Towards Problem-dependent Optimal Learning Rates

Yunbei Xu
Columbia University
New York, NY 10027
yunbei.xu@gsb.columbia.edu

Assaf Zeevi
Columbia University
New York, NY 10025
assaf@gsb.columbia.edu

Abstract

We study problem-dependent rates, i.e., generalization errors that scale tightly with the variance or the effective loss at the "best hypothesis." Existing uniform convergence and localization frameworks, the most widely used tools to study this problem, often fail to simultaneously provide parameter localization and optimal dependence on the sample size. As a result, existing problem-dependent rates are often rather weak when the hypothesis class is "rich" and the worst-case bound of the loss is large. In this paper we propose a new framework based on a "uniform localized convergence" principle. We provide the first (moment-penalized) estimator that achieves the optimal variance-dependent rate for general "rich" classes; we also establish improved loss-dependent rate for standard empirical risk minimization.

1 Introduction

Problem Statement. Consider the following statistical learning setting. Assume that a random sample z follows an unknown distribution $\mathbb{P}$ with support $\mathcal{Z}$. For each realization of z, let $\ell(\cdot; z)$ be a real-valued *loss function*, defined over the *hypothesis class* $\mathcal{H}$. Let $h^* \in \mathcal{H}$ be the optimal hypothesis that minimizes the *population risk*

$$\mathbb{P}\ell(h; z) := \mathbb{E}[\ell(h; z)].$$

Given n i.i.d. samples $\{z_i\}_{i=1}^n$ drawn from $\mathbb{P}$, our goal, roughly speaking, is to "learn" a hypothesis $\hat{h} \in \mathcal{H}$ that makes the *generalization error*

$$\mathscr{E}(\hat{h}) := \mathbb{P}\ell(\hat{h}; z) - \mathbb{P}\ell(h^*; z)$$

as small as possible. This pursuit is ubiquitous in machine learning, statistics and stochastic optimization.

Let $\mathcal{V}^*$ and $\mathcal{L}^*$ be the variance and the "effective loss" at the best hypothesis h^*:

$$\mathcal{V}^* := \mathrm{Var}[\ell(h^*; z)], \quad \mathcal{L}^* := \mathbb{P}[\ell(h^*; z) - \inf_{\mathcal{H}} \ell(h; z)].$$

We study finite-sample generalization errors that scale tightly with $\mathcal{V}^*$ or $\mathcal{L}^*$, which we call *problem-dependent rates*, without invoking strong convexity or margin conditions. While the direct dependence of $\mathscr{E}(\hat{h})$ on the sample size n is often well-understood, it typically only reflects an "asymptotic" perspective, placing less emphasis on the scale of problem-dependent parameters $\mathcal{V}^*$ and $\mathcal{L}^*$.

Main challenges. In absence of strong convexity and margin conditions, perhaps the most popular framework to study problem-dependent rates is the traditional "local Rademacher complexity" analysis [2, 6, 16], which has become a standard tool in learning theory. However, as we will discuss later, this analysis makes the "direct dependence" on the sample size (n) sub-optimal for all "rich" classes with the exception of parametric classes.

The absence of more precise localization analysis also challenges the design of more refined estimation procedures. For example, designing estimators to achieve variance-dependent rates requires penalizing

34th Conference on Neural Information Processing Systems (NeurIPS 2020), Vancouver, Canada.

the empirical second moment to achieve the "right" bias-variance trade-off. Most antecedent work is predicated on either the traditional "local Rademacher complexity" analysis [12, 4] or coarser approaches [8, 14]. Thus, to the best of our knowledge, the question of optimal variance-dependent rates for general rich classes is still open.

When assuming suitable curvature or margin conditions, much progress on problem-dependent rates has been made under particular formulations, such as supervised learning with strong convexity [10, 11, 7]. Methods tailored for these settings can not directly adopt the general setting we study.

Contributions. We introduce a new framework to study localization in statistical learning, dubbed "uniform localized convergence," which simultaneously provides optimal "direct dependence" on the sample size, and correct scaling with problem-dependent parameters. This framework resolves some fundamental limitations of existing localization analysis.

We employ the above ideas to design the first estimator that achieves optimal variance-dependent rates for general function classes. The derivation is based on a novel two-stage procedure that optimally penalizes the empirical (centered) second moment. We also establish improved loss-dependent rates for standard empirical risk minimization, which has computational advantages.

Organization. Section 2 introduces our proposed "uniform localized convergence" principle. Section 3 provides preliminaries. Section 4 presents the loss-dependent rate. Section 5 presents the variance-dependent rate. Section 6 illustrates our findings in two examples: non-parametric classes and VC classes.

2 The "uniform localized convergence" principle

2.1 The current blueprint

Denote the *empirical risk*

$$\mathbb{P}_n \ell(h; z) := \frac{1}{n} \sum_{i=1}^{n} \ell(h; z_i),$$

and consider the following straightforward decomposition of the generalization error

$$\mathscr{E}(\hat{h}) = (\mathbb{P} - \mathbb{P}_n)\ell(\hat{h}; z) + \big(\mathbb{P}_n \ell(\hat{h}; z) - \mathbb{P}_n \ell(h^*; z)\big) + (\mathbb{P}_n - \mathbb{P})\ell(h^*; z). \tag{2.1}$$

The main difficulty in studying $\mathscr{E}(\hat{h})$ comes from bounding the first term $(\mathbb{P} - \mathbb{P}_n)\ell(\hat{h}; z)$, since $\hat{h}$ depends on the n samples. The simplest approach, which does not achieve problem-dependent rates, is to bound the uniform error

$$\sup_{h \in \mathcal{H}} (\mathbb{P} - \mathbb{P}_n)\ell(h; z)$$

over the *entire* hypothesis class $\mathcal{H}$. In order to obtain problem-dependent rates, a natural modification is to consider uniform convergence over *localized* subsets of $\mathcal{H}$.

We first give an overview of the traditional "local Rademacher complexity" analysis [2, 6, 16]. Consider a generic function class $\mathcal{F}$ that we wish to concentrate, which consists of real-valued functions defined on $\mathcal{Z}$ (e.g., one can set $f(z) = \ell(h; z)$). Denote

$$\mathbb{P}f := \mathbb{E}[f(z)], \quad \mathbb{P}_n f := \frac{1}{n} \sum_{i=1}^{n} f(z_i),$$

and denote by $\psi(r; \delta)$ a surrogate function that upper bounds the uniform error within a localized region $\{f \in \mathcal{F} : T(f) \le r\}$, where we call $T : \mathcal{F} \to \mathbb{R}_+$ the "measurement functional". Formally, let ψ be a function that maps $[0, \infty) \times (0, 1)$ to $[0, \infty)$, which possibly depends on the observed samples $\{z_i\}_{i=1}^n$. Assume ψ satisfies for arbitrary fixed δ, r, with probability at least $1 - \delta$,

$$\sup_{f \in \mathcal{F} : T(f) \le r} (\mathbb{P} - \mathbb{P}_n) f \le \psi(r; \delta). \tag{2.2}$$

By default we ask $\psi(r; \delta)$ to be a non-decreasing and non-negatove function. The main result of the traditional "local Rademacher complexity" analysis can be stated as follows (adapted from [2, Section 3.2]).

Statement 1 (current blueprint). *Assume that ψ is a sub-root function, i.e., $\psi(r;\delta)/\sqrt{r}$ is non-increasing with respect to $r \in \mathbb{R}_+$. Assume the Bernstein condition $T(f) \le B_e \mathbb{P}[f]$, $B_e > 0$, $\forall f \in \mathcal{F}$. Then with probability at least $1-\delta$, for all $f \in \mathcal{F}$ and $K > 1$,*

$$(\mathbb{P} - \mathbb{P}_n) f \le \frac{1}{K}\mathbb{P} f + \frac{100(K-1)r^*}{B_e}, \tag{2.3}$$

where r^* is the "fixed point" solution of the equation $r = B_e \psi(r;\delta)$.

Since its inception, Statement 1 has become a standard tool in learning theory. However, it requires a rather technical proof, and it appears to be loose when compared with the original assumption (2.2)—ideally, we would like to directly extend (2.2) to hold uniformly without sacrificing any accuracy. Moreover, some assumptions in the statement are restrictive and might not be necessary.

2.2 Key ideas of the "uniform localized convergence" principle.

We provide a surprisingly simple approach which greatly improves and simplifies the current blueprint. While Statement 1 relies heavily on restrictive assumptions like the "sub-root" property of ψ and the Bernstein condition, the following proposition holds essentially without any restrictions.

Proposition 1 (uniform localized convergence). *For function class $\mathcal{F}$ and functional $T : \mathcal{F} \to [0, R]$, assume there is a function $\psi(r;\delta)$ (possibly depending on the samples), which is non-decreasing with respect to r and satisfies that $\forall \delta \in (0,1)$, $\forall r \in (0, R]$, with probability at least $1-\delta$,*

$$\sup_{f \in \mathcal{F}: T(f) \le r} (\mathbb{P} - \mathbb{P}_n) f \le \psi(r;\delta). \tag{2.4}$$

Then, given any $\delta \in (0,1)$ and $r_0 \in (0, R]$, with probability at least $1-\delta$, for all $f \in \mathcal{F}$, either $T(f) \le r_0$ or

$$(\mathbb{P} - \mathbb{P}_n) f \le \psi\bigg(2T(f); \delta(\log_2 \frac{2R}{r_0})^{-1}\bigg). \tag{2.5}$$

The key intuition behind Proposition 1 is that the uniform restatement of the "localized" argument (2.4) is nearly cost-free, because the deviations $(\mathbb{P} - \mathbb{P}_n) f$ can be controlled solely by the real valued functional $T(f)$. As a result, we essentially only require uniform convergence over an interval $[r_0, R]$. The "cost" of this uniform convergence, namely, the additional $\log_2(\frac{2R}{r_0})$ term in (2.5), will only appear in the form $\log(\delta/\log_2(\frac{2R}{r_0}))$ in high-probability bounds, which is of a negligible $O(\log\log n)$ order in general.

Formally, we apply a "peeling" technique: we take $r_k = 2^k r_0$, where $k = 1, 2, \ldots, \lceil \log_2 \frac{R}{r_0} \rceil$, and we use the union bound to extend (2.4) to hold for all these r_k. Then for any $f \in \mathcal{F}$ such that $T(f) > r_0$ is true, there exists a non-negative integer k such that $2^k r_0 < T(f) \le 2^{k+1} r_0$. By the non-decreasing property of the ψ function, we then have

$$(\mathbb{P} - \mathbb{P}_n) f \le \psi\bigg(r_{k+1}; \delta(\log_2 \frac{2R}{r_0})^{-1}\bigg) \le \psi\bigg(2T(f); \delta(\log_2 \frac{2R}{r_0})^{-1}\bigg),$$

which is exactly (2.5). Interestingly, the proof of the classical result (Statement 1) relies on a relatively heavy machinery that includes more complicated peeling and re-weighting arguments (see [2, Section 3.1.4]). However, that analysis obscures the key intuition that we elucidate under inequality (2.5).

The results presented in this paper essentially originate from the noticeable gap between Proposition 1 and Statement 1, illustrated by the following (informal) conclusion:

Statement 2 (improvements over the current blueprint (informal statement)). *Under the assumptions of Statement 1, Proposition 1 provides a strict improvement. In particular, the slower ψ grows, the larger the gap between the bounds in the two results, and the bounds become identical only when ψ is proportional to $\sqrt{r}$, i.e., when the function class $\mathcal{F}$ is parametric and not "rich."*

Formalizing as well as providing rigorous justification for this conclusion is relatively straightforward: taking the "optimal choice" of K in Statement 1, we can re-write its conclusion as

$$(\mathbb{P} - \mathbb{P}_n) f \le 20\sqrt{\frac{r^* \mathbb{P} f}{B_e}} - \frac{r^*}{B_e} \quad \text{[Statement 1]},$$

where the right hand side is of order $\sqrt{r^*\mathbb{P}f/B_e}$ when $\mathbb{P}f > r^*/B_e$, and order r^*/B_e when $\mathbb{P}[f] \le r^*/B_e$. Our result (2.5) is also of order r^*/B_e when $\mathbb{P}f \le r^*/B_e$. However, for every f such that $\mathbb{P}f > r^*/B_e$, it is straightforward to verify that under the assumptions in Statetment 1,

$$\begin{aligned}\psi(2T(f);\delta) &\le \psi(2B_e\mathbb{P}f;\delta) \quad \text{[Bernstein condition: } T(f) \le B_e\mathbb{P}f\text{]} \\ &\le \frac{\sqrt{2B_e\mathbb{P}f}}{\sqrt{r^*}}\psi(r^*;\delta) \quad \text{[}\psi(r;\delta)\text{ is sub-root]} \\ &\le \sqrt{\frac{2r^*\mathbb{P}f}{B_e}} \quad \text{[}r^*\text{ is the fixed point of } B\psi(r;\delta)\text{]}. \end{aligned} \tag{2.6}$$

Therefore, the argument $\psi(2T(f);\delta) \le \sqrt{2r^*\mathbb{P}f/B_e}$ established by (2.6) shows that the "uniform localized convergence" argument (2.5) strictly improves over Statement 1.

3 Preliminaries

Our results on problem-dependent rates essentially only require the loss function to be uniformly bounded by $[-B, B]$, i.e., $|\ell(h;z)| \le B$ for all $h \in \mathcal{H}$ and $z \in \mathcal{Z}$. This is a standard assumption used in almost all previous works that do not invoke curvature conditions or rely on other problem-specific structure. Extensions to unbounded targets can be obtained via truncation techniques (see, e.g. [5]), and our problem-dependent results allow B to be very large, potentially scaling with n.

We represent the complexity through a surrogate function $\psi(r;\delta)$ that satisfies for all $\delta \in (0,1)$,

$$\sup_{f\in\mathcal{F}:\mathbb{P}[f^2]\le r} (\mathbb{P}-\mathbb{P}_n)f \le \psi(r;\delta), \tag{3.1}$$

with probability at least $1-\delta$, where $\mathcal{F}$ is taken to be the *excess loss class*

$$\ell\circ\mathcal{H} - \ell\circ h^* := \{z \mapsto \ell(h;z) - \ell(h^*;z) : h \in \mathcal{H}\}. \tag{3.2}$$

To achieve non-trivial complexity control (and ensure existence of the fixed point), we only consider "meaningful" surrogate functions stated below.

Definition 1 (meaningful surrogate function). A bivariate function $\psi(r;\delta)$ defined over $[0,\infty)\times(0,1)$ is called a meaningful surrogate function if it is non-decreasing, non-negative and bounded with respect to r for every fixed $\delta \in (0,1)$.

We note that the above does not place significant restrictions on the choice of the surrogate function: the left hand side of (3.1) is itself non-decreasing and non-negative; and the boundedness requirement can always be met by setting $\psi(r;\delta) = \psi(4B^2;\delta)$ for all $r \ge 4B^2$. We now give the formal definition of fixed points.

Definition 2 (fixed point). Given a non-decreasing, non-negative and bounded function $\varphi(r)$ defined over $[0,\infty)$, we define the fixed point of $\varphi(r)$ to be $\sup\{r>0 : \varphi(r) \le r\}$. Equivalently, the fixed point of $\varphi(r)$ is the maximal solution to the equation $\varphi(r) = r$.

Given a bounded class $\mathcal{F}$, empirical process theory provides a general way to construct surrogate function by upper bounding the "local Rademacher complexity" $\mathfrak{R}\{f\in\mathcal{F} : \mathbb{P}[f^2] \le r\}$ (see Lemma 4 in Appendix H). We give the definition of Rademacher complexity for completeness.

Definition 3 (Rademacher complexity). For a function class $\mathcal{F}$ that consists of mappings from $\mathcal{Z}$ to $\mathbb{R}$, define

$$\mathfrak{R}\mathcal{F} := \mathbb{E}_{z,\upsilon}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \upsilon_i f(z_i), \quad \mathfrak{R}_n\mathcal{F} := \mathbb{E}_{\upsilon}\sup_{f\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n \upsilon_i f(z_i),$$

as the *Rademacher complexiy* and the *empirical Rademacher complexity* of $\mathcal{F}$, respectively, where $\{\upsilon_i\}_{i=1}^n$ are i.i.d. Rademacher variables for which $\text{Prob}(\upsilon_i = 1) = \text{Prob}(\upsilon_i = -1) = \frac{1}{2}$. $\mathbb{E}_z$ means taking expectations over $\{z_i\}_{i=1}^n$ and $\mathbb{E}_\upsilon$ means taking expectations over $\upsilon_{i\,i=1}^n$.

Furthermore, Dudley's integral bound (Lemma 3 in Appendix H) provides one general solution to construct a *computable* upper bound of local Rademacher complexity via the covering number of $\mathcal{F}$. We give the definition of covering number.

Definition 4 (covering number and metric entropy). A $\varepsilon-$cover of a function class $\mathcal{F}$ with the $L_2(\mathbb{P}_n)$ metric is a set $\{f_1,\dots,f_m\} \subseteq \mathcal{F}$ that satisfies for each $f\in\mathcal{F}$, there exists $i\in\{1,\dots,m\}$ such that $\sqrt{\mathbb{P}_n(f(z)-f_i(z))^2} \le \varepsilon$. The covering number $\mathcal{N}(\varepsilon,\mathcal{F},L_2(\mathbb{P}_n))$ is the cardinality of the smallest $\varepsilon-$cover. We call $\log\mathcal{N}(\varepsilon,\mathcal{F},L_2(\mathbb{P}_n))$ the metric entropy.

4 Loss-dependent rates via empirical risk minimization

In this section we are interested in loss-dependent rates, which should scale tightly with $\mathcal{L}^* := \mathbb{P}[\ell(h^*; z) - \inf_{\mathcal{H}} \ell(h; z)]$; the best achievable "effective loss" on $\mathcal{H}$. The following theorem characterizes the loss-dependent rate of empirical risk minimization (ERM) via a surrogate function ψ, its fixed point r^*, the effective loss $\mathcal{L}^*$ and B.

Theorem 1 (**loss-dependent rate of ERM).** *For the excess loss class $\mathcal{F}$ in* (3.2)*, assume there is a meaningful surrogate function $\psi(r; \delta)$ that satisfies $\forall \delta \in (0, 1)$ and $\forall r > 0$, with probability at least $1 - \delta$,*

$$\sup_{f \in \mathcal{F}: \mathbb{P}[f^2] \le r} (\mathbb{P} - \mathbb{P}_n) f \le \psi(r; \delta).$$

Then the empirical risk minimizer $\hat{h}_{\text{ERM}} \in \arg\min_{\mathcal{H}} \{\mathbb{P}_n \ell(h; z)\}$ satisfies for any fixed $\delta \in (0, 1)$ and $r_0 \in (0, 4B^2)$, with probability at least $1 - \delta$,

$$\mathscr{E}(\hat{h}_{\text{ERM}}) \le \psi\left(24B\mathcal{L}^*; \frac{\delta}{C_{r_0}}\right) \vee \frac{r^*}{6B} \vee \frac{r_0}{24B},$$

where $C_{r_0} = 2\log_2 \frac{8B^2}{r_0}$, and r^* is the fixed point of $6B\psi\left(8r; \frac{\delta}{C_{r_0}}\right)$.

Remarks. 1) The term r_0 is negligible since it can be arbitrarily small. One can simply set $r_0 = B^2/n^4$, which will much smaller than r^* in general (r^* is at least of order $B^2 \log \frac{1}{\delta}/n$ in the traditional "local Rademacher complexity" analysis). In high-probability bounds, C_{r_0} will only appear in the form $\log(C_{r_0}/\delta))$, which is of a negligible $O(\log \log n)$ order, so C_{r_0} can be viewed an absolute constant for all practical purposes. As a result, our generalization error bound can be viewed to be of the order

$$\mathscr{E}(\hat{h}_{\text{ERM}}) \le O\left(\psi(B\mathcal{L}^*; \delta) \vee \frac{r^*}{B}\right). \tag{4.1}$$

2) By using the empirical "effective loss," $\mathbb{P}_n[\ell(\hat{h}_{\text{ERM}}; z) - \inf_{\mathcal{H}} \ell(h; z)]$, to estimate $\mathcal{L}^*$, the loss-dependent rate can be estimated from data without knowledge of $\mathcal{L}^*$. We defer the details to Appendix A.

Comparison to existing results. Under additional restrictions (to be explained later), the traditional analysis (2.3) leads to a loss-dependent rate of the order [2]

$$\mathscr{E}(\hat{h}_{\text{ERM}}) \le O\left(\sqrt{\frac{\mathcal{L}^* r^*}{B}} \vee \frac{r^*}{B}\right), \tag{4.2}$$

which is strictly worse than our result (4.1) due to reasoning following Statement 2. When $B\mathcal{L}^* \le O(r^*)$, both (4.1) and (4.2) are dominated by the order r^*/B so there is no difference between them. However, when $B\mathcal{L}^* \ge \Omega(r^*)$, our result (4.1) will be of order $\psi(B\mathcal{L}^*; \delta)$ and the previous result (4.2) will be of order $\sqrt{\mathcal{L}^* r^*/B}$. In this case, the square-root function $\sqrt{\mathcal{L}^* r^*/B}$ is only a coarse relaxation of $\psi(B\mathcal{L}^*; \delta)$: as the traditional analysis requires ψ to be sub-root, we can compare the two orders by

$$\psi\left(B\mathcal{L}^*; \delta\right) \overset{\text{sub-root}}{\le} \sqrt{\frac{B\mathcal{L}^*}{r^*}} \psi(r^*; \delta) \overset{\text{fixed point}}{=} O\left(\sqrt{\frac{\mathcal{L}^* r^*}{B}}\right). \tag{4.3}$$

The "sub-root" inequality (the first inequality in (4.3)) becomes an equality when $\psi(r; \delta) = O(\sqrt{dr/n})$ in the parametric case, where d is the parametric dimension. However, when $\mathcal{F}$ is rich, $\psi(r; \delta)/\sqrt{r}$ will be strictly decreasing so that the "sub-root" inequality can become quite loose. For example, when $\mathcal{F}$ is a non-parametric class we often have $\psi(r; \delta) = O(\sqrt{r^{1-\rho}/n})$ for some $\rho \in (0, 1)$. The richer $\mathcal{F}$ is (e.g., the larger ρ is), the looser the "sub-root" inequality. This intuition will be validated via examples in Section 6.

Theorem 1 also applies to broader settings than previous results. For example, in [2] it is assumed that the loss is non-negative, and their original result only adapts to $\mathbb{P}\ell(h^*; z)$ rather than the "effective loss" $\mathcal{L}^*$. Our proof (see Appendix D) is quite different as we bypass the Bernstein condition (which is traditionally implied by non-negativity, but not satisfied by the class used here), bypass the sub-root assumption on ψ, and adapt to the "better" parameter $\mathcal{L}^*$.

5 Variance-dependent rates via moment penalization

The loss-dependent rate proved in Theorem 1 contains a complexity parameter $B\mathcal{L}^*$ within its ψ function, which may still be much larger than the optimal variance $\mathcal{V}^*$. Despite its prevalent use in practice, standard empirical risk minimization is unable to achieve variance-dependent rates in general. An example is given in [12] where $\mathcal{V}^* = 0$ and the optimal rate is at most $O(\log n/n)$, while $\mathscr{E}(\hat{h}_{\text{ERM}})$ is proved to be slower than $n^{-\frac{1}{2}}$.

We follow the path of penalizing empirical second moments (or variance) [8, 14, 12, 4] to design an estimator that achieves the "right" bias-variance trade-off for general, potentially "rich," classes. Our proposed estimator simultaneously achieves correct scaling on $\mathcal{V}^*$, along with minimax-optimal sample dependence (n). Besides empirical first and second moments, it only depends on the boundedness parameter B, a computable surrogate function ψ, and the confidence parameter δ. All of these quantities are essentially assumed known in previous works: e.g., [8, 14] require covering number of the loss class, which implies a computable surrogate ψ via Dudley's integral bound; and estimators in [12, 4] rely on the fixed point r^* of a computable surrogate ψ.

In order to adapt to $\mathcal{V}^*$, we use a sample-splitting two-stage estimation procedure (this idea is built on the prior work [4]). Without loss of generality, we assume access to a data set of size $2n$. We split the data set into the "primary" data set S and the "auxiliary" data set S', both of which are of size n. We denote $\mathbb{P}_n$ the empirical distribution of the "primary" data set, and $\mathbb{P}_{S'}$ the empirical distribution of the "auxiliary" data set.

Strategy 1 (the two-stage sample-splitting estimation procedure.). *At the first-stage, we derive a preliminary estimate of $\mathcal{L}_0^* := \mathbb{P}\ell(h^*; z)$ via the "auxiliary" data set S', which we refer to as $\mathcal{L}_{S'}^*$. Then, at the second stage, we perform regularized empirical risk minimization on the "primal" data set S, which penalizes the centered second moment $\mathbb{P}_n[(\ell(h; z) - \mathcal{L}_{S'}^*)^2]$.*

As we will present later, it is rather trivial to obtain a qualified preliminary estimate $\mathcal{L}_{S'}^*$ via empirical risk minimization. Therefore, we firstly introduce the second-stage moment-penalized estimator, which is more crucial and interesting.

Strategy 2 (the second-stage moment-penalized estimator.). *Consider the excess loss class $\mathcal{F}$ in* (3.2)*. Let $\psi(r; \delta)$ be a meaningful surrogate function that satisfies $\forall \delta \in (0, 1)$, $\forall r > 0$, with probability at least $1 - \delta$,*

$$4\mathfrak{R}_n\{f \in \mathcal{F} : \mathbb{P}_n[f^2] \le 2r\} + \sqrt{\frac{2r \log \frac{8}{\delta}}{n}} + \frac{9B \log \frac{8}{\delta}}{n} \le \psi(r; \delta).$$

Denote $C_n = 4 \log_2 n + 10$. Given a fixed $\delta \in (0, 1)$, let the estimator $\hat{h}_{\text{MP}}$ be

$$\hat{h}_{\text{MP}} \in \arg\min_{\mathcal{H}} \left\{ \mathbb{P}_n \ell(h; z) + \psi\left(16\mathbb{P}_n[(\ell(h; z) - \mathcal{L}_{S'}^*)^2]; \frac{\delta}{C_n} \right) \right\}. \tag{5.1}$$

Given an arbitrary preliminary estimate $\mathcal{L}_{S'}^* \in [-B, B]$, we can prove that the generalization error of the moment-penalized estimator $\hat{h}_{\text{MP}}$ is at most

$$\mathscr{E}(\hat{h}_{\text{MP}}) \le 2\psi\left(c_0 \left[\mathcal{V}^* \vee (\mathcal{L}_{S'}^* - \mathcal{L}_0^*)^2 \vee r^* \right]; \frac{\delta}{C_n} \right), \tag{5.2}$$

with probability at least $1 - \delta$, where c_0 is an absolute constant, and r^* is the fixed point of $16B\psi(r; \frac{\delta}{C_n})$. Moreover, the first-stage estimation error will be negligible if

$$(\mathcal{L}_{S'}^* - \mathcal{L}_0^*)^2 \le O(r^*). \tag{5.3}$$

It is rather elementary to show that performing the standard empirical risk minimization on S' suffices to satisfy (5.3), provided an additional assumption that ψ is a "sub-root" function. We now give our theorem on the generalization error following this two-stage procedure.

Theorem 2 (variance-dependent rate). *Let $\mathcal{L}_{S'}^* = \inf_{\mathcal{H}} \mathbb{P}_{S'} \ell(h; z)$ be attained via empirical risk minimization on the auxiliary data set S'. Assume that the meaningful surrogate function $\psi(r; \delta)$ is*

"sub-root," i.e. $\frac{\psi(r;\delta)}{\sqrt{r}}$ *is non-increasing over* $r \in [0, 4B^2]$ *for all fixed* δ*. Then for any* $\delta \in (0, \frac{1}{2})$*, by performing the moment-penalized estimator in Strategy 2, with probability at least* $1 - 2\delta$*,*

$$\mathscr{E}(\hat{h}_{\text{MP}}) \leq 2\psi\left(c_1 \mathcal{V}^*; \frac{\delta}{C_n}\right) \vee \frac{c_1 r^*}{8B},$$

where r^* *is the fixed point of* $B\psi(r; \frac{\delta}{C_n})$ *and* c_1 *is an absolute constant.*

Remarks. 1) In high-probability bounds, C_n will only appear in the form $\log(C_n/\delta))$, which is of a negligible $O(\log\log n)$ order, so there is no much difference to view C_n as an absolute constant.

2) The "sub-root" assumption in Theorem 2 is only used to to bound the first-stage estimation error (see (5.3)). This assumption is not needed for the result (5.2) on the second-stage moment penalized estimator.

3) Replacing $\mathcal{V}^*$ by an empirical centered second moment, we can prove a fully data-dependent generalization error bound that is computable from data without the knowledge of $\mathcal{V}^*$. We leave the full discussion to Appendix A.

Comparison to existing results. The best variance-dependent rate attained by existing estimators is of the order [4]

$$\sqrt{\frac{\mathcal{V}^* r^*}{B^2}} \vee \frac{r^*}{B},$$

which is strictly worse than the rate proved in Theorem 2. The reasoning is similar to Statement 2 and the explanation after Theorem 1: when $\mathcal{V}^* \leq O(r^*)$ the two results are essentially identical, but our estimator can perform much better when $\mathcal{V}^* \geq \Omega(r^*)$. Because ψ is sub-root and r^* is the fixed point, we can compare the orders of the rates

$$\psi(\mathcal{V}^*; \delta) \overset{\text{sub-root}}{\leq} \sqrt{\frac{\mathcal{V}^*}{r^*}}\psi(r^*; \delta) \overset{\text{fixed point}}{=} O\left(\sqrt{\frac{\mathcal{V}^* r^*}{B^2}}\right).$$

Since variance-dependent rates are generally used in applications that require robustness or exhibit large worst-case boundedness parameter, $\mathcal{V}^* \geq r^*$ is the more critical regime where one wants to ensure the estimation performance will not degrade.

Discussion. Per our "uniform localized convergence" principle, the most obvious difficulty in proving Theorem 2 is in establishing (5.2): the empirical second moment is sample-dependent, whereas our Proposition 1 crucially depends on a "measurement functional" (the T functional in Proposition 1) that is unrelated to the samples. The core techniques in the proof essentially overcome this difficulty, and may be of independent interest. We defer details to Appendix E.

The tightness of our variance-dependent rates depend on tightness of the computable surrogate function ψ. When covering numbers of the excess loss class are given, a direct choice is Dudley's integral bound (Lemma 3 in Appendix H), which is known to be rate-optimal for many important classes.

Previous approaches usually take a simper regularization term [8, 4] that is proportional to the square root of the empirical second moment (or empirical variance). That type of penalization is "too aggressive" for rich classes from our viewpoint. [12] propose a regularization term that preserves convexity of empirical risk. However, based on an equivalence proved in their paper, they have similar limitations to the approaches that penalizes the square root of the empirical variance.

6 Discussion and illustrative examples

6.1 Discussion

Recall that our loss-dependent rates and variance-dependent (moment-penalized) rates are of the orders

$$\mathscr{E}(\hat{h}_{\text{ERM}}) \leq O\left(\psi(B\mathcal{L}^*; \delta) \vee \frac{r^*}{B}\right) \quad \text{and} \quad \mathscr{E}(\hat{h}_{\text{MP}}) \leq O\left(\psi(\mathcal{V}^*; \delta) \vee \frac{r^*}{B}\right), \tag{6.1}$$

respectively. In contrast, the best known loss-dependent rates [2] and variance-dependent rates [4] are of the orders

$$\mathscr{E}(\hat{h}_{\text{ERM}}) \le O\left(\sqrt{\frac{\mathcal{L}^* r^*}{B}} \vee \frac{r^*}{B}\right) \quad \text{and} \quad \mathscr{E}(\hat{h}_{\text{previous}}) \le O\left(\sqrt{\frac{\mathcal{V}^* r^*}{B^2}} \vee \frac{r^*}{B}\right), \tag{6.2}$$

respectively (we use $\hat{h}_{\text{previous}}$ to denote the previous best known moment-penalized estimator proposed in [4]). To illustrate the noticeable gaps between our new results and previous known ones, we compare the two different variance-dependent rates in (6.1) and (6.2) on two important families of "rich" classes: non-parametric classes of polynomial growth and VC classes. The implications of this comparison will similarly apply to loss-dependent rates.

Before presenting the advantages of the new problem-dependent rates, we would like to discuss how to compute them. In Theorem 1 and Theorem 2, the class of concentrated functions, $\mathcal{F}$, is the excess loss class $\ell \circ \mathcal{H} - \ell \circ h^*$ in (3.2). As we have mentioned in earlier sections, a general solution for the ψ function is to use Dudley's integral bound (Lemma 3 in Appendix H). Knowledge of the metric entropy of the excess loss class can be used to calculate Dudley's integral bound and construct the surrogate function ψ needed in our theorems. Note that there is no difference between the metric entropy of the excess loss class and metric entropy of the loss class itself: from the definition of covering number and metric entropy, one has

$$\log \mathcal{N}(\varepsilon, \ell \circ \mathcal{H} - \ell \circ h^*, L_2(\mathbb{P}_n)) = \log \mathcal{N}(\varepsilon, \ell \circ \mathcal{H}, L_2(\mathbb{P}_n)).$$

We comment that almost all existing *theoretical* works that discuss general function classes and losses [2, 8, 14, 4] impose metric entropy conditions on the loss class/excess loss class rather than the hypothesis class, and we follows that line as well to allow for a seamless comparison of the results. As a complement, we will discuss how to obtain such metric entropy conditions for practical applications in Appendix B.

6.2 Non-parametric classes of polynomial growth

Example 1 (non-parametric classes of polynomial growth). Consider a loss class $\ell \circ \mathcal{H}$ with the metric entropy condition

$$\log \mathcal{N}(\varepsilon, \ell \circ \mathcal{H}, L_2(\mathbb{P}_n)) \le O\left(\varepsilon^{-2\rho}\right), \tag{6.3}$$

where $\rho \in (0, 1)$ is a constant. Using Dudley's integral bound to find ψ and solving $r \le O\left(B\psi(r; \delta)\right)$, it is not hard to verify that

$$\psi(r; \delta) \le O\left(\sqrt{\frac{r^{1-\rho}}{n}}\right), \quad r^* \le O\left(\frac{B^{\frac{2}{1+\rho}}}{n^{\frac{1}{1+\rho}}}\right).$$

As a result, our variance-dependent rate (6.1) is of the order

$$\mathscr{E}(\hat{h}_{\text{MP}}) \le O\left(\mathcal{V}^{*\frac{1-\rho}{2}} n^{-\frac{1}{2}} \vee \frac{r^*}{B}\right), \tag{6.4}$$

which is $O\left(\mathcal{V}^{*\frac{1-\rho}{2}} n^{-\frac{1}{2}}\right)$ when $\mathcal{V}^* \ge \Omega(r^*)$. In contrast, the previous best-known result (6.2) is of the order

$$\mathscr{E}(\hat{h}_{\text{previous}}) \le O\left(\sqrt{\mathcal{V}^*} B^{-\frac{\rho}{1+\rho}} n^{-\frac{1}{2+2\rho}} \vee \frac{r^*}{B}\right), \tag{6.5}$$

which is $O\left(\sqrt{\mathcal{V}^*} B^{-\frac{\rho}{1+\rho}} n^{-\frac{1}{2+2\rho}}\right)$ when $\mathcal{V}^* \ge \Omega\left(r^*\right)$. Therefore, for arbitrary choice of $n, \mathcal{V}^*, B$, the "sub-optimality gap" is

$$\text{ratio between (6.5) and (6.4)} := \frac{\sqrt{\mathcal{V}^*} B^{-\frac{\rho}{1+\rho}} n^{-\frac{1}{2+2\rho}} \vee \frac{r^*}{B}}{\mathcal{V}^{*\frac{1-\rho}{2}} n^{-\frac{1}{2}} \vee \frac{r^*}{B}} = 1 \vee (\mathcal{V}^* (\frac{n}{B^2})^{\frac{1}{1+\rho}})^{\frac{\rho}{2}}, \tag{6.6}$$

which can be arbitrary large and grows polynomially with n.

We consider two stylized regimes as follows (we use the notation $\approx$ when the left hand side and the right hand side are of the same order).

- The more "traditional" regime: $B \approx 1$, $\mathcal{V}^* \approx n^{-a}$ where $a > 0$ is a fixed constant. This regime captures the traditional supervised learning problems where B is not large, but one wants to use the relatively small order of $\mathcal{V}^*$ to achieve "faster" rates.
- The "high-risk" regime: $B \approx n^b$ where $b > 0$ is a fixed constant, and $\mathcal{V}^* \ll B^2$ (i.e., $\mathcal{V}^*$ is much smaller than order n^{2b}). This regime captures modern "high-risk" learning problems such as counterfactual risk minimization [14], policy learning [1], and supervised learning with limited number of samples. In those settings, the worst-case boundedness parameter is considered to scale with n so that one wants to avoid (or reduce) the dependence on B.

In both the two regimes, generalization errors via naive (non-localized) uniform convergence arguments will be worse than our approach by orders polynomial in n, so we only need to compare with previous variance-dependent rates.

The "traditional" regime. The "sub-optimality gap" (6.6) is $1 \vee (\mathcal{V}^* n^{\frac{1}{1+\rho}})^{\frac{\rho}{2}}$. It is quite clear that when $\mathcal{V}^* \approx n^{-a}$ where $0 < a < \frac{1}{1+\rho}$, our variance-dependent rate improves over all previous generalization error rates by orders polynomial in n.

The "high-risk" regime. We restrict our attention to the simple case $B^{\frac{2}{1+\rho}} \leq \mathcal{V}^* \ll 4B^2$ to gain some insight, where our result exhibits an improvement of order $O(n^{\frac{\rho}{2}(\frac{1}{1+\rho})})$ relative to the previous result. Clearly the larger ρ, the more improvement we provide. By letting $\rho \to 1$ our improvement can be as large as $O(n^{\frac{1}{4}})$.

6.3 VC-type classes

Our next example considers VC-type classes. Although this classical example has been extensively studied in learning theory, our results provide strict improvements over antecedents.

Example 2 (VC-type classes). One general definition of VC-type classes (which is not necessarily binary) uses the metric entropy condition. Consider a loss class $\ell \circ \mathcal{H}$ that satisfies

$$\log \mathcal{N}(\varepsilon, \ell \circ \mathcal{H}, L_2(\mathbb{P}_n)) \leq O\left(d \log \frac{1}{\varepsilon}\right),$$

where d is th so-called the Vapnik–Chervonenkis (VC) dimension [15]. Using Dudley's integral bound to find the surrogate ψ and solving $r \leq O(B\psi(r;\delta))$, it can be proven [6] that

$$\psi(r;\delta) \leq O\left(\sqrt{\frac{dr}{n} \log \frac{8B^2}{r}} \vee \frac{Bd}{n} \log \frac{8B^2}{r}\right), \quad r^* \leq O\left(\frac{B^2 d \log n}{n}\right).$$

Recently, [4] proposed a moment-penalized estimator whose generalization error is of the rate

$$\mathscr{E}(\hat{h}_{\text{previous}}) \leq O\left(\sqrt{\frac{d\mathcal{V}^* \log n}{n}} + \frac{Bd \log n}{n}\right),$$

in the worst case without invoking other assumptions. This result has a $O(\log n)$ gap compared with the $\Omega(\sqrt{\frac{d\mathcal{V}^*}{n}})$ lower bound [3], which holds for arbitrary sample size. There is much recent interest focused on when the sub-optimal $\log n$ factor can be removed [1, 4].

By applying Theorem 2, our refined moment-penalized estimator gives a generalization error bound of tighter rate

$$\mathscr{E}(\hat{h}_{\text{MP}}) \leq O\left(\sqrt{\frac{d\mathcal{V}^* \log \frac{8B^2}{\mathcal{V}^*}}{n}} \vee \frac{Bd \log n}{n}\right). \tag{6.7}$$

This closes the $O(\log n)$ gap in the regime $\mathcal{V}^* \geq \Omega(\frac{B^2}{(\log n)^\alpha})$, where $\alpha > 0$ is an arbitrary positive constant. Though this is not the central regime, it is the first positive result that closes the notorious $O(\log n)$ gap without invoking any additional assumptions on the loss/hypothesis class (e.g., the rather complex "capacity function" assumption introduced in [4]). We anticipate additional improvements are possible under further assumptions on the hypothesis class and the loss function.

Broader Impact

This work is theoretical and does not present any foreseeable ethical consequence to the society.

References

[1] Susan Athey and Stefan Wager. Efficient policy learning. *arXiv preprint arXiv:1702.02896*, 2017.

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

[3] Luc Devroye and Gábor Lugosi. Lower bounds in pattern recognition and learning. *Pattern recognition*, 28(7):1011–1018, 1995.

[4] Dylan J Foster and Vasilis Syrgkanis. Orthogonal statistical learning. *arXiv preprint arXiv:1901.09036*, 2019.

[5] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[6] Vladimir Koltchinskii and Dmitriy Panchenko. Rademacher processes and bounding the risk of function learning. In *High dimensional probability II*, pages 443–457. Springer, 2000.

[7] Tengyuan Liang, Alexander Rakhlin, and Karthik Sridharan. Learning with square loss: Localization through offset rademacher complexity. In *Conference on Learning Theory*, pages 1260–1285, 2015.

[8] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

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

[10] Shahar Mendelson. Learning without concentration. In *Conference on Learning Theory*, pages 25–39, 2014.

[11] Shahar Mendelson. Learning without concentration for general loss functions. *Probability Theory and Related Fields*, 171(1-2):459–502, 2018.

[12] Hongseok Namkoong and John C Duchi. Variance-based regularization with convex objectives. In *Advances in Neural Information Processing Systems*, pages 2971–2980, 2017.

[13] Karthik Sridharan. Note on refined dudley integral covering number bound. *Unpublished Manuscript, https://www.cs.cornell.edu/ sridharan/dudley.pdf*, 2010.

[14] Adith Swaminathan and Thorsten Joachims. Counterfactual risk minimization: Learning from logged bandit feedback. In *International Conference on Machine Learning*, pages 814–823, 2015.

[15] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[16] Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

A Estimating the loss-dependent and variance-dependent rates from data

In this section we present problem-dependent bounds that can be computed from data, where the unknown quantity $\mathcal{L}^*$ and $\mathcal{V}^*$ are replaced by some empirical estimates.

Corollary 3 (estimate of the loss-dependent rate from data). *Recall the term $\mathcal{L}^*$ is $\mathbb{P}[\ell(h^*;z) - \inf_{\mathcal{H}} \ell(h^*;z)]$ and denote $\widehat{\mathcal{L}^*} = \mathbb{P}_n[\ell(\hat{h}_{\text{ERM}};z) - \inf_{\mathcal{H}} \ell(h;z)]$. Under the conditions of Theorem 1, setting $C_n = 2\log_2 n + 6$, then for any fixed $\delta \in (0, \frac{1}{2})$, with probability at least $1-2\delta$, we have*

$$\mathscr{E}(\hat{h}_{\text{ERM}}) \le \psi\left(cB\widehat{\mathcal{L}^*}; \frac{\delta}{C_n}\right) \vee \frac{cr^*}{B} \vee \frac{cB\log\frac{2}{\delta}}{n} \tag{A.1}$$

and

$$\mathcal{L}^* \le c_1\left(\widehat{\mathcal{L}^*} \vee \frac{r^*}{B} \vee \frac{B\log\frac{2}{\delta}}{n}\right) \le c_2\left(\mathcal{L}^* \vee \frac{r^*}{B} \vee \frac{B\log\frac{2}{\delta}}{n}\right), \tag{A.2}$$

where c, c_1, c_2 are absolute constants.

Remarks. 1) The $B\log\frac{2}{\delta}/n$ terms (A.1) and (A.2) are negligible, because r^* is at least of order $B^2\log\frac{1}{\delta}/n$ for most practical applications. This order is unavoidable in traditional "local Rademacher complexity" analysis and two-sided concentration inequalities.

2) The generalization error bound (A.1) shows that without knowledge of $\mathcal{L}^*$, one can estimate the order of our loss-dependent rate by using $\widehat{\mathcal{L}^*} = \mathbb{P}_n[\ell(\hat{h}_{\text{ERM}};z) - \inf_{\mathcal{H}} \ell(h;z)]$ as a proxy. Despite replacing $\mathcal{L}^*$ by $\widehat{\mathcal{L}^*}$, other quantities in the bound remain unchanged in order.

3) The inequality (A.2) shows that the estimation of $\mathcal{L}^*$ is tight.

Corollary 4 (estimate of the variance-dependent rate from data). *Consider the empirical centered second moment*

$$\widehat{\mathcal{V}^*} := \mathbb{P}_n\left[\ell(\hat{h}_{\text{NMP}};z) - \widehat{\mathcal{L}_0^*})^2\right],$$

where $\mathcal{L}^*_{S'} \in [-B, B]$ is the preliminary estimate of $\mathcal{L}^*$ obtained in the first-stage, ψ is defined in Strategy 2, and

$$\hat{h}_{\text{NMP}} \in \arg\min_{\mathcal{H}} \mathbb{P}_n\ell(h;z) - 2\psi\left(16\mathbb{P}_n\left[(\ell(h;z) - \widehat{\mathcal{L}_0^*})^2\right]\right).$$

For any fixed $\delta \in (0,1)$, by performing the moment-penalized estimator in Strategy 2, with probability at least $1-\delta$,

$$\mathscr{E}(\hat{h}_{\text{MP}}) \le 4\psi\left(16\widehat{\mathcal{V}^*}; \frac{\delta}{C_n}\right) \vee \frac{r^*}{8B}, \tag{A.3}$$

where r^* is the fixed point of $16B\psi(r; \frac{\delta}{C_n})$.

Remarks. 1) The subscript "NMP" within $\hat{h}_{\text{NMP}}$ means "negative moment penalization." Note that $\hat{h}_{\text{NMP}}$ may not have good generalization performance, it is only used to compute $\widehat{\mathcal{V}^*}$ so that we can evaluate the estimator $\hat{h}_{\text{MP}}$ proposed in Strategy 2.

2) While the fully data-dependent generalization error bound (A.3) provides a way to evaluate the moment-penalized estimator in Strategy 2 from training data, it seems that $\widehat{\mathcal{V}^*}$ and $\mathcal{V}^*$ are not necessarily of the same order. Therefore, (A.3) may not be as tight as the original variance-dependent rate in Theorem 2. One should view (A.3) as a relaxation of the original variance-dependent rate in Theorem 2.

3) We also comment that the "sub-root" assumption in Theorem 2 is not needed here as we do not discuss the precision of $\widehat{\mathcal{L}_0^*}$. It is easy to combine Corollary 4 with the guarantee on $\widehat{\mathcal{L}_0^*}$ proved in Appendix E.2.

B Application areas of problem-dependent rates

In Section 6 we illustrate the advantages of our problem-dependent rates in "rich" non-parametric and VC classes, where we use metric entropy conditions of the loss/excess loss class. In practical applications it is more standard to consider metric entropy conditions of the hypothesis class $\mathcal{H}$. In view of this, we introduce three important settings where the metric entropy on the loss/excess loss class can be obtained from metric entropy conditions on the hypothesis class $\mathcal{H}$.

Supervised learning with Lipchitz continuous cost. In supervised learning, the data z is a feature-label pair (x, y), and the loss $\ell(h; z)$ is of the form

$$\ell(h; z) = \ell_{\mathtt{sv}}(h(x), y),$$

where $\ell_{\mathtt{sv}}$ is a fixed cost function that is $L_{\mathtt{sv}}-$Lipchitz continuous with respect to its first argument, namely, Lipchitz with parameter $L_{\mathtt{sv}}$. For hypothesis classes characterized by metric entropy conditions, properties are preserved because

$$\log \mathcal{N}(\varepsilon, \ell \circ \mathcal{H}, L_2(\mathbb{P}_n)) \leq \log \mathcal{N}(\frac{\varepsilon}{L_{\mathtt{sv}}}, \mathcal{H}, L_2(\mathbb{P}_n)).$$

Note that $L_{\mathtt{sv}}$ only depends on the cost function and is usually of constant order. Our theory naturally applies to supervised learning problems where the cost function is Lipchitz continuous and not strongly-convex (for example, the ℓ_1 cost, the hinge cost, the ramp cost, etc.).

Counterfactual risk minimization. Denote $x \in \mathcal{X}$ the feature and $t \in \mathcal{T}$ the treatment (e.g. $\mathcal{T} = \{0, 1\}$ in binary treatment experimental design), and $c(x, t)$ the unknown cost function. A hypothesis (policy) h is a map from $\mathcal{X} \times \mathcal{T}$ to $[0, 1]$ such that $\sum_{t\in\mathcal{T}} h(x, t) = 1$. Thus, a hypothesis (policy) essentially maps features to a distribution over treatments. We consider the standard formulation of "learning with logged bandit feedback," dubbed "counterfactual risk minimization" [14]: a batch of samples $\{(x_i, t_i, c_i)\}_{i=1}^n$ are obtained by applying a known policy h_0, so that t_i is sampled from $h_0(x_i, \cdot)$ and one can only observe the cost c_i associated with t_i. We write $z = (x, t, c)$ and let

$$\ell(h; z_i) = \frac{c_i}{h_0(x_i, t_i)} h(x_i, t_i), \tag{B.1}$$

be the "constructed loss" using importance sampling. It is straightforward to show that the population risk $\mathbb{P}\ell(h; z)$ is equal to the expected cost of policy h, so determining good policies requires one to minimize the generalization error $\mathscr{E}(\hat{h})$. It is usually convenient to obtain metric entropy condition of the loss/excess loss class by using the linearity structure of (B.1). In particular, from the Cauchy-Schwartz inequality we can prove that

$$\log \mathcal{N}(\varepsilon, \ell \circ \mathcal{H}, L_2(\mathbb{P}_n)) \leq \mathcal{N}(\frac{\varepsilon}{\gamma_n}, \mathcal{H}, L_4(\mathbb{P}_n)), \tag{B.2}$$

where $\gamma_n := \sqrt[4]{\mathbb{P}_n \left[(\frac{c(x,t)}{h_0(x,t)})^4 \right]}$ only depends on the functions c, h_0 in the given problem, and the samples rather than the worst-case parameters. A systematical challenge in counterfactual risk minimization is that the worst-case boundedness parameter, $\sup_{h,z} |\ell(h; z)|$, is typically very large, since the inverse probability term $\frac{1}{h_0(x_i,t_i)}$ in (B.1) is typically large in the worst case.

Causal inference with observational data. When one only observes the samples $\{(x_i, t_i, c_i\}_{i=1}^n$ but the policy h_0 used to generate them is not known, counterfactual risk minimization becomes more challenging. This task is referred to as "policy learning with observational data" [1]. More broadly, the recent work [4] establishes a unified framework that covers policy learning and other causal inference problems under the framework "orthogonal statistical learning." Consider the problem to minimize the generalization error of the "unknown loss" $\beta(z, g^*)h(z)$ without knowledge of the nuisance function g^*. For example, g^* is the unknown past policy h_0 in policy learning. The universal principle here is sample splitting and "plug-in" estimation: the learner uses a part of the samples to obtain an estimate $\hat{g}$, and use the remaining samples to learn the best policy through the "constructed loss"

$$\ell(h; z) = \beta(z, \hat{g})h(z).$$

In [4, Theorem 2], it is shown that any generalization error on the "constructed loss" $\ell(h;z)$ can be converted to the generalization error on the "unknown loss" $\beta(z,g^*)h(z)$, under certain regularity conditions. Hence, the generalization error results in our paper are directly applicable to many causal inference problems covered in [4]. Similar to (B.2), it is usually convenient to obtain metric entropy conditions of the loss/excess loss class by using the linearity of the loss function. Again, a central challenge in these applications is that the worst-case boundedness parameter is typically very large since $\ell(h;z)$ is the counterfactual outcome constructed by importance sampling techniques.

In the following sections, we will present proofs of the theoretical results. In all the proofs we consider a fixed sample size n. In order to distinguish "probability of events" and "expectation with respect to $\mathbb{P}$," we will use the notation $\text{Prob}(\mathscr{A})$ to denote the probability of the event $\mathscr{A}$ (as a substitute to $\mathbb{P}(\mathscr{A})$).

C Proof of Proposition 1

Given any $r_0 \in (0, R]$, take $r_k = 2^k r_0$, $k = 1, \cdots, \lceil \log_2 \frac{R}{r_0} \rceil$. Note that $\lceil \log_2 \frac{R}{r_0} \rceil \le \log_2 \frac{2R}{r_0}$.

We use a union bound to establish that $\sup_{T(f)\le r}(\mathbb{P}-\mathbb{P}_n)f \le \psi(r;\delta)$ holds for all these r_k simultaneously: $\forall \delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{T(f)\le r_k} (\mathbb{P}-\mathbb{P}_n)f \le \psi\left(r_k; \frac{\delta}{\log_2 \frac{2R}{r_0}}\right), \quad k = 1, \cdots, \left\lceil \log_2 \frac{R}{r_0} \right\rceil.$$

For any fixed $f \in \mathcal{F}$, if $T(f) \le r_0$ is false, then let k be the non-negative integer such that $2^k r_0 < T(f) \le 2^{k+1} r_0$, we further know that $r_{k+1} = 2^{k+1} r_0 \le 2T(f)$. Therefore, with probability at least $1-\delta$,

$$\begin{aligned}(\mathbb{P}-\mathbb{P}_n)f &\le \sup_{\tilde{f}\in\mathcal{F}: T(\tilde{f})\le r_{k+1}} (\mathbb{P}-\mathbb{P}_n)\tilde{f} \\ &\le \psi\left(r_{k+1}; \frac{\delta}{\log_2 \frac{2R}{r_0}}\right) \\ &\le \psi\left(2T(f); \frac{\delta}{\log_2 \frac{2R}{r_0}}\right).\end{aligned}$$

Therefore, with probability at least $1-\delta$, $\forall f \in \mathcal{F}$, either $T(f) \le r_0$ or

$$(\mathbb{P}-\mathbb{P}_n)f \le \psi\left(2T(f); \frac{\delta}{\log_2 \frac{2R}{r_0}}\right).$$

This completes the proof. □

D Proof of Theorem 1

Let $\mathcal{F}$ be the excess loss class in (3.2). Clearly, its members f are uniformly bounded in $[-2B, 2B]$. Let $T(f) = \mathbb{P}[f^2]$. Define $\hat{f}$ by $\hat{f}(z) = \ell(\hat{h}_{\text{ERM}}; z) - \ell(h^*; z), \forall z \in \mathcal{Z}$.

For a fixed $r_0 \in (0, 4B^2)$, Denote $C_{r_0} = 2\log_2 \frac{8B^2}{r_0}$. Then from Proposition 1 we know with probability at least $1 - \frac{\delta}{2}$, either $T(\hat{f}) \le r_0$ or

$$(\mathbb{P}-\mathbb{P}_n)\hat{f} \le \psi\left(2T(\hat{f}); \frac{\frac{\delta}{2}}{\log_2 \frac{8B^2}{r_0}}\right) = \psi\left(2T(\hat{f}); \frac{\delta}{C_{r_0}}\right). \tag{D.1}$$

We denote the events $\mathscr{A}_1 = \{T(\hat{f}) \le r_0\}$ and $\mathscr{A}_2 = \{\text{inequality (D.1) holds true}\}$, then we have

$$\text{Prob}(\mathscr{A}_1) + \text{Prob}(\mathscr{A}_1) \ge 1 - \frac{\delta}{2}.$$

Consider the event $\mathscr{A}_1$. From the surrogate property of ψ, we have

$$\text{Prob}\left(\mathscr{A}_1 \cap \left\{(\mathbb{P}-\mathbb{P}_n)\hat{f} \le \psi(2r_0; \frac{\delta}{C_{r_0}})\right\}\right) \ge \text{Prob}(\mathscr{A}_1) - \frac{\delta}{C_{r_0}} \ge \text{Prob}(\mathscr{A}_1) - \frac{\delta}{2}.$$

Combine the events $\mathscr{A}_1$ and $\mathscr{A}_2$, we have

$$\text{Prob}\left(\left\{(\mathbb{P}-\mathbb{P}_n)\hat{f} \le \psi\left(2T(\hat{f}) \vee 2r_0; \frac{\delta}{C_{r_0}}\right)\right\}\right) \ge \text{Prob}(\mathscr{A}_1) - \frac{\delta}{2} + \text{Prob}(\mathscr{A}_2) \ge 1-\delta.$$

From the property of ERM we have $\mathbb{P}_n\hat{f} \le 0$, so with probability at least $1-\delta$,

$$\mathscr{E}(\hat{h}_{\text{ERM}}) \le (\mathbb{P}-\mathbb{P}_n)\hat{f} \le \psi\left(2T(\hat{f}) \vee 2r_0; \frac{\delta}{C_{r_0}}\right). \tag{D.2}$$

From now to the end of this proof, we will prove the generalization error bound on the event

$$\mathscr{A} = \{\text{the inequality (D.2) holds true}\}, \tag{D.3}$$

whose measure is at least $1-\delta$. Define $\hat{g}$ by $\hat{g}(z) = \ell(\hat{h}_{\text{ERM}}; z) - \inf_{\mathcal{H}} \ell(h; z), \forall z \in \mathcal{Z}$. Let $T(\hat{g}) = \mathbb{P}[\hat{g}^2]$. We have $\hat{f}(z) = \hat{g}(z) - (\ell(h^*; z) - \inf_{\mathcal{H}} \ell(h; z)), \forall z$ so that

$$\begin{aligned}\mathbb{P}[\hat{f}^2] &\le 2\mathbb{P}[\hat{g}^2] + 2\mathbb{P}[(\ell(h^*; z) - \inf_{\mathcal{H}} \ell(h; z))^2] \\ &\le 2\mathbb{P}[\hat{g}^2] + 4B\mathcal{L}^* \le 4\mathbb{P}[g^2] \vee 8B\mathcal{L}^*.\end{aligned}$$

That is,

$$T(\hat{f}) \le 4T(\hat{g}) \vee 8B\mathcal{L}^*. \tag{D.4}$$

From (D.2) and (D.4) we have

$$\mathbb{P}\hat{g} - \mathcal{L}^* = \mathscr{E}(\hat{h}_{\text{ERM}}) \le \psi\left(8T(\hat{g}) \vee 16B\mathcal{L}^* \vee 2r_0; \frac{\delta}{C_{r_0}}\right). \tag{D.5}$$

Since $\hat{g}(z) \in [0, 2B]$ for all z, we have $T(\hat{g}) \le 2B\mathbb{P}\hat{g}$. From this fact and (D.5) we obtain

$$\begin{aligned}T(\hat{g}) &\le 2B\mathbb{P}\hat{g} \\ &\le 2B\left(\mathcal{L}^* + \psi\left(8T(\hat{g}) \vee 16B\mathcal{L}^* \vee 2r_0; \frac{\delta}{C_{r_0}}\right)\right) \\ &= 2B\mathcal{L}^* + 2B\psi\left(8T(\hat{g}) \vee 16B\mathcal{L}^* \vee 2r_0; \frac{\delta}{C_{r_0}}\right).\end{aligned}$$

Whether $B\mathcal{L}^*$ is less than $2B\psi\left(8T(\hat{g}) \vee 16B\mathcal{L}^* \vee 2r_0; \frac{\delta}{C_{r_0}}\right)$, or $B\mathcal{L}^*$ is greater or equal to $2B\psi\left(8T(\hat{g}) \vee 16B\mathcal{L}^* \vee 2r_0; \frac{\delta}{C_{r_0}}\right)$, the above inequality always implies that

$$\begin{aligned}T(\hat{g}) &\le 3B\mathcal{L}^* \vee 6B\psi\left(8T(\hat{g}) \vee 16B\mathcal{L}^* \vee 2r_0; \frac{\delta}{C_{r_0}}\right) \\ &\le 3B\mathcal{L}^* \vee 6B\psi\left(8T(\hat{g}); \frac{\delta}{C_{r_0}}\right) \vee 6B\psi\left(16B\mathcal{L}^* \vee 2r_0; \frac{\delta}{C_{r_0}}\right).\end{aligned} \tag{D.6}$$

Let r^* be the fixed point of $6B\psi(8r; \frac{\delta}{C_n})$. From the definition of fixed points whether $2B\mathcal{L}^* \vee \frac{r_0}{4} \le r^*$ or $2B\mathcal{L}^* \vee \frac{r_0}{4} > r^*$, we always have

$$6B\psi\left(16B\mathcal{L}^* \vee 2r_0; \frac{\delta}{C_{r_0}}\right) \le r^* \vee 2B\mathcal{L}^* \vee \frac{r_0}{4}.$$

Combine the above inequality with (D.6), we have

$$T(\hat{g}) \le 3B\mathcal{L}^* \vee 6B\psi\left(8T(\hat{g}); \frac{\delta}{C_{r_0}}\right) \vee r^* \vee \frac{r_0}{4}.$$

From the above inequality and again the definition of fixed points, it is straightforward to prove that

$$T(\hat{g}) \leq 3B\mathcal{L}^* \vee r^* \vee \frac{r_0}{4}.$$

Combining the above inequality with (D.4), we have

$$T(\hat{f}) \leq 12B\mathcal{L}^* \vee 4r^* \vee r_0.$$

From the above inequality and (D.2) we have

$$\mathscr{E}(\hat{h}_{\text{ERM}}) \leq (\mathbb{P} - \mathbb{P}_n)\hat{f} \leq \psi\left(24B\mathcal{L}^* \vee 8r^* \vee 2r_0; \frac{\delta}{C_{r_0}}\right), \tag{D.7}$$

which implies that

$$\mathscr{E}(\hat{h}_{\text{ERM}}) \leq \psi\left(24B\mathcal{L}^*; \frac{\delta}{C_{r_0}}\right) \vee \psi\left(8r^* \vee 2r_0; \frac{\delta}{C_{r_0}}\right).$$

Recall that r^* is the fixed point of $6B\psi(8r; \frac{\delta}{C_{r_0}})$. Since $r^* \vee \frac{r_0}{4} \geq r^*$, from the definition of fixed points we have

$$6B\psi(8r^* \vee 2r_0; \frac{\delta}{C_{r_0}}) \leq r^* \vee \frac{r_0}{4}.$$

So we finally obtain

$$\mathscr{E}(\hat{h}_{\text{ERM}}) \leq \psi\left(24B\mathcal{L}^*; \frac{\delta}{C_{r_0}}\right) \vee \frac{r^*}{6B} \vee \frac{r_0}{24B}.$$

Recall that the generalization error bound holds true on the event $\mathscr{A}$ defined in (D.3), whose measure is at least $1 - \delta$. This completes the proof. □

E Proof of Theorem 2

The main goal of this subsection is to prove Theorem 2. We first prove Theorem 5 (the bound (5.2) in the main paper), a guarantee for the second-stage moment penalized estimator $\hat{h}_{\text{MP}}$. In order to prove Theorem 2, we then combine Theorem 5 with a guarantee for the first-stage empirical risk minimization (ERM) estimator.

E.1 Analysis for the second-stage moment-penalized estimator

Theorem 5 (variance-dependent rate of the second-stage estimator). *Given arbitrary preliminary estimate $\mathcal{L}^*_{S'} \in [-B, B]$, the generalization error of the moment-penalized estimator $\hat{h}_{\text{MP}}$ in Strategy 2 is bounded by*

$$\mathscr{E}(\hat{h}_{\text{MP}}) \leq 2\psi\left(c_0\left[\mathcal{V}^* \vee (\mathcal{L}^*_{S'} - \mathcal{L}^*_0)^2 \vee r^*\right]; \frac{\delta}{C_n}\right),$$

with probability at least $1 - \delta$, where c_0 is an absolute constant and r^* is the fixed point of $16B\psi(r; \frac{\delta}{C_n})$.

Proof of Theorem 5: the proof of Theorem 5 consist of four parts.

Part I: use ψ to upper bound localized empirical processes

Lemma 1 (bound on localized empirical processes). *Given a fixed $\delta_1 \in (0, 1)$, let $r_1^*(\delta_1)$ be the fixed point of $16B\psi(r; \delta_1)$ where ψ is defined in Strategy 2. Then with probability at least $1 - \delta_1$, for all $r > 0$,*

$$\sup_{\mathbb{P}[f^2] \leq r} (\mathbb{P} - \mathbb{P}_n)f \leq \psi\left(r \vee r_1^*(\delta_1); \delta_1\right). \tag{E.1}$$

Proof of Lemma 1: Recall that $\mathcal{F}$ is the excess loss class in (3.2). Clearly, its members f are uniformly bounded in $[-2B, 2B]$. When $\mathbb{P}[f^2] \le r$, we have $\mathbb{P}[f^4] \le 4B^2 r$. From Lemma 4 (the two-sided version of its second inequality), with probability at least $1 - \frac{\delta_1}{2}$,

$$\begin{aligned}
&\sup_{\mathbb{P}[f^2]\le r} \left|(\mathbb{P} - \mathbb{P}_n) f^2\right| \\
&\le 4\mathfrak{R}_n\{f^2 : \mathbb{P}[f^2] \le r\} + 2B\sqrt{\frac{2r\log\frac{8}{\delta_1}}{n}} + \frac{18B^2\log\frac{8}{\delta_1}}{n} \\
&\le 16B\mathfrak{R}_n\{f : \mathbb{P}[f^2] \le r\} + 2B\sqrt{\frac{2r\log\frac{8}{\delta_1}}{n}} + \frac{18B^2\log\frac{8}{\delta_1}}{n},
\end{aligned}$$

where the last inequality follows from the Lipchitz contraction property of Rademahcer complexity (see, e.g., [9, Theorem 7]), and the fact that for all $f_1, f_2 \in \mathcal{F}$, $|f_1^2(z) - f_2^2(z)| \le 4B|f_1(z) - f_2(z)|$. We conclude that with probability at least $1 - \frac{\delta_1}{2}$,

$$\sup_{\mathbb{P}[f^2]\le r} \left|(\mathbb{P} - \mathbb{P}_n) f^2\right| \le \varphi_{\delta_1}(r), \tag{E.2}$$

where $\varphi_{\delta_1}(r) := 16B\mathfrak{R}_n\{f : \mathbb{P}[f^2] \le r\} + 2B\sqrt{\frac{2r\log\frac{8}{\delta_1}}{n}} + \frac{18B^2\log\frac{8}{\delta_1}}{n}$.

Denote $r_2^*(\delta_1)$ the fixed point of $4\varphi_{\delta_1}(r)$ (the fixed point must exist as $4\varphi_{\delta_1}(r)$ is a non-decreasing, non-negative and bounded function). From (E.2) and the fact that $r_2^*(\delta_1)$ is the fixed point of $4\varphi_{\delta_1}(r)$, if $r > r_2^*(\delta_1)$, then with probability at least $1 - \frac{\delta_1}{2}$,

$$\sup_{\mathbb{P}[f^2]\le r} \left|(\mathbb{P} - \mathbb{P}_n) f^2\right| \le \frac{r}{4}. \tag{E.3}$$

(E.3) implies that with probability at least $1 - \frac{\delta_1}{2}$, for all $r > r_2^*(\delta_1)$, $\mathbb{P}[f^2] \le r$ implies that

$$\mathbb{P}_n[f^2] \le \frac{5}{4}r \le 2r. \tag{E.4}$$

Again from the two-sided version of the second inequality in Lemma 4, we know that with probability at least $1 - \frac{\delta_1}{2}$,

$$\sup_{\mathbb{P}[f^2]\le r} |(\mathbb{P} - \mathbb{P}_n) f| \le 4\mathfrak{R}_n\{f : \mathbb{P}[f^2] \le r\} + \sqrt{\frac{2r\log\frac{8}{\delta_1}}{n}} + \frac{9B\log\frac{8}{\delta_1}}{n}.$$

Combining the above inequality and (E.4) using a union bound, we know that with probability at least $1 - \frac{\delta_1}{2} - \frac{\delta_1}{2} = 1 - \delta_1$, if $r > r_2^*(\delta_1)$, then

$$\begin{aligned}
\sup_{\mathbb{P}[f^2]\le r} (\mathbb{P} - \mathbb{P}_n) f &\le 4\mathfrak{R}_n\{f : \mathbb{P}[f^2] \le r\} + \sqrt{\frac{2r\log\frac{8}{\delta_1}}{n}} + \frac{9B\log\frac{8}{\delta_1}}{n} \\
&\le 4\mathfrak{R}_n\{f : \mathbb{P}_n[f^2] \le 2r\} + \sqrt{\frac{2r\log\frac{8}{\delta_1}}{n}} + \frac{9B\log\frac{8}{\delta_1}}{n}.
\end{aligned} \tag{E.5}$$

Recall that the ψ function satisfies that $\forall r > 0$,

$$4\mathfrak{R}_n\{f : \mathbb{P}_n[f^2] \le 2r\} + \sqrt{\frac{2r\log\frac{8}{\delta_1}}{n}} + \frac{9B\log\frac{8}{\delta_1}}{n} \le \psi(r; \delta_1).$$

From this fact and (E.5), we see that with probability at least $1 - \delta_1$, for all $r > 0$,

$$\sup_{\mathbb{P}[f^2]\le r} (\mathbb{P} - \mathbb{P}_n) f \le \psi\left(r \vee r_2^*(\delta_1); \delta_1\right). \tag{E.6}$$

From (E.6), in order to prove the result (E.1) in Lemma 1, we only need to prove that

$$r_2^*(\delta_1) \le r_1^*(\delta_1). \tag{E.7}$$

Assume this is not true, i.e. $r_2^*(\delta_1) > r_1^*(\delta_1)$. Since $r_1^*(\delta_1)$ is the fixed point of $16B\psi(r;\delta_1)$, from the definition of fixed points we have

$$r_2^*(\delta_1) > 16B\psi(r_2^*(\delta_1);\delta_1).$$

From the definitions of ψ and φ_{δ_1}, for all $r > r_1^*(\delta_1)$,

$$4\varphi_{\delta_1}(r) \leq 16B\psi(r;\delta_1).$$

From the above two inequalities and $r_2^*(\delta_1) > r_1^*(\delta_1)$, we have

$$r_2^*(\delta_1) > 16B\psi(r_2^*(\delta_1);\delta_1) \geq 4\varphi_{\delta_1}(r_2^*(\delta_1)). \tag{E.8}$$

From the fact that $r_2^*(\delta_1)$ is the fixed point of $4\varphi_{\delta_1}$, we have

$$4\varphi_{\delta_1}(r_2^*(\delta_1)) = r_2^*(\delta_1). \tag{E.9}$$

The above two inequalities (E.8) and (E.9) result in a contradiction. So the assumption $r_2^*(\delta_1) > r_1^*(\delta_1)$ is false. Therefore $r_2^*(\delta_1) \leq r_1^*(\delta_1)$, and this completes the proof of Lemma 1. □

Part II: a "uniform localized convergence" argument with data-dependent measurement.

Based on Lemma 1, we will modify the proof of Proposition 1 to obtain a "uniform localized convergence" argument with the data-dependent "measurement" functional $\mathbb{P}_n[f^2]$.

Lemma 2 (a "uniform localized convergence" argument with the data-dependent "measurement" functional). *Given a fixed $\delta_1 \in (0,1)$, let $r_1^*(\delta_1)$ be the fixed point of $16B\psi(r;\delta_1)$ where ψ is defined in Strategy 2. Then with probability at least $1 - 2\left(\log_2 \frac{8B^2 \vee 2r_1^*(\delta_1)}{r_1^*(\delta_1)}\right)\delta_1$, for all $f \in \mathcal{F}$ either $\mathbb{P}[f^2] \leq r_1^*(\delta_1)$, or*

$$(\mathbb{P} - \mathbb{P}_n)f \leq \psi\left(4\mathbb{P}_n[f^2];\delta_1\right). \tag{E.10}$$

Proof of Lemma 2: from the definition of ψ and the fact that $r_1^*(\delta_1)$ is the fixed point of $16B\psi(r;\delta_1)$, we know that $r_1^*(\delta_1) \geq \frac{144B^2 \log \frac{8}{\delta_1}}{n} > 0$. Take $r_0 = r_1^*(\delta_1)$.

Take $R = 4B^2 \vee r_0$ to be a uniform upper bound for $\mathbb{P}f^2$, and take $r_k = 2^k r_0, k = 1, \cdots, \lceil \log_2 \frac{R}{r_0} \rceil$. Note that $\lceil \log_2 \frac{R}{r_0} \rceil \leq \log_2 \frac{2R}{r_0}$. We use the union bound to establish that $\sup_{\mathbb{P}[f^2]\leq r}(\mathbb{P} - \mathbb{P}_n)f \leq \psi(r;\delta_1)$ holds for all $\{r_k\}$ simultaneously: with probability at least $1 - \log_2 \frac{2R}{r_0}\delta_1$,

$$\sup_{\mathbb{P}[f^2]\leq r_k} (\mathbb{P} - \mathbb{P}_n)f \leq \psi(r_k;\delta_1), \quad k = 1, \cdots, \left\lceil \log_2 \frac{R}{r_0} \right\rceil.$$

For any fixed $f \in \mathcal{F}$, if $\mathbb{P}[f^2] \leq r_0$ is false, let k be the non-negative integer such that $2^k r_0 < \mathbb{P}[g(h;z)^2] \leq 2^{k+1} r_0$. We further have that $r_{k+1} = 2^{k+1} r_0 \leq 2\mathbb{P}[f^2]$. Therefore, with probability at least$1 - \log_2 \frac{2R}{r_0}\delta_1$,

$$\begin{aligned} \mathbb{P}f &\leq \mathbb{P}_n f + \sup_{\tilde{f}\in\mathcal{F}:\mathbb{P}[\tilde{f}^2]\leq r_{k+1}} (\mathbb{P} - \mathbb{P}_n)\tilde{f} \\ &\leq \mathbb{P}_n f + \psi(r_{k+1};\delta_1) \end{aligned} \tag{E.11}$$

By (E.2) we know that with probability at least $1 - \frac{\delta_1}{2}$,

$$\sup_{\mathbb{P}[f^2]\leq r} \left(\mathbb{P}[f^2] - \mathbb{P}_n[f^2]\right) \leq \frac{r}{4}$$

for all $r > r_0$ (here we have used the fact $r_0 = r_1^*(\delta_1) \geq r_2^*(\delta_1)$, which is the result (E.7) in the proof of Lemma 1). From the union bound, with probability at least $1 - (\log_2 \frac{2R}{r_0} + \frac{1}{2})\delta_1 \geq 1 - 2(\log_2 \frac{2R}{r_0})\delta_1$, the condition $r_{k+1} \geq \mathbb{P}[f^2] > r_k$ will imply

$$\mathbb{P}_n[f^2] \geq \mathbb{P}[f^2] - \frac{1}{4}r_{k+1} \geq \frac{1}{4}r_{k+1},$$

so

$$r_{k+1} \leq 4\mathbb{P}_n[f^2].$$

Combining this result with (E.11), we have that for all f such that $T(f) > r_0$, with probability at least $1 - 2(\log_2 \frac{2R}{r_0})\delta_1$,

$$\begin{aligned}\mathbb{P}f &\leq \mathbb{P}_n f + \psi(r_{k+1}; \delta_1) \\ &\leq \mathbb{P}_n f + \psi\bigg(4\mathbb{P}_n[f^2]; \delta_1\bigg).\end{aligned}$$

We conclude that with probability at least $1 - 2(\log_2 \frac{2R}{r_0})\delta_1$, for all $f \in \mathcal{F}$, either $\mathbb{P}[f^2] \leq r_1^*(\delta_1)$, or

$$(\mathbb{P} - \mathbb{P}_n)f \leq \psi\bigg(4\mathbb{P}_n[f^2]; \delta_1\bigg).$$

This completes the proof of Lemma 2. □

Part III: specify the moment-penalized estimator and its error bound. We specify the moment-penalized estimator to be

$$\hat{h}_{\text{MP}} = \arg\min_{\mathcal{H}} \left\{ \mathbb{P}_n \ell(h; z) + \psi\bigg(16\mathbb{P}_n[(\ell(h; z) - \mathcal{L}^*_{S'})^2]; \delta_1\bigg)\right\}. \tag{E.12}$$

Define $\hat{f}$ by $\hat{f}(z) = \ell(\hat{h}_{\text{MP}}; z) - \ell(h^*; z), \forall z \in \mathcal{Z}$, We define the event

$$\mathscr{A}_1 = \{\mathbb{P}[\hat{f}^2] \leq r_1^*(\delta_1)\},$$

and event

$$\mathscr{A}_2 = \{\text{the inequality (E.10) holds true at } \hat{f}\}.$$

Lemma 2 has proven that

$$\text{Prob}(\mathscr{A}_1) + \text{Prob}(\mathscr{A}_2) \geq 1 - 2\left(\log_2 \frac{8B^2 \vee 2r_1^*(\delta_1)}{r_1^*(\delta_1)}\right)\delta_1.$$

Consider the event $\mathscr{A}_1$ where $\mathbb{P}[\hat{f}^2] \leq r_1^*(\delta_1)$ holds true. Due to the surrogate property of ψ,

$$\text{Prob}\left(\mathscr{A}_1 \cap \left\{(\mathbb{P} - \mathbb{P}_n)\hat{f} \leq \psi(r_1^*(\delta_1); \delta_1)\right\}\right) \geq \text{Prob}(\mathscr{A}_1) - \delta_1.$$

Combining events $\mathscr{A}_1$ and $\mathscr{A}_2$, we conclude that with probability at least $1 - 2\left(\log_2 \frac{8B^2 \vee 2r_1^*(\delta_1)}{r_1^*(\delta_1)} + 1\right)\delta_1$, we have

$$(\mathbb{P} - \mathbb{P}_n)\hat{f} \leq \psi\bigg(4\mathbb{P}_n[\hat{f}^2] \vee r_1^*(\delta_1); \delta_1\bigg). \tag{E.13}$$

Denote $w(h; z) = \ell(h; z) - \mathcal{L}^*_{S'}$. Then $\hat{f}(z) = w(\hat{h}_{\text{MP}}; z) - w(h^*; z), \forall z \in \mathcal{Z}$, and we have that

$$\begin{aligned}4\mathbb{P}_n[\hat{f}^2] &\leq 8\mathbb{P}_n[w(\hat{h}_{\text{MP}}; z)^2] + 8\mathbb{P}_n[w(h^*; z)^2] \\ &\leq 16\mathbb{P}_n[w(\hat{h}_{\text{MP}}; z)^2] \vee 16\mathbb{P}_n[w(h^*; z)^2].\end{aligned}$$

From the above conclusion and (E.13) we obtain with probability at least $1 - 2\left(\log_2 \frac{8B^2 \vee 2r_1^*(\delta_1)}{r_1^*(\delta_1)} + 1\right)\delta_1$,

$$\begin{aligned}&\mathscr{E}(\hat{h}_{\text{MP}}) + \mathbb{P}_n \ell(h^*; z) \leq \mathbb{P}_n \ell(\hat{h}_{\text{MP}}; z) + \psi(4\mathbb{P}_n[\hat{f}^2] \vee r_1^*(\delta_1); \delta_1) \\ &\leq \mathbb{P}_n(\hat{h}_{\text{MP}}; z) + \psi\bigg(16\mathbb{P}_n[w(\hat{h}_{\text{MP}}; z)^2] \vee 16\mathbb{P}_n[w(h^*; z)^2] \vee r_1^*(\delta_1); \delta_1\bigg) \\ &\leq \mathbb{P}_n(\hat{h}_{\text{MP}}; z) + \psi\bigg(16\mathbb{P}_n[w(\hat{h}_{\text{MP}}; z)^2]\delta_1\bigg) + \psi\bigg(16\mathbb{P}_n[w(h^*; z)^2] \vee r_1^*(\delta_1); \delta_1\bigg).\end{aligned} \tag{E.14}$$

From the definition (E.12) of $\hat{h}_{\text{MP}}$, we have

$$\mathbb{P}_n\ell(\hat{h}_{\text{MP}};z)+\psi\bigg(16\mathbb{P}_n[w(\hat{h}_{\text{MP}};z)^2];\delta_1\bigg)\le\mathbb{P}_n\ell(h^*;z)+\psi\bigg(16\mathbb{P}_n[w(h^*;z)^2];\delta_1\bigg) \tag{E.15}$$

Therefore, with probability at least $1-2\left(\log_2\frac{8B^2\vee 2r_1^*(\delta_1)}{r_1^*(\delta_1)}+1\right)\delta_1$,

$$\begin{aligned}\mathscr{E}(\hat{h}_{\text{MP}})&\le\mathbb{P}_n\ell(\hat{h}_{\text{MP}};z)+\psi\bigg(16\mathbb{P}_n[w(\hat{h}_{\text{MP}};z)^2];\delta_1\bigg)+\psi\bigg(16\mathbb{P}_n[w(h^*;z)^2]\vee r_1^*(\delta_1);\delta_1\bigg)-\mathbb{P}_n\ell(h^*;z)\\&=\arg\min_{\mathcal{H}}\left\{\mathbb{P}_n\ell(h;z)+\psi\bigg(16\mathbb{P}_n[w(h;z)];\delta_1\bigg)\right\}-\mathbb{P}_n\ell(h^*;z)+\psi\bigg(16\mathbb{P}_n[w(h^*;z)^2]\vee r_1^*(\delta_1);\delta_1\bigg)\\&\le\psi\bigg(16\mathbb{P}_n[w(h^*;z)^2];\delta_1\bigg)+\psi\bigg(16\mathbb{P}_n[w(h^*;z)^2]\vee r_1^*(\delta_1);\delta_1\bigg)\\&\le 2\psi\bigg(16\mathbb{P}_n[w(h^*;z)^2]\vee r_1^*(\delta_1);\delta_1\bigg),\end{aligned} \tag{E.16}$$

where the first inequality is due to (E.14) and the second inequality is due to (E.15).

From Bernstein's inequality at the single element h^*, for any fixed $\delta_2\in(0,1)$, with probability at least $1-\delta_2$,

$$\begin{aligned}\mathbb{P}_n[w(h^*;z)^2]&\le\mathbb{P}[w(h^*;z)^2]+2B\sqrt{\frac{2\mathbb{P}[w(h^*;z)^2]\log\frac{2}{\delta_2}}{n}}+\frac{4B^2\log\frac{2}{\delta_2}}{n}\\&\le 2\mathbb{P}[w(h^*;z)^2]+\frac{6B^2\log\frac{2}{\delta_2}}{n}.\end{aligned}$$

Therefore, we conclude that with probability at least $1-2\left(\log_2\frac{8B^2\vee 2r_1^*(\delta_1)}{r_1^*(\delta_1)}+1\right)\delta_1-\delta_2$,

$$\begin{aligned}\mathscr{E}(\hat{h}_{\text{MP}})&\le 2\psi\left(16\mathbb{P}_n[w(h^*;z)]\vee r_1^*(\delta_1)\vee\frac{B^2}{n};\delta_1\right)\\&\le 2\psi\left(\left(32\mathbb{P}[w(h^*;z)^2]+\frac{96B^2\log\frac{2}{\delta_2}}{n}\right)\vee r_1^*(\delta_1)\vee\frac{B^2}{n};\delta_1\right).\end{aligned} \tag{E.17}$$

Part IV: final steps.

From the definition of ψ and the fact that $r_1^*(\delta_1)$ is the fixed point of $16B\psi(r;\delta_1)$, we know that

$$r_1^*(\delta_1)\ge\frac{144B^2\log\frac{8}{\delta_1}}{n}. \tag{E.18}$$

Denote $C_n:=4\log_2 n+10\ge 4\log_2$ and take

$$\delta_1=\frac{\delta}{C_n}, \tag{E.19}$$

then we have

$$\begin{aligned}4\log_2\frac{8B^2\vee 2r_1^*(\delta_1)}{r_1^*(\delta_1)}+6&\le\max\left\{4\log_2\frac{8n}{144\log 8},4+6\right\}\\&\le\max\{4\log_2 n,10\}\le C_n,\end{aligned}$$

so

$$\left(2\log_2\frac{8B^2\vee 2r_1^*(\delta_3)}{r_1^*(\delta_3)}+3\right)\delta_1\le\frac{\delta}{2}.$$

Set $r^* = r_1^*(\delta_1)$ and take $\delta_2 = \frac{\delta}{2}$. From (E.17), we obtain that with probability at least $1-\delta$, the generalization error of $\hat{h}_{\text{MP}}$ is upper bounded by

$$\mathscr{E}(\hat{h}_{\text{MP}}) \leq 2\psi\bigg(c\left[\mathbb{P}[w(h^*;z)^2] \vee r^* \vee \frac{B^2 \log\frac{4}{\delta}}{n}\right]; \frac{\delta}{C_n}\bigg), \tag{E.20}$$

where c is an absolute constant. From (E.18) we have $r_1^*(\delta_1) \geq \frac{144B^2\log\frac{8C_n}{\delta}}{n} \geq \frac{B^2\log\frac{4}{\delta}}{n}$. Combine this fact with the inequality (E.20), we obtain that

$$\begin{aligned}\mathscr{E}(\hat{h}_{\text{MP}}) &\leq 2\psi\left(c\left[\mathbb{P}[(\ell(h^*;z)-\mathcal{L}_{S'}^*)^2] \vee r^*\right]; \frac{\delta}{C_n}\right) \\ &\leq 2\psi\left(c_0\left[\mathcal{V}^* \vee r^* \vee (\mathcal{L}_{S'}^* - \mathcal{L}_0^*)^2\right]; \frac{\delta}{C_n}\right). \end{aligned} \tag{E.21}$$

where c_0 is an absolute constant. This completes the proof of Theorem 5. □

E.2 Analysis of the first-stage ERM estimator

After proving Theorem 5, the remaining part needed to prove Theorem 2 is to bound $(\mathcal{L}_{S'}^* - \mathcal{L}_0^*)^2$—the error of the first-stage ERM estimator.

The remaining steps in the proof of Theorem 2: We will give a guarantee on the first-stage ERM estimator, and combine this guarantee with Theorem 5 to prove Theorem 2. Recall that $\mathbb{P}_{S'}$ is the empirical distribution of the "auxiliary" data set. Denote $\hat{h}_{\text{ERM}} \in \arg\min_{\mathcal{H}} \mathbb{P}_{S'}\ell(h;z)$.

From Part I in the proof of Theorem 5, $\forall \delta \in (0, \frac{1}{2})$, with probability at least $1-\delta$,

$$\sup_{\mathcal{F}} |(\mathbb{P}-\mathbb{P}_n)f| \leq \psi(4B^2;\delta) \leq \psi\left(4B^2; \frac{\delta}{C_n}\right).$$

Since ψ is sub-root with respect to its first argument, we have

$$\frac{\psi(4B^2; \frac{\delta}{C_n})}{\sqrt{4B^2}} \leq \frac{\psi(r^*; \frac{\delta}{C_n})}{\sqrt{r^*}} = \frac{\sqrt{r^*}}{16B},$$

where r^* is the fixed point of $16B\psi(r; \frac{\delta}{C_n})$. So we have proved that $\psi(4B^2; \frac{\delta}{C_n}) \leq \frac{\sqrt{r^*}}{8}$. Therefore,

$$\sup_{\mathcal{F}} |(\mathbb{P}-\mathbb{P}_n)f| \leq \frac{\sqrt{r^*}}{8}.$$

Because $\hat{h}_{\text{ERM}} \in \arg\min_{\mathcal{H}} \mathbb{P}_{S'}\ell(h;z)$ and $\mathbb{P}_{S'}\ell(\hat{h}_{\text{ERM}};z) = \mathcal{L}_{S'}^*$, we have

$$\begin{aligned}\mathcal{L}_{S'}^* - \mathcal{L}_0^* &= (\mathbb{P}_{S'}\ell(\hat{h}_{\text{ERM}};z) - \mathbb{P}_{S'}\ell(h^*;z)) + (\mathbb{P}_{S'}\ell(h^*;z) - \mathbb{P}\ell(h^*;z)) \\ &\leq \mathbb{P}_{S'}\ell(h^*;z) - \mathbb{P}\ell(h^*;z) \leq \sup_{\mathcal{F}} |(\mathbb{P}-\mathbb{P}_n)f|,\end{aligned}$$

and

$$\begin{aligned}\mathcal{L}_{S'}^* - \mathcal{L}_0^* &= (\mathbb{P}_{S'}\ell(\hat{h}_{\text{ERM}};z)) - \mathbb{P}\ell(\hat{h}_{\text{ERM}};z)) + (\mathbb{P}\ell(\hat{h}_{\text{ERM}};z) - \mathbb{P}\ell(h^*;z)) \\ &\geq \mathbb{P}_{S'}\ell(\hat{h}_{\text{ERM}};z)) - \mathbb{P}\ell(\hat{h}_{\text{ERM}};z) \geq -\sup_{\mathcal{F}} |(\mathbb{P}-\mathbb{P}_n)f|.\end{aligned}$$

Hence we have

$$(\mathcal{L}_{S'}^* - \mathcal{L}_0^*)^2 \leq (\sup_{\mathcal{F}} |(\mathbb{P}-\mathbb{P}_n)f|)^2 \leq \frac{r^*}{64}.$$

Combine this result with (E.21), we have that $\forall \delta \in (0, \frac{1}{2})$, with probability $1-2\delta$,

$$\begin{aligned}\mathscr{E}(\hat{h}_{\text{MP}}) &\leq 2\psi\left(c_1\left(\mathcal{V}^* \vee r^*\right); \frac{\delta}{C_n}\right) \\ &\leq 2\left(\psi\left(c_1\mathcal{V}^*; \frac{\delta}{C_n}\right) \vee \psi\left(c_1 r^*; \frac{\delta}{C_n}\right)\right) \\ &\leq 2\psi\left(c_1\mathcal{V}^*; \frac{\delta}{C_n}\right) \vee \frac{c_1 r^*}{8B},\end{aligned}$$

where $c_1 = \max\{c_0, 16\}$ is an absolute constant, and the last inequality follows from the fact that $\frac{c_1 r^*}{16} > r^*$ and the definition of fixed points. This completes the proof of Theorem 2. □

F Proof of Corollary 3

From the definitions, we know that $\mathcal{L}^* = \mathbb{P}[\ell(h^*;z) - \inf_{\mathcal{H}} \ell(h^*;z)]$, $\widehat{\mathcal{L}^*} = \mathbb{P}_n[\ell(\hat{h}_{\text{ERM}};z) - \inf_{\mathcal{H}} \ell(h;z)]$ and $\mathbb{P}\ell(h^*;z) \le \mathbb{P}\ell(\hat{h}_{\text{ERM}};z)$. As a result, we have

$$\begin{aligned}\mathcal{L}^* - \widehat{\mathcal{L}^*} &= \mathbb{P}\ell(h^*;z) - \mathbb{P}_n\ell(\hat{h}_{\text{ERM}};z) - (\mathbb{P}-\mathbb{P}_n)[\inf_{\mathcal{H}} \ell(h;z)] \\ &\le (\mathbb{P}-\mathbb{P}_n)\ell(\hat{h}_{\text{ERM}};z) - (\mathbb{P}-\mathbb{P}_n)[\inf_{\mathcal{H}} \ell(h;z)] \\ &= (\mathbb{P}-\mathbb{P}_n)\hat{f} + (\mathbb{P}-\mathbb{P}_n)[\ell(h^*;z) - \inf_{\mathcal{H}} \ell(h;z)], \end{aligned} \tag{F.1}$$

where $\hat{f}$ is defined by $\hat{f}(z) = \ell(\hat{h}_{\text{ERM}};z) - \ell(h^*;z), \forall z \in \mathcal{Z}$.

We take $r_0 = \frac{B^2}{n}$ in Theorem 1, and denote $C_n := C_{r_0} = 2\log_2 n + 6$. From (D.7) in the proof of Theorem 1, on the event $\mathscr{A}$ defined in (D.3) (whose measure is at least $1-\delta$),

$$\mathscr{E}(\hat{h}_{\text{ERM}}) \le (\mathbb{P}-\mathbb{P}_n)\hat{f} \le \psi(24B\mathcal{L}^* \vee 8r^* \vee \frac{2B^2}{n}; \frac{\delta}{C_n}). \tag{F.2}$$

Since $3B\mathcal{L}^* \vee r^* \vee \frac{B^2}{4n} \ge r^*$, from the definition of fixed points we have

$$\begin{aligned}(\mathbb{P}-\mathbb{P}_n)\hat{f} &\le \psi\left(8\left(3B\mathcal{L}^* \vee r^* \vee \frac{B^2}{4n}\right); \frac{\delta}{C_n}\right) \\ &\le \frac{3B\mathcal{L}^* \vee r^* \vee \frac{B^2}{4n}}{6B} \le \frac{\mathcal{L}^*}{2} + \frac{r^*}{6B} + \frac{B}{24n}. \end{aligned} \tag{F.3}$$

This result holds together with the result of Theorem 1 on the event $\mathscr{A}$.

The random variable $\ell(h^*;z) - \inf_{\mathcal{H}} \ell(h;z)$ is uniformly bounded by $[0, 2B]$. From Bernstein's inequality and the fact $\text{Var}[\ell(h^*;z) - \inf_{\mathcal{H}} \ell(h;z)] \le 2B\mathcal{L}^*$, with probability at least $1-\delta$,

$$\left|(\mathbb{P}-\mathbb{P}_n)[\ell(h^*;z) - \inf_{\mathcal{H}} \ell(h;z)]\right| \le \sqrt{\frac{4B\mathcal{L}^* \log\frac{2}{\delta}}{n}} + \frac{2B\log\frac{2}{\delta}}{n} \le \frac{\mathcal{L}^*}{4} + \frac{3B\log\frac{2}{\delta}}{n}. \tag{F.4}$$

Consider the event

$$\mathscr{A}_3 = \mathscr{A} \cup \{\text{inequality (F.4) holds true}\},$$

whose measure is at least $1-2\delta$. On the event $\mathscr{A}_3$, from inequalities (F.1) (F.3) (F.4), it is straightforward to show that

$$\mathcal{L}^* - \widehat{\mathcal{L}^*} \le \frac{3}{4}\mathcal{L}^* + \frac{r^*}{6B} + \frac{4B\log\frac{2}{\delta}}{n},$$

which implies

$$\mathcal{L}^* \le 4\widehat{\mathcal{L}^*} + \frac{2r^*}{3B} + \frac{16B\log\frac{2}{\delta}}{n}. \tag{F.5}$$

From this result and (F.2), it is straightforward to show that

$$\mathscr{E}(\hat{h}_{\text{ERM}}) \le \psi\left(cB\widehat{\mathcal{L}^*}; \frac{\delta}{C_n}\right) \vee \frac{cr^*}{n} \vee \frac{cB\log\frac{2}{\delta}}{n},$$

where c is an absolute constant.

We also have

$$\begin{aligned}\widehat{\mathcal{L}^*} - \mathcal{L}^* &= \mathbb{P}_n\ell(\hat{h}_{\text{ERM}}) - \mathbb{P}\ell(h^*;z) - (\mathbb{P}_n-\mathbb{P})[\inf_{\mathcal{H}} \ell(h;z)] \\ &\le (\mathbb{P}_n-\mathbb{P})\ell(h^*;z) - (\mathbb{P}_n-\mathbb{P})[\inf_{\mathcal{H}} \ell(h;z)] \\ &= (\mathbb{P}_n-\mathbb{P})[\ell(h^*;z) - \inf_{\mathcal{H}} \ell(h;z)]. \end{aligned}$$

From this result and (F.4), on the event $\mathscr{A}_3$,

$$\widehat{\mathcal{L}^*} \le \frac{5}{4}\mathcal{L}^* + \frac{3B\log\frac{2}{\delta}}{n}. \tag{F.6}$$

Combine (F.5) and (F.6) we obtain

$$\mathcal{L}^* \le c_1\left(\widehat{\mathcal{L}^*} \vee \frac{r^*}{B} \vee \frac{B\log\frac{2}{\delta}}{n}\right) \le c_2\left(\mathcal{L}^* \vee \frac{r^*}{B} \vee \frac{B\log\frac{2}{\delta}}{n}\right),$$

where c_1 and c_2 are absolute constants. This completes the proof. □

G Proof of Corollary 4

Define $\hat{f}_{\text{NMP}}$ by $\hat{f}_{\text{NMP}}(z) = \ell(\hat{h}_{\text{NMP}}; z) - \ell(h^*; z), \forall z \in \mathcal{Z}$, and $w(h; z) = \ell(h; z) - \widehat{\mathcal{L}_0^*}$. In the proof of Theorem 5, the result (E.16) and the specification of δ_1 in (E.19) show that with probability at least $1 - \frac{\delta}{2}$,

$$\mathscr{E}(\hat{h}_{\text{MP}}) \leq 2\psi\left(16\mathbb{P}_n[w(h^*; z)^2] \vee r^*; \frac{\delta}{C_n}\right). \tag{G.1}$$

We also refer to another implication of the proof of Theorem 5. Note that the proof of the result (E.13) does not depends on any property of the estimator $\hat{h}_{\text{MP}}$. By the repeating the lines between (E.12) and (E.13) for the estimator $\hat{h}_{\text{NMP}}$, and use the specification of δ_1 in (E.19), it is straightforward to show that with probability at least $1 - \frac{\delta}{2}$,

$$(\mathbb{P} - \mathbb{P}_n)\hat{f}_{\text{NMP}} \leq \psi\left(4\mathbb{P}_n[\hat{f}_{\text{NMP}}^2] \vee r^*; \frac{\delta}{C_n}\right). \tag{G.2}$$

We continue the proof on the event

$$\mathscr{A} := \{\text{the inequalities (G.2) and (G.1) hold true}\},$$

whose measure is at least $1 - \delta$.

From the definition of $\hat{h}_{\text{NMP}}$,

$$\mathbb{P}_n\ell(\hat{h}_{\text{NMP}}; z) - 2\psi\left(16\mathbb{P}_n[w(\hat{h}_{\text{NMP}}; z)^2]; \frac{\delta}{C_n}\right) \leq \mathbb{P}_n\ell(h^*; z) - 2\psi\left(16\mathbb{P}_n[w(h^*; z)^2]; \frac{\delta}{C_n}\right). \tag{G.3}$$

Therefore, we have

$$\begin{aligned}
&2\psi\left(16\mathbb{P}_n[w(h^*; z)^2]; \frac{\delta}{C_n}\right) \\
&\leq 2\psi\left(16\mathbb{P}_n[w(\hat{h}_{\text{NMP}}; z)^2]; \frac{\delta}{C_n}\right) + \mathbb{P}_n\ell(h^*; z) - \mathbb{P}_n\ell(\hat{h}_{\text{NMP}}; z) \\
&= 2\psi\left(16\mathbb{P}_n[w(\hat{h}_{\text{NMP}}; z)^2]; \frac{\delta}{C_n}\right) + \mathbb{P}[\ell(h^*; z) - \ell(\hat{h}_{\text{NMP}}; z)] + (\mathbb{P}_n - \mathbb{P})[\ell(h^*; z) - \ell(\hat{h}_{\text{NMP}}; z)] \\
&\leq 2\psi\left(16\mathbb{P}_n[w(\hat{h}_{\text{NMP}}; z)^2]; \frac{\delta}{C_n}\right) + (\mathbb{P} - \mathbb{P}_n)\hat{f}_{\text{NMP}} \\
&\leq 2\psi\left(16\mathbb{P}_n[w(\hat{h}_{\text{NMP}}; z)^2]; \frac{\delta}{C_n}\right) + \psi\left(4\mathbb{P}_n[\hat{f}_{\text{NMP}}^2]; \frac{\delta}{C_n}\right),
\end{aligned} \tag{G.4}$$

where the first inequality is due to (G.3), the second inequality is due to the fact that h^* minimizes the population risk; and the last inequality is due to (G.2).

Note that

$$\begin{aligned}
4\mathbb{P}_n[\hat{f}_{\text{NMP}}^2] &\leq 8\mathbb{P}_n[w(\hat{h}_{\text{NMP}}; z)^2] + 8\mathbb{P}_n[w(h^*; z)^2] \\
&\leq 16\mathbb{P}_n[w(\hat{h}_{\text{NMP}}; z)^2] \vee 16\mathbb{P}_n[w(h^*; z)^2].
\end{aligned}$$

From the above inequality and (G.4), we have

$$\begin{aligned}
&2\psi\left(16\mathbb{P}_n[w(h^*; z)^2]; \frac{\delta}{C_n}\right) \\
&\leq\ 2\psi\left(16\mathbb{P}_n[w(\hat{h}_{\text{NMP}}; z)^2]; \frac{\delta}{C_n}\right) + \psi\left(16\mathbb{P}_n[w(\hat{h}_{\text{NMP}}; z)^2]; \frac{\delta}{C_n}\right) \vee \psi\left(16\mathbb{P}_n[w(h^*; z)^2]; \frac{\delta}{C_n}\right).
\end{aligned} \tag{G.5}$$

Whether $\mathbb{P}_n[w(h^*; z)^2] \leq 16\mathbb{P}_n[w(\hat{h}_{\text{NMP}}; z)^2]$ or $\mathbb{P}_n[w(h^*; z)^2] > 16\mathbb{P}_n[w(\hat{h}_{\text{NMP}}; z)^2]$, the inequality (G.5) always implies

$$\psi\left(16\mathbb{P}_n[w(h^*; z)^2]; \frac{\delta}{C_n}\right) \leq 2\psi\left(16\mathbb{P}_n[w(\hat{h}_{\text{NMP}}; z)^2]; \frac{\delta}{C_n}\right) = 2\psi\left(16\widehat{\mathcal{V}^*}; \frac{\delta}{C_n}\right). \tag{G.6}$$

(Note that $\widehat{\mathcal{V}^*} := \mathbb{P}_n[w(\hat{h}_{\text{NMP}}; z)^2]$.) We conclude that with probability at least $1-\delta$,

$$\begin{aligned}
\mathscr{E}(\hat{h}_{\text{MP}}) &\leq 2\psi\left(16\mathbb{P}_n[w(h^*;z)^2] \vee r^*; \frac{\delta}{C_n}\right) \\
&= 2\psi\left(16\mathbb{P}_n[w(h^*;z)^2]; \frac{\delta}{C_n}\right) \vee 2\psi(r^*; \frac{\delta}{C_n}\Big) \\
&\leq 4\psi\left(16\widehat{\mathcal{V}^*}; \frac{\delta}{C_n}\right) \vee \frac{r^*}{8B},
\end{aligned}$$

where the first inequality is due to (G.1) and the last inequality is due to (G.6). This completes the proof. □

H Auxiliary lemmas

Lemma 3 (Dudley's integral bound, [13]). *Given $r > 0$ and a class $\mathcal{F}$ that consists of functions defined on $\mathcal{Z}$,*

$$\mathfrak{R}_n\{f \in \mathcal{F} : \mathbb{P}_n[f^2] \leq r\} \leq \inf_{\varepsilon_0 > 0}\left\{4\varepsilon_0 + 12\int_{\varepsilon_0}^{\sqrt{r}} \sqrt{\frac{\log \mathcal{N}(\varepsilon, \mathcal{F}, L_2(\mathbb{P}_n))}{n}} d\varepsilon\right\}.$$

Lemma 4 (Talagrand's concentration inequality for empirical processes, [2]). *Let $\mathcal{F}$ be a class of functions that map $\mathcal{Z}$ into $[B_1, B_2]$. Assume that there is some $r > 0$ such that for every $f \in \mathcal{F}$, $\text{Var}[f(z_i)] \leq r$. Then, for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\sup_{f\in\mathcal{F}}(\mathbb{P} - \mathbb{P}_n)f \leq 3\mathfrak{R}\mathcal{F} + \sqrt{\frac{2r\log\frac{1}{\delta}}{n}} + (B_2 - B_1)\frac{\log\frac{1}{\delta}}{n},$$

and with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}(\mathbb{P} - \mathbb{P}_n)f \leq 4\mathfrak{R}_n\mathcal{F} + \sqrt{\frac{2r\log\frac{2}{\delta}}{n}} + \frac{9}{2}(B_2 - B_1)\frac{\log\frac{2}{\delta}}{n}.$$

Moreover, the same results hold for the quantity $sup_{f\in\mathcal{F}}(\mathbb{P}_n - \mathbb{P})f$.

Lemma 5 (Bernstein's inequality). *Let $X_1, \cdots, X_n$ be real-valued, independent, mean-zero random variables and suppose that for some constants $\sigma, B > 0$,*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}|X_i|^k \leq \frac{k!}{2}\sigma^2 B^{k-2}, \quad k = 2, 3, \cdots$$

Then, $\forall\delta \in (0,1)$, with probability at least $1-\delta$

$$\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| \leq \sqrt{\frac{2\sigma^2\log\frac{2}{\delta}}{n}} + \frac{B\log\frac{2}{\delta}}{n}. \tag{H.1}$$