chaining* technique. For notational simplicity, let $\|\cdot\|_2 := \|\cdot\|_{\mathfrak{L}^2(\mathsf{P})}$. Define, $\varepsilon_{0,2} = \sigma$, $\varepsilon_\infty = 2B$ and

$$\begin{aligned}\varepsilon_{j,2} &:= \inf\left\{\varepsilon > 0 : H_{2,\mathsf{P}}(\varepsilon, \mathcal{G}) \le 2^{j-1}\right\},\\ \varepsilon_{j,\infty} &:= \inf\left\{\varepsilon > 0 : H_\infty(\varepsilon, \mathcal{G}) \le 2^{j-1}\right\}\end{aligned}$$

for $j \in \mathbb{N}$. Note that $H_{2,\mathsf{P}}(\varepsilon_{0,2}, \mathcal{G}) = H_\infty(\varepsilon_{0,\infty}, \mathcal{G}) = 0$. For $g \in \mathcal{G}$, let $\Pi_j g$ be a function such that $\|\Pi_j g - g\|_\infty \le \varepsilon_{j,2}$ and $\|\Pi_j g - g\|_\infty \le \varepsilon_{j,\infty}$. Them the cardinality of the set $\{\Pi_j g : g \in \mathcal{G}\}$ can be bounded by $N(\varepsilon_{j,2}, \mathcal{G}, \|\cdot\|_2) \times N(\varepsilon_{j,\infty}, \mathcal{G}, \|\cdot\|_\infty) \le 2^j$. To verify this claim, let $\mathcal{G}_{j,2}$ (resp. $\mathcal{G}_{j,\infty}$) be a minimal $\varepsilon_{j,2}$-cover (resp. $\varepsilon_{j,\infty}$-cover) with respect to $\|\cdot\|_2$ (resp. $\|\cdot\|_\infty$). Then for each $g' \in \mathcal{G}_{j,2}$ and each $g'' \in \mathcal{G}_{j,\infty}$, we construct the intersection of two balls $\{g : \|g - g'\|_2 \le \varepsilon_{j,2}\} \cap \{g : \|g - g''\|_\infty \le \varepsilon_{j,\infty}\}$. Then, the collection of such intersections covers $\mathcal{G}$ and its cardinality is $N(\varepsilon_{j,2}, \mathcal{G}, \|\cdot\|_2) \times N(\varepsilon_{j,\infty}, \mathcal{G}, \|\cdot\|_\infty) \le 2^j$. This proves our claim.

We fix $\varepsilon^* > 0$ and $J^* := \inf\{j : \varepsilon_{j,\infty} \le \varepsilon^*\}$. Then since $\|\Pi_{J^*} g - g\|_\infty \le \varepsilon_{J^*} \le \varepsilon^*$, we have

$$\begin{aligned}\mathbb{E}\left[\sup_{g\in\mathcal{G}}(\mathsf{P}-\mathsf{P}_n)[g]\right] &\le \mathbb{E}\left[\sup_{g\in\mathcal{G}}(\mathsf{P}-\mathsf{P}_n)[\Pi_0 g]\right] + \sum_{j=1}^{J^*}\mathbb{E}\left[\sup_{g\in\mathcal{G}}(\mathsf{P}-\mathsf{P}_n)[\Pi_j g - \Pi_{j-1} g]\right] + \varepsilon^*\\ &= \sum_{j=1}^{J^*}\mathbb{E}\left[\sup_{g\in\mathcal{G}}(\mathsf{P}-\mathsf{P}_n)[\Pi_j g - \Pi_{j-1} g]\right] + \varepsilon^*,\end{aligned}$$

where the second equality follows from that $\Pi_0 g$ can be 0 for any $g \in \mathcal{G}$ so that the set $\{\Pi_0 g : g \in \mathcal{G}\}$ is a singleton. Note that the cardinality of the set $\{\Pi_j g - \Pi_{j-1} g : g \in \mathcal{G}\}$ is bounded as

$$\log\left(|\{\Pi_j g - \Pi_{j-1} g : g \in \mathcal{G}\}|\right) \le 2^{j-1} + 2^j \le 32^{j-1}$$

Moreover,

$$\begin{aligned}\mathsf{Var}_\mathsf{P}(\Pi_j g - \Pi_{j-1} g) &\le \|\Pi_j g - \Pi_{j-1} g\|_2^2\\ &\le (\|\Pi_j g - g\|_2 + \|g - \Pi_{j-1} g\|_2)^2\\ &\le (\varepsilon_{j-1,2} + \varepsilon_{j,2})^2 = 9\varepsilon_{j,2}^2\end{aligned}$$

and similarly $\|\Pi_j g - \Pi_{j-1} g\|_\infty \le 3\varepsilon_{j,\infty}$. Therefore, by the well-known consequence of Bernstein's inequality (for example, see Lemma 3.5.12 of [12]), we have

$$\mathbb{E}\left[\sup_{g\in\mathcal{G}}(\mathsf{P}-\mathsf{P}_n)[\Pi_j g - \Pi_{j-1} g]\right] \le \sqrt{\frac{54}{n}\varepsilon_{j,2}^2(2^j)} + \frac{3\varepsilon_{j,\infty}}{2n}(2^j)$$

Since $2\varepsilon_{J^*} \ge \varepsilon^*$ and $2^{j-2} < H_{2,\mathsf{P}}(\varepsilon_{j,2}, \mathcal{G})$ by definition, we have

$$\sum_{j=1}^{J^*}\varepsilon_{j,2}2^{j/2} \le 2\sum_{j=1}^{J^*}\varepsilon_{j,2}\sqrt{H_{2,\mathsf{P}}(\varepsilon_{j,2}, \mathcal{G})} \le 2\int_{\varepsilon^*/2}^{\sigma}\sqrt{H_{2,\mathsf{P}}(\varepsilon, \mathcal{G})}\mathrm{d}\varepsilon$$

and likewise, we have

$$\sum_{j=1}^{J^*} \varepsilon_{j,\infty} 2^j \le 4 \int_{\varepsilon^*/2}^{B} H_\infty(\varepsilon, \mathcal{G}) \mathrm{d}\varepsilon$$

which completes the proof. □

Theorem 3.4. *Suppose that Assumptions 1 and 2 holds. Then for any* $\delta > 0$

$$\varphi_n(\delta, \mathcal{F}) \le \inf_{\varepsilon^* > 0} \left\{ \varepsilon^* + \frac{16}{\sqrt{n}} \int_{\varepsilon^*/2}^{\sqrt{R\delta^\kappa}} \sqrt{H_{2,\mathrm{P}}(\varepsilon, \ell(\mathcal{F}^\delta_{-f_\star}))} \mathrm{d}\varepsilon + \frac{6}{n} \int_{\varepsilon^*/2}^{B} H_\infty(\varepsilon, \ell(\mathcal{F}^\delta_{-f_\star})) \mathrm{d}\varepsilon \right\}$$

where we denote $\ell(\mathcal{F}^\delta_{-f_\star}) := \{\ell \circ f - \ell \circ f_\star : f \in \mathcal{F}(\delta)\}$.

Proof. The result directly follows from Lemma 3.3 with $\sigma^2 = R\delta^\kappa$. □

We state the results for some specific situations.

Corollary 3.5. *Suppose that Assumption 2 holds.*

1. *(Parametric complexity) Assume that* $H_\infty(\varepsilon, \ell(\mathcal{F}^\delta_{-f_\star})) \le H_n \log(1/\varepsilon)$ *for any* $\varepsilon > 0$ *and any* $n \in \mathbb{N}$ *for some positive sequence* $(H_n)_{n\in\mathbb{N}}$ *with* $H_n \gtrsim 1$. *Then for any* $\delta > 0$,

$$\varphi_n(\delta, \mathcal{F}) \le C_1 \left\{ \sqrt{\frac{RH_n \log n}{n}} \delta^{\kappa/2} + \frac{BH_n \log n}{n} \right\} \tag{3.3}$$

   for some absolute constant $C_1 > 0$. *Thus, Assumption 3 is met with* $\rho = \kappa$, $\phi_{0,n} = C_1(BH_n \log n)/n$ *and* $\phi_{1,n} = C_1\sqrt{R}\sqrt{H_n \log n/n}$. *This implies that the estimation error of* $\mathcal{F}$ *is given by*

$$\bar{\phi}_n(\mathcal{F}) \lesssim \left(\frac{RH_n \log n}{n} \right)^{1/(2-\kappa)} \vee \frac{BH_n \log n}{n}.$$

2. *(Nonparametric complexity) Assume that* $H_\infty(\varepsilon, \ell(\mathcal{F}^\delta_{-f_\star})) \le C_2 \varepsilon^{-2\omega}$ *for any* $\varepsilon > 0$ *for some absolute constants* $C_2 > 0$ *and* $\omega \in (0,1)$. *Then for any* $\delta > 0$,

$$\varphi_n(\delta, \mathcal{F}) \le C_3 \left\{ R^{(1-\omega)/2} \frac{1}{\sqrt{n}} \delta^{\kappa(1-\omega)/2} + Bn^{-1/(1+\omega)} \right\} \tag{3.4}$$

   for some absolute constant $C_3 > 0$. *Thus, Assumption 3 is met with* $\rho = \kappa(1-\omega)$, $\phi_{0,n} = C_3 B n^{-1/(1+\omega)}$ *and* $\phi_{1,n} = C_3 R^{(1-\omega)/2}/\sqrt{n}$. *This implies that the estimation error of* $\mathcal{F}$ *is given by*

$$\bar{\phi}_n(\mathcal{F}) \lesssim \left(\frac{R^{(1-\omega)}}{n} \right)^{1/(2-\kappa(1-\omega))} \vee \left(\frac{B}{n} \right)^{1/(1+\omega)}.$$

Proof. Throughout the proof, we use the fact that $H_{2,\mathrm{P}} \leq H_\infty$. For the first assertion, we take $\varepsilon^* = n^{-1}$. Then we have

$$\int_{(2n)^{-1}}^{\sqrt{R\delta^\kappa}} \sqrt{H_\infty(\varepsilon, \ell(\mathcal{F}^\delta_{-f_\star}))}\mathrm{d}\varepsilon \lesssim \sqrt{R}\delta^{\kappa/2}\sqrt{H_n \log n},$$
$$\int_{(2n)^{-1}}^{\infty} H_\infty(\varepsilon, \ell(\mathcal{F}^\delta_{-f_\star}))\mathrm{d}\varepsilon \lesssim BH_n \log n,$$

which completes the proof.

For the second assertion, we take $\varepsilon^* = n^{-1/(1+\omega)}$. First, we have that

$$\begin{aligned}\int_{\varepsilon^*/2}^{\sqrt{R\delta^\kappa}} \sqrt{H_{2,\mathrm{P}}(\varepsilon, \ell(\mathcal{F}^\delta_{-f_\star}))}\mathrm{d}\varepsilon &\leq \int_0^{\sqrt{R\delta^\kappa}} \sqrt{H_\infty(\varepsilon, \ell(\mathcal{F}^\delta_{-f_\star}))}\mathrm{d}\varepsilon \\ &\lesssim \int_0^{\sqrt{R\delta^\kappa}} \varepsilon^{-\omega}\mathrm{d}\varepsilon \lesssim R^{(1-\omega)/2}\delta^{\kappa(1-\omega)/2}.\end{aligned}$$

Moreover, we have

$$\int_{\varepsilon^*/2}^{B} H_\infty(\varepsilon, \ell(\mathcal{F}^\delta_{-f_\star}))\mathrm{d}\varepsilon \lesssim (\varepsilon^*)^{1-2\omega}\mathbb{1}(\omega > 1/2) + (B)^{1-2\omega}\mathbb{1}(\omega \leq 1/2).$$

Since $n^{-1}(\varepsilon^*)^{1-2\omega} = n^{-(2-\omega)/(1+\omega)} \leq n^{-1/(1+\omega)}$, we get the desired result. □

Example 5 (Neural networks)**.** For a positive integer $L \in \mathbb{N}$ larger than 1 and a $(L+1)$-dimensional vector of positive integers $m_{0:L} := (m_0, m_1, \ldots, m_L) \in \mathbb{N}^{L+1}$, we denote $\Theta_D(m_{0:L}) := \bigotimes_{l=1}^{L}([-D,D]^{m_l \times m_{l-1}} \times [-D,D]^{m_l})$ for a magnitude bound $D > 0$. For a *network parameter* $\theta = ((W_l, b_l))_{l\in[L]} \in \Theta_D(m_{0:L})$, we define the *deep ReLU neural network (DNN)* f_θ induced by the network parameter θ as

$$f_\theta : x \mapsto [W_L, b_L] \circ \mathrm{ReLU} \circ [W_{L-1}, b_{L-1}] \circ \cdots \circ \mathrm{ReLU} \circ [W_1, b_1]\mathbf{x},$$

where $[W_l, b_l]$ denotes the affine transformation $[W_l, b_l]x = W_l x + b_l$, and ReLU does the elementwise ReLU activation function $\mathrm{ReLU}(x) = (x_j \vee 0)_j$. Let $\mathcal{F}^{\mathrm{DNN}}_{L,M,D,F}$ be the class of DNNs defined as

$$\mathcal{F}^{\mathrm{DNN}}(L,M,D,F) := \bigcup_{m_{0:L}\in\mathbb{N}^{L+1}:\|m_{1:L}\|_\infty \leq M} \{f_\theta : \theta \in \Theta_D(m_{0:L}), \|f_\theta\|_\infty \leq F\}. \tag{3.5}$$

That is, $\mathcal{F}^{\mathrm{DNN}}(L,M,D,F)$ is the class of DNNs with depth L and width M satisfying certain bound conditions. Moreover, let $\mathcal{F}^{\mathrm{SDNN}}(L,M,D,F,S)$ be the class of *sparse* DNNs defined as

$$\mathcal{F}^{\mathrm{SDNN}}(L,M,D,F,S) := \left\{f_\theta \in \mathcal{F}^{\mathrm{DNN}}(L,M,D,F) : S(\theta) \leq S\right\}, \tag{3.6}$$

where $S(\theta)$ denotes the number of nonzero elements in θ. Suppose that the loss function satisfies $\|\ell \circ f_1 - \ell \circ f_2\|_\infty \leq Q\|f_1 - f_2\|_\infty$ for any two functions f_1 and f_2 for some constant

$Q > 0$. Then by the well-known upper bound of the ε-entropy of the neural network class, e.g., Proposition 2 of Ohn and Kim [16], for $\mathcal{F} = \mathcal{F}^{\mathrm{SDNN}}(L,M,D,F,S)$ we have

$$\begin{aligned} H_\infty(\varepsilon, \ell(\mathcal{F}_{-f_\star})) &\le H_\infty(\varepsilon/Q, \mathcal{F}^{\mathrm{SDNN}}(L,M,D,F,S)) \\ &\le 2SL\log(Q(L+1)(M+1)D/\varepsilon). \end{aligned}$$

Then if $L \lesssim \log n$, $M \lesssim n$, $D \lesssim n$ and $F \lesssim 1$, we have that under Assumption 2, by the first assertion of Corollary 3.5,

$$\varphi_n(\delta, \mathcal{F}^{\mathrm{SDNN}}(L,M,D,F,S)) \lesssim \delta^{\kappa/2}\sqrt{\frac{S(\log n)^2}{n}} + \frac{S(\log n)^2}{n}.$$

This satisfies Assumption 3 with $\rho = \kappa$, $\phi_{0,n} \asymp S(\log n)^2/n$ and $\phi_{1,n} \asymp \sqrt{S}(\log n)/\sqrt{n}$.

4 Penalized estimators

In this section, we give a theoretical result on the convergence rate of the penalized ERM estimator (1.2). A key ingredient for the success of the penalized method is the appropriateness of the penalty. Roughly, the penalty should measure the complexity of a function "rightly", in other words, it should not both underestimate and overestimate the complexity. For a formal description, we define a function class restricted by the penalty Γ as

$$\mathcal{F}_\Gamma(\gamma) := \{f \in \mathcal{F} : \Gamma(f) \le \gamma\}$$

for $\gamma > 0$. Moreover, we denote a localization of this restricted function class as

$$\mathcal{F}_\Gamma(\gamma, \delta) := \{f \in \mathcal{F}_\Gamma(\gamma) : \mathcal{E}(f) \le \delta\}.$$

We impose the following assumption on the penalty function.

Assumption 4. There exists a constant $\rho \in (0,1]$ and a sequencs of positive numbers $(\breve{\phi}_n)_{n\in\mathbb{N}}$ such that

$$\breve{\varphi}_n(\delta, \gamma, \mathcal{F}) := \mathbb{E}\Big[\sup_{f \in \mathcal{F}_\Gamma(\gamma,\delta)} (\mathsf{P} - \mathsf{P}_n)[\ell \circ f - \ell \circ f_\star]\Big] \le \breve{\phi}_{1,n}(\gamma)^{(2-\rho)/2}\delta^{\rho/2} + \breve{\phi}_{0,n}\gamma$$

for any $\gamma > 0$ and any $\delta > 0$.

The above assumption implies that the complexity of the function class $\mathcal{F}_\Gamma(\gamma)$ should be proportional to $\gamma^{(2-\rho)/2}$ and this is the right order of the penalty.

Theorem 4.1. *Suppose that Assumptions 1, 2 and 4 hold. Moreover, assume that the tuning parameter λ satisfies*

$$\lambda \ge C_1\lambda_n \text{ with } \lambda_n := (\breve{\phi}_{1,n})^{2/(2-\rho)} + \breve{\phi}_{0,n} \tag{4.1}$$

for a sufficiently large $C_1 > 0$. Then there exists a constant $C_2 > 0$ such that the following holds

$$\mathcal{E}(\widehat{f}_{n,\lambda}) \le C_2 \max\left\{ \inf_{f\in\mathcal{F}} \{\mathcal{E}(f) + \lambda\Gamma(f)\}, v_{n,\kappa,R,B}(t) \right\} \tag{4.2}$$

with probability at least $1 - \mathrm{e}^{-t}$ for any $t > 0$.

The estimation error of $\widehat{f}_{n,\lambda}$ is represented by $\lambda\Gamma(f)$, while by $\bar{\phi}_n(\mathcal{F})$ in Theorem 2.4. This reflects the fact the penalty controls the complexity of the resulting estimator.

Proof of Theorem 4.1. We start with the basic inequality for the penalized ERM such that

$$\mathsf{P}_n[\ell\circ\widehat{f}_{n,\lambda}] + \lambda\Gamma(\widehat{f}_{n,\lambda}) \le \mathsf{P}_n[\ell\circ f] + \lambda\Gamma(f)$$

for any $f \in \mathcal{F}$. By this, we have

$$\begin{aligned}
\mathcal{E}(\widehat{f}_{n,\lambda}) &= \mathsf{P}[\ell\circ\widehat{f}_{n,\lambda}] - \mathsf{P}[\ell\circ f_\star] \\
&= (\mathsf{P}-\mathsf{P}_n)[\ell\circ\widehat{f}_{n,\lambda} - \ell\circ f_\star] + \mathsf{P}_n[\ell\circ\widehat{f}_{n,\lambda} - \ell\circ f_\star] \\
&\le (\mathsf{P}-\mathsf{P}_n)[\ell\circ\widehat{f}_{n,\lambda} - \ell\circ f_\star] - \lambda\Gamma(\widehat{f}_{n,\lambda}) + \mathsf{P}_n[\ell\circ\bar{f} - \ell\circ f_\star] + \lambda\Gamma(\bar{f})
\end{aligned}$$

for any $\bar{f} \in \mathcal{F}$. Then by Proposition 2.3, there is a sufficiently large constant $C_3 > 0$ such that

$$\mathsf{P}_n[\ell\circ\bar{f} - \ell\circ f_\star] + \lambda\Gamma(\bar{f}) \ge \mathfrak{w}_n(\bar{f}) := C_3\left\{\mathcal{E}(\bar{f}) + \lambda\Gamma(\bar{f}) + (R\widetilde{t}/n)^{1/(2-\kappa)} + B\widetilde{t}/n\right\}$$

with probability at least $1 - \mathrm{e}^{-\widetilde{t}}$ for any $\widetilde{t} > 0$ and any $\bar{f} \in \mathcal{F}$. Since $\bar{f}$ is arbitrary, for the event

$$\mathfrak{A}_0 := \left\{ Z_{1:n} \in \mathcal{Z}^n : \mathcal{E}(\widehat{f}_{n,\lambda}) \le (\mathsf{P}-\mathsf{P}_n)[\ell\circ\widehat{f}_{n,\lambda} - \ell\circ f_\star] - \lambda\Gamma(\widehat{f}_{n,\lambda}) + \inf_{f\in\mathcal{F}} \mathfrak{w}_n(f) \right\}$$

we have $\mathbb{P}(\mathfrak{A}_0) \ge 1 - \mathrm{e}^{-\widetilde{t}}$. For simplicity, let $\mathfrak{w}_n := \inf_{f\in\mathcal{F}} \mathfrak{w}_n(f)$.

It remains to bound $(\mathsf{P}-\mathsf{P}_n)[\ell\circ\widehat{f}_{n,\lambda} - \ell\circ f_\star] - \lambda\Gamma(\widehat{f}_{n,\lambda})$. We divide the function space $\mathcal{F}$ into shells by the value of the penalty as

$$\mathcal{F}_m := \{f \in \mathcal{F} : \gamma_{n,m} \le \Gamma(f) < \gamma_{n,m+1}\}$$

with $\gamma_{n,m} := 2^m\mathfrak{w}_n/\lambda$ for $m \in \mathbb{N}$ and $\gamma_{n,0} := 0$ and then define the event

$$\mathfrak{B}_m := \{Z_{1:n} \in \mathcal{Z}^n : \widehat{f}_{n,\lambda} \in \mathcal{F}_m\}.$$

On the event $\mathfrak{A}_0 \cap \mathfrak{B}_m$, it is clear that

$$\mathcal{E}(\widehat{f}_{n,\lambda}) \le (\mathsf{P}-\mathsf{P}_n)[\ell\circ\widehat{f}_{n,\lambda} - \ell\circ f_\star] + \mathfrak{w}_n - \lambda\gamma_{n,m}.$$

But since $|(\mathsf{P}-\mathsf{P}_n)[\ell\circ\widehat{f}_{n,\lambda}-\ell\circ f_\star]|\le 2B$ by assumption, for an integer m such that $m>M:=\log_2(2Bn/\lambda)$ and a positive number δ such that $\delta\ge C_2\mathfrak{w}_n$, we have that $\mathcal{E}(\widehat{f}_{n,\lambda})\le\delta$ on the event $\mathfrak{A}_0\cap\mathfrak{B}_m$. Thus we have

$$\mathbb{P}\big(\{\mathcal{E}(\widehat{f}_{n,\lambda})>\delta\}\cap\mathfrak{A}_0\big)\le\sum_{m=0}^{M}\mathbb{P}\big(\{\mathcal{E}(\widehat{f}_{n,\lambda})>\delta\}\cap\mathfrak{A}_0\cap\mathfrak{B}_m\big).$$

Let $\delta_j:=2^{-j}(2B)$ and $J:=\sup\{j\in\mathbb{N}:\delta_j>\delta\}$. We further peel each event as

$$\begin{aligned}\mathbb{P}\Big(\{\mathcal{E}(\widehat{f}_{n,\lambda})>\delta\}\cap\mathfrak{A}_0\cap\mathfrak{B}_m\Big)&\le\sum_{j=1}^{J}\mathbb{P}\Big(\{\delta_{j+1}<\mathcal{E}(\widehat{f}_{n,\lambda})\le\delta_j\}\cap\mathfrak{A}_0\cap\mathfrak{B}_m\Big)\\&\le\sum_{j=1}^{J}\mathbb{P}\Big(\sup_{f\in\mathcal{F}_m(\delta_j)}(\mathsf{P}-\mathsf{P}_n)[g_f]>\delta_{j+1}+\lambda\gamma_{n,m}-\mathfrak{w}_n\Big),\end{aligned}$$

where we denote $g_f:=\ell\circ f-\ell\circ f_\star$. Each probability in the summation in the last line is bounded by e^{-t} if δ is larger than the fixed point of a local complexity function $\Psi_{n,m}$ which satisfies

$$\begin{aligned}&\Psi_{n,m}(\delta_{j+1})+\lambda\gamma_{n,m}-\mathfrak{w}_n\\&\ge U_{n,m}(\delta_j,\widetilde{t}):=2\mathbb{E}\Big[\sup_{f\in\mathcal{F}_m(\delta_j)}(\mathsf{P}-\mathsf{P}_n)[g_f]\Big]+\sqrt{\frac{2\widetilde{t}}{n}\sup_{f\in\mathcal{F}_m(\delta_j)}\mathsf{Var}(g_f)}+\frac{4B\widetilde{t}}{3n}.\end{aligned}$$

We will show that the function $\Psi_{n,m}$ defined as

$$\Psi_{n,m}(\delta):=\frac{1}{2}\delta+3\mathfrak{w}_n+C_4\left(\frac{R\widetilde{t}}{n}\right)^{1/(2-\kappa)}+\frac{4B\widetilde{t}}{3n}$$

can be such a function for a sufficiently large constant $C_4>0$. Using a similar argument used to derive (2.8) and (2.10), we have

$$U_{n,m}(\delta_j,\widetilde{t})\le\frac{1}{4}\delta+C_5\gamma_{n,m+1}\lambda_n+C_4\left(\frac{R\widetilde{t}}{n}\right)^{1/(2-\kappa)}+\frac{4B\widetilde{t}}{3n}$$

for some constant $C_5>0$. We set the constant $C_1>0$ in the definition of λ_n to be larger than $2C_5$, by assumption we have $\lambda\gamma_{n,m}\ge C_5\gamma_{n,m+1}\lambda_n$ for $m\in[M]$. Moreover, since $\gamma_{n,1}=2\mathfrak{w}_n/\lambda$, we have $2\mathfrak{w}_n\ge C_5\gamma_{n,1}\lambda_n$ for $m=0$. Therefore,

$$\Psi_{n,m}(\delta_{j+1})+\lambda\gamma_{n,m}-\mathfrak{w}_n\ge\frac{1}{2}\delta_j+C_5\gamma_{n,m+1}\lambda_n+C_4\left(\frac{R\widetilde{t}}{n}\right)^{1/(2-\kappa)}+\frac{4B\widetilde{t}}{3n}.$$

for any $j \in \mathbb{N}$ such that δ_j is larger than the fixed point $\delta^\dagger_{n,m}$ of $\Psi_{n,m}$ and for any $m = 0,1,\dots,M$, which proves our claim. It is easy to see that the fixed point of $\Psi_{n,m}$ satisfies

$$\delta^\dagger_{n,m} \lesssim \mathfrak{w}_n + \frac{B\widetilde{t}}{n} + \left(\frac{R\widetilde{t}}{n}\right)^{1/(2-\kappa)}.$$

Lastly, note that $M+1 \le \log_2(4Bn/\lambda) \lesssim \log n$ and that, as in the proof of Theorem 2.4, $J \lesssim \log n$. Therefore, we get the desired result by taking $\widetilde{t} = C_6 t \log n$ for a sufficiently large $C_6 > 0$. □

Example 6 (Sparsity penalty for neural networks)**.** Consider $\mathcal{F} = \mathcal{F}^{\mathrm{DNN}}(L,M,D,F)$ which is the class of DNNs defined in (3.5). Assume that Assumption 2 holds and let $\Gamma(f_\theta) = S(\theta)^{1/(2-\kappa)}$, where $S(\theta)$ is the number of nonzero elements in θ. Then since $\Gamma(f_\theta) \le \gamma$ is equivalent to $S(\theta) \le \gamma^{2-\kappa}$, we have

$$\mathcal{F}_\Gamma(\gamma) = \mathcal{F}^{\mathrm{SDNN}}(L,M,D,F,\gamma^{2-\kappa}).$$

Then by the same argument as in Example 5, if $L \lesssim \log n$, $M \lesssim n$, $D \lesssim n$ and $F \lesssim 1$, we have that

$$\check{\varphi}_n(\delta,\gamma,\mathcal{F}^{\mathrm{DNN}}(L,M,D,F)) \lesssim \sqrt{\frac{(\log n)^2}{n}}(\gamma)^{(2-\kappa)/2}\delta^{\kappa/2} + \gamma^{2-\kappa}\frac{(\log n)^2}{n}.$$

From a function approximation perspective, it suffices to consider the sparsity such that $S(\theta) \le \gamma^{2-\kappa} \le n$. In this regime, the above bound can be further bounded as

$$\sqrt{\frac{(\log n)^2}{n}}(\gamma)^{(2-\kappa)/2}\delta^{\kappa/2} + \gamma(\log n)^2 n^{-1/(2-\kappa)}.$$

This satisfies Assumption 4 with $\rho = \kappa$, $\check{\phi}_{1,n} \asymp \log n/\sqrt{n}$ and $\check{\phi}_{0,n} \asymp (\log n)^2 n^{-1/(2-\kappa)}$

5 Applications

5.1 Classification

In this subsection, we consider a binary classification problem where each sample point is given by $Z_i = (X_i, Y_i)$ with binary label $Y_i \in \{-1,1\}$ and input $X_i \in [0,1]^d$. Assume that Ξ is a class of real-valued functions on $[0,1]^d$ with $\sup_{f\in\Xi}\|f\|_\infty \le F$ for some $F > 0$. When we use a function $f \in \Xi$ for prediction, the label Y associated with an input X is estimated by the sign of $f(X)$. The performance of f is usually evaluated by the *misclassification error* $\mathsf{P}(Yf(X) < 0)$, which is the expectation of the 0-1 loss $\ell_{0/1}$ such that $\ell_{0/1} \circ f(Y,X) = \mathbb{1}(Yf(X) < 0)$. A natural approach to finding a good estimator is optimizing the empirical risk $\mathsf{P}_n[\ell_{0/1} \circ f]$ given with the 0-1 loss. However, this optimization is computationally infeasible due to the discrete nature of the 0-1 loss.

In practice, computationally feasible surrogate losses can used to overcome the computational issue. They fall into a class of *margin-based* loss functions, which are given by

$\ell \circ f(X,Y) = \ell(Yf(X))$. Examples include the logistic loss $\ell_{\text{logit}} : z \mapsto \log(1+\exp(-z))$, exponential loss $\ell_{\exp} : z \mapsto \exp(-z)$ and the hinge loss $\ell_{\text{hinge}} : z \mapsto (1-z) \vee 0$. We may use the (penalized) ERM estimator that minimizes the empirical risk given with such a surrogate loss. Then we can provide a convergence rate of the excess risk defined with the surrogate loss we use. But typically, a theoretically interesting quantity is the 0-1 excess risk, which is defined as

$$\mathcal{E}_{0/1}(f) := \mathsf{P}[\mathbb{1}(Yf(X) < 0)] - \min_{\tilde{f} \in \Xi} \mathsf{P}[\mathbb{1}(Y\tilde{f}(X) < 0)]$$

for $f \in \Xi$. Motivated by this, the relationship between the excess risks with respect to the 0-1 and surrogate loss was investigated. Zhang [17], Bartlett et al. [18] proved that when the surrogate loss function is strictly convex, the following calibration inequality

$$\mathcal{E}_{0/1}(f) \leq \mathcal{E}_{\text{scvx}}(f)^{1/2}$$

holds, where $\mathcal{E}_{\text{scvx}}$ denotes the excess risk with respect to a strongly convex loss. This inequality is also sharp, that is, the exponent cannot be larger than the current one $1/2$, see the discussion below Theorem 2.2 of Zhang et al. [19]. The calibration inequality implies that even if we have a fate convergence rate of $\mathcal{E}_{\text{scvx}}(f)$, we are not able to get a faster rate of the 0-1 excess risk $\mathcal{E}_{0/1}(f)$ than $n^{-1/2}$. In contrast, the hinge loss is promising since it allows a shaper calibration inequality

$$\mathcal{E}_{0/1}(f) \leq \mathcal{E}_{\text{hinge}}(f)$$

where $\mathcal{E}_{\text{hinge}}$ denotes the excess risk with respect to the hinge loss. A problem of the hinge loss is that it does not satisfy the Bernstein assumption in general. However, there is a reasonable assumption that overcome this issue.

Assumption 5 (Tsybakov's noise condition)**.** The conditional class probability function $\eta(X) := \mathsf{P}(Y = 1|X)$ satisfies

$$\mathsf{P}(|\eta(X) - 1/2| \leq t) \leq C_1 t^{\alpha}$$

for any $t > 0$ for some absolute constants $\alpha > 0$ and $C_1 > 0$.

The *noise* condition was introduced by Mammen and Tsybakov [20], Tsybakov [21]. This is related to the behavior of the distribution of the input X near the decision boundary $\{X \in \mathcal{X} : \eta(X) = 1/2\}$. The exponent α determines the "easiness" of the problem. A large α means that there is a small mass near the decision boundary, and so the chance that we encounter data points difficult to classify is low.

Under the noise condition, the hinge loss leads to a Bernstein class.

Lemma 5.1 (Lemma 6.1 of Steinwart and Scovel [22])**.** *Under Assumption 5 with $\alpha > 0$, the class $\ell_{\text{hinge}}(\Xi_{-f_\star}) := \{\ell_{\text{hinge}} \circ f - \ell_{\text{hinge}} \circ f_\star : f \in \Xi\}$ is a $(\alpha/(\alpha+1), R)$-Bernstein class for some absolute constant $R > 0$, where $f_\star = \operatorname{argmin}_{f \in \Xi} \mathsf{P}[\ell_{\text{hinge}} \circ f]$.*

Thanks to the variance bound in the above lemma, we can attain a fast convergence rate for the (penalized) ERM estimator with the hinge loss. Several specific examples were studied in the literature, for example, the support vector machine [22] and deep neural network classifier [23].

Theorem 5.2. *Consider the classification setup described in this subsection with Assumption 5 being met. Then we have the following.*

1. *Suppose that Assumption 3 holds. Then the ERM estimator with the hinge loss satisfies*

$$\mathcal{E}_{0/1}(\widehat{f}_n) \le C_1 \max\left\{ \inf_{f\in\mathcal{F}} \mathcal{E}(f), \bar{\phi}_n(\mathcal{F}), \left(\frac{t\log n}{n}\right)^{(\alpha+1)/(\alpha+2)} \right\}$$

   with probability at least $1-\mathrm{e}^{-t}$ *for any* $t>0$ *for some absolute constant* $C_1>0$.
2. *Suppose that Assumption 4 holds and the tuning parameter satisfoes (4.1). Then the penalized ERM estimator with the hinge loss satisfies*

$$\mathcal{E}_{0/1}(\widehat{f}_{n,\lambda}) \le C_2 \max\left\{ \inf_{f\in\mathcal{F}} \{\mathcal{E}(f)+\lambda\Gamma(f)\}, \left(\frac{t\log n}{n}\right)^{(\alpha+1)/(\alpha+2)} \right\}$$

   with probability at least $1-\mathrm{e}^{-t}$ *for any* $t>0$ *for some absolute constant* $C_2>0$.

Example 7 (Neural network classifier)**.** We consider a deep neural network classifier that minimizes the penalized hinge empirical risk plus the sparsity-inducing penalty as

$$\widehat{f}^{\mathrm{DNN}}_{n,\lambda_n} = \operatorname*{argmin}_{f_\theta\in\mathcal{F}^{\mathrm{DNN}}_n} \left\{ \mathsf{P}_n[\ell_{\mathrm{hinge}}\circ f_\theta] + \lambda_n S(\theta)^{(\alpha+1)/(\alpha+2)} \right\}$$

with $\lambda_n := C_1(\log n)^2 n^{-(\alpha+1)/(\alpha+2)}$ for a sufficiently large constant $C_1>0$, where we denote $\mathcal{F}^{\mathrm{DNN}}_n := \mathcal{F}^{\mathrm{DNN}}(L\asymp\log n, M\asymp n, D\asymp n, F\lesssim 1)$. To the best of our knowledge, such an estimator has not been studied yet. The ERM procedure with the hinge loss and a deep neural network model was extensively studied by [23]. Assume that η is β-Hölder continuous. Then there exists a positive constant $C_2>0$ such that

$$\mathcal{E}_{0/1}(\widehat{f}^{\mathrm{DNN}}_{n,\lambda_n}) \le C_2 \max\left\{ n^{-\frac{\beta(\alpha+1)}{\beta(\alpha+2)+d}} (\log n)^{(\alpha+1)(\frac{\beta}{d}\vee\frac{2}{\alpha+2})}, \left(\frac{t\log n}{n}\right)^{\frac{\alpha+1}{\alpha+2}} \right\}$$

holds with probability at least $1-\mathrm{e}^{-t}$ for any $t>0$. This rate is minimax optimal up to a logarithmic factor [24].

To prove this, we apply Theorem 4.1 with $\kappa=\alpha/(\alpha+1)$. Since $\Gamma(f_\theta) = S(\theta)^{(\alpha+1)/(\alpha+2)} = S(\theta)^{1/(2-\kappa)} \le \gamma$ is equivalent to $S(\theta)\le\gamma^{2-\kappa}$, by the same argument as Example 6, we know that Assumption 4 holds with $\rho=\kappa=\alpha/(\alpha+1)$. Also by Example 6, $\check{\phi}_{0,n} = (\log n)^2 n^{-1/(2-\kappa)} = (\log n)^2 n^{-(\alpha+1)/(\alpha+2)}$ and $(\check{\phi}_{1,n})^{2/(2-\rho)} \asymp ((\log n)^2/n)^{-(\alpha+1)/(\alpha+2)}$. Thus, the tuning parameter condition (4.1) is met. By a similar argument as the proof of Theorem 3.3 of Kim et al. [23], which utilizes the approximation power of the neural network architecture, we have

$$\inf_{f_\theta\in\mathcal{F}^{\mathrm{DNN}}_n} \{\mathcal{E}_{\mathrm{hinge}}(f_\theta) + \lambda_n S(\theta)^{(\alpha+1)/(\alpha+2)}\}$$

$$
\begin{aligned}
&\leq \inf_{f_\theta \in \mathcal{F}_n^{\mathrm{DNN}}} \left\{ \|f_\theta - \eta\|_\infty^{\alpha+1} + \lambda_n S(\theta)^{(\alpha+1)/(\alpha+2)} \right\} \\
&\lesssim \min_{S \in \mathbb{N}: S \lesssim n} \left\{ (S/\log n)^{-\beta(\alpha+1)/d} + \lambda_n S^{(\alpha+1)/(\alpha+2)} \right\} \\
&\leq (\log n)^{(\alpha+1)(\frac{\beta}{d} \vee \frac{2}{\alpha+2})} \min_{S \in \mathbb{N}: S \lesssim n} \left\{ S^{-\beta(\alpha+1)/d} + (S/n)^{(\alpha+1)/(\alpha+2)} \right\}
\end{aligned}
$$

We take $S_n \asymp n^{d/(\beta(\alpha+2)+d)}$ that attains the balance $S_n^{-\beta(\alpha+1)/d} \asymp (S_n/n)^{(\alpha+1)/(\alpha+2)}$. Thus, the above display can be bounded by

$$
n^{-\frac{\beta(\alpha+1)}{\beta(\alpha+2)+d}} (\log n)^{(\alpha+1)(\frac{\beta}{d} \vee \frac{2}{\alpha+2})}.
$$

which completes the proof.

5.2 Optimal transport map estimation

Let P and Q be two probability measures on $\Omega \subset \mathbb{R}^d$. For a *transport map* $T : \Omega \mapsto \Omega$, the pushforward measure $T_\sharp \mathsf{P}$ of P through T is a measure satisfying $T_\# \mathsf{P}(E) = \mathsf{P}(T_\#^{-1}(E))$ for any Borel measurable subset $E \subset \Omega$. The optimal transport map $T_\star$ is the transport map from P to Q that solves the *Monge problem*

$$
T_\star = \operatorname*{argmin}_{T: T_\# \mathsf{P} = \mathsf{Q}} \int |T(x) - x|^2 \mathrm{d}\mathsf{P}(x).
$$

The Monge problem can be solved indirectly by solving the so-called *semi-dual problem*. Let $\mathfrak{L}^1(\mathsf{P}) := \{f : \int |f| \mathrm{d}\mathsf{P} < \infty\}$. For a *potential* $f \in \mathfrak{L}^1(\mathsf{P})$, define the semi-dual objective function

$$
\mathcal{S}(f) := \mathsf{P}[f] + \mathsf{Q}[f^*],
$$

where f^* denotes the convex conjugate of f defined as $f^*(y) = \sup_{x \in \mathbb{R}^d} \{x^\top y - f(x)\}$. Then Brenier's theorem [25] states that the optimal transport map is given by

$$
T^\star = \nabla f_\star \text{ with } f_\star = \operatorname*{argmin}_{f \in \mathfrak{L}^1(\mathsf{P})} \mathcal{S}(f)
$$

and the optimal potential $f_\star$ is a convex function.

Suppose that we have the two-sample data $X_{1:n} := (X_1, \ldots, X_n)$ and $Y_{1:n} = (Y_1, \ldots, Y_n)$ which are n i.i.d. random variables following P and Q, respectively. We assume that $X_{1:n}$ and $Y_{1:n}$ are independent. Our aim is to estimate the optimal transport map from P to Q based on the data $X_{1:n}$ and $Y_{1:n}$. Motivated by Brenier's theorem, our strategy is to first obtain the ERM estimator of the optimal potential such that

$$
\widehat{f}_n = \operatorname*{argmin}_{f \in \mathcal{F}} \mathcal{S}_n(f) := \mathsf{P}_n[f] + \mathsf{Q}_n[f^*]
$$

for some function class $\mathcal{F} \subset \mathfrak{L}^1(\mathsf{P})$ satisfying the assumption given below, and estimate $T_\star$ by its gradient $\widehat{T}_n := \nabla \widehat{f}_n$.

Assumption 6. The optimal transport map $f_\star$ and the function class $\mathcal{F} \subset \mathfrak{L}^1(\mathsf{P})$ satisfy the following.

1. There exists a constant $F > 0$ such that $\|f_\star\| \vee \sup_{f \in \mathcal{F}} \|f\|_\infty \leq F$.
2. Any member f in $\mathcal{F}$ is twice continuously differentiable on $\mathbb{R}^d$ and satisfies

$$\frac{1}{2} I \preceq \nabla^2 f(x) \preceq 2I$$

for any $x \in \Omega$, where I denotes the identity matrix and $A \preceq B$ means that $B - A$ is non-negative definite.

Lemma 5.3. *Under Assumption 6, we have*

$$\frac{1}{4} \|\nabla f - \nabla f_\star\|^2_{\mathfrak{L}^2(\mathsf{P})} \leq \mathcal{S}(f) - \mathcal{S}(f_\star) \leq 2 \|\nabla f - \nabla f_\star\|^2_{\mathfrak{L}^2(\mathsf{P})}$$

for any $f \in \mathcal{F}$.

Proof. See Theorem 3.1 of [26]. □

To guarantee the Bernstein condition, we need some additional conditions on the probability measures P and Q.

Definition 5. A probability measure P satisfies the Poincaré inequality if there exists an absolute constant $K > 0$ such that

$$\mathsf{Var}_\mathsf{P}(f) \leq K \|\nabla f\|^2_{\mathfrak{L}^2(\mathsf{P})}$$

for every function $f : \mathbb{R}^d \mapsto \mathbb{R}$ with $\mathsf{P}[f^2] < \infty$.

Assumption 7. Both P and Q satisfy the Poincaré inequality.

Lemma 5.4. *Under Assumptions 6 and 7, the function class $\{(f - f_\star) + (f^* - f_\star^*) : f \in \mathcal{F}\}$ is a $(1, 16K)$-Bernstain class.*

Proof. By Assumption 7 and Lemma 5.3,

$$\mathsf{Var}(f - f_\star) \leq K \|\nabla f - \nabla f_\star\|^2_{\mathfrak{L}^2(\mathsf{P})} \leq 4K \{\mathcal{S}(f) - \mathcal{S}(f_\star)\}.$$

Define $\mathcal{S}^*(g) := \mathsf{Q}[g] + \mathsf{P}[g^*]$, that is, $\mathcal{S}^*$ is the semi-dual objective function which interchanges the roles of P and Q. By the properties of the convex conjugate (c.f. Lemma A.9 of [26]), f^* satisfies Assumption 6. Hence, by Assumption 7 and Lemma 5.3 again

$$\mathsf{Var}(f^* - f_\star^*) \leq K \|\nabla f^* - \nabla f_\star^*\|^2_{\mathfrak{L}^2(\mathsf{Q})} \leq 4K \{\mathcal{S}^*(f^*) - \mathcal{S}^*(f_\star^*)\}.$$

Since $(f^*)^* = f$ for all convex and lower semicontinuous function f, we have $\mathcal{S}(f) = \mathcal{S}^*(f^*)$. This together with the simple fact $\mathsf{Var}(X+Y) \leq 2\mathsf{Var}(X) + 2\mathsf{Var}(Y)$, completes the proof. □

Thanks to the above lemma, Theorem 2.4 gives the following result.

Theorem 5.5. *Suppose that Assumptions 3, 6 and 7 hold. Then the transport map estimator* $\widehat{T}_n := \nabla \widehat{f}_n$ *with the ERM* $\widehat{f}_n$ *satisfies*

$$\|\widehat{T}_n - T_\star\|^2_{\mathfrak{L}^2(\mathsf{P})} \le C_1 \max\left\{ \inf_{f\in\mathcal{F}} \|\nabla f - T_\star\|^2_{\mathfrak{L}^2(\mathsf{P})}, \bar{\phi}_n(\mathcal{F}), \frac{t\log n}{n} \right\}$$

with probability at least $1-\mathrm{e}^{-t}$ *for any* $t>0$ *for some absolute constant* $C_1>0$.

5.3 Density estimation

We introduce our density estimation setup and additional notation. Suppose that we have the sample $X_{1:n} = (X_1,\dots,X_n)$ which are $\mathcal{X}$-valued i.i.d. random variables generated from the distribution with density $f_\star$ with respect to the reference measure μ, that is, $f_\star = \mathrm{d}\mathsf{P}/\mathrm{d}\mu$, where $\mathcal{X}$ is a compact subset of $\mathbb{R}^d$. Let Ξ be the class of densities on $\mathcal{X}$. We assume that the true density function satisfies the following.

Assumption 8. There exists a constant $c_0>0$ such that $f_\star(x)\ge c_0$ for all $x\in\mathcal{X}$.

Let Hel and KL denote the Hellinger distance and Kullback-Leibler divergence, respectively, that is, for two densities f_1 and f_2,

$$\mathrm{Hel}^2(f_1,f_2) := \frac{1}{2}\int_{\mathcal{X}} (\sqrt{f_1}-\sqrt{f_2})^2 \mathrm{d}\mu,$$
$$\mathrm{KL}(f_1,f_2) := \int_{\mathcal{X}} \log\left(\frac{f_1}{f_2}\right) f_1 \mathrm{d}\mu.$$

We estimate the true density function by the maximum likelihood estimator (MLE) given as

$$\widehat{f}_n = \operatorname*{argmin}_{f\in\mathcal{F}} \mathsf{P}_n[-\log f]$$

for some class of densities $\mathcal{F}\subset\Xi$ which satisfies $\sup_{f\in\Xi}\|f\|_\infty \le F$ for some $F>0$. This can be viewed as the ERM with the negative log loss $\ell_{-\log}: x\mapsto -\log(z)$. But we cannot apply Theorem 2.4 or Theorem 4.1 since $\ell_{-\log}\circ f$ is not uniformly bounded, as $\ell_{-\log}\circ f(x)$ diverges when $f(x)$ is close to 0.

A classical idea to overcome this issue is to consider the transformed function class

$$\mathcal{H} := \mathcal{H}(\mathcal{F}) = \left\{ h_f := \sqrt{\frac{f+f_\star}{2f_\star}} : f\in\mathcal{F} \right\}.$$

Then, under Assumption 8, since every f in $\mathcal{F}$ is bounded by F, we have $\|h_f\|_\infty \le \sqrt{(F+c_0)/2c_0}$ and $h_f(x)\ge\sqrt{1/2}$ for any $x\in\mathcal{X}$. Then the class $\{\ell_{-\log}\circ h_f - \ell_{-\log}\circ h_{f_\star} : f\in\mathcal{F}\}$ is a uniformly bounded class. Furthermore, it turns out that this is a $(1,4)$-Bernstein class. To see this, note that

$$\begin{aligned}\mathsf{Var}_\mathsf{P}(\ell_{-\log}\circ h_f - \ell_{-\log}\circ h_{f_\star}) &\le \mathsf{P}[(\ell_{-\log}\circ h_f - \ell_{-\log}\circ h_{f_\star})^2] \\ &\le 2\mathsf{P}[(h_f - h_{f_\star})^2],\end{aligned}$$

where the second inequality follows from that the negative log function is $\sqrt{2}$-Lipschiz on the interval $[\sqrt{1/2},\infty)$. Then by the well-known relationship $\mathrm{KL}(f_1,f_2) \geq 2\mathrm{Hel}^2(f_1,f_2)$, we further have

$$\begin{aligned}\mathsf{P}[(h-h_{f_\star})^2] = 2\mathrm{Hel}^2\left(f_\star, \frac{f+f_\star}{2}\right) &\leq \mathrm{KL}\left(f_\star, \frac{f+f_\star}{2}\right) \\ &= 2\mathsf{P}[\ell_{-\log}\circ h_f - \ell_{-\log}\circ h_{f_\star}],\end{aligned}$$

where the last equality follows from that $h_{f_\star} = 1$.

But we still have something to check to apply Theorem 2.4. First, we need the following basic inequality

$$\mathrm{Hel}^2\left(\frac{\widehat{f}_n + f_\star}{2}, f_\star\right) \leq (\mathsf{P}_n - \mathsf{P})[\ell_{-\log}\circ h_f - \ell_{-\log}\circ h_{f_\star}]$$

which replaces the standard basic inequality (2.2). The proof of the above can be found in Lemma 4.1 of [7]. Thus, we attain a high-probability upper bound of the excess risk $\mathcal{E}_{\mathcal{H}}(f) := \mathcal{E}(h_f) := \mathsf{P}[\ell_{-\log}\circ h_f - \ell_{-\log}\circ h_{f_\star}]$ for the transformed function class. Since we have, by Lemma 4.2 of [7],

$$\mathrm{Hel}^2(f,f_\star) \leq 16\mathrm{Hel}^2\left(f_\star, \frac{f+f_\star}{2}\right) \leq 8\mathcal{E}_{\mathcal{H}}(f),$$

such a upper bound automatically implies a high-probability upper bound of the squared Hellinger distance $\mathrm{Hel}^2(\widehat{f}_n, f_\star)$. Combining these results, we get the next theorem.

Theorem 5.6. *Suppose that Assumption 8 holds and that $\mathcal{H}$ satisfies Assumption 3. Then the MLE satisfies*

$$\mathrm{Hel}^2(\widehat{f}_n, f_\star) \leq C_1 \max\left\{\inf_{f\in\mathcal{F}} \mathcal{E}_{\mathcal{H}}(f), \bar{\phi}_n(\mathcal{H}(\mathcal{F})), \frac{t\log n}{n}\right\}$$

with probability at least $1-\mathrm{e}^{-t}$ for any $t>0$ for some absolute constant $C_1>0$.

Example 8 (Log density model)**.** We consider a density function of the form

$$f_\theta(x) = \exp(b_\theta(x) - A(\theta)), \text{ with } A(\theta) = \log\left(\int_{\mathcal{X}} \exp(b_\theta(x))\mathrm{d}\mu(x)\right)$$

for some function $b_\theta : \mathcal{X} \mapsto \mathbb{R}$ parameterized by $\theta \in \Theta$ where Θ denotes a parameter space and let $\mathcal{F}_\Theta := \{f_\theta : \theta \in \Theta\}$. That is, we model the log density function by a certain (non)-parametric model. Then for any $\theta_1 \in \Theta$ and $\theta_2 \in \Theta$, we have

$$\begin{aligned}\left\|-\log\left(\sqrt{\frac{f_{\theta_1}+f_\star}{2f_\star}}\right) + \log\left(\sqrt{\frac{f_{\theta_2}+f_\star}{2f_\star}}\right)\right\|_\infty &= \frac{1}{2}\left\|-\log\left(f_{\theta_1}+f_\star\right) + \log\left(f_{\theta_2}+f_\star\right)\right\|_\infty \\ &\leq \frac{1}{2}\left\|-\log(f_{\theta_1}) + \log\left(f_{\theta_2}\right)\right\|_\infty\end{aligned}$$

$$\leq \left\| b_{\theta_1} - b_{\theta_2} \right\|_\infty,$$

where the first inequality follows from the inequality $\log(x+b) - \log(x+a) > \log(b) - \log(a)$ for any $x > 0$ when $b > a$ and the second inequality follows from that

$$|A(\theta_1) - A(\theta_2)| = \left| \log \left(\frac{\int_{\mathcal{X}} \exp(b_{\theta_1}(x)) \mathrm{d}\mu(x)}{\int_{\mathcal{X}} \exp(b_{\theta_2}(x)) \mathrm{d}\mu(x)} \right) \right| \leq \|b_{\theta_1} - b_{\theta_2}\|_\infty.$$

This implies that

$$H_\infty(\varepsilon, \mathcal{H}(\mathcal{F}_\Theta)) \leq H_\infty(\varepsilon, \{b_\theta : \theta \in \Theta\}).$$

Hence, we can bound the estimation error in terms of the metric entropy of the class $\{b_\theta : \theta \in \Theta\}$. For the approximation error, we may employ the fact that

$$\begin{aligned}
\mathcal{E}_{\mathcal{H}}(f_\theta) &= \frac{1}{2} \mathsf{P} \left[-\log \left(\frac{f_\theta + f_\star}{2} \right) + \log(f_\star) \right] \\
&\leq = \frac{1}{2} \mathsf{P} \left[-\log \left(\frac{f_\theta + f_\star}{2} \right) + \log(f_\star) \right] \\
&\leq \frac{1}{4} \mathsf{P} \left[-\log(f_\theta) + \log(f_\star) \right] \\
&\leq \frac{1}{4} \| \log(f_\theta) - \log(f_\star) \|_\infty \leq \frac{1}{2} \| b_\theta - b_\star \|_\infty
\end{aligned}$$

and find a good approximation b_θ of $b_\star := \log(f_\star)$.

5.4 Sub-Gaussian regression

In this subsection, we consider a nonparametric regression problem where we can access the sample $Z_{1:n} = (Z_1, \ldots, Z_n)$ where each Z_i is generated as

$$Y_i = f_\star(X_i) + \zeta_i, \quad X_i \overset{\text{iid}}{\sim} \mathsf{P}_X$$

where P_X is a distribution on a compact subset of $[0,1]^d$ and $\zeta_1, \ldots, \zeta_n$ are i.i.d. mean-zero errors which are independent to $X_1, \ldots, X_n$. We assume that ε_i is a sub-Gaussian random variable with parameter $\sigma_\zeta > 0$, i.e., $\mathsf{P}[\mathrm{e}^{u\zeta_i}] \leq \mathrm{e}^{u^2\sigma_\zeta^2/2}$ for any $u \in \mathbb{R}$. Assume that Ξ is a class of real-valued functions on $[0,1]^d$ with $\sup_{f \in \Xi} \|f\|_\infty \leq F$ for some $F > 0$. We consider the square loss function ℓ_{sq} such that $\ell_{\text{sq}} \circ f(Y, X) = (Y - f(X))^2$. The corresponding excess risk is given by

$$\begin{aligned}
\mathcal{E}(f) &= \mathsf{P}[(Y - f(X))^2] - \mathsf{P}[(Y - f_\star(X))^2] \\
&= \|f - f_\star\|^2_{\mathfrak{L}^2(\mathsf{P}_X)} := \int |f(X) - f_\star(X)|^2 \mathrm{d}\mathsf{P}_X(x).
\end{aligned}$$

A technical problem here is that $\ell_{\text{sq}} \circ f - \ell_{\text{sq}} \circ f_\star$ is not bounded, and thus the theoretical results cannot be applied. But sub-Gaussian random variables have very thin tails, so we can

deal with them as "almost bounded" random variables. A truncation argument used in [16, 27] utilizes this idea.

Theorem 5.7. *Consider the sub-Gaussian regression setup described in this subsection. Then we have the following.*

1. *Suppose that Assumption 3 holds. Then there exists a constant $C_1 > 0$ such that*

$$\|\widehat{f}_n - f_\star\|^2_{\mathfrak{L}^2(\mathsf{P}_X)} \le C_1 \max\left\{ \inf_{f\in\mathcal{F}} \mathcal{E}(f), \bar{\phi}_n(\mathcal{F}), \frac{t(\log n)^3}{n} \right\}$$

   with probability at least $1-\mathrm{e}^{-t}$ for any $t>0$.
2. *Suppose that Assumption 4 holds and the tuning parameter satisfies (4.1). Then there exists a constant $C_2 > 0$ such that*

$$\|\widehat{f}_{n,\lambda} - f_\star\|^2_{\mathfrak{L}^2(\mathsf{P}_X)} \le C_3 \max\left\{ \inf_{f\in\mathcal{F}} \{\mathcal{E}(f) + \lambda\Gamma(f)\}, \frac{t(\log n)^3}{n} \right\}$$

   with probability at least $1-\mathrm{e}^{-t}$ for any $t>0$.

Proof. We only prove the case of the penalized ERM. The ERM case is can be proved similarly. For a real-valued random variable Y, we define $Y^{\le B} := \mathrm{sign}(Y)(|Y| \wedge B)$, which is a truncation of Y at level $B>0$. Furthermore, we let $\ell_{\mathrm{sq}}^{\le B} \circ f$ denote the square loss with the truncated input as $\ell_{\mathrm{sq}}^{\le B} \circ f(Y,X) = (Y^{\le B} - f(X))^2$ and let $f_\star^{\le B}$ be the minimizer of the truncated risk such that

$$f_\star^{\le B} = \operatorname*{argmin}_{f\in\Xi} \mathsf{P}[\ell_{\mathrm{sq}}^{\le B} \circ f].$$

Note that $f_\star^{\le B}(X) = \mathsf{P}(Y^{\le B}|X)$ where $\mathsf{P}(\cdot|X)$ denotes the conditional expectation given X. For notational convenience, we write $g_f := \ell_{\mathrm{sq}} \circ f - \ell_{\mathrm{sq}} \circ f_\star$ and $g_f^{\le B} := \ell_{\mathrm{sq}}^{\le B} \circ f - \ell_{\mathrm{sq}}^{\le B} \circ f_\star^{\le B}$. Then for any $f \in \mathcal{F}$, we have

$$\begin{aligned}
\left| g_f(Y,X) - g_f^{\le B}(Y,X) \right| &\le \left| 2\{f(X) - f_\star(X)\}(Y^{\le B} - Y) + (f_\star^{\le B}(X) - Y^{\le B})^2 - (f_\star(X) - Y^{\le B})^2 \right| \\
&\le 4F \left| Y^{\le B} - Y \right| + \left| f_\star^{\le B}(X) - f_\star(X) \right| \left| f_\star^{\le B}(X) + f_\star(X) - 2Y^{\le B} \right| \\
&\le 4F|Y| \mathbb{1}(|Y| > B) + 2(B+F)|f_\star^{\le B}(X) - f_\star(X)|.
\end{aligned}$$

We denote the two terms in the last line as

$$W_1 := 4F|Y|\mathbb{1}(|Y|>B), \quad W_2 := 2(B+F)|f_\star^{\le B}(X) - f_\star(X)|.$$

Note that W_1 and W_2 do not depend on the choice of f. With the truncation error $W := W_1 + W_2$, we can decompose the excess risk as

$$\begin{aligned}
\mathcal{E}(\widehat{f}_{n,\lambda}) &\le \mathsf{P}[W] + \mathsf{P}_n[g_{\widehat{f}_{n,\lambda}}^{\le B}] \\
&\le \mathsf{P}[W] + (\mathsf{P} - \mathsf{P}_n)[g_{\widehat{f}_{n,\lambda}}^{\le B}] + \mathsf{P}_n[W] + \mathsf{P}_n[g_{\widehat{f}_{n,\lambda}}]
\end{aligned}$$

$$\leq \mathsf{P}[W] + \mathsf{P}_n[W] + (\mathsf{P}-\mathsf{P}_n)[g^{\leq B}_{\widehat{f}_{n,\lambda}}] - \lambda\Gamma(\widehat{f}_{n,\lambda}) + \mathsf{P}_n[g_{\bar{f}}] + \lambda\Gamma(\bar{f})$$
$$\leq \mathsf{P}[W] + 2\mathsf{P}_n[W] + (\mathsf{P}-\mathsf{P}_n)[g^{\leq B}_{\widehat{f}_{n,\lambda}}] - \lambda\Gamma(\widehat{f}_{n,\lambda}) + \mathsf{P}_n[g^{\leq B}_{\bar{f}}] + \lambda\Gamma(\bar{f})$$

for any $\bar{f} \in \mathcal{F}$, where the last inequality follows from the basic inequality for the penalized ERM. Hence, if we succeed in bounding $\mathsf{P}[W]+2\mathsf{P}_n[W]$, we get the desired result by applying Theorem 4.1 to the rest of the terms, which leads to the bound

$$\inf_{f\in\mathcal{F}} \left\{ \mathsf{P}[g^{\leq B}_f] + \lambda\Gamma(f) \right\} \vee \frac{B^2 t \log n}{n}. \tag{5.1}$$

This is due to that the class $\{g^{\leq B}_f : f \in \mathcal{F}\}$ is a bounded $(1, 16B^2)$-Bernstein class as shown in Example 3. Now, we take $B = B_n := C_3 \log n$ for a sufficiently large $C_3 > 0$. Then by Lemma A.2, we have $\mathsf{P}_n[W_1] \lesssim t/n$ and $\mathsf{P}_n[W_2] \lesssim Bt/n \asymp t\log n/n$ with probability at least $1 - 2\mathrm{e}^{-t}$. Lastly, we need to handle the term $\mathsf{P}[g^{\leq B}_{\bar{f}}]$ term in (5.1) associated with the artificially introduced function $f^{\leq B}_\star$. But we will show that this term is very close to $\mathcal{E}(f)$ for any f. Note that

$$\begin{aligned}
|\mathsf{P}[g^{\leq B}_f] - \mathcal{E}(f)| &\leq (4F)^2 \mathsf{P}[(f_\star - f^{\leq B}_\star)^2] \\
&= (4F)^2 \mathsf{P}[|\mathsf{P}(Y|X) - \mathsf{P}(Y^{\leq B}|X)|^2] \\
&\leq (4F)^2 \mathsf{P}[(Y - Y^{\leq B})^2],
\end{aligned}$$

where the last inequality follows from Jensen's inequality. Furthermore, since Y is sub-Gaussian and $B \asymp \log n$, we have

$$\mathsf{P}[(Y - Y^{\leq B})^2] = \mathsf{P}[|Y|^2 \mathbb{1}(|Y| > B)] \leq 2\mathsf{P}[\exp(2Y - B)] \lesssim n^{-1},$$

which completes the proof. □

Appendix A Technical lemmas

Lemma A.1. *Let $p > 1$ and $q > 1$ be conjugate indices such that $1/p + 1/q = 1$. Then we have*

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}$$

for any $a \geq 0$ and $b \geq 0$.

Proof. See Lemma 7.1 of Steinwart and Christmann [28]. □

Lemma A.2. *Let $Y_1, \ldots, Y_n$ be i.i.d. sub-Gaussian random variables. Let $F > 0$ and $W_i := F|Y_i|\mathbb{1}(|Y_i| > \widetilde{B}\log n)$ for $i \in [n]$ for a sufficiently large $\widetilde{B} > 0$. Then there exist constants $C_1 > 0$*

such that

$$\mathbb{P}\Big(\frac{1}{n}\sum_{i=1}^{n} W_i \geq C_1 \frac{t}{n}\Big) \leq \mathrm{e}^{-t}.$$

Proof. Since Y is sub-Gaussian, by Proposition 2.5.2 of Vershynin [29], there is a constant $C_2 > 0$ such that $\mathbb{E}[\exp(Y^2/C_2^2)] \leq 2$. Hence, taking $C_3 = C_2 F$, we have $\mathbb{E}[\exp(W^2/C_3^2)] \leq 2$ which implies W is sub-Gaussian. Thus, by Bernstein's inequality, there exists a constant $C_4 > 0$ such that

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} W_i \geq \mathsf{P}[W_1] + \sqrt{\frac{2t}{n}\mathsf{Var}(W_1)} + C_4\frac{t}{n}\right) \leq \mathrm{e}^{-t}.$$

Since $\mathbb{1}(|Y_1| > B_2) \leq \exp((|Y_1| - B_2)/(2C_2^2))$, and $|Y|/(2C_2^2) \leq \exp(|Y|/(2C_2^2))$ we have

$$\mathsf{P}[W_1] \leq F\mathsf{P}[\exp(Y^2/C_2^2 - \widetilde{B}\log n/(2C_2^2)] \leq 2F\exp(-\widetilde{B}\log n/(2C_2^2))$$

So if $\widetilde{B}$ is larger than $2C_2^2$, the last display is less than n^{-1} up to a constant. By a similar argument, we can see that $\mathsf{P}[W_1^2] \lesssim n^{-1}$, which completes the proof. □

References

[1] Vapnik, V., Chervonenkis, A.: Theory of pattern recognition. Nauka, Moscow (1974)

[2] Koltchinskii, V.: Rademacher penalties and structural risk minimization. IEEE Transactions on Information Theory **47**(5), 1902–1914 (2001)

[3] Koltchinskii, V., Panchenko, D.: Empirical margin distributions and bounding the generalization error of combined classifiers. The Annals of Statistics **30**(1), 1–50 (2002)

[4] Bartlett, P.L., Boucheron, S., Lugosi, G.: Model selection and error estimation. Machine Learning **48**, 85–113 (2002)

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

[6] Dudley, R.M.: The sizes of compact subsets of hilbert space and continuity of gaussian processes. Journal of Functional Analysis **1**(3), 290–330 (1967)

[7] Geer, S.A.: Empirical Processes in M-estimation vol. 6. Cambridge university press, ??? (2000)

[8] Shen, X., Wong, W.H.: Convergence rate of sieve estimates. The Annals of Statistics **22**(2), 580–615 (1994)

[9] Bartlett, P., Bousquet, O., Mendelson, S.: Local Rademacher complexities. The Annals of Statistics **33**(4), 1497–1537 (2005)

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

[11] Massart, P., Nédélec, É.: Risk bounds for statistical learning. The Annals of Statistics **34**(5), 2326–2366 (2006)

[12] Giné, E., Nickl, R.: Mathematical Foundations of Infinite-dimensional Statistical Models vol. 40. Cambridge university press, ??? (2015)

[13] Bousquet, O.: A bennett concentration inequality and its application to suprema of empirical processes. Comptes Rendus Mathematique **334**(6), 495–500 (2002)

[14] Mendelson, S.: Geometric parameters of kernel machines. In: International Conference on Computational Learning Theory, pp. 29–43 (2002). Springer

[15] Wainwright, M.J.: High-dimensional Statistics: A Non-asymptotic Viewpoint vol. 48. Cambridge university press, ??? (2019)

[16] Ohn, I., Kim, Y.: Nonconvex sparse regularization for deep neural networks and its optimality. Neural Computation **34**(2), 476–517 (2022)

[17] Zhang, T.: Statistical behavior and consistency of classification methods based on convex risk minimization. The Annals of Statistics **32**(1), 56–85 (2004)

[18] Bartlett, P.L., Jordan, M.I., McAuliffe, J.D.: Convexity, classification, and risk bounds. Journal of the American Statistical Association **101**(473), 138–156 (2006)

[19] Zhang, Z., Shi, L., Zhou, D.-X.: Classification with deep neural networks and logistic loss. Journal of Machine Learning Research **25**(125), 1–117 (2024)

[20] Mammen, E., Tsybakov, A.B.: Smooth discrimination analysis. The Annals of Statistics **27**(6), 1808–1829 (1999)

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

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

[23] Kim, Y., Ohn, I., Kim, D.: Fast convergence rates of deep neural networks for classification. Neural Networks **138**, 179–197 (2021)

[24] Audibert, J.-Y., Tsybakov, A.B.: Fast learning rates for plug-in classifiers. The Annals of Statistics **35**(2), 608–633 (2007)

[25] Brenier, Y.: Polar factorization and monotone rearrangement of vector-valued functions. Communications on Pure and Applied Mathematics **44**, 375–417 (1991)

[26] Chewi, S., Niles-Weed, J., Rigollet, P.: Statistical optimal transport. arXiv preprint arXiv:2407.18163 (2024)

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

[28] Steinwart, I., Christmann, A.: Support Vector Machines. Springer, ??? (2008)

[29] Vershynin, R.: High-dimensional Probability: An Introduction with Applications in Data Science vol. 47. Cambridge university press, ??? (2018)