Guide for reading supplementary materials

The supplementary materials are organized as shown in the table of contents. In particular,

- Appendix A presents additional results of our general framework related to the topic of "learning without concentration" (proofs are deferred to Appendix D).
- Appendix B and C contain the proofs for all theoretical claims in the main paper. These can be read in a selective manner; key steps of each proof will be indicated shortly.
- Appendix B.3 and B.5 discuss how to compute loss-dependent and variance-dependent rates directly from data, filling in details that were omitted from the main paper for space consideration.

For convenience of readers, we will present a guide for reading the supplementary materials.

Table of Contents

Overview for Appendix A

Appendix A applies our general framework to study supervised learning problems with structured convex cost. The relationship between this setting and other parts of the paper has been explained in Section 2.4 and Section 5.1. In essence, the treatment presented here allows for non-parametric and heavy-tailed hypothesis classes, whreas the main paper (Section 5-7) focuses on parametric models and assumes a sub-exponential type assumption (Assumption 1). Unlike the main paper which focuses primally on iterative optimization algorithms, popular non-convex learning problems, and generic-form stochastic optimization problems, the focus here is on better studied "classical" problem to best illustrate the key points.

The appendix presents work related to a stream of work pioneered by Mendelson and others referred to as "learning without concentration" (Mendelson (2014, 2018)). This line of work has motivated by the need to find new localization approaches to replace the traditional "local Rademacher complexity" analysis. An interesting open question is whether one can achieve the same goals by directly strengthening the traditional concentration-based approaches. Our investigation shows that one of the core limitations of traditional localization approaches in this paradigm is its requirement of "sub-root" surrogate functions. Since our new approach completely remove this requirement, we are able to answer this question in the affirmative; the results from Mendelson (2014, 2018) can be recovered via a concentration-based analysis. Moreover, we are able to show some technical improvements—our approach does not require the "star-hull" of the hypothesis class that may increase complexity, and there are concrete examples showing that the improvement may be meaningful for non-convex classes. Appendix A.3 focuses on these new findings.

Overview of Appendix B and C

Appendix B and C provide proofs for all the theoretical claims in the main paper. Readers may read them in a selective manner, and to that end, we present a high level guide here.

Per our "uniform localized convergence" principle, a proof for problem-dependent generalization error bounds contain two steps: 1) obtaining "localized uniform convergence" arguments; and 2) subsequent analysis that is customized to the problem setting. Among the major theoretical results in the main paper, the following are dedicated to the first step:

- Proposition 1 provides a general tool to prove "uniform localized convergence" arguments. The proof of this result is "one-shot" via a surprisingly simple observation explained in Section 2.2. The formal proof is given in Appendix B.1 (we actually prove a more general version, Proposition 3), which is succinct and straightforward to verify.

- Proposition 2, the "uniform localized convergence of gradients" argument, is the foundation for all our results in the parametric "fast rate" regime. The proof (which is presented in Appendix C.2) crucially relies on a careful choice of concentrated function (5.3), a novel chaining analysis (see lines from (C.1) to (C.3)), and the application of Proposition 1.

- Theorem 2 (using empirical moment penalization to achieve optimal variance-dependent rate) crucially requires a "uniform localized convergence" argument where the measurement functional is data-dependent. The argument, Lemma 4 in Appendix B.4.1, uses tools from empirical processes theory and is somewhat technical in nature.

As for the second step (subsequent analysis that is customized to the problem setting and the learning algorithm), the paper presents three different approaches: 1) using the definition of the

estimator to establish an inequality and then calculating the fixed point (this is used in most traditional approaches); 2) adding a regularization term and then directly using the definition of the regularized estimator; and 3) coupling the statistical error with analysis of an iterative optimization algorithm. Below are some of the key points.

- Theorem 1 (using empirical risk minimization to achieve optimal loss-dependent rate) uses the "fixed point analysis" approach. The core step is to establish an inequality where a "measurement" functional of $\hat{h}_{\mathrm{ERM}}$ appears in both sides of the inequality (see inequality (B.8)). Then one can use the definition of the fixed point to prove loss-dependent generalization error bounds. The proof is presented in Appendix B.2.

- For Theorem 2 (using empirical moment penalization to achieve optimal variance-dependent rate), the core message is that the definition of the proposed moment-regularized estimator directly leads to variance-dependent generalization error bounds. The proof is somewhat lengthy, but the readers may focus on "Part III" in Appendix B.4.1 (in particular, the lines from (B.32) to (B.35)) for the main message.

- Theorem 3 ("fast rate" of approximate stationary points) uses the "fixed point analysis" approach, and the core step is to establish the inequality (C.10). The readers can parse (C.10) in a rather simple manner: the right hand side is mostly due to Proposition 2 (the "uniform localized convergence of gradients" argument); and the left hand side is due to the property of the Polyak-Lojasiewisz (PL) condition. The full proof is presented in Appendix C.3.

- The proofs for Theorem 4 (Appendix C.4) and Theorem 6 (Appendix C.6) are very similar. The core idea is to couple the statistical error from Proposition 2 to the optimization analysis of an iterative algorithm. The major difference is that Theorem 4 discusses sample-based gradient descent (before coupling the statistical error, its optimization analysis leads to (C.11)); and Theorem 6 discusses sample-based first-order Expectation-Maximization (before coupling the statistical error, its optimization analysis leads to (C.15)).

Additional corollaries. Besides the above, there are two corollaries in the main paper. Corollary 7 is the application of Theorem 6 to Example 4 (mixture of two Gaussians) and Example 5 (Mixture of two component linear regression). The explicit calculation of problem-dependent parameters here is quite novel (Appendix C.7), but the informal explanation at the end of Section 7.3 (from (7.11) to (7.12)) should serve as a better source to understand the main message. Corollary 5 is the application of Theorem 3 and Theorem 4 to Example 3 (non-convex regression with non-linear activation), and the verification of the assumptions here are mostly technical in nature (Appendix C.5).

Data-dependent bounds. Lastly but importantly, we would like to highlight Appendix B.3 and Appendix B.5—they discuss how to estimate the loss-dependent and variance-dependent rates from data, which are mentioned in the main paper (see remarks after Theorem 1 and Theorem 2) but details are omitted there. A central challenge is to replace the loss $\mathcal{L}^*$ and the variance $\mathcal{V}^*$ (which depends on the unknown "best hypothesis" h^*) by suitable empirical estimates. Readers who are interested in fully data-dependent generalization error bounds may find this of interest.

A Fast rates in supervised learning with structured convex cost

The main purpose of this section is to recover the problem-dependent rates in Mendelson (2018, 2014) for (possibly non-parametric and heavy-tailed) supervised learning problems with structured convex cost functions. While Mendelson (2018, 2014) propose an approach they call "learning without concentration," our approach emphasizes the use of surrogate functions that are not "sub-root," and relates one-sided uniform inequalities to two-sided concentration of "truncated" functions. Besides providing a unification, there are some technical improvements as well. For example, our approach does not require the "star-hull" of the hypothesis class that may increase complexity, and there are concrete examples showing that the improvement may be meaningful for non-convex classes. See Appendix A.3 for contributions of our method, and detailed comparison with existing approaches.

A.1 Background

Problem formulation and assumptions. Let the data z be a feature-label pair (x, y) where $x \in \mathcal{X}$ and $y \in \mathcal{Y} \subseteq \mathbb{R}$. We assume every hypothesis h in the hypothesis class $\mathcal{H}$ is a mapping from $\mathcal{X}$ to $\mathbb{R}$. In supervised learning, the loss function is of the form $\ell(h;(x,y)) = \ell_{\mathtt{sv}}(h(x), y)$ where the deterministic bivariate function $\ell_{\mathtt{sv}} : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ is called the *cost function*. We assume that the cost function is differentiable, globally convex with respect to its first argument, and the population risk is smooth.

Assumption 7 (differentiability, convexity and smoothness). *The partial derivative of ℓ_{su} with respect to its first argument, denoted $\partial_1 \ell_{\mathrm{su}}$, exists and is continuous everywhere, and $\ell_{\mathtt{sv}}$ is a convex function with respect to its first argument, i.e., $\forall u_1, u_2, y \in \mathbb{R}$,*

$$\ell_{\mathrm{su}}(u_1, y) - \ell_{\mathrm{su}}(u_2, y) - \partial_1 \ell_{\mathrm{sv}}(u_2, y)(u_1 - u_2) \geq 0.$$

In addition, the population risk is smooth, i.e., there exists a constant $\beta_{\mathtt{sv}} > 0$ such that $\forall h_1, h_2 \in \mathcal{H}$,

$$\mathbb{P}\ell_{\mathrm{sv}}(h_1(x), y) - \mathbb{P}\ell_{\mathrm{sv}}(h_2(x), y) - \mathbb{P}[\partial_1 \ell_{\mathrm{sv}}[(h_2(x), y)(h_1(x) - h_2(x))] \leq \frac{\beta_{\mathtt{sv}}}{2} \mathbb{P}[(h_1(x) - h_2(x))^2].$$

Given a cost function that is globally convex and locally strongly convex, we define $\{\alpha(v)\}_{v \geq 0}$ as follows.

Definition 5 (strong convexity parameter). *For a fixed $v > 0$, let $\alpha(v)$ be the largest constant such that for all $y \in \mathcal{Y}$, $\ell_{\mathtt{sv}}(u + y, y)$ is $\alpha(v)-$strongly convex with respect to u when $u \in [-v, v]$. That is,*

$$\ell_{\mathtt{sv}}(u_1 + y, y) - \ell_{\mathtt{sv}}(u_2 + y, y) - \partial_1 \ell_{\mathtt{sv}}(u_2 + y, y)(u_1 - u_2) \geq \frac{\alpha(v)}{2}(u_1 - u_2)^2, \quad \forall u_1, u_2 \in [-v, v], \forall y \in \mathcal{Y}.$$

Clearly $\{\alpha(v)\}_{v \geq 0}$ is non-increasing with respect to v, and we denote $\alpha(0) = \limsup_{v \to 0} \alpha(v)$.

When $\ell_{\mathtt{sv}}$ is second-order continuously differentiable, we have the simple relation

$$\alpha(v) = \sup_{|u| \leq v, y \in \mathcal{Y}} \partial_{1,1}^2 \ell_{\mathtt{sv}}(u + y, y), \quad , \forall v \geq 0,$$

where $\partial_{1,1}^2 \ell_{\mathtt{sv}}$ is the second order partial derivative of $\ell_{\mathtt{sv}}$ with respect to its first argument. Moreover, to accommodate popular choices of robust costs, Definition 5 also allows $\partial_1 \ell_{\mathtt{sv}}$ to be non-differentiable at certain points in its domain. We list three widely used cost functions, their strong convexity parameters $\{\alpha(v)\}_{v \geq 0}$, and the smoothness parameters $\beta_{\mathtt{sv}}$ of the corresponding population risks.

- Square cost: consider the regression setting $\mathbb{E}[y|x] = h_{\text{true}}(x)$, where h_{true} is the function we want to estimate (not necessarily in $\mathcal{H}$). It is natural to consider the square cost function

$$\ell_{\texttt{sv}}(h(x), y) = \frac{1}{2}(h(x) - y)^2.$$

  Here $\ell_{\texttt{sv}}(u+y, y) = u^2$. Thus $\alpha(v) = \frac{1}{2}, \forall v \geq 0$. The smoothness parameter of the population risk is $\beta_{\texttt{sv}} = \frac{1}{2}$. In this example, one does not need to localize the strong convexity parameter $\alpha(v)$ as it is a constant.

- Huber cost: consider the regression setting $\mathbb{E}[y|x] = h_{\text{true}}(x)$, where h_{true} is the function we want to estimate (not necessarily in $\mathcal{H}$). When the conditional distribution of y is "heavy tailed," one often considers the Huber cost function as follows. For $\gamma > 0$, let

$$\ell_{\texttt{sv},\gamma}(h(x), y) = \begin{cases} \frac{1}{2}(h(x) - y)^2 & \text{for } |h(x) - y| \leq \gamma, \\ \gamma|h(x) - y| - \frac{\gamma^2}{2} & \text{for } |h(x) - y| > \gamma. \end{cases} \tag{A.1}$$

  Here $\alpha(v) = \frac{1}{2}$ whenever $v \leq \gamma$ but $\alpha(v) = 0$ for all $v > \gamma$. The smoothness parameter of the population risk is $\beta_{\texttt{sv}} = \frac{1}{2}$. Localization analysis of $\alpha(v)$ is required for this loss, and the key is to avoid its inverse diverging to infinity.

- Logistic cost: consider the standard logistic regression setting, where $y \in \{-1, 1\}$ and one models the "log odd ratio" as

$$\log\left(\text{Prob}(y = 1|x)/\text{Prob}(y = -1|x)\right) = h_{\text{true}}(x). \tag{A.2}$$

  Here h_{true} is the discriminant function to be estimated (not necessarily in $\mathcal{H}$). The maximum likelihood estimation problem corresponds to using the cost function

$$\ell_{\texttt{sv}}(h(x), y) = \log\Big(1 + \exp(-yh(x))\Big).$$

  Here $\partial^2_{1,1}\ell_{\texttt{sv}}(u + y, y) = \frac{\exp(1+uy)}{(1+\exp(1+uy))^2}$, so we have $\alpha(v) = \frac{\exp(v+1)}{(\exp(v+1)+1)^2}$, $\forall v \geq 0$, and the smoothness parameter of the population risk is $\beta_{\texttt{sv}} = \frac{1}{4}$. The issue is that $\frac{1}{\alpha(v)}$, a complexity constant that will appear in the generalization error bound, grows exponentially with v (Hazan et al. (2014); Marteau-Ferey et al. (2019)). This issue strongly motivate us to localize the parameter v within $\alpha(v)$ to avoid large exponential constants.

The following assumption is usually invoked in the most representative literature on this topic (Mendelson (2014, 2018)).

Assumption 8 (optimality condition). *Recall that $h^* \in \mathbb{P}\ell_{\texttt{sv}}(h(x), y)$ is the population risk minimizer. Assume for all $h \in \mathcal{H}$,*

$$\mathbb{P}[\partial_1\ell_{\texttt{sv}}(h^*(x), y)(h(x) - h^*(x))] \geq 0.$$

We summarize the two primary settings where Assumption 8 holds true.

- Well-specified models: for certain problems, as long as the model is well-specified, then $\partial_1\ell_{\texttt{sv}}(h^*(x), y)$ is independent of x and $\mathbb{E}\partial_1\ell_{\texttt{sv}}(h^*(x), y) = 0$. Thus Assumption 8 will hold. Examples include 1) the settings studied in Mendelson (2018) where $\ell_{\texttt{sv}}$ is a univariate function of $(h(x) - y)$ and $\partial_1\ell_{\texttt{sv}}(h^*(x), y)$ is odd with respect to y, such as applications that use the square cost or the Huber cost; and 2) generalized linear models where the conditional distribution of y belongs to the exponential family, such as the the logistic regression problem (A.2).

- $\mathcal{H}$ is a convex class of functions: in this case, we verify Assumption 8 as follows. If there exists some $h_1 \in \mathcal{H}$ such that Assumption 8 is not true, then by considering $h_\lambda = \lambda h_1 + (1-\lambda)h^* \in \mathcal{H}$ with λ sufficiently close to 0, we find $\mathbb{P}\ell_{\mathtt{sv}}(h_\lambda(x), y) < \mathbb{P}\ell_{\mathtt{sv}}(h^*(x), y)$, in contradiction, as h^* is the population risk minimizer.

We call the random variable $\partial_1 \ell_{\mathtt{sv}}(h^*(x), y)$ the "noise multiplier" as it often characterizes the "effective noise" of the learning problem when using a particular cost function. We define another random variable $\xi := h^*(x) - y$. In some applications, ξ is closely related to the "noise multiplier" (e.g., they are equivalent when one uses the square cost). And the notation ξ is useful in other applications as well, because one always seeks to localize the parameter v in $\alpha(v)$ to the order of $\|\xi\|_{L_2}$. We denote $\Delta = \sup_{h\in\mathcal{H}} \|h(x) - y\|_{L_2}$ and $\Delta_\infty = \sup_{h,x,y} |h(x) - y|$ as the worst-case L_2 distance and L_∞ distance between $h(x)$ and y. respectively. It is clear that we typically have $\|\xi\|_{L_2} \ll \Delta \ll \Delta_\infty$ in practical applications.

Our analysis requires a very weak distributional assumption:

Assumption 9 ("small ball" property). *There exist constants $\kappa > 0$ and $c_\kappa \in (0, 1)$ such that for all $h \in \mathcal{H}$,*

$$\text{Prob}\left(|h(x) - h^*(x)| \geq \kappa \|h - h^*\|_{L_2}\right) \geq c_\kappa.$$

Assumption 9 is often referred to as "minimal" in the literature, and there are many examples in which it can be verified for κ and c_κ that are absolute constants (Mendelson (2014, 2018); Lecué and Mendelson (2014); Koltchinskii and Mendelson (2015); Rudelson and Vershynin (2015); Lecué and Mendelson (2018)). The scope of Assumption 9 subsumes and is much broader than the "sub-Gaussian" setting. For example, it is naturally satisfied when the class $\{h - h^* : h \in \mathcal{H}\}$ satisfies any sort of moment equivalence (see, e.g., Lemma 4.1 in Mendelson (2014)).

Main challenges. Let us first examine limitations of the results obtained using the traditional "local Rademacher complexity" analysis (Statement 1), which includes the results from Bartlett et al. (2005); Wainwright (2019); Foster and Syrgkanis (2019) in the fast-rate regime. Assuming the cost function to be $L_{\mathtt{sv}}-$Lipchitz continuous with respect to its first argument and setting $f(z) = \ell_{\mathtt{sv}}(h(x), y) - \ell_{\mathtt{sv}}(h^*(x), y)$, $T(f) = \mathbb{P}[f^2]$, and $B_e = L_{\mathtt{sv}}^2/\alpha(\Delta_\infty)$, following Statement 1, one can prove that the empirical risk minimizer $\hat{h}$ satisfies

$$\mathscr{E}(\hat{h}) \leq O\left(\frac{r^*}{B_e}\right), \tag{A.3}$$

where r^* is the fixed point of $B_e\psi$, and ψ is a sub-root surrogate function that governs $\sup_{\mathbb{P}[f^2]\leq r} (\mathbb{P} - \mathbb{P}_n)f$. Denote by r_1^* the fixed point of ψ. From the sub-root property of ψ we know that $r^* \geq B_e^2 r_1^*$, so the generalization error bound (A.3) is at least of order

$$\frac{r^*}{B_e} \geq B_e r_1^* = \frac{L_{\mathtt{sv}}^2}{\alpha(\Delta_\infty)} r_1^*. \tag{A.4}$$

The main message here is that the traditional result (A.3) is often loose and not problem-dependent. As indicated by Mendelson in a series of papers (Mendelson (2014, 2018)), the traditional result (A.3) has the following limitations.

- The global Lipchitz constant $L_{\mathtt{sv}}$ is not problem-dependent and potentially unbounded. $L_{\mathtt{sv}}$ is effectively the worst-case value $\sup_{h,x,y} |\partial_1 \ell(h(x), y)|$. For the square cost, this is

$\Delta_\infty = \sup_{h,x,y} |h(x) - y|$ and is unbounded when either the hypothesis class or noise are unbounded. It would be beneficial to have a bound that only scales with a measure related to the "noise multiplier" $\partial_1 \ell(h^*(x), y)$, because we usually have $|\partial_1 \ell(h^*(x), y)| \ll L_{\mathtt{sv}}$ in practical applications.

- The global strong convexity parameter $\alpha(\Delta_\infty)$ is often very small for the logistic cost and the Huber cost, so its inverse is often large (and potentially unbounded). The challenge here is to sharpen this to the inverse of a localized strongly convex parameter $\alpha(O(\|\xi\|_{L_2}))$. Since we usually have $\sigma \ll \Delta_\infty$, the inverse of $\alpha(O(\|\xi\|_{L_2}))$ can be much smaller than the inverse of $\alpha(\Delta_\infty)$.

The "small ball method" and beyond. The breakthrough papers Mendelson (2014, 2018) propose the "small ball method" (also referred to as "learning without concentration") to provide problem-dependent rates that overcome the limitations mentioned above. Their proofs builds on structural results of $0-1$ valued indicator functions under the small-ball condition, whose connection to the traditional localization analysis may not be completely obvious. Moving the focal point from indicator functions to "truncated" functions, we provide the following perspectives.

1) A simple interpretation to the "small-ball" condition is that, suitably "truncated" quadratic forms are of the same scale as the original quadratic forms. Under the "small-ball" condition, one can trivially show that uniformly over all $h \in \mathcal{H}$,

$$\begin{aligned}\mathbb{P}[\min\{(h(x) - h^*(x))^2, \kappa^2 \|h - h^*\|_{L_2}^2\}] &\geq \text{Prob}\left(|h(x) - h^*(x)| \geq \kappa \|h - h^*\|_{L_2}\right) \kappa^2 \|h - h^*\|_{L_2}^2 \\ &\geq c_\kappa \kappa^2 \mathbb{P}[(h(x) - h^*(x))^2].\end{aligned}$$

This suggests that one only needs to concentrate simple "truncated" functions to derive generalization error bounds.

2) One-sided uniform inequalities are contained in the "uniform localized convergence" framework and are often derived from concentration of truncated functions. Many one-sided uniform inequalities can be equivalently written as "uniform localized convergence" arguments. Consider the uniform "lower isomorphic bound" (which plays a central role in the "small-ball" method): for some constant $c > 0$, with high probability, uniformly over all $h \in \mathcal{H}$,

$$\mathbb{P}_n[(h(x) - h^*(x))^2] \geq c\mathbb{P}[(h(x) - h^*(x)^2].$$

The above argument is equivalent with the following "uniform localized convergence" argument:

$$(\mathbb{P} - \mathbb{P}_n)[(h(x) - h^*(x))^2] \leq (1 - c)T(h), \quad \forall h \in \mathcal{H}$$

where the measurement functional $T(h)$ is set to be $\|h - h^*\|_{L_2}^2$. A more flexible perspective may directly view the truncated quadratic forms as the concentrated functions, making traditional two-sided uniform convergence tools applicable in a straightforward manner.

Motivated by the above observations, an interesting open question is to recover the results in Mendelson (2014, 2018) by directly strengthening the traditional concentration framework, explicitly figuring out which component of the excess loss contributes to which part of the surrogate function. In what follows, we will present such an analysis. While our error bounds roughly follow the same form as the results in Mendelson (2014, 2018), we obtain several technical improvements; see Appendix A.3 for the novel implications and methodological contributions of our approach.

A.2 Main results and illustrative examples

We assume some regularity conditions that hold for non-pathological choices of surrogate functions.

Assumption 10 (regularity conditions on surrogate functions). *Assume there is a non-decreasing, non-negative and bounded function $\varphi(r)$ such that $\forall r > 0$,*

$$\mathfrak{R}\{h - h^* : h \in \mathcal{H}, \|h - h^*\|_{L_2}^2 \le r\} \le \varphi(r); \tag{A.5}$$

and there is a meaningful surrogate function $\varphi_{\text{noise}}(r, \delta)$ that is non-decreasing w.r.t. r, and satisfies that $\forall \delta \in (0,1)$, with probability at least $1 - \delta$,

$$\sup_{h \in \mathcal{H}, \|h - h^*\|_{L_2}^2 \le r} \{(\mathbb{P} - \mathbb{P}_n)[\partial_1 \ell_{\text{sv}}(h^*(x), y)(h - h^*)]\} \le \varphi_{\text{noise}}(r, \delta). \tag{A.6}$$

Given any fixed $\delta \in (0,1)$ and $r_0 \in (0, 4\Delta^2)$, denote $C_{r_0} = 2 + \left(\frac{16}{c_\kappa} + 2\right) \log \frac{4\Delta^2}{r_0}$. Assume there is a positive integer $\bar{N}_{\delta, r_0}$ such that for all $n \ge \bar{N}_{\delta, r_0}$,

$$\varphi_{\text{noise}}\left(8\Delta^2; \frac{\delta}{C_{r_0}}\right) \le \frac{\alpha(4\|\xi\|_{L_2}/\sqrt{c_\kappa})\|\xi\|_{L_2}^2}{2} \quad \text{and} \quad \varphi\left(8\Delta^2\right) \le \frac{\sqrt{2c_\kappa}\|\xi\|_{L_2}^2}{16\Delta}. \tag{A.7}$$

We note that the requirements do not place meaningful restrictions on the choice of surrogate function. The main requirement, condition (A.7), asks for uniform errors over $\mathcal{H}$ to be smaller than some fixed values that are independent of n. For non-pathological choices of surrogate functions, this will always be satisfied as long as the sample size n is larger than some positive integer $\bar{N}_{\delta, r_0}$. The boundedness requirement for φ (and φ_{noise}) can always be met by setting $\varphi(r) = \varphi(4\Delta^2)$ (and $\varphi_{\text{noise}}(r; \delta) = \varphi_{\text{noise}}(4\Delta^2; \delta)$) for all $r \ge 4\Delta^2$, because $\|h - h^*\|_{L_2} \le 2\Delta$ for all $h \in \mathcal{H}$.

Theorem 8 (supervised learning with structured convex cost). *Let Assumptions 7, 8, 9, 10 hold and $\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right) > 0$. Let r_{ver}^* be the fixed point of the function*

$$\frac{4}{c_\kappa \kappa^2 \cdot \alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)} \varphi_{noise}\left(2r; \frac{\delta}{C_{r_0}}\right). \tag{A.8}$$

Given any fixed $\delta \in (0,1)$ and $r_0 \in (0, 4\Delta^2)$, let r_{noise}^* be the fixed point of the function

$$\frac{8}{c_\kappa \kappa} \sqrt{2r} \varphi(2r). \tag{A.9}$$

Then with probability at least $1 - \delta$, the empirical risk minimizer $\hat{h}$ satisfies

$$\|\hat{h} - h^*\|_{L_2(\mathbb{P})}^2 \le \max\left\{r_{\text{noise}}^*,\ r_{\text{ver}}^*,\ r_0\right\} \quad \text{and}$$
$$\mathscr{E}(\hat{h}) \le \frac{\beta_{\text{sv}}}{2} \max\left\{r_{\text{noise}}^*,\ r_{\text{ver}}^*,\ r_0\right\},$$

provided that $n > \max\left\{\bar{N}_{\delta, r_0}, \frac{72}{c_\kappa^2} \log \frac{C_{r_0}}{\delta}\right\}$.

Remarks. 1) The term r_0 is negligible since it can be arbitrarily small. One can simply set $r_0 = 1/n^4$, which will be much smaller than r^*_{noise} for typical applications. 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. In the subsequent discussion, we will hide parameters that only depend on κ and c_κ in the big O notation, as they are often absolute constants in practical applications.
2) The two fixed points r^*_{noise} and r^*_{ver} correspond to the two sources of complexities: the uniform errors characterized by the two surrogate functions in (A.8) and (A.9). Recall that a fundamental limitation of the traditional "local Rademacher complexity analysis" is that it requires a "sub-root" surrogate function that can not differentiate the two sources of complexity. In contrast, the surrogate function in (A.8) (which we write as $O(\sqrt{r}\varphi(r))$ for simplicity) is obviously a "super-root" function, thus our analysis overcomes that limitation and provides more precise upper bounds. The key point is that $O(\sqrt{r}\varphi(r))$ is a benign "super-root" surrogate function, in the sense that its fixed point r^*_{ver} is "very small" when the sample size is large enough; in other words, when the problem is learnable. For example, for a $d-$dimensional linear classes, where $\varphi = O(\sqrt{dr/n})$, r^*_{ver} will be the fixed point of $O(dr/n)$. Thus r^*_{ver} will be 0 as long as the sample size n is larger than $O(d)$. Therefore, the typical generalization error derived by Theorem 8 is of order

$$\mathscr{E}(\hat{h}) \leq \frac{\beta_{\mathtt{sv}}}{2} r^*_{\text{noise}},$$

where r^*_{noise} is the fixed point of the function in (A.9). Clearly, r^*_{noise} only depends on the noise multiplier at h^* and the local strong convexity parameter, and it is independent of the worst-case parameters of the cost function.

At a high level, the subscripts "ver" and "noise" have the the meaning of "version space" and "noise multiplier," respectively. Intuitively, r_{ver} is the estimation error of the noise-free realizable problem, which reflects the complexity of version space—the random subset of $\mathcal{H}$ that consists of all h that agree with h^* on $\{x_i\}_{i=1}^n$. On the other hand, r_{noise} is the estimation error induced by the interaction of $\mathcal{H}$ and noise multiplier $\partial_1 \ell_{\text{sv}}(h^*(x), y)$. We refer to Mendelson (2014) for a more detailed discussion on the source of these two fixed points.

Now we present some representative applications of Theorem 8.

Example 6 (localization of unfavorable parameters). In practical applications, one often wants to avoid the global Lipchitz constant and the inverse of the global strong convexity parameter. For example, in regression with square cost, the global Lipchitz constant is equal to Δ_∞ and is often unbounded, so it is desirable to convert it to $\|\xi\|_{L_2}$; and in logistic regression, the inverse of global strong convexity parameter is an exponential constant $e^{O(\Delta_\infty)}$, which we hope to convert to $e^{O(\|\xi\|_{L_2})}$. These goals are achieved in Theorem 8: since the right hand side of (A.8) contains an extra $\sqrt{r}$ factor, r^*_{ver} is typically much smaller than r^*_{noise} for sufficiently large n (see remark 2 after Theorem 8). Therefore, the generalization error bound will be determined by the fixed point r^*_{noise}, which only depends on the noise multiplier at h^* and the local strong convexity parameter.

Example 7 (regression with heavy-tailed noise). We consider the problem of predicting y using $h(x)$, and allow the "noise" $\xi = h^*(x) - y$ to be heavy-tailed. To illustrate the main message of this example, we consider the $d-$dimensional linear class with sub-Gaussian features. That is, $h(x) = \theta^T x$ where $\theta \in \mathbb{R}^d$, and the random feature $x \in \mathbb{R}^d$ is sub-Gaussian. In this setting, the Huber cost is preferred to the square cost.

- For the Huber cost and truncation parameter $\gamma = O(\|\xi\|_{L_2})$ in the definition (A.1), Theorem 8 implies that the parameter v will be localized to the region where the strong convexity parameter $\alpha(v)$ is non-zero. As a result, the strong convexity parameter in the generalization

error bound will be $\frac{1}{2}$ rather than the problematic value 0 (since the generalization error scales with the inverse of $\alpha(v)$, the value 0 will make the bound vacuous). For the $d-$dimensional linear class, r^*_{ver} will be 0 as long as $n \geq O(d)$. Since $\partial_1 \ell_{\text{sv}}(h^*(x), y)$ will be uniformly bounded by $O(\sigma)$, we obtain

$$r^*_{\text{noise}} \leq O\left(\frac{\|\xi\|^2_{L_2}(d + \log\frac{1}{\delta})}{n}\right),$$

which recovers the problem-dependent rate in Mendelson (2018).

- For the square cost, the fixed point r_{noise} will often cause the generalization error to be sub-optimal. For the $d-$dimensional linear class, r_{noise} will have a polynomial dependence on $1/\delta$ as explained in Mendelson (2018). The reason is that in the definition of $\varphi_{\text{noise}}(r, \delta)$ in A.6, the noise multiplier $\partial_1 \ell_{\text{sv}}(h^*(x), y)$ is equal to ξ for the square cost. For "heavy-tailed" ξ, this will cause the rate r_{noise} to be sub-optimal.

We note that the condition that $\hat{h}$ is the empirical risk minimizer is not essential to the proof of Theorem 8. Similar to the prior work Lecué and Mendelson (2018), we can extend the result to more general learning rules that are based on regularization (e.g., LASSO (Tibshirani (1996)), SLOPE (Bogdan et al. (2015)), etc.) as follows.

Corollary 9 (extension to general regularized learning rules). *Let Assumptions 7 8, 9 hold. Let $\hat{h}$ be the solution of*

$$\min_{\mathcal{H}} \mathbb{P}_n \ell_{\text{sv}}(h(x), y) + \Psi(h), \tag{A.10}$$

where $\Psi(h)$ is a non-negative regularization term. Let $\mathcal{H}_0$ be a subset of $\mathcal{H}$ that is independent of the samples. If inequality (A.5) is modified to

$$\mathfrak{R}\{h - h^* : h \in \mathcal{H}_0, \|h - h^*\|^2_{L_2} \leq r\} \leq \varphi(r),$$

and inequality (A.6) is modified to

$$\sup_{h \in \mathcal{H}_0, \|h-h^*\|^2_{L_2} \leq r} \left\{(\mathbb{P} - \mathbb{P}_n)[\partial_1 \ell_{\text{su}}(h^*(x), y)(h - h^*)]\right\} + \Psi(h^*) \leq \varphi_{\text{noise}}(r; \delta),$$

then under Assumption 10, conditioned on the event $\{\hat{h} \in \mathcal{H}_0\}$, the conclusion of Theorem 8 remains true.

As illustrated in the following example, Corollary 9 is able to recover several important results in the high-dimensional statistics literature.

Example 8 (high-dimensional estimation and LASSO). Consider the linear regression set-up $\mathbb{E}[y|x] = x^T\theta^*$ where $\theta \in \Theta \subseteq \mathbb{R}^d$, $d \gg n$ and $\|\theta^*\|_0 \leq s \ll d$. Consider the LASSO estimator $\hat{\theta}$, which is the solution of the ℓ_1-norm regularized risk minimization problem, where the regularization term is $\Phi(h) = \lambda\|\theta\|_1$ and $\lambda > 0$ is the regularization parameter, i.e.,

$$\hat{\theta} \in \arg\min_{\Theta} \mathbb{P}_n \ell_{\text{sv}}(\theta^T x, y) + \lambda\|\theta\|_1.$$

Assume ℓ_{sv} is the square cost and ξ is $\sigma-$sub-Gaussian, or ℓ_{sv} is the Huber cost with truncation parameter $\gamma = O(\sigma)$. Assume the feature $x \in \mathbb{R}^d$ is sub-Gaussian. Following standard analysis (see,

e.g., Lemma 1 in Negahban et al. (2012)), by setting λ to be of order $\sqrt{\sigma^2 \log(d/\delta)/n}$, the Lasso estimator $\hat{\theta}$ will lie in a sparse cone Θ_S (with high probability), where it can be proven (Loh and Wainwright (2013)) that $\varphi(r) = O(\sqrt{rs \log d/n})$ and $\varphi_{\text{noise}}(r;\delta) = O(\sqrt{r\sigma^2 s \log(d/\delta)/n})$ (ignoring dependence on the parameters C and p described in Assumption 9). Applying Corollary 9 with $\mathcal{H}_0 = \{x \mapsto \theta^T x : \theta \in \Theta_S\}$ and $n \geq \Omega(s \log d)$, we have $r^*_{\text{ver}} = 0$ and

$$r^*_{\text{noise}} \leq O\left(\frac{\sigma^2 s \log \frac{d}{\delta}}{n}\right).$$

A.3 Contributions relative to previous approaches

So far we have recovered the main results in the prior works Mendelson (2014, 2018), which are valid for unbounded regression problems and thus improve the traditional "local Rademacher complexity" analysis. Now we would like to illustrate how Theorem 8 improves the results in Mendelson (2014, 2018) by removing a "star-shape" requirement. That is, we do not need to assume the hypothesis class is star-shaped/convex, or consider the star-hull of it which may increase complexity.

To be specific, Mendelson (2014, 2018) assumes that $\mathcal{H}$ is a convex class (and thus star-shaped). When $\mathcal{H}$ is not star-shaped, the results in Mendelson (2014, 2018) are still valid by taking the star-hull of $\mathcal{F}$ and considering the local Rademacher/Gaussian complexity of the star-hull. The increase in complexity is quite moderate for traditional hypothesis classes (e.g., those characterized by covering number conditions, see Lemma 4.6 in Mendelson (2002) for details). However, taking the star-hull may significantly increase the local Rademacher complexity of modern non-convex and overparameterized classes. Here we show that, even for very simple function classes (e.g., linear classes with non-convex support), our approach improves on what can be achieved using the star-hulls.

Note that the improvement brought by our approach is systematic and may carry over to more complicated learning procedures as well. A more comprehensive comparison with existing localization approaches will be presented after the following example.

Example 9 (overparameterized linear class with growing sparsity). Consider the linear regression model

$$y \sim N(x^T\theta^*, \sigma^2), \quad x \sim N(0, I_{d\times d}),$$

where $\theta^* \in \Theta \subseteq \mathbb{R}^d$ and $d \gg n$ (i.e., the model is overoarameterized). Assume the feasible parameter set θ satisfies that for all $\theta \in \Theta$,

$$\|\theta - \theta^*\|_0 \leq \lfloor \|\theta - \theta^*\|_2^2 \rfloor. \tag{A.11}$$

In other words, the sparsity of θ increases the more θ deviates from θ^*. The maximum likelihood estimation problem corresponds to minimize the empirical average of the square cost with respect to $\mathcal{H} = \{x \mapsto x^T\theta : \theta \in \Theta\}$. For this problem, the surrogate function φ_{noise} need to satisfy (with probability at $1-\delta$)

$$\sup_{\theta\in\Theta, \|\theta-\theta^*\|_2^2 \leq r} (\mathbb{P} - \mathbb{P}_n)[\xi \cdot x^T(\theta - \theta^*)] \leq \varphi_{\text{noise}}(r;\delta), \tag{A.12}$$

where the left hand side of (A.12) is the localized Gaussian complexity of $\mathcal{H}$. Thanks to the sparsity

condition (A.11), it can be tightly controlled by

$$\varphi_{\text{noise}}(r;\delta) = O\left(\sqrt{\frac{\sigma^2(\|\theta^*\|_0 + r)r\log\frac{d}{\delta}}{n}}\right) = \underbrace{O\left(\sqrt{\frac{\sigma^2\|\theta^*\|_0 r\log\frac{d}{\delta}}{n}}\right)}_{\text{problem-dependent component}} + \underbrace{O\left(\sqrt{\frac{\sigma^2\log\frac{d}{\delta}}{n}\cdot r}\right)}_{\text{benign ``super-root'' component}}. \tag{A.13}$$

Here, the benign "super-root" component" in $\varphi_{\text{noise}}(r;\delta)$ does not affect the order of its fixed point r^*_{noise}: when $n \geq \Omega(4\sigma^2\log\frac{d}{\delta})$, the "super-root" component" in (A.13) will be less than $\frac{1}{2}r$ so that r^*_{noise} is of order $\sigma^2\|\theta^*\|_0\log\frac{d}{\delta}/n$. In other words, only the problem-dependent component in $\varphi_{\text{noise}}(r;\delta)$ matters.

In contrast, if one takes the star-hull (e.g., expanding Θ to $\text{star}(\Theta) = \{\theta^* + \lambda(\theta - \theta^*) : \theta \in \Theta, \lambda \in [0,1]\}$, then it is straightforward to verify that φ_{noise} has to be a "sub-root" function. A sub-root surrogate function that governs (A.13) will be at least of order

$$\overline{\varphi_{\text{noise}}}(r;\delta) = O\left(\sqrt{\frac{\sigma^2(\|\theta^*\|_0 + \Delta)r\log\frac{d}{\delta}}{n}}\right),$$

whose fixed point unavoidably scales with the worst-case L_2 distance Δ. Here we do not consider computational issues, and the key message is that if the complexity (e.g., the "effective dimension") of an overparameterized non-convex class grows very rapidly with respect to its localization scale, then some "fast growing components" may still be benign and they may not necessarily increase the complexity. It is an open question whether such phenomena manifests in more practical applications.

Comparison with the "small ball method." In a series of pioneering works, Mendelson (Mendelson (2014, 2018, 2017a,b)) proposes the "small ball method" as an alternative approach to the traditional "concentration-contraction" framework. Under the "small ball" condition, that approach establishes one-sided uniform inequalities through structural results on binary valued indicator functions. Motivated by these works, we seek to refine the traditional concentration framework. Our approach brings added flexibility to concentration by emphasizing the use of surrogate functions that are not "sub-root," and relates one-sided uniform inequalities to two-sided concentration of simple "truncated" functions. Following are the main contributions relative to the "small ball method."

First, our approach does not require the hypothesis class to be star-shaped/convex (or to consider the star-hull of the hypothesis class). This improvement is particularly relevant for non-convex hypothesis classes whose complexity can grow rapidly when "away" from the optimal hypothesis. In Example 9 (and its discussion) we show that the improvement may be meaningful for some non-convex, overparametrized classes; and the phenomenon of "benign fast growing" components in overparameterized models may be of independent interest.

To the best of our knowledge, the "small ball method" cannot overcome the star-shape requirement in a straightforward manner, without additional uniform convergence arguments. The "small ball method" is able to prove one-sided inequalities that hold uniformly over a fixed sphere $\{h \in \mathcal{H} : \|h - h^*\|_{L_2}^2 = r\}$, and by assuming the class $\mathcal{H}$ to be star-shaped around h^*, it circumvents the need to have a uniform bound that holds simultaneously for all possible values of r. However, without the star-shape assumption and additional uniform convergence arguments, it is not clear how to uniformly extend the bound to all r using peeling. In our analysis, we introduce some new

tricks to address this issue. In particular, we use "adaptive truncation levels" and concentration over "rings." Combining these with the "uniform localized convergence" procedure, we completely circumvent the need for star-hulls (see "Part II" in Appendix D.1 for details).

The discussion here is orthogonal to lifting the star-shape/convexity assumptions using aggregation (Liang et al. (2015)), whose primary goal is to remove Assumption 8 (recall that this assumption implicitly asks the hypothesis class to be convex/star-shaped when the model is misspecified). When using aggregation and improper learning procedures, it is natural to consider the complexity of the enlarged class. Still, we suspect that taking the star-hull may be unnecessary if the enlarged class need not to be star-shaped (Mendelson (2017a,b)), and our analysis may be useful there as well. We note in passing that aggregation procedures are often computationally demanding.

Lastly, the formulation of supervised costs is slightly broader here compared with Mendelson (2018). In that paper, the loss is assumed to be a univariate function of $(h(x) - y)$, so costs involving the term $yh(x)$ (e.g., the canonical logistic cost and the costs in some other generalized linear models) are not permitted (Mendelson (2018) instead analyzes a modified version of the logistic cost).

Comparison with offset Rademacher complexity. Under the square cost and assuming the so-called "lower isometry bound" as an a priori condition (see Definition 5 in Liang et al. (2015)), offset Rademahcer complexity (Liang et al. (2015)) is also able to provide problem-dependent rates. However, establishing such a "lower isometry bound" is typically challenging, so this approach may still need to rely on the "small ball method" (or our analysis) for unbounded regression problems. Moreover, this tool is tailored to the setting of supervised learning with square cost, and it is unclear how to extend the analysis to more general losses.

Comparison with the "restricted strong convexity" framework in high-dimensional statistics. In the high-dimensional statistics literature, the "restricted strong convexity" framework (Negahban et al. (2012); Wainwright (2019)) provides analytical tools to prove problem-dependent rates, but only when such condition is assumed as an a priori (see Definition 2 in Negahban et al. (2012)). To achieve this, Raskutti et al. (2012); Negahban et al. (2012); Loh and Wainwright (2013) develop a truncation-based analysis that can establish "restricted strong convexity" for sparse kernel regression and sparse generalized linear models. Those works also indicate that one-sided uniform inequalities can be established by two-sided concentration of the "truncated" functions. There are several differences between their analysis and ours. First, those proofs rely on linearity/star-shape of the hypothesis class and thus only need to prove the "restricted strong convexity" on a fixed sphere (similar to what we have discussed in comparison with the "small-ball method"). In contrast, our framework does not put any geometric restriction on the hypothesis class, by passing this through the use of "adaptive truncation levels" and concentration over "rings," tools that may be of independent interest from a technical perspective. Second, when seeking problem-dependent generalization error bounds, the proposed $L_2 - L_4$ moment equivalence condition (Negahban et al. (2012); Loh and Wainwright (2013)) is stronger than the "small ball" condition used in our analysis. Third, the analysis does not fully localize the strong convexity parameter, and does not cover interesting supervised costs that may have zero curvature, e.g., the Huber cost.

B Proofs for Section 2 and Section 3

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 $\mathscr{A}$.

B.1 Proofs for Proposition 1 and its variants

We prove a more general version of of Proposition 1. The differences are that 1) here we use a more general "peeling scale" λ which can be any value larger than 1, while in Proposition 1 we simply set λ to be 2; and 2) we only ask $\psi(r;\delta)$ to be a high-probability surrogate function of the uniform error over the "ring" $\{f \in \mathcal{F} : r/\lambda \le T(f) \le r\}$ rather than the "bigger" localized area $\{f \in \mathcal{F} : 0 \le T(f) \le r\}$.

Proposition 3 (a more general "uniform localized convergence" argument). *For a function class $\mathcal{G} = \{g_f : f \in \mathcal{F}\}$ 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 $1-\delta$,*

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

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

$$(\mathbb{P} - \mathbb{P}_n) g_f \le \psi\left(\lambda T(f); \delta \left(\log_\lambda \frac{\lambda R}{r_0}\right)^{-1}\right).$$

Proof of Proposition 3: we apply a "peeling" technique. Given any $r_0 \in (0,R]$, take $r_k = \lambda^k r_0$, $k = 1, \cdots, \lceil \log_\lambda \frac{R}{r_0} \rceil$. Note that $\lceil \log_\lambda \frac{R}{r_0} \rceil \le \log_\lambda \frac{\lambda R}{r_0}$.

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

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

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 $\lambda^k r_0 < T(f) \le \lambda^{k+1} r_0$, and we further know that $r_{k+1} = \lambda^{k+1} r_0 \le \lambda T(f)$. Therefore, with probability at least $1-\delta$,

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

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

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

This completes the proof of Proposition 3. □

Clearly, Proposition 1 can be viewed as a corollary of Proposition 3. We now present an implication of Proposition 1, which may be more convenient to use for some problems.

Proposition 4 (a variant of the "uniform localized convergence" argument). *For a function class $\mathcal{G} = \{g_f : f \in \mathcal{F}\}$ 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 $1-\delta$,*

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

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

$$(\mathbb{P}-\mathbb{P}_n)g_f \le \psi\left(2T(f) \vee r_0; \frac{\delta}{C_{r_0}}\right),$$

where $C_{r_0} = 2\log_2 \frac{2R}{r_0}$.

Proof of Proposition 4: From Proposition 1 we know that with probability at least $1-\frac{\delta}{2}$, for all $f \in \mathcal{F}$, either $T(f) \le r_0$ or

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

We denote the event

$$\mathscr{A}_1 = \left\{\text{there exists } f \in \mathcal{F} \text{ such that } T(f) \ge r_0 \text{ and } (\mathbb{P}-\mathbb{P}_n)g_f > \psi\left(2T(f); \frac{\delta}{C_{r_0}}\right)\right\}.$$

Then from (B.1), we have

$$\text{Prob}(\mathscr{A}_1) \le \frac{\delta}{2}. \tag{B.2}$$

We denote the event

$$\mathscr{A}_2 = \left\{\text{there exists } f \in \mathcal{F} \text{ such that } T(f) > r_0 \text{ and } (\mathbb{P}-\mathbb{P}_n)g_f > \psi\left(r_0; \frac{\delta}{C_{r_0}}\right)\right\}.$$

Then from the surrogate property of ψ and the fact $C_{r_0} \ge 2$, we have

$$\text{Prob}(\mathscr{A}_2) \le \frac{\delta}{C_{r_0}} \le \frac{\delta}{2}. \tag{B.3}$$

Combining (B.2) and (B.3) by an union bound, we have

$$\text{Prob}(\mathscr{A}_1 \cup \mathscr{A}_2) \le \text{Prob}(\mathscr{A}_1) + \text{Prob}(\mathscr{A}_2) \le \delta.$$

From the above argument, it is straightforward to prove that with probability at least $1-\delta$, for all $f \in \mathcal{F}$,

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

This completes the proof of Proposition 4.

B.2 Proof of Theorem 1

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

For a fixed $r_0 \in (0, 4B^2)$, Denote $C_{r_0} = 2\log_2 \frac{8B^2}{r_0}$. From now to the end of this proof, we will prove the generalization error bound on the event

$$\mathscr{A} = \left\{ \text{for all } f \in \mathcal{F},\ (\mathbb{P} - \mathbb{P}_n) f \le \psi\left(2T(f) \vee r_0; \frac{\delta}{C_{r_0}}\right) \right\}. \tag{B.4}$$

From Proposition 4 we know that

$$\mathrm{Prob}(\mathscr{A}) \ge 1 - \delta.$$

This means that proving the generalization error bound on the event $\mathscr{A}$ suffices to prove the theorem.

Denote $g(z) = \ell(h;z) - \inf_{\mathcal{H}} \ell(h;z)$ and $\hat{g}(z) = \ell(\hat{h}_{\mathrm{ERM}};z) - \inf_{\mathcal{H}} \ell(h;z)$. Let $T(g) = \mathbb{P}[g^2]$. We have

$$f(z) = g(z) - (\ell(h^*;z) - \inf_{\mathcal{H}} \ell(h;z)), \quad \forall z,$$

which implies that

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

Therefore, we have

$$T(\hat{f}) \le 4T(\hat{g}) \vee 8B\mathcal{L}^*. \tag{B.5}$$

From the property of ERM, we have $\mathbb{P}_n \hat{f} \le 0$, which implies that

$$\mathscr{E}(\hat{h}_{\mathrm{ERM}}) \le (\mathbb{P} - \mathbb{P}_n)\hat{f} \le \psi\left(2T(\hat{f}) \vee r_0; \frac{\delta}{C_{r_0}}\right). \tag{B.6}$$

From (B.5) and (B.6) we have

$$\mathbb{P}\hat{g} - \mathcal{L}^* = \mathscr{E}(\hat{h}_{\mathrm{ERM}}) \le \psi\left(8T(\hat{g}) \vee 16B\mathcal{L}^* \vee r_0; \frac{\delta}{C_{r_0}}\right). \tag{B.7}$$

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 (B.7) we obtain

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

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

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

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

$$6B\psi\left(16B\mathcal{L}^* \vee r_0; \frac{\delta}{C_{r_0}}\right) \leq r^* \vee 2B\mathcal{L}^* \vee \frac{r_0}{8}.$$

Combining the above inequality with (B.8), we have

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

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

$$T(\hat{g}) \leq 3B\mathcal{L}^* \vee r^* \vee \frac{r_0}{8}.$$

Combining the above inequality with (B.5), we have

$$T(\hat{f}) \leq 12B\mathcal{L}^* \vee 4r^* \vee \frac{r_0}{2}.$$

From the above inequality and (B.6) we have

$$\mathscr{E}(\hat{h}_{\mathrm{ERM}}) \leq (\mathbb{P} - \mathbb{P}_n)\hat{f} \leq \psi\left(24B\mathcal{L}^* \vee 8r^* \vee r_0; \frac{\delta}{C_{r_0}}\right), \tag{B.9}$$

which implies that

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

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

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

So we finally obtain

$$\mathscr{E}(\hat{h}_{\mathrm{ERM}}) \leq \psi\left(24B\mathcal{L}^*; \frac{\delta}{C_{r_0}}\right) \vee \frac{r^*}{6B} \vee \frac{r_0}{48B}.$$

Recall that the generalization error bound holds true on the event $\mathscr{A}$ defined in (B.4), whose measure is at least $1 - \delta$. This completes the proof. □

B.3 Estimating loss-dependent rates from data

In the remarks following Theorem 1, we comment that fully data-dependent loss-dependent bounds can be derived using the empirical "effective loss," $\mathbb{P}_n[\ell(\hat{h}_{\mathrm{ERM}}; z) - \inf_{\mathcal{H}} \ell(h; z)]$ to estimate the unknown parameter $\mathcal{L}^*$. Here we present the full details and some discussion of this approach.

Theorem 10 (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* $\widehat{\mathcal{L}^*} = \mathbb{P}_n[\ell(\hat{h}_{\mathrm{ERM}};z) - \inf_{\mathcal{H}} \ell(h;z)]$. *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*

$$\mathscr{E}(\hat{h}_{\mathrm{ERM}}) \le \psi\left(cB\widehat{\mathcal{L}^*}; \frac{\delta}{C_n}\right) \vee \frac{cr^*}{B} \vee \frac{cB\log\frac{2}{\delta}}{n} \tag{B.10}$$

and

$$\mathcal{L}^* \le c_1\left(\widehat{\mathcal{L}^*} \vee \frac{r^*}{B} \vee \frac{B\log\frac{2}{\delta}}{n}\right) \le c_2\left(\mathcal{L}^* \vee \frac{r^*}{B} \vee \frac{B\log\frac{2}{\delta}}{n}\right), \tag{B.11}$$

where c, c_1, c_2 *are absolute constants.*

Remarks. 1) The $B\log\frac{2}{\delta}/n$ terms (B.10) and (B.11) 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 (B.10) shows that without knowledge of $\mathcal{L}^*$, one can estimate the order of our loss-dependent rate by using $\widehat{\mathcal{L}^*} = \mathbb{P}_n[\ell(\hat{h}_{\mathrm{ERM}};z) - \inf_{\mathcal{H}} \ell(h;z)]$ as a proxy. Despite replacing $\mathcal{L}^*$ by $\widehat{\mathcal{L}^*}$, other quantities in the bound remain unchanged in order.
3) The inequality (B.11) shows that the estimation of $\mathcal{L}^*$ is tight.

Proof of Theorem 10: from the definitions, we know that $\mathcal{L}^* = \mathbb{P}[\ell(h^*;z) - \inf_{\mathcal{H}} \ell(h^*;z)]$, $\widehat{\mathcal{L}^*} = \mathbb{P}_n[\ell(\hat{h}_{\mathrm{ERM}};z) - \inf_{\mathcal{H}} \ell(h;z)]$ and $\mathbb{P}\ell(h^*;z) \le \mathbb{P}\ell(\hat{h}_{\mathrm{ERM}};z)$. As a result, we have

$$\begin{aligned}
\mathcal{L}^* - \widehat{\mathcal{L}^*} &= \mathbb{P}\ell(h^*;z) - \mathbb{P}_n\ell(\hat{h}_{\mathrm{ERM}};z) - (\mathbb{P}-\mathbb{P}_n)[\inf_{\mathcal{H}} \ell(h;z)] \\
&\le (\mathbb{P}-\mathbb{P}_n)\ell(\hat{h}_{\mathrm{ERM}};z) - (\mathbb{P}-\mathbb{P}_n)[\inf_{\mathcal{H}} \ell(h;z)] \\
&= (\mathbb{P}-\mathbb{P}_n)\hat{f} + (\mathbb{P}-\mathbb{P}_n)[\ell(h^*;z) - \inf_{\mathcal{H}} \ell(h;z)],
\end{aligned} \tag{B.12}$$

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

We take $r_0 = \frac{B^2}{n}$ in Theorem 1, and denote $C_n := C_{r_0} = 2\log_2 n + 6$. From (B.9) in the proof of Theorem 1, on the event $\mathscr{A}$ defined in (B.4) (whose measure is at least $1-\delta$),

$$\mathscr{E}(\hat{h}_{\mathrm{ERM}}) \le (\mathbb{P}-\mathbb{P}_n)\hat{f} \le \psi(24B\mathcal{L}^* \vee 8r^* \vee \frac{B^2}{n}; \frac{\delta}{C_n}), \tag{B.13}$$

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

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

$$\begin{aligned}
(\mathbb{P}-\mathbb{P}_n)\hat{f} &\le \psi\left(8\left(3B\mathcal{L}^* \vee r^* \vee \frac{B^2}{8n}\right); \frac{\delta}{C_n}\right) \\
&\le \frac{3B\mathcal{L}^* \vee r^* \vee \frac{B^2}{8n}}{6B} \le \frac{\mathcal{L}^*}{2} + \frac{r^*}{6B} + \frac{B}{48n}.
\end{aligned} \tag{B.14}$$

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 $\mathrm{Var}[\ell(h^*;z) - \inf_{\mathcal{H}} \ell(h;z)] \le 2B\mathcal{L}^*$, with probability at least $1-\delta$,

$$\left|(\mathbb{P} - \mathbb{P}_n)[\ell(h^*;z) - \inf_{\mathcal{H}} \ell(h;z)]\right| \le \sqrt{\frac{4B\mathcal{L}^* \log\frac{2}{\delta}}{n}} + \frac{2B\log\frac{2}{\delta}}{n} \le \frac{\mathcal{L}^*}{4} + \frac{3B\log\frac{2}{\delta}}{n}. \tag{B.15}$$

Consider the event

$$\mathscr{A}_3 = \mathscr{A} \cup \{\text{inequality (B.15) holds true}\},$$

whose measure is at least $1-2\delta$. On the event $\mathscr{A}_3$, from inequalities (B.12) (B.14) (B.15), 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

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

From this result and (B.13), it is straightforward to show that

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

where c is an absolute constant.

We also have

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

From this result and (B.15), on the event $\mathscr{A}_3$,

$$\widehat{\mathcal{L}^*} \le \frac{5}{4}\mathcal{L}^* + \frac{3B\log\frac{2}{\delta}}{n}. \tag{B.17}$$

Combine (B.16) and (B.17) 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. □

B.4 Proof of Theorem 2

The main goal of this subsection is to prove Theorem 2. We first prove Theorem 11 (the bound (3.6) in the main paper), a guarantee for the second-stage moment penalized estimator $\hat{h}_{\mathrm{MP}}$. In order to prove Theorem 2, we then combine Theorem 11 with a guarantee for the first-stage empirical risk minimization (ERM) estimator.

B.4.1 Analysis for the second-stage moment-penalized estimator

Theorem 11 (variance-dependent rate of the second-stage estimator). *Given arbitrary preliminary estimate $\widehat{\mathcal{L}_0^*} \in [-B, B]$, the generalization error of the moment-penalized estimator $\hat{h}_{\mathrm{MP}}$ in Strategy 2 is bounded by*

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

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 11: the proof of Theorem 11 consist of four parts.

Part I: use ψ to upper bound localized empirical processes. Let $\mathcal{F}$ be the excess loss class in (3.2), and define its member f is defined by $f(z) = \ell(h; z) - \ell(h^*; z), \forall z \in \mathcal{Z}$. We have the following lemma.

Lemma 3 (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 $r > 0$,*

$$\sup_{\mathbb{P}[f^2] \leq r} (\mathbb{P} - \mathbb{P}_n) f \leq \psi\left(r \vee r_1^*(\delta_1); \delta_1\right). \tag{B.18}$$

Proof of Lemma 3: clearly, $\mathcal{F}$ is uniformly bounded in $[-2B, 2B]$. When $\mathbb{P}[f^2] \leq r$, we have $\mathbb{P}[f^4] \leq 4B^2 r$. From Lemma 5 (the two-sided version of its second inequality), with probability at least $1 - \frac{\delta_1}{2}$,

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

where the last inequality follows from the Lipchitz contraction property of Rademahcer complexity (see, e.g., Theorem 7 in Meir and Zhang (2003)), and the fact that for all $f_1, f_2 \in \mathcal{F}$, $|f_1^2(z) - f_2^2(z)| \leq 4B|f_1(z) - f_2(z)|$. We conclude that with probability at least $1 - \frac{\delta_1}{2}$,

$$\sup_{\mathbb{P}[f^2] \leq r} \left|(\mathbb{P} - \mathbb{P}_n) f^2\right| \leq \varphi_{\delta_1}(r), \tag{B.19}$$

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

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 (B.19) 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 $1 - \frac{\delta_1}{2}$,

$$\sup_{\mathbb{P}[f^2] \leq r} \left|(\mathbb{P} - \mathbb{P}_n) f^2\right| \leq \frac{r}{4}. \tag{B.20}$$

(B.20) 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

$$\mathbb{P}_n[f^2] \le \frac{5}{4} r \le 2r. \tag{B.21}$$

Again from the two-sided version of the second inequality in Lemma 5, 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 (B.21) 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

$$\begin{aligned}\sup_{\mathbb{P}[f^2]\le r} (\mathbb{P}-\mathbb{P}_n)f &\le 4\mathfrak{R}_n\{f : \mathbb{P}[f^2] \le r\} + \sqrt{\frac{2r\log\frac{8}{\delta_1}}{n}} + \frac{9B\log\frac{8}{\delta_1}}{n} \\ &\le 4\mathfrak{R}_n\{f : \mathbb{P}_n[f^2] \le 2r\} + \sqrt{\frac{2r\log\frac{8}{\delta_1}}{n}} + \frac{9B\log\frac{8}{\delta_1}}{n}. \end{aligned} \tag{B.22}$$

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 (B.22), we see that with probability at least $1-\delta_1$, for all $r > 0$,

$$\sup_{\mathbb{P}[f^2]\le r} (\mathbb{P}-\mathbb{P}_n)f \le \psi\left(r \vee r_2^*(\delta_1);\delta_1\right). \tag{B.23}$$

From (B.23), in order to prove the result (B.18) in Lemma 3, we only need to prove that

$$r_2^*(\delta_1) \le r_1^*(\delta_1). \tag{B.24}$$

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 $r > r_1^*(\delta_1)$,

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

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

$$r_2^*(\delta_1) > 16B\psi(r_2^*(\delta_1);\delta_1) \ge 4\varphi_{\delta_1}(r_2^*(\delta_1)). \tag{B.25}$$

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

$$4\varphi_{\delta_1}(r_2^*(\delta_1)) = r_2^*(\delta_1). \tag{B.26}$$

The above two inequalities (B.25) and (B.26) result in a contradiction. So the assumption $r_2^*(\delta_1) > r_1^*(\delta_1)$ is false. Therefore $r_2^*(\delta_1) \le r_1^*(\delta_1)$, and this completes the proof of Lemma 3. □

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

Based on Lemma 3, 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 4 (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 $1-\left(\log_2 \frac{8B^2 \vee 2r_1^*(\delta_1)}{r_1^*(\delta_1)} + \frac{1}{2}\right)\delta_1$, for all $f \in \mathcal{F}$ either $\mathbb{P}[f^2] \le r_1^*(\delta_1)$, or*

$$(\mathbb{P}-\mathbb{P}_n)f \le \psi\left(4\mathbb{P}_n[f^2];\delta_1\right). \tag{B.27}$$

Proof of Lemma 4: 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 $r_1^*(\delta_1) \ge \frac{144B^2 \log\frac{8}{\delta_1}}{n} > 0$. 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 $\lceil \log_2 \frac{R}{r_0} \rceil \le \log_2 \frac{2R}{r_0}$. We use the union bound to establish that $\sup_{\mathbb{P}[f^2]\le r}(\mathbb{P}-\mathbb{P}_n)f \le \psi(r;\delta_1)$ holds for all $\{r_k\}$ simultaneously: with probability at least $1-\log_2\frac{2R}{r_0}\delta_1$,

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

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

$$\begin{aligned}\mathbb{P}f &\le \mathbb{P}_n f + \sup_{\tilde{f}\in\mathcal{F}:\mathbb{P}[\tilde{f}^2]\le r_{k+1}}(\mathbb{P}-\mathbb{P}_n)\tilde{f} \\ &\le \mathbb{P}_n f + \psi(r_{k+1};\delta_1)\end{aligned} \tag{B.28}$$

By (B.19) we know that with probability at least $1-\frac{\delta_1}{2}$,

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

for all $r > r_0$ (here we have used the fact $r_0 = r_1^*(\delta_1) \ge r_2^*(\delta_1)$, which is the result (B.24) in the proof of Lemma 3). From the union bound, with probability at least $1-(\log_2\frac{2R}{r_0}+\frac{1}{2})\delta_1$, the condition $r_{k+1} \ge \mathbb{P}[f^2] > r_k$ will imply

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

so

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

Combining this result with (B.28), we have that for all f such that $T(f) > r_0$, with probability at least $1-\left(\log_2\frac{2R}{r_0}+\frac{1}{2}\right)\delta_1$,

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

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

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

This completes the proof of Lemma 4. □

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

We define the event

$$\mathscr{A}_1=\left\{\text{there exists } f\in\mathcal{F} \text{ such that } \mathbb{P}[f^2]\ge r_0 \text{ and } (\mathbb{P}-\mathbb{P}_n)f>\psi\left(4\mathbb{P}_n[f^2];\delta_1\right)\right\}.$$

Lemma 4 has proven that

$$\text{Prob}(\mathscr{A}_1)\le\left(\log_2\frac{8B^2\vee 2r_1^*(\delta_1)}{r_1^*(\delta_1)}+\frac{1}{2}\right)\delta_1. \tag{B.29}$$

We denote the event

$$\mathscr{A}_2=\{\text{there exists } f\in\mathcal{F} \text{ such that } \mathbb{P}[f^2]\le r_0 \text{ and } (\mathbb{P}-\mathbb{P}_n)f>\psi(r_0;\delta_1)\}.$$

Due to the surrogate property of ψ, we have

$$\text{Prob}\left(\mathscr{A}_2\right)\le\delta_1. \tag{B.30}$$

Denote the event

$$\mathscr{A}=\left\{\text{for all } f\in\mathcal{F},\ (\mathbb{P}-\mathbb{P}_n)f\le\psi\left(4\mathbb{P}_n[f^2]\vee r_1^*(\delta_1);\delta_1\right)\right\}.$$

From (B.29) and (B.30), it is straightforward to prove that

$$\begin{aligned}\text{Prob}(\mathscr{A}) &\ge 1-\text{Prob}(\mathscr{A}_1)-\text{Prob}(\mathscr{A}_2)\\ &\ge 1-\left(\log_2\frac{8B^2\vee 2r_1^*(\delta_1)}{r_1^*(\delta_1)}+\frac{1}{2}\right)\delta_1-\delta_1\\ &\ge 1-\left(\log_2\frac{8B^2\vee 2r_1^*(\delta_1)}{r_1^*(\delta_1)}+\frac{3}{2}\right)\delta_1.\end{aligned} \tag{B.31}$$

Since f is the excess loss. we can equivalently write the event $\mathscr{A}$ as

$$\mathscr{A}=\left\{\text{for all } h\in\mathcal{H},\ \mathscr{E}(h)\le\mathbb{P}_n[\ell(h;z)-\ell(h^*z)]+\psi\left(4\mathbb{P}_n[(\ell(h;z)-\ell(h^*z))^2]\vee r_1^*(\delta_1);\delta_1\right)\right\}. \tag{B.32}$$

Denote $w(h;z)=\ell(h;z)-\widehat{\mathcal{L}_0^*}$, we have

$$\begin{aligned}4\mathbb{P}_n[(\ell(h;z)-\ell(h^*z))^2] &\le 8\mathbb{P}_n[w(h;z)^2]+8\mathbb{P}_n[w(h^*;z)^2]\\ &\le 16\mathbb{P}_n[w(h;z)^2]\vee 16\mathbb{P}_n[w(h^*;z)^2].\end{aligned}$$

From the above conclusion and (B.32), we obtain that on the event $\mathscr{A}$,

$$\begin{aligned}\mathscr{E}(h) + \mathbb{P}_n\ell(h^*;z) &\le \mathbb{P}_n\ell(h;z) + \psi(4\mathbb{P}_n[(\ell(h;z)-\ell(h^*;z))^2] \vee r_1^*(\delta_1);\delta_1) \\ &\le \mathbb{P}_n(h;z) + \psi\bigg(16\mathbb{P}_n[w(h;z)^2] \vee 16\mathbb{P}_n[w(h^*;z)^2] \vee r_1^*(\delta_1);\delta_1\bigg) \\ &\le \mathbb{P}_n(h;z) + \psi\bigg(16\mathbb{P}_n[w(h;z)^2]\delta_1\bigg) + \psi\bigg(16\mathbb{P}_n[w(h^*;z)^2] \vee r_1^*(\delta_1);\delta_1\bigg). \end{aligned} \tag{B.33}$$

We specify the moment-penalized estimator to be

$$\hat{h}_{\mathrm{MP}} = \arg\min_{\mathcal{H}} \left\{ \mathbb{P}_n\ell(h;z) + \psi\bigg(16\mathbb{P}_n[(\ell(h;z)-\widehat{\mathcal{L}_0^*})^2];\delta_1\bigg)\right\}.$$

Then we have

$$\mathbb{P}_n\ell(\hat{h}_{\mathrm{MP}};z) + \psi\bigg(16\mathbb{P}_n[w(\hat{h}_{\mathrm{MP}};z)^2];\delta_1\bigg) \le \mathbb{P}_n\ell(h^*;z) + \psi\bigg(16\mathbb{P}_n[w(h^*;z)^2];\delta_1\bigg) \tag{B.34}$$

Therefore, on the event $\mathscr{A}$,

$$\begin{aligned}\mathscr{E}(\hat{h}_{\mathrm{MP}}) &\le \mathbb{P}_n\ell(\hat{h}_{\mathrm{MP}};z) + \psi\bigg(16\mathbb{P}_n[w(\hat{h}_{\mathrm{MP}};z)^2];\delta_1\bigg) + \psi\bigg(16\mathbb{P}_n[w(h^*;z)^2] \vee r_1^*(\delta_1);\delta_1\bigg) - \mathbb{P}_n\ell(h^*;z) \\ &= \arg\min_{\mathcal{H}} \left\{ \mathbb{P}_n\ell(h;z) + \psi\bigg(16\mathbb{P}_n[w(h;z)];\delta_1\bigg)\right\} - \mathbb{P}_n\ell(h^*;z) + \psi\bigg(16\mathbb{P}_n[w(h^*;z)^2] \vee r_1^*(\delta_1);\delta_1\bigg) \\ &\le \psi\bigg(16\mathbb{P}_n[w(h^*;z)^2];\delta_1\bigg) + \psi\bigg(16\mathbb{P}_n[w(h^*;z)^2] \vee r_1^*(\delta_1);\delta_1\bigg) \\ &\le 2\psi\bigg(16\mathbb{P}_n[w(h^*;z)^2] \vee r_1^*(\delta_1);\delta_1\bigg), \end{aligned} \tag{B.35}$$

where the first inequality is due to (B.33) and the second inequality is due to (B.34).

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

$$\begin{aligned}\mathbb{P}_n[w(h^*;z)^2] &\le \mathbb{P}[w(h^*;z)^2] + 2B\sqrt{\frac{2\mathbb{P}[w(h^*;z)^2]\log\frac{2}{\delta_2}}{n}} + \frac{4B^2\log\frac{2}{\delta_2}}{n} \\ &\le 2\mathbb{P}[w(h^*;z)^2] + \frac{6B^2\log\frac{2}{\delta_2}}{n}. \end{aligned} \tag{B.36}$$

From (B.31) (B.35) (B.36), with probability at least

$$\mathrm{Prob}(\mathscr{A}) - \delta_2 \ge 1 - \left(\log_2\frac{8B^2 \vee 2r_1^*(\delta_1)}{r_1^*(\delta_1)} + \frac{3}{2}\right)\delta_1 - \delta_2,$$

we have

$$\begin{aligned}\mathscr{E}(\hat{h}_{\mathrm{MP}}) &\le 2\psi\left(16\mathbb{P}_n[w(h^*;z)] \vee r_1^*(\delta_1) \vee \frac{B^2}{n};\delta_1\right) \\ &\le 2\psi\left(\left(32\mathbb{P}[w(h^*;z)^2] + \frac{96B^2\log\frac{2}{\delta_2}}{n}\right) \vee r_1^*(\delta_1) \vee \frac{B^2}{n};\delta_1\right), \end{aligned} \tag{B.37}$$

where the first inequality is due to (B.35) and the second inequality is due to (B.36).

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

$$r_1^*(\delta_1) \geq \frac{144B^2 \log \frac{8}{\delta_1}}{n}. \tag{B.38}$$

Denote $C_n := 2\log_2 n + 5$ and take

$$\delta_1 = \frac{\delta}{C_n},$$

then we have

$$\begin{aligned} 2\log_2 \frac{8B^2 \vee 2r_1^*(\delta_1)}{r_1^*(\delta_1)} + 3 &\leq \max\left\{2\log_2 \frac{8n}{144\log 8}, 2+3\right\} \\ &\leq \max\{2\log_2 n, 5\} \leq C_n, \end{aligned}$$

so

$$\left(\log_2 \frac{8B^2 \vee 2r_1^*(\delta_1)}{r_1^*(\delta_1)} + \frac{3}{2}\right)\delta_1 \leq \frac{\delta}{2}. \tag{B.39}$$

Set $r^* = r_1^*(\delta_1)$ and take $\delta_2 = \frac{\delta}{2}$. From (B.37), we obtain that with probability at least $1-\delta$, the generalization error of $\hat{h}_{\mathrm{MP}}$ is upper bounded by

$$\mathscr{E}(\hat{h}_{\mathrm{MP}}) \leq 2\psi\left(c\left[\mathbb{P}[w(h^*;z)^2] \vee r^* \vee \frac{B^2 \log \frac{4}{\delta}}{n}\right]; \frac{\delta}{C_n}\right), \tag{B.40}$$

where c is an absolute constant. From (B.38) 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 (B.40), we obtain that

$$\begin{aligned} \mathscr{E}(\hat{h}_{\mathrm{MP}}) &\leq 2\psi\left(c\left[\mathbb{P}[(\ell(h^*;z) - \widehat{\mathcal{L}}_0^*)^2] \vee r^*\right]; \frac{\delta}{C_n}\right) \\ &\leq 2\psi\left(c_0\left[\mathcal{V}^* \vee r^* \vee (\widehat{\mathcal{L}}_0^* - \mathcal{L}_0^*)^2\right]; \frac{\delta}{C_n}\right). \end{aligned} \tag{B.41}$$

where c_0 is an absolute constant. This completes the proof of Theorem 11. □

B.4.2 Analysis of the first-stage ERM estimator

After proving Theorem 11, the remaining part needed to prove Theorem 2 is to bound $(\widehat{\mathcal{L}}_0^* - \mathcal{L}_0^*)^2$—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 11 to prove Theorem 2. Recall that $\mathbb{P}_{S'}$ is the empirical distribution of the "auxiliary" data set. Denote $\hat{h}_{\mathrm{ERM}} \in \arg\min_{\mathcal{H}} \mathbb{P}_{S'}\ell(h;z)$.

From Part I in the proof of Theorem 11, $\forall \delta \in (0, \frac{1}{2})$, 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 $\hat{h}_{\mathrm{ERM}} \in \arg\min_{\mathcal{H}} \mathbb{P}_{S'} \ell(h; z)$ and $\mathbb{P}_{S'} \ell(\hat{h}_{\mathrm{ERM}}; z) = \widehat{\mathcal{L}_0^*}$, we have

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

and

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

Hence we have

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

Combine this result with (B.41), we have with probability $1 - 2\delta$,

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

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. □

B.5 Estimating variance-dependent rates from data

In the remark following Theorem 2, we comment that fully data-dependent variance-dependent bounds can be derived by employing an empirical estimate to the unknown parameter $\mathcal{V}^*$. Here we present the full details and some discussion of this approach.

Theorem 12 (estimate of the variance-dependent rate from data). *Consider the empirical centered second moment*

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

where $\widehat{\mathcal{L}_0^*} \in [-B, B]$ *is the preliminary estimate of* $\mathcal{L}^*$ *obtained in the first-stage,* ψ *is defined in Strategy 2, and*

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

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

$$\mathscr{E}(\hat{h}_{\mathrm{MP}}) \leq 4\psi\left(16\widehat{\mathcal{V}^*}; \frac{\delta}{C_n}\right) \vee \frac{r^*}{8B}, \tag{B.42}$$

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

Remarks. 1) The subscript "NMP" within $\hat{h}_{\mathrm{NMP}}$ means "negative moment penalization." Note that $\hat{h}_{\mathrm{NMP}}$ may not have good generalization performance, it is only used to compute $\widehat{\mathcal{V}^*}$ so that we can evaluate the estimator $\hat{h}_{\mathrm{MP}}$ proposed in Strategy 2.
2) While the fully data-dependent generalization error bound (B.42) 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, (B.42) may not be as tight as the original variance-dependent rate in Theorem 2. One should view (B.42) 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 Theorem 12 with the guarantee on $\widehat{\mathcal{L}_0^*}$ proved in Appendix B.4.2.

Proof of Theorem 12: define $\hat{f}_{\mathrm{NMP}}$ by $\hat{f}_{\mathrm{NMP}}(z) = \ell(\hat{h}_{\mathrm{NMP}}; z) - \ell(h^*; z), \forall z \in \mathcal{Z}$, and $w(h; z) = \ell(h; z) - \widehat{\mathcal{L}_0^*}$. From the the results (B.31) (B.35) (B.39) in the proof of Theorem 11, we have with probability at least $1 - \frac{\delta}{2}$,

$$(\mathbb{P} - \mathbb{P}_n) f \leq \psi\left(4\mathbb{P}_n[f^2] \vee r^*; \frac{\delta}{C_n}\right), \quad \forall f \in \mathcal{F} \tag{B.43}$$

and

$$\mathscr{E}(\hat{h}_{\mathrm{MP}}) \leq 2\psi\left(16\mathbb{P}_n[w(h^*; z)^2] \vee r^*; \frac{\delta}{C_n}\right). \tag{B.44}$$

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

$$\mathbb{P}_n \ell(\hat{h}_{\mathrm{NMP}}; z) - 2\psi\left(16\mathbb{P}_n[w(\hat{h}_{\mathrm{NMP}}; z)^2]; \frac{\delta}{C_n}\right) \leq \mathbb{P}_n \ell(h^*; z) - 2\psi\left(16\mathbb{P}_n[w(h^*; z)^2]; \frac{\delta}{C_n}\right). \tag{B.45}$$

Therefore, with probability at least $1 - \frac{\delta}{2}$, we have

$$
\begin{aligned}
&2\psi\left(16\mathbb{P}_n[w(h^*;z)^2];\frac{\delta}{C_n}\right)\\
&\le 2\psi\left(16\mathbb{P}_n[w(\hat{h}_{\mathrm{NMP}};z)^2];\frac{\delta}{C_n}\right)+\mathbb{P}_n\ell(h^*;z)-\mathbb{P}_n\ell(\hat{h}_{\mathrm{NMP}};z)\\
&= 2\psi\left(16\mathbb{P}_n[w(\hat{h}_{\mathrm{NMP}};z)^2];\frac{\delta}{C_n}\right)+\mathbb{P}[\ell(h^*;z)-\ell(\hat{h}_{\mathrm{NMP}};z)]+(\mathbb{P}_n-\mathbb{P})[\ell(h^*;z)-\ell(\hat{h}_{\mathrm{NMP}};z)]\\
&\le 2\psi\left(16\mathbb{P}_n[w(\hat{h}_{\mathrm{NMP}};z)^2];\frac{\delta}{C_n}\right)+(\mathbb{P}-\mathbb{P}_n)\hat{f}_{\mathrm{NMP}}\\
&\le 2\psi\left(16\mathbb{P}_n[w(\hat{h}_{\mathrm{NMP}};z)^2];\frac{\delta}{C_n}\right)+\psi\left(4\mathbb{P}_n[\hat{f}^2_{\mathrm{NMP}}];\frac{\delta}{C_n}\right), \qquad \text{(B.46)}
\end{aligned}
$$

where the first inequality is due to (B.45), the second inequality is due to the fact that h^* minimizes the population risk; and the last inequality is due to (B.43).

Note that

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

From the above inequality and (B.46), with probability at least $1-\frac{\delta}{2}$, we have

$$
\begin{aligned}
&2\psi\left(16\mathbb{P}_n[w(h^*;z)^2];\frac{\delta}{C_n}\right)\\
&\le\ 2\psi\left(16\mathbb{P}_n[w(\hat{h}_{\mathrm{NMP}};z)^2];\frac{\delta}{C_n}\right)+\psi\left(16\mathbb{P}_n[w(\hat{h}_{\mathrm{NMP}};z)^2];\frac{\delta}{C_n}\right)\vee\psi\left(16\mathbb{P}_n[w(h^*;z)^2];\frac{\delta}{C_n}\right). \qquad \text{(B.47)}
\end{aligned}
$$

Whether $\mathbb{P}_n[w(h^*;z)^2] \le 16\mathbb{P}_n[w(\hat{h}_{\mathrm{NMP}};z)^2]$ or $\mathbb{P}_n[w(h^*;z)^2] > 16\mathbb{P}_n[w(\hat{h}_{\mathrm{NMP}};z)^2]$, the inequality (B.47) always implies

$$
\psi\left(16\mathbb{P}_n[w(h^*;z)^2];\frac{\delta}{C_n}\right)\le 2\psi\left(16\mathbb{P}_n[w(\hat{h}_{\mathrm{NMP}};z)^2];\frac{\delta}{C_n}\right)=2\psi\left(16\widehat{\mathcal{V}^*};\frac{\delta}{C_n}\right). \qquad \text{(B.48)}
$$

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

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

where the first inequality is due to (B.44) and the last inequality is due to (B.48). This completes the proof. □

B.6 Auxiliary lemmata

Lemma 5 (**Talagrand's concentration inequality for empirical processes, Bartlett et al. (2005)**). *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 $f \in \mathcal{F}$, $\mathrm{Var}[f(z_i)] \leq r$. 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 6 (**Bernstein's inequality, Dirksen (2015)**). *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 $1-\delta$

$$\left|\frac{1}{n}\sum_{i=1}^{n}X_i\right| \leq \sqrt{\frac{2\sigma^2\log\frac{2}{\delta}}{n}} + \frac{B\log\frac{2}{\delta}}{n}. \tag{B.49}$$

C Proofs for Section 5, Section 6 and Section 7

C.1 Proof of Lemma 2

Fix $u \in \mathcal{B}^d(0,1)$ and $\theta_1, \theta_2 \in \Theta$, then we have

$$u^T(\nabla\ell(\theta_1;z) - \nabla\ell(\theta_2;z))^T = \int_0^1 u^T[\nabla^2\ell(\theta_2 + v(\theta_1-\theta_2);z)](\theta_1-\theta_2)dv.$$

By Jensen's inequality,

$$\begin{aligned}\exp\left(\frac{u^T(\nabla\ell(\theta_1;z)-\nabla\ell(\theta_2;z))}{\beta\|\theta_1-\theta_2\|}\right) &= \exp\left(\int_0^1 u^T[\nabla^2\ell(\theta_2+v(\theta_1-\theta_2);z)]\frac{(\theta_1-\theta_2)}{\|\theta_1-\theta_2\|}dv\right)\\ &\leq \int_0^1 \exp\left(u^T[\nabla^2\ell(\theta_2+v(\theta_1-\theta_2);z)]\frac{(\theta_1-\theta_2)}{\|\theta_1-\theta_2\|}\right)dv.\end{aligned}$$

It is then straightforward to prove the lemma by taking expectation with respect to z in the above inequality and using the condition (5.2). □

C.2 Proof of Proposition 2

Take $V = \{v \in \mathbb{R}^d : \|v\| \le \max\{\Delta_M, \frac{1}{n}\}\}$. We will first prove a "uniform localized convergence" argument over all $\theta \in \Theta$ and $v \in V$.

Proposition 5 (directional "uniform localized convergence" of gradient). *Under Assumption 1, $\forall \delta \in (0,1)$, with probability at least $1-\delta$, for all $\theta \in \Theta, v \in V$, either $\|\theta-\theta^*\|^2 + \|v\|^2 \le \frac{2}{n^2}$, or*

$$
\begin{aligned}
&(\mathbb{P}-\mathbb{P}_n)\left[(\nabla \ell(\theta;z) - \nabla \ell(\theta^*;z))^T v\right] \\
&\le c_1 \beta \max\left\{\|\theta-\theta^*\|^2 + \|v\|^2, \frac{2}{n^2}\right\}\left(\sqrt{\frac{d+\log\frac{2\log_2(2n^2\Delta_M^2+2)}{\delta}}{n}} + \frac{d+\log\frac{2\log_2(2n^2\Delta_M^2+2)\}}{\delta}}{n}\right),
\end{aligned}
$$

where c_1 is an absolute constant.

Proof of Proposition 5: for $(\theta, v) \in \Theta \times V$, let $g_{(\theta,v)} = (\nabla \ell(\theta;z) - \nabla \ell(\theta^*;z))^T v$. For (θ_1, v_1) and $(\theta_2, v_2) \in \Theta \times v$, define the norm on the product space $\Theta \times V$ as

$$
\|(\theta_1, v_1) - (\theta_2, v_2)\|_{\text{pro}} = \sqrt{\|\theta_1-\theta_2\|^2 + \|v_1-v_2\|^2}. \tag{C.1}
$$

Denote $\mathcal{B}(\sqrt{r}) := \{(\theta, v) \in \Theta \times V : \|\theta-\theta^*\|^2 + \|v\|^2 \le r\}$. Given $(\theta_1, v_1), (\theta_2, v_2) \in \mathcal{B}(\sqrt{r})$, we perform the following re-arrangement and decomposition steps:

$$
\begin{aligned}
&g_{(\theta_1,v_1)}(z) - g_{(\theta_2,v_2)}(z) \\
&= (\nabla \ell(\theta_1;z) - \nabla \ell(\theta^*;z))^T v_1 - (\nabla \ell(\theta_2;z) - \nabla \ell(\theta^*;z))^T v_2 \\
&= (\nabla \ell(\theta_1;z) - \nabla \ell(\theta^*;z))^T (v_1-v_2) + (\nabla \ell(\theta_1;z) - \nabla \ell(\theta^*;z))^T v_2 + (\nabla \ell(\theta^*;z) - \nabla \ell(\theta_2;z))^T v_2 \\
&= (\nabla \ell(\theta_1;z) - \nabla \ell(\theta^*;z))^T (v_1-v_2) + (\nabla \ell(\theta_1;z) - \nabla \ell(\theta_2;z))^T v_2
\end{aligned} \tag{C.2}
$$

When $(\theta_1, v_1), (\theta_2, v_2) \in \mathcal{B}(\sqrt{r})$, we have

$$
\|\theta_1-\theta^*\|\|v_1-v_2\| \le \sqrt{r}\|v_1-v_2\| \le \sqrt{r}\|(\theta_1, v_1) - (\theta_2, v_2)\|_{\text{pro}},
$$

so from Assumption 1, $(\nabla \ell(\theta_1;z) - \nabla \ell(\theta^*;z))^T(v_1-v_2)$ is $\beta\sqrt{r}\|(\theta_1, v_1)-(\theta_2, v_2)\|_{\text{pro}}$−sub-exponential. Similarly, we can prove $(\nabla \ell(\theta_1;z) - \nabla \ell(\theta_2;z))^T v_2$ to be $\beta\sqrt{r}\|(\theta_1, v_1)-(\theta_2, v_2)\|_{\text{pro}}$−sub-exponential. From the decomposition (C.2) and Jensen's inequality, for all $(\theta_1, v_1), (\theta_2, v_2) \in \mathcal{B}(\sqrt{r})$, we have

$$
\begin{aligned}
&\exp\left(\frac{g_{(\theta_1,v_1)}(z) - g_{(\theta_2,v_2)}(z)}{2\beta\sqrt{r}\|(\theta_1, v_1) - (\theta_2, v_2)\|_{\text{pro}}}\right) \\
&\le \frac{1}{2}\exp\left(\frac{(\nabla \ell(\theta_1;z) - \nabla \ell(\theta^*;z))^T(v_1-v_2)}{\beta\sqrt{r}\|(\theta_1, v_1) - (\theta_2, v_2)\|_{\text{pro}}}\right) + \frac{1}{2}\exp\left(\frac{(\nabla \ell(\theta_1;z) - \nabla \ell(\theta_2;z))^T v_2}{\beta\sqrt{r}\|(\theta_1, v_1) - (\theta_2, v_2)\|_{\text{pro}}}\right).
\end{aligned}
$$

By taking expectation with respect to z in the above inequality, we prove that $g_{(\theta_1,v_1)}(z) - g_{(\theta_2,v_2)}(z)$ is a $2\beta\sqrt{r}\|(\theta_1, v_1) - (\theta_2, v_2)\|_{\text{pro}}$−sub-exponential random variable, i.e.,

$$
\|g_{(\theta_1,v_1)}(z) - g_{(\theta_2,v_2)}(z)\|_{\text{Orlicz}_1} \le 2\beta\sqrt{r}\|(\theta_1, v_1) - (\theta_2, v_2)\|_{\text{pro}}.
$$

From Bernstein inequality for sub-exponential variables (Lemma 10), for any fixed $u \ge 0$ and $(\theta_1, v_1), (\theta_2, v_2) \in \Theta \times V$,

$$
\begin{aligned}
\text{Prob}\Bigg\{|(\mathbb{P}-\mathbb{P}_n)[g_{(\theta_1,v_1)}(z) - g_{(\theta_2,v_2)}(z)]| \ge 2\beta\sqrt{r}\|(\theta_1, v_1) - (\theta_2, v_2)\|_{\text{pro}}\sqrt{\frac{2}{n}}\sqrt{u} \\
+\frac{2\beta\sqrt{r}\|(\theta_1, v_1) - (\theta_2, v_2)\|_{\text{pro}}}{n}u\Bigg\} \le 2\exp(-u).
\end{aligned}
$$

The above inequality implies that the empirical process $(\mathbb{P}-\mathbb{P}_n)g_{(\theta,v)}$ has a mixed sub-Gaussian-sub-exponential increments with respect to the metrics $(\frac{2\beta\sqrt{r}}{n}\|\cdot\|_{\text{pro}}, \frac{2\sqrt{2}\beta\sqrt{r}}{\sqrt{n}}\|\cdot\|_{\text{pro}})$ (see Definition 8).

From Lemma 13, there exists an absolute constants C such that $\forall\delta\in(0,1)$, with probability at least $1-\delta$,

$$\sup_{\|\theta-\theta^*\|^2+\|v\|^2\le r}(\mathbb{P}-\mathbb{P}_n)g_{(\theta,v)} \le C\left(\gamma_2\left(\mathcal{B}(\sqrt{r}), \frac{2\sqrt{2}\beta\sqrt{r}}{\sqrt{n}}\|\cdot\|_{\text{pro}}\right)+\gamma_1\left(\mathcal{B}(\sqrt{r}), \frac{2\beta\sqrt{r}}{n}\|\cdot\|_{\text{pro}}\right)\right.$$
$$\left.+\beta r\sqrt{\frac{\log\frac{1}{\delta}}{n}}+\beta r\frac{\log\frac{1}{\delta}}{n}\right).$$

Using Dudley's integral (Lemma 12) to bound the γ_1 functional and the γ_2 functional, we obtain that there exist absolute constant c_1 such that $\forall\delta\in(0,1)$, with probability at least $1-\delta$,

$$\sup_{\|(\theta,\theta')-(\theta^*,\theta^*)\|^2\le r}\left|(\mathbb{P}-\mathbb{P}_n)g_{(\theta,v)}\right| \le c_1\beta r\left(\sqrt{\frac{d+\log\frac{1}{\delta}}{n}}+\frac{d+\log\frac{1}{\delta}}{n}\right). \tag{C.3}$$

We set

$$\psi(r;\delta)=c_1\beta r\left(\sqrt{\frac{d+\log\frac{1}{\delta}}{n}}+\frac{d+\log\frac{1}{\delta}}{n}\right).$$

Denote $R=2(\Delta_M^2+\frac{1}{n^2})$ and $r_0=\frac{2}{n^2}$. Since V is a $d-$dimensional ball centered at the origin with radius $\max\{\Delta_M,\frac{1}{n}\}$, we know that $\|\theta-\theta^*\|^2+\|v\|^2\le 2\Delta_M^2+\frac{1}{n^2}\le R$. We apply Proposition 4 and obtain: for any fixed $\delta\in(0,1)$, with probability at least $1-\delta$, for all $\theta\in\Theta$ and $v\in V$,

$$\begin{aligned}
&(\mathbb{P}-\mathbb{P}_n)\left[(\nabla\ell(\theta;z)-\nabla\ell(\theta^*;z))^Tv\right]=(\mathbb{P}-\mathbb{P}_n)g_{(\theta,v)}\\
&\le\psi\left(\max\left\{\|\theta-\theta^*\|^2+\|v\|^2,\frac{2}{n^2}\right\};\frac{\delta}{2\log_2(2R/\frac{2}{n^2})}\right)\\
&=c_1\beta\max\left\{\|\theta-\theta^*\|^2+\|v\|^2,\frac{2}{n^2}\right\}\left(\sqrt{\frac{d+\log\frac{2\log_2(n^2R)}{\delta}}{n}}+\frac{d+\log\frac{2\log_2(n^2R)}{\delta}}{n}\right).
\end{aligned}$$

This completes the proof of Proposition 5. □

Proof of Proposition 2: in order to uniformly bound $\|(\mathbb{P}-\mathbb{P}_n)(\nabla\ell(\theta;z)-\nabla\ell(\theta^*;z))\|$ for all $\theta\in\Theta$, we take

$$v=\max\left\{\|\theta-\theta^*\|,\frac{1}{n}\right\}\cdot\frac{(\mathbb{P}-\mathbb{P}_n)(\nabla\ell(\theta;z)-\nabla\ell(\theta^*;z))}{\|(\mathbb{P}-\mathbb{P}_n)(\nabla\ell(\theta;z)-\nabla\ell(\theta^*;z))\|}$$

in Proposition 5. Clearly $\|v\| = \max\{\|\theta - \theta^*\|, \frac{1}{n}\}$. From Proposition 2, we can prove that there exists an absolute constant c such that $\forall \delta \in (0,1)$, with probability at least $1-\delta$, for all $\theta \in \Theta$,

$$\begin{aligned}
&\|(\mathbb{P} - \mathbb{P}_n)(\nabla \ell(\theta; z) - \nabla \ell(\theta^*; z))\| \\
&\leq c\beta \max\left\{\|\theta - \theta^*\|, \frac{1}{n}\right\} \left(\sqrt{\frac{d + \log \frac{2\log_2(2n^2\Delta_M^2+2)}{\delta}}{n}} + \frac{d + \log \frac{2\log_2(2n^2\Delta_M^2+2)}{\delta}}{n}\right) \\
&\leq c\beta \max\left\{\|\theta - \theta^*\|, \frac{1}{n}\right\} \left(\sqrt{\frac{d + \log \frac{4\log_2(2n\Delta_M+2)}{\delta}}{n}} + \frac{d + \log \frac{4\log_2(2n\Delta_M+2)}{\delta}}{n}\right).
\end{aligned}$$

This completes the proof of Proposition 2.

□

C.3 Proof of Theorem 3

We first prove a proposition on the uniform localized convergence of gradients under Assumption 1 and Assumption 2.

Proposition 6 (uniform localized convergence of gradients). *Let Assumption 1, Assumption 2 hold along with the optimality condition $\mathbb{P}\nabla \ell(\theta^*; z) = 0$. Given $\delta \in (0,1)$, denote*

$$\begin{aligned}
\text{term I} &:= \sqrt{\frac{2\mathbb{P}[\|\nabla \ell(\theta^*; z)\|^2] \log \frac{4}{\delta}}{n}} + \frac{G_* \log \frac{4}{\delta}}{n}, \\
\text{term II} &:= \sqrt{\frac{d + \log \frac{8\log_2(2n\Delta_M+2)}{\delta}}{n}} + \frac{d + \log \frac{\log_2(2n\Delta_M+2)}{\delta}}{n}.
\end{aligned}$$

Then with probability at least $1-\delta$, we have the following:

$$\|(\mathbb{P}_n - \mathbb{P})\nabla \ell(\theta^*; z)\| \leq \text{term I}, \tag{C.4}$$

and

$$\|(\mathbb{P}_n - \mathbb{P})\nabla \ell(\theta; z)\| \leq \text{term I} + c_0\beta \max\left\{\frac{1}{n}, \|\theta - \theta^*\|\right\} \cdot \text{term II}, \quad \forall \theta \in \Theta, \tag{C.5}$$

where c_0 is an absolute constant.

Proof of Proposition 6: from Proposition 2, there exists an absolute constant c_0 such that $\forall \delta_1 > 0$, with probability at least $1 - \frac{\delta}{2}$, for all $\theta \in \Theta$,

$$\begin{aligned}
&\|(\mathbb{P}_n - \mathbb{P})\nabla \ell(\theta; z)\| \\
&\leq \|(\mathbb{P}_n - \mathbb{P})\nabla \ell(\theta^*; z)\| + c_0\beta \max\left\{\|\theta - \theta^*\|, \frac{1}{n}\right\} \cdot \text{term II}.
\end{aligned} \tag{C.6}$$

From Bernstein's inequality for vectors (Lemma 11), we have with probability at least $1 - \frac{\delta}{2}$,

$$\|\mathbb{P}\nabla \ell(\theta^*; z) - \mathbb{P}_n \nabla \ell(\theta^*; z)\| \leq \sqrt{\frac{2\mathbb{P}[\|\nabla \ell(\theta^*; z)\|^2] \log \frac{4}{\delta}}{n}} + \frac{G_* \log \frac{4}{\delta}}{n} = \text{term I}, \tag{C.7}$$

Combining (C.6) and (C.7) by a union bound, we complete the proof of Proposition 6. □

We first present the following lemma.

Lemma 7 (relationship between curvature conditions). *For a function F, consider the following conditions:*

1. *Strong convexity (SC): for all $\theta_1, \theta_2 \in \Theta$ we have*

$$F(\theta_1) \geq F(\theta_2) + \nabla F(\theta_2)^T(\theta_1 - \theta_2) + \frac{\mu}{2}\|\theta_1 - \theta_2\|^2.$$

2. *Polyak-Lojasiewisz (PL): for all $\theta \in \Theta$ we have*

$$F(\theta) - F(\theta^*) \leq \frac{1}{2\mu}\|\nabla F(\theta)\|^2.$$

3. *Error Bound (EB): for all $\theta \in \Theta$ we have*

$$\|\nabla F(\theta)\| \geq \mu\|\theta - \theta^*\|.$$

4. *Quadratic Growth (QG): for all $\theta \in \Theta$ we have*

$$F(\theta) - F(\theta^*) \geq \frac{\mu}{2}\|\theta - \theta^*\|^2.$$

Then, the following hold:

$$(SC) \Longrightarrow (PL) \Longrightarrow (EB) \Longrightarrow (QG).$$

Proof of Lemma 7 can be adapted from Appendix A in Karimi et al. (2016). Note that some parameters in the original statements in Karimi et al. (2016) have typos though the proof ideas are correct. In Lemma 7 we fix those typos on the parameters. As argued in Karimi et al. (2016), (PL) and the equivalent (QG) (under the smoothness condition and change of parameters) are the most general conditions that allow linear convergence to a global minimizer.

We now prove Theorem 3.

Proof of Theorem 3: we prove the results on the event

$$\mathscr{A} := \{\text{the results (C.4) (C.5) in Proposition 6 hold true}\},$$

whose measure is at least $1 - \delta$. We keep the notations "term I" and "term II" used in Proposition 6, which are defined by

$$\text{term I} := \sqrt{\frac{2\mathbb{P}[\|\nabla \ell(\theta^*; z)\|^2] \log \frac{4}{\delta}}{n}} + \frac{G_* \log \frac{4}{\delta}}{n},$$

$$\text{term II} := \sqrt{\frac{d + \log \frac{8 \log_2(2n\Delta_M + 2)}{\delta}}{n}} + \frac{d + \log \frac{\log_2(2n\Delta_M + 2)}{\delta}}{n}.$$

The PL condition (Assumption 3) implies that $\mathbb{P}\nabla \ell(\theta^*; z) = 0$. From the result (C.4) in Proposition 6,

$$\|\mathbb{P}_n \nabla \ell(\theta^*; z)\| = \|(\mathbb{P}_n - \mathbb{P})\nabla \ell(\theta^*; z)\| \leq \text{term I}.$$

So we know that the equation

$$\|\mathbb{P}_n \nabla \ell(\theta; z)\| \leq \text{term I}. \tag{C.8}$$

must have a solution within Θ.

The result (C.5) implies that for all $\theta \in \Theta$ such that $\|\theta - \theta^*\| \leq \frac{1}{n}$,

$$\|\mathbb{P}\nabla \ell(\theta; z)\| \leq \|\mathbb{P}_n \nabla \ell(\theta; z)\| + \text{term I} + c_0 \beta \|\theta - \theta^*\| \cdot \text{term II}.$$

Since the PL condition implies (see Lemma 7)

$$\|\mathbb{P}\nabla \ell(\theta; z)\| \geq \mu \|\theta - \theta^*\|,$$

for all $\theta \in \Theta$ such that $\|\theta - \theta^*\| \leq \frac{1}{n}$, we have

$$\mu\|\theta - \theta^*\| \leq \|\mathbb{P}\nabla \ell(\theta; z)\| \leq \|\mathbb{P}_n \nabla \ell(\theta; z)\| + \text{term I} + c_0 \beta \|\theta - \theta^*\| \cdot \text{term II},$$

where c is an absolute constant. Therefore, for all $\theta \in \Theta$, we must have

$$\mu\|\theta - \theta^*\| \leq \|\mathbb{P}\nabla \ell(\theta; z)\| \leq \|\mathbb{P}_n \nabla \ell(\theta; z)\| + \text{term I} + c_0 \beta \|\theta - \theta^*\| \cdot \text{term II} + \frac{\mu}{n}. \tag{C.9}$$

Let $\hat{\theta} \in \Theta$ be an arbitrary solution that satisfies (C.8). From (C.9), we obtain the inequalities for $\|\hat{\theta} - \theta^*\|$:

$$\mu\|\hat{\theta} - \theta^*\| \leq \|\mathbb{P}\nabla \ell(\hat{\theta}; z)\| \leq 2 \cdot \text{term I} + c_0 \beta \cdot \text{term II} \cdot \|\theta - \theta^*\| + \frac{\mu}{n}. \tag{C.10}$$

Let $c = \max\{4c_0^2, 1\}$. When

$$n \geq \frac{c\beta^2 (d + \log \frac{4 \log(2n\Delta_M + 1)}{\delta})}{\mu^2},$$

we have $c_0 \beta \cdot \text{term II} \leq \frac{\mu}{2}$ so that from (C.10),

$$\|\hat{\theta} - \theta^*\| \leq \frac{2}{\mu}(2 \cdot \text{term I} + \frac{\mu}{n})$$

and the event $\mathscr{A}$. Plugging in "$c_0 \beta \cdot \text{term II} \leq \frac{\mu}{2}$" and "$\|\hat{\theta} - \theta^*\| \leq \frac{2}{\mu}(2 \cdot \text{term I} + \frac{\mu}{n})$" into the second inequality within (C.10), we further have

$$\begin{aligned} \|\mathbb{P}\nabla \ell(\hat{\theta}; z)\| &\leq 2 \cdot \text{term I} + \frac{\mu}{n} + \frac{\mu}{2}\|\hat{\theta} - \theta^*\| \\ &\leq 4 \cdot \text{term I} + \frac{2\mu}{n}. \end{aligned}$$

Lastly, since the PL condition implies (see Lemma 7)

$$\mathbb{P}\ell(\hat{\theta}; z) - \mathbb{P}\ell(h^*; z) \leq \frac{\|\mathbb{P}\ell(\hat{\theta}; z)\|^2}{2\mu},$$

by plugging in "$\|\mathbb{P}\nabla \ell(\hat{\theta}; z)\| \leq 4 \cdot \text{term I} + \frac{2\mu}{n}$" we have

$$\begin{aligned} \mathbb{P}\ell(\hat{\theta}; z) - \mathbb{P}\ell(h^*; z) &\leq \frac{\|\mathbb{P}\ell(\hat{\theta}; z)\|^2}{2\mu} \\ &\leq \frac{16}{\mu}(\text{term I})^2 + \frac{4\mu}{n^2} \\ &\leq \frac{64\mathbb{P}[\|\nabla \ell(\theta^*; z)\|^2] \log \frac{4}{\delta}}{\mu n} + \frac{32 G_*^2 \log^2 \frac{4}{\delta} + 4\mu^2}{\mu n^2}. \end{aligned}$$

This completes the proof of Theorem 3. □

C.4 Proof of Theorem 4

We first prove a simple proposition, which studies how the accumulation of sample approximation errors at every step influences the convergence of the algorithm.

Proposition 7 (localized statistical error of a linearly convergent iterative algorithm). *Consider a function F (for which we call the "Lyapunov function") and a parameter $\gamma \in (0,1)$. Assume an algorithm satisfies for all $t = 0, 1, \ldots$*

$$\begin{aligned} F(\theta^{t+1}) &\leq (1-\gamma)F(\theta^t) + \varepsilon^t(n), \\ \varepsilon^t(n) &\leq \alpha(n)F(\theta^t) + \varepsilon^*(n), \\ &\text{and} \quad \theta^t \in \Theta. \end{aligned}$$

When the sample size n is large enough such that $\alpha(n) \leq \frac{\gamma}{2}$, we have

$$F(\theta^t) \leq \left(1 - \frac{\gamma}{2}\right)^t F(\theta^0) + \frac{2}{\gamma}\varepsilon^*(n), \quad t = 0, 1, \cdots.$$

Proof of Proposition 7: we have

$$\begin{aligned} F(\theta^{t+1}) &\leq (1 - \gamma + \alpha(n))F(\theta^t) + \varepsilon^*(n) \\ &\leq \left(1 - \frac{\gamma}{2}\right) F(\theta^*) + \varepsilon^*(n). \end{aligned}$$

Then by induction we have

$$F(\theta^t) \leq \left(1 - \frac{\gamma}{2}\right)^t F(\theta^0) + \frac{2}{\gamma}\varepsilon^*(n), \quad t = 0, 1, \cdots.$$

This completes the proof of Proposition 7. □

We now prove Theorem 4.

Proof of Theorem 4: Assumption 1 implies that the population risk is $\beta-$smooth. Consider the gradient descent algorithm (5.8) with fixed step size $\frac{1}{\beta}$. We have for all $t = 0, 1, \cdots,$

$$\theta^{t+1} = \theta^t - \frac{1}{\beta}\mathbb{P}_n \nabla \ell(\theta^t; z).$$

So we have

$$\begin{aligned} \mathbb{P}\ell(\theta^{t+1}; z) - \mathbb{P}\ell(\theta^t; z) &\leq (\mathbb{P}\nabla\ell(\theta^t; z))^T(\theta^{t+1} - \theta^t) + \frac{\beta}{2}\|\theta^{t+1} - \theta^t\|^2 \\ &= -\frac{1}{\beta}(\mathbb{P}\nabla\ell(\theta^t; z))^T(\mathbb{P}_n\nabla\ell(\theta^t; z)) + \frac{1}{2\beta}\|\mathbb{P}_n\nabla\ell(\theta^t; z)\|^2 \\ = -\frac{1}{\beta}\|\mathbb{P}\nabla\ell(\theta^t; z)\|^2 &- \frac{1}{\beta}(\mathbb{P}\nabla\ell(\theta^t; z))^T(\mathbb{P}_n\nabla\ell(\theta^t; z) - \mathbb{P}\nabla\ell(\theta^t; z)) \\ &+ \frac{1}{2\beta}\|\mathbb{P}\nabla\ell(\theta^t; z) + (\mathbb{P}_n\nabla\ell(\theta^t; z) - \mathbb{P}\nabla\ell(\theta^t; z))\|^2 \\ &= -\frac{1}{2\beta}\|\mathbb{P}\nabla\ell(\theta^t; z)\|^2 + \frac{1}{2\beta}\|\mathbb{P}_n\nabla\ell(\theta^t; z) - \mathbb{P}\nabla\ell(\theta^t; z)\|^2 \\ \leq -\frac{\mu}{\beta}(\mathbb{P}\ell(\theta^t; z) - \mathbb{P}\ell(\theta^*; z)) &+ \frac{1}{2\beta}\|\mathbb{P}_n\nabla\ell(\theta^t; z) - \mathbb{P}\nabla\ell(\theta^t; z)\|^2. \end{aligned}$$

Rearranging the above inequality, and subtracting $\mathbb{P}\ell(\theta^*;z)$ from both sides, we obtain

$$\mathbb{P}\ell(\theta^{t+1};z)-\mathbb{P}\ell(\theta^*;z)\le\left(1-\frac{\mu}{\beta}\right)\left(\mathbb{P}\ell(\theta^t;z)-\mathbb{P}\ell(\theta^*;z)\right)+\frac{1}{2\beta}\|\mathbb{P}_n\nabla\ell(\theta^t;z)-\mathbb{P}\nabla\ell(\theta^t;z)\|^2. \tag{C.11}$$

Applying Proposition 6, we continue the proof on the event

$$\mathscr{A}:=\{\text{the results (C.4) (C.5) in Proposition 6 hold true}\},$$

whose measure is at least $1-\delta$. We keep the notations "term I" and "term II" used in Proposition 6, which are defined by

$$\text{term I}:=\sqrt{\frac{2\mathbb{P}[\|\nabla\ell(\theta^*;z)\|^2]\log\frac{4}{\delta}}{n}}+\frac{G_*\log\frac{4}{\delta}}{n},$$
$$\text{term II}:=\sqrt{\frac{d+\log\frac{8\log_2(2n\Delta_M+2)}{\delta}}{n}}+\frac{d+\log\frac{\log_2(2n\Delta_M+2)}{\delta}}{n}.$$

The result (C.5) in Proposition 6 implies that $\forall\theta\in\Theta$,

$$\begin{aligned}\|\mathbb{P}_n\nabla\ell(\theta;z)-\mathbb{P}\nabla\ell(\theta;z)\|&\le\text{term I}+c_0\beta\max\left\{\|\theta-\theta^*\|,\frac{1}{n}\right\}\cdot\text{term II}\\&\le\left(\text{term I}+\frac{c_0\beta}{n}\cdot\text{term II}\right)+c_0\beta\cdot\text{term II}\cdot\|\theta-\theta^*\|,\end{aligned}$$

where c_0 is an absolute constant. Since the PL condition implies (see Lemma 7) that

$$\mathbb{P}\ell(\theta;z)-\mathbb{P}\ell(\theta^*;z)\ge\frac{\mu}{2}\|\theta-\theta^*\|^2,\quad\forall\theta\in\Theta,$$

we have

$$\begin{aligned}\|\mathbb{P}_n\nabla\ell(\theta;z)-\mathbb{P}\nabla\ell(\theta;z)\|^2\le{}&2\left(\text{term I}+\frac{c_0\beta}{n}\cdot\text{term II}\right)^2\\&+\frac{4c_0^2\beta^2}{\mu}\left(\mathbb{P}\ell(\theta;z)-\mathbb{P}\ell(\theta^*;z)\right)(\text{term II})^2.\end{aligned}\tag{C.12}$$

Combining (C.11) and (C.12), we have that for all $t=0,1,\ldots,$

$$\begin{aligned}\mathscr{E}(\theta^{t+1})&\le\left(1-\frac{\mu}{\beta}\right)\mathscr{E}(\theta^t)+\varepsilon^t(n),\\\varepsilon^t(n)&\le\alpha(n)\mathscr{E}(\theta^t)+\varepsilon^*(n),\end{aligned}$$

where

$$\begin{aligned}\varepsilon^t(n)&=\frac{1}{2\beta}\|\mathbb{P}_n\nabla\ell(\theta^t;z)-\mathbb{P}\nabla\ell(\theta^t;z)\|^2,\\\alpha(n)&=\frac{2c_0^2\beta}{\mu}(\text{term II})^2,\\\varepsilon^*(n)&=\frac{1}{\beta}\left(\text{term I}+\frac{c_0\beta}{n}\cdot\text{term II}\right)^2.\end{aligned}$$

Consider the following two conditions on the sample size (note that they will be satisfied as long as n is large enough):

$$\alpha(n) \le \frac{\mu}{2\beta}, \tag{C.13}$$

$$\varepsilon^*(n) \le \frac{\mu^2}{4\beta}\Delta_m^2. \tag{C.14}$$

Now consider the condition

$$n \ge \frac{c\beta^2}{\mu^2}\left(d + \log\frac{8\log_2(2n\Delta_M + 2)}{\delta}\right),$$

where $c = \max\{16c_0^2, 1\}$ is an absolute constant. Then (C.13) holds. Since we also require the sample size n to be large enough such that the "statistical error" term in (5.9) is smaller than $\frac{\mu}{2}\Delta_m^2$, the condition (C.14) is also true because

$$\frac{\mu}{2}\Delta_m^2 \ge \frac{16\mathbb{P}[\|\nabla\ell(\theta^*;z)\|^2]\log\frac{4}{\delta}}{\mu n} + \frac{8G_*^2\log^2\frac{4}{\delta} + \mu^2}{\mu n^2} \ge \frac{2}{\mu}\left(\text{term II}^2 + \frac{\mu}{2n}\right)^2 \ge \frac{2\beta}{\mu}\cdot\varepsilon^*(n).$$

Therefore, both condition (C.13) and condition (C.14) hold true under Theorem 4's requirement on the sample size.

Since both (C.13) and (C.14) are true, we can use induction to prove that with probability at least $1-\delta$, for all $t = 0, 1, \ldots,$

$$\mathscr{E}(\theta^t) \le \frac{\mu}{2}\Delta_m^2.$$

Therefore, for all $t = 0, 1, \ldots,$

$$\theta^t \in \mathcal{B}^d(\theta^*, \Delta_m) \subseteq \Theta.$$

We choose the "Lyapunov function" in Proposition 7 to be the excess risk function $\mathscr{E}(\theta)$. Applying Proposition 7, we obtain that: when the sample size n is large enough such that the conditions (C.13) and (C.14) hold ture, we have

$$\begin{aligned}
\mathbb{P}\ell(\theta^t;z) - \mathbb{P}\ell(\theta^*;z) &\le \left(1 - \frac{\mu}{2\beta}\right)^t \mathscr{E}(\theta^0) + \frac{2\beta}{\mu}\cdot\varepsilon^*(n) \\
&\le \left(1 - \frac{\mu}{2\beta}\right)^t \mathscr{E}(\theta^0) + \frac{2}{\mu}\left(\text{term I} + \frac{\mu}{2n}\right)^2, \\
&\le \left(1 - \frac{\mu}{2\beta}\right)^t \mathscr{E}(\theta^0) + \frac{16\mathbb{P}[\|\nabla\ell(\theta^*;z)\|^2]\log\frac{8}{\delta}}{\mu n} + \frac{8G_*^2\log^2\frac{4}{\delta} + \mu^2}{\mu n^2}.
\end{aligned}$$

This completes the proof of Theorem 4. □

C.5 Proof of Corollary 5

We first verify Assumption 1. We have

$$\nabla^2\ell(\theta;z) = 2\left(\eta'(\theta^T x)^2 + (\eta(\theta^T x) - y)\eta''(\theta^T x)]\right)xx^T.$$

Since x is $\tau-$sub-Gaussian, xx^T is a τ^2-sub-exponential. From the fact

$$\left|2(\eta'(\theta^T x)^2 + (\eta(\theta^T x) - y)\eta''(\theta^T x))\right| \le C_\eta(B + C_\eta),$$

Assumption 1 holds with $\beta = 2C_\eta(C_\eta + \sqrt{B})\tau^2$.

We then verify Assumption 2. We know

$$\nabla \ell(\theta^*; z) = 2(\eta(x^T\theta^*) - y)\eta'(x^T\theta^*)x.$$

So we have for all z,

$$\|\nabla \ell(\theta^*; z)\| \le \sqrt{d}\|\nabla \ell(\theta^*; z)\|_\infty \le 2C_\eta\sqrt{Bd}.$$

So Assumption 2 holds with $G_* = 2C_\eta\sqrt{Bd}$.

Lastly, by inequality (16) of Lemma 5 in Foster et al. (2018), Assumption 3 holds with $\mu = \frac{2c_\eta^3\tau^2\gamma}{C_\eta}$. This completes the proof. □

C.6 Proof of Theorem 6

Before proving Theorem 6, we refer to Theorem 1 in Balakrishnan et al. (2017) for the following result on the population-based first-order EM update.

Lemma 8 (linear convergence of population-based first-order EM). *Under Assumption 5, Assumption 6 and the condition that $\mathbb{P}\ell(\theta; z)$ is $\beta-$smooth, the following update,*

$$\theta^+ = \theta - \frac{2}{\beta + \mu_1}\mathbb{P}\nabla \ell_\theta(\theta; z)$$

satisfies that

$$\|\theta^+ - \theta^*\| \le \left(1 - \frac{2\mu_1 - \mu_2}{\beta + \mu_1}\right)\|\theta - \theta^*\|.$$

We now prove Theorem 6.

Proof of Theorem 6: Assumption 1 implies that $\mathbb{P}\ell(\theta; z)$ is $\beta-$smooth, so Lemma 8 holds under the assumptions of Theorem 6. Now we turn to analyze the sample-based first-order EM. Consider the update of sample-based first-order EM,

$$\theta^{t+1} = \theta^t - \frac{2}{\beta + \mu_1}\mathbb{P}_n\nabla \ell_{\theta^t}(\theta^t; z), \quad t = 0, 1, \ldots$$

Fix $t \ge 0$. We have

$$\begin{aligned}\|\theta^{t+1} - \theta^*\| &\le \|\theta^t - \frac{2}{\beta + \mu_1}\mathbb{P}\nabla \ell_{\theta^t}(\theta^t; z)\| + \frac{2}{\beta + \mu_1}\|(\mathbb{P} - \mathbb{P}_n)\nabla \ell_{\theta^t}(\theta^t; z)\| \\ &\le \left(1 - \frac{2\mu_1 - \mu_2}{\beta + \mu_1}\right)\|\theta^t - \theta^*\| + \frac{2}{\beta + \mu_1}\|(\mathbb{P} - \mathbb{P}_n)\nabla \ell(\theta^t; z)\|. \qquad \text{(C.15)}\end{aligned}$$

Applying Proposition 6, we continue the proof on the event

$$\mathcal{A} := \{\text{the results (C.4) (C.5) in Proposition 6 hold true}\},$$

whose measure is at least $1-\delta$. We keep the notations "term I" and "term II" used in Proposition 6, which are defined by

$$\text{term I} := \sqrt{\frac{2\mathbb{P}[\|\nabla \ell(\theta^*; z)\|^2] \log \frac{4}{\delta}}{n}} + \frac{G_* \log \frac{4}{\delta}}{n},$$
$$\text{term II} := \sqrt{\frac{d + \log \frac{8 \log_2(2n\Delta_M + 2)}{\delta}}{n}} + \frac{d + \log \frac{\log_2(2n\Delta_M+2)}{\delta}}{n}.$$

Note that we have the optimality condition $\nabla \ell(\theta^*; z) = 0$, because the true parameter θ^* is assumed to minimizes the population risk over $\mathbb{R}^d$ in the problem setting. The result (C.5) in Proposition 6 implies that $\forall \theta \in \Theta$,

$$\|\mathbb{P}_n \nabla \ell(\theta; z) - \mathbb{P} \nabla \ell(\theta; z)\| \le \text{term I} + c_0 \beta \max\left\{\|\theta - \theta^*\|, \frac{1}{n}\right\} \cdot \text{term II} \tag{C.16}$$
$$\le \left(\text{term I} + \frac{c_0 \beta}{n} \cdot \text{term II}\right) + c_0 \beta \cdot \text{term II} \cdot \|\theta - \theta^*\|, \tag{C.17}$$

where c_0 is an absolute constant. Therefore, we have that for all $t = 0, 1, \ldots,$

$$\mathscr{E}(\theta^{t+1}) \le \left(1 - \frac{2\mu_1 - \mu_2}{\beta + \mu_1}\right) \mathscr{E}(\theta^t) + \varepsilon^t(n),$$
$$\varepsilon^t(n) \le \alpha(n) \mathscr{E}(\theta^t) + \varepsilon^*(n),$$

where

$$\varepsilon^t(n) = \frac{2}{\beta + \mu_1} \|(\mathbb{P} - \mathbb{P}_n) \nabla \ell(\theta^t; z)\|,$$
$$\alpha(n) = \frac{2 c_0 \beta}{\beta + \mu_1} \cdot \text{term II},$$
$$\varepsilon^*(n) = \frac{2}{\beta + \mu_1} \left(\text{term I} + \frac{c_0 \beta}{n} \cdot \text{term II}\right).$$

Consider the following two conditions on the sample size (note that they will be satisfied as long as n is large enough):

$$\alpha(n) \le \frac{2\mu_1 - \mu_2}{2(\beta + \mu_1)}, \tag{C.18}$$
$$\varepsilon^*(n) \le \frac{2\mu_1 - \mu_2}{2(\beta + \mu_1)} \Delta_m. \tag{C.19}$$

When the sample size n is large enough so that both (C.18) and (C.19) are true, we can use induction to prove that with probability at least $1 - \delta$, for all $t = 0, 1, \ldots,$

$$\|\theta^t - \theta^*\| \le \Delta_m^2.$$

Therefore, for all $t = 0, 1, \ldots,$

$$\theta^t \in \mathcal{B}^d(\theta^*, \Delta_m) \subseteq \Theta.$$

We choose the "Lyapunov function" in Proposition 7 to be $\|\theta - \theta^*\|$. Applying Proposition 7, we obtain: when the sample size n is large enough such that the conditions (C.18) and (C.19) hold ture, we have

$$\begin{aligned}\|\theta^t - \theta^*\| &\leq \left(1 - \frac{2\mu_1 - \mu_2}{2(\beta + \mu_1)}\right)^t \|\theta^0 - \theta^*\| + \frac{2(\beta + \mu_1)}{2\mu_1 - \mu_2} \cdot \varepsilon^*(n) \\ &\leq \left(1 - \frac{2\mu_1 - \mu_2}{2(\beta + \mu_1)}\right)^t \|\theta^0 - \theta^*\| + \frac{4}{2\mu_1 - \mu_2} \cdot \text{term I} + \frac{2}{n}. \end{aligned} \tag{C.20}$$

When the sample size is large enough such that

$$n \geq \frac{c\beta^2}{(2\mu_1 - \mu_2)^2}\left(d + \log\frac{8\log_2(2n\Delta_M + 2)}{\delta}\right) \quad \text{and} \quad \text{term I} + \frac{2\mu_1 - \mu_2}{2n} \leq \frac{(2\mu_1 - \mu_2)\Delta_m}{4}, \tag{C.21}$$

where $c = \max\{64c_0^2, 1\}$ is an absolute constant, we have

$$\text{term I} \leq \frac{2\mu_1 - \mu_2}{4}\left(\Delta_m - \frac{2}{n}\right) \quad \text{and} \quad \text{term II} \leq \frac{2\mu_1 - \mu_2}{4c_0\beta},$$

which further guarantee that both the condition (C.18) and the condition (C.19) are true. We conclude that when the sample size condition (C.21) is true, we have the bound (C.20).

Now we use the fact $\mu_1 \leq 2\mu_1 - \mu_2 \leq 2\mu_1$ to simplify the sample size condition (C.21) and the bound (C.20). It is straightforward to verify that the sample size condition (C.21) will be satisfied when

$$n \geq \max\left\{\frac{c\beta^2}{\mu_1^2}\left(d + \log\frac{8\log_2(2n\Delta_M + 2)}{\delta}\right), \quad \frac{128\mathbb{P}[\|\nabla\ell(\theta^*; z)\|^2]\log\frac{4}{\delta}}{\mu_1\Delta_M}, \quad \frac{8G_*\log\frac{4}{\delta} + 8\mu_1}{\mu_1\Delta_M}\right\}; \tag{C.22}$$

and the bound (C.20) implies

$$\|\theta^t - \theta^*\| \leq \frac{4}{\mu_1}\left(\sqrt{\frac{2\mathbb{P}[\|\nabla\ell(\theta^*; z)\|^2]\log\frac{4}{\delta}}{n}} + \frac{G_*\log\frac{4}{\delta} + \mu_1}{n}\right) + \left(1 - \frac{2\mu_1 - \mu_2}{2(\beta + \mu_1)}\right)^t \|\theta^0 - \theta^*\|. \tag{C.23}$$

Since we always have

$$\mathscr{E}(\theta^t) \leq \frac{\beta}{2}\|\theta^t - \theta^*\|^2,$$

the bound (C.23) will imply

$$\mathscr{E}(\theta^t) \leq \frac{16\beta}{\mu_1^2}\left(\sqrt{\frac{2\mathbb{P}[\|\nabla\ell(\theta^*; z)\|^2]\log\frac{4}{\delta}}{n}} + \frac{G_*\log\frac{4}{\delta} + \mu_1}{n}\right)^2 + \left(1 - \frac{2\mu_1 - \mu_2}{2(\beta + \mu_1)}\right)^{2t} \beta\|\theta^0 - \theta^*\|^2. \tag{C.24}$$

Clearly, the sample size condition (C.22) and the bounds (C.23) (C.24) are identical to those presented in the statement of the theorem. This completes the proof. □

C.7 Proof of Corollary 7

For both examples, verification of Assumptions 1, 2 and 5 is trivial. The parameters can be specified as $\beta = 1$, $G_* = \sigma\sqrt{d}$, and $\mu_1 = 1$.

As for verification of Assumption 6, we refer to the following results that are direct consequence of Balakrishnan et al. (2017).

Lemma 9 (verification of Assumption 6). *(a)* Lemma 2 in Balakrishnan et al. (2017): *Consider Example 4 under the SNR condition (7.6), where η is a sufficiently large constant such that $\eta > \frac{4\sqrt{3}}{3}$ and $c_1(1+\frac{1}{\eta^2}+\eta^2)e^{-c_2\eta^2} < 1$. Then Assumption 6 holds with $\mu_2 = c_1(1+\frac{1}{\eta^2}+\eta^2)e^{-c_2\eta^2}$. Here c_1 and c_2 denote the same absolute constants as in the proof of Lemma 2 in Balakrishnan et al. (2017). Clearly, we can verify Assumption 6 for all η larger than a certain absolute constant.*

(b) Lemma 3 in Balakrishnan et al. (2017): *Consider Example 5 under the SNR condition (7.6), where η is a sufficiently large constant such that*

$$c\eta^{1-\frac{c_\tau^2}{2}} + c_\tau^2\frac{\log\eta}{\eta} + \frac{2}{\eta} \le \frac{1}{8},$$
$$\sqrt{\frac{\|\theta^*\|}{8\eta} + (4+\frac{2}{31})\frac{C_\tau^2\log\eta}{\eta} + 3\eta^{1-\frac{C_\tau^2}{2}}} \le \frac{1}{8}$$

hold true for some sufficiently large constants c_τ, C_τ and an absolute constant c. Then Assumption 6 holds with $\mu_2 = \frac{1}{4}$. Here c, c_τ, C_τ denote the same quantity as in the proof of Lemma 3 in Balakrishnan et al. (2017). Clearly, we can verify Assumption 6 for all η larger than a certain absolute constant.

To prove the generalization error bound in this corollary, we need to upper bound the problem-dependent parameter $\mathbb{P}[\|\nabla\ell(\theta^*;z)\|^2]$ for the two examples.

Bounding $\mathbb{P}[\|\nabla\ell(\theta^*;z)\|^2]$ for Example 4: we define the function $g:\mathbb{R}^d\to(0,2)$ as

$$g(u) = \frac{2e^{-\frac{\|2\theta^*-u\|^2}{2\sigma^2}}}{e^{-\frac{\|u\|^2}{2\sigma^2}}+e^{-\frac{\|2\theta^*-u\|^2}{2\sigma^2}}} = \frac{2}{e^{\frac{2\|\theta^*\|^2-2u^T\theta^*}{\sigma^2}}+1}.$$

In Example 4, when conditioned on $w=1$ (i.e., when z is drawn from $N(\theta^*,\sigma^2 I_{d\times d})$), the random gradient $\nabla\ell(\theta^*;z)$ at θ^* can be shown to be equal to

$$(\nabla\ell(\theta^*;z)|w=1) = u\underbrace{\left(\frac{e^{-\frac{\|u\|^2}{2\sigma^2}}-e^{-\frac{\|2\theta^*-u\|^2}{2\sigma^2}}}{e^{-\frac{\|u\|^2}{2\sigma^2}}+e^{-\frac{\|2\theta^*-u\|^2}{2\sigma^2}}}\right)}_{1-g(u)} + \theta^*\underbrace{\left(\frac{2e^{-\frac{\|2\theta^*-u\|^2}{2\sigma^2}}}{e^{-\frac{\|u\|^2}{2\sigma^2}}+e^{-\frac{\|2\theta^*-u\|^2}{2\sigma^2}}}\right)}_{g(u)},$$

where $u=\theta^*-z$ is a random vector drawn from $N(0,\sigma^2 I_{d\times d})$. And when conditioned on $w=-1$ (i.e., when z is drawn from $N(0,\sigma^2 I_{d\times d})$), $\nabla\ell(\theta^*;z)$ can be shown to be equal to

$$(\nabla\ell(\theta^*;z)|w=-1) = v\underbrace{\left(\frac{e^{-\frac{\|v\|^2}{2\sigma^2}}-e^{-\frac{\|2\theta^*-v\|^2}{2\sigma^2}}}{e^{-\frac{\|v\|^2}{2\sigma^2}}+e^{-\frac{\|2\theta^*-v\|^2}{2\sigma^2}}}\right)}_{1-g(v)} + \theta^*\underbrace{\left(\frac{2e^{-\frac{\|2\theta^*-v\|^2}{2\sigma^2}}}{e^{-\frac{\|v\|^2}{2\sigma^2}}+e^{-\frac{\|2\theta^*-v\|^2}{2\sigma^2}}}\right)}_{g(v)},$$

where $v = \theta^* + z$ is a random vector drawn from $N(0, \sigma^2 I_{d\times d})$.

Therefore, we have

$$\begin{aligned}
&\mathbb{P}[\|\nabla \ell(\theta^*; z)\|^2] \\
&= \frac{1}{2}\mathbb{E}\left[\|\nabla \ell(\theta^*; z)\|^2 | w = 1\right] + \frac{1}{2}\mathbb{E}\left[\|\nabla \ell(\theta^*; z)\|^2 | w = -1\right] \\
&= \frac{1}{2}\mathbb{E}_u[\|u \cdot (1 - g(u)) + \theta^* \cdot g(u)\|^2] + \frac{1}{2}\mathbb{E}_v[\|v \cdot (1 - g(v)) + \theta^* \cdot g(v)\|^2] \\
&= \mathbb{E}_u[\|u \cdot (1 - g(u)) + \theta^* \cdot g(u)\|^2], \qquad \text{(C.25)}
\end{aligned}$$

where the notation $\mathbb{E}_u$ means taking expectation with respect to $u \sim N(0, \sigma^2 I_{d\times d})$, and the notation $\mathbb{E}_v$ means taking expectation with respect to $v \sim N(0, \sigma^2 I_{d\times d})$.

Since $0 < g(u) < 2$, we have $|1 - g(u)| \le 1$. Thus

$$\begin{aligned}
&\|u \cdot (1 - g(u)) + \theta^* \cdot g(u)\|^2 \\
&\le 2\|u\|^2 \cdot |1 - g(u)|^2 + 2\|\theta^*\|^2 \cdot |g(u)|^2 \\
&= 2\|u\|^2 + 2\|\theta^*\|^2 \cdot g(u)^2. \qquad \text{(C.26)}
\end{aligned}$$

From (C.25) and (C.26), we have

$$\begin{aligned}
\mathbb{P}[\|\nabla \ell(\theta^*; z)\|^2] &\le 2\mathbb{E}_u[\|u\|^2] + 2\|\theta^*\|^2\mathbb{E}_u[g(u)^2] \\
&= 2\sigma^2 d + 2\|\theta^*\|^2\mathbb{E}_u[g(u)^2]. \qquad \text{(C.27)}
\end{aligned}$$

Now we know that $u^T\theta^*$ is a $\|\theta^*\|\sigma-$sub-Gaussian vector with mean 0. From Markov's inequality,

$$\begin{aligned}
\text{Prob}\left(|u^T\theta^*| > \frac{1}{2}\|\theta^*\|^2\right) &\le 2\exp(-\frac{\frac{1}{4}\|\theta^*\|^4}{\|\theta^*\|^2\sigma^2}) \\
&= \frac{2}{\exp(\frac{\|\theta^*\|^2}{4\sigma^2})} \le \frac{8\sigma^2}{\|\theta^*\|^2}. \qquad \text{(C.28)}
\end{aligned}$$

When $|u^T\theta^*| \le \frac{1}{2}\|\theta^*\|^2$, we have

$$g(u) = \frac{2}{e^{\frac{\|\theta^*\|^2 - u^T\theta^*}{\sigma^2}} + 1} \le \frac{2}{e^{\frac{\|\theta^*\|^2}{2\sigma^2}}} \le \frac{4\sigma^2}{\|\theta^*\|^2}.$$

Since $0 < g(u) < 2$, when $|u^T\theta^*| \le \frac{1}{2}\|\theta^*\|^2$, we have

$$g(u)^2 \le \frac{8\sigma^2}{\|\theta^*\|^2}. \qquad \text{(C.29)}$$

As a result,

$$\begin{aligned}
&\mathbb{E}_u[g(u)^2] \\
&\le \text{Prob}\left(|u^T\theta^*| > \frac{1}{2}\|\theta^*\|^2\right)\mathbb{E}\left[g(u)^2 \middle| |u^T\theta^*| > \frac{1}{2}\|\theta^*\|^2\right] \\
&\quad + \text{Prob}\left(|u^T\theta^*| \le \frac{1}{2}\|\theta^*\|^2\right)\mathbb{E}\left[g(u)^2 \middle| |u^T\theta^*| \le \frac{1}{2}\|\theta^*\|^2\right] \\
&\le 4 \cdot \text{Prob}\left(|u^T\theta^*| > \frac{1}{2}\|\theta^*\|^2\right) + \frac{8\sigma^2}{\|\theta^*\|^2} \\
&\le \frac{40\sigma^2}{\|\theta^*\|^2},
\end{aligned}$$

where the second inequality is due to the fact $0 < g(u) < 2$ and (C.29), and the last inequality is due to (C.28). Combining the above result with (C.27), we have

$$\mathbb{P}[\|\nabla \ell(\theta^*; z)\|^2] \leq (2d + 40)\sigma^2. \tag{C.30}$$

Bounding $\mathbb{P}[\|\nabla \ell(\theta^*; z)\|^2]$ for Example 5: we define the function $g : \mathbb{R} \times \mathbb{R}^d \to (0, 2)$ as

$$g(u, x) = \frac{2e^{-\frac{(2x^T\theta^* - u)^2}{2\sigma^2}}}{e^{-\frac{u^2}{2\sigma^2}} + e^{-\frac{(2x^T\theta^* - u)^2}{2\sigma^2}}} = \frac{2}{e^{\frac{2(x^T\theta^*)^2 - 2u(x^T\theta^*)}{\sigma^2}} + 1}.$$

In Example 5, we have

$$(\nabla \ell(\theta^*; z)|w = 1, x) = \left[u \underbrace{\left(\frac{e^{-\frac{u^2}{2\sigma^2}} - e^{-\frac{(2x^T\theta^* - u)^2}{2\sigma^2}}}{e^{-\frac{u^2}{2\sigma^2}} + e^{-\frac{(2x^T\theta^* - u)^2}{2\sigma^2}}} \right)}_{1 - g(u,x)} + (x^T\theta^*) \underbrace{\left(\frac{2e^{-\frac{(2x^T\theta^* - u)^2}{2\sigma^2}}}{e^{-\frac{u^2}{2\sigma^2}} + e^{-\frac{(2x^T\theta^* - u)^2}{2\sigma^2}}} \right)}_{g(u,x)} \right] x,$$

where $u = x^T\theta^* - y$ is a random vector drawn from $N(0, \sigma^2)$. And we have

$$(\nabla \ell(\theta^*; z)|w = -1, x) = \left[v \underbrace{\left(\frac{e^{-\frac{v^2}{2\sigma^2}} - e^{-\frac{(2x^T\theta^* - v)^2}{2\sigma^2}}}{e^{-\frac{v^2}{2\sigma^2}} + e^{-\frac{(2x^T\theta^* - v)^2}{2\sigma^2}}} \right)}_{1 - g(v,x)} + (x^T\theta^*) \underbrace{\left(\frac{2e^{-\frac{(2x^T\theta^* - v)^2\|^2}{2\sigma^2}}}{e^{-\frac{v^2}{2\sigma^2}} + e^{-\frac{(2x^T\theta^* - v)^2}{2\sigma^2}}} \right)}_{g(v,x)} \right] x,$$

where $v = x^T\theta^* + y$ is a random vector drawn from $N(0, \sigma^2)$.

Therefore, we have

$$\begin{aligned}
&\mathbb{P}[\|\nabla \ell(\theta^*; z)\|^2] \\
&= \frac{1}{2}\mathbb{E}\left[\|\nabla \ell(\theta^*; z)\|^2 | w = 1\right] + \frac{1}{2}\mathbb{E}\left[\|\nabla \ell(\theta^*; z)\|^2 | w = -1\right] \\
&= \frac{1}{2}\mathbb{E}_x\left[\|x\|^2 \mathbb{E}_u[(u \cdot (1 - g(u, x)) + \theta^* \cdot g(u, x))^2 | x]\right] + \frac{1}{2}\mathbb{E}_x\left[\|x\|^2 \mathbb{E}_v[(v \cdot (1 - g(v, x)) + \theta^* \cdot g(v, x))^2]\right] \\
&= \mathbb{E}_x\left[\|x\|^2 \mathbb{E}_u[(u \cdot (1 - g(u, x)) + \theta^* \cdot g(u, x))^2 | x]\right],
\end{aligned} \tag{C.31}$$

where the notation $\mathbb{E}_u$ means taking expectation with respect to $u \sim N(0, \sigma^2)$, the notation $\mathbb{E}_v$ means taking expectation with respect to $v \sim N(0, \sigma^2)$, and the notation $\mathbb{E}_x$ means taking expectation with respect to $x \sim N(0, I_{d\times d})$.

Similar to the last part (i.e., the proof of (C.30)), we can prove

$$\mathbb{E}_u[(u \cdot (1 - g(u, x)) + \theta^* \cdot g(u, x))^2 | x] \leq 42\sigma^2, \quad \forall x.$$

Combine this result with (C.31), we obtain

$$\mathbb{P}[\|\nabla \ell(\theta^*; z)\|^2] \leq 42\sigma^2 \mathbb{E}_x[\|x\|^2] = 42\sigma^2 d.$$

This gives an upper bound on $\mathbb{P}[\|\nabla \ell(\theta^*; z)\|^2]$ in Example 5.

Given that we have upper bounded $\mathbb{P}[\|\nabla \ell(\theta^*; z)\|^2]$ by $42\sigma^2 d$ in both Example 4 and Example 5, it is straightforward to prove the generalization error bound in Corollary 7. □

C.8 Auxiliary definitions and lemmata

Definition 6 (**Orlicz norms, sub-Gaussian, sub-exponential**). *For every $\alpha \in (0, +\infty)$ we define the Orlicz-α norm of a random u:*

$$\|u\|_{\text{Orlicz}_\alpha} = \inf\{K > 0 : \mathbb{E}\exp\left((|u|/K)^\alpha\right) \le 2\}.$$

A random variable/vector $X \in \mathbb{R}^d$ is $K-$sub-Gaussian if $\forall \lambda \in \mathbb{R}^d$, we have

$$\|\lambda^T X\|_{\text{Orlicz}_2} \le K\|\lambda\|_2.$$

A random variable/vector $X \in \mathbb{R}^d$ is $K-$sub-exponential if $\forall \lambda \in \mathbb{R}^d$, we have

$$\|\lambda^T X\|_{\text{Orlicz}_1} \le K\|\lambda\|_2.$$

Lemma 10 (**Bernstein's inequality for sub-exponential random variables**). *If $X_1, \cdots, X_m$ are sub-exponential random variables, then Bernstein's inequality (the inequality (B.49) in Lemma 6 holds with*

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^n \|X_i\|^2_{\text{Orlicz}_1}, \quad B = \max_{1\le i\le n} \|X_i\|_{\text{Orlicz}_1}.$$

Lemma 11 (**vector Bernstein's inequality, Pinelis (1994, 1999)**). *Let $\{X_i\}_{i=1}^n$ be a sequence of i.i.d. random variables taking values in a real separable Hilbert space. Assume that $\mathbb{E}[X_i] = \mu$, $\mathbb{E}[\|X_i - \mu\|^2] = \sigma^2$, $\forall 1 \le i \le n$. We say that* vector Bernstein's condition *with parameter B holds if for all $1 \le i \le n$,*

$$\mathbb{E}[\|X_i - \mu\|^k] \le \frac{1}{2}k!\sigma^2 B^{k-2}, \quad \forall 2 \le k \le n.$$

If this condition holds, then for all $\delta \in (0, 1)$, with probability at least $1 - \delta$ we have

$$\left\|\frac{1}{n}\sum_{i=1}^n X_i - \mu\right\| \le \sqrt{\frac{2\sigma^2 \log\frac{2}{\delta}}{n}} + \frac{B\log\frac{2}{\delta}}{n}.$$

The following definitions and lemmata provide some background on generic chaining.

Definition 7 (**Orlicz-α processes**). *Let $\{X_f\}_{f\in\mathcal{F}}$ be a sequence of random variables. $\{X_f\}_{f\in\mathcal{F}}$ is called an Orlicz-α process for a metric $\texttt{metr}(\cdot,\cdot)$ on $\mathcal{F}$ if*

$$\|X_{f_1} - X_{f_2}\|_{\text{Orlicz}_\alpha} \le \texttt{metr}(f_1, f_2), \forall f_1, f_2 \in \mathcal{F}.$$

In particular, Orlicz-2 process is called "process with sub-Gaussian increments" and Orlicz-1 process is called "process with sub-exponential increments".

Definition 8 (**mixed sub-Gaussian-sub-exponential increments, Dirksen (2015)**). *We say a process $(X_\theta)_{\theta\in\Theta}$ has mixed sub-Gaussian-sub-exponential increments with respect to the pair $(\texttt{metr}_1, \texttt{metr}_2)$ if for all $\theta_1, \theta_2 \in \Theta$,*

$$\text{Prob}\left(\|X_{\theta_1} - X_{\theta_2}\| \ge \sqrt{u}\cdot\texttt{metr}_2(\theta_1,\theta_2) + u\cdot\texttt{metr}_1(\theta_1,\theta_2)\right) \le 2e^{-u}, \forall u \ge 0.$$

Definition 9 (**Talagrand's γ_α−functional**). *A sequence $F = (\mathcal{F}_n)_{n\geq 0}$ of subsets of $\mathcal{F}$ is called admissible if $|\mathcal{F}_0| = 1$ and $|\mathcal{F}_n| \leq 2^{2^n}$ for all $n \geq 1$. For any $0 < \alpha < \infty$, the γ_α−functional of $(\mathcal{F}, \mathtt{metr})$ is defined by*

$$\gamma_\alpha(F, d) = \inf_F \sup_{f\in\mathcal{F}} \sum_{n=0}^{\infty} 2^{\frac{n}{\alpha}} \mathtt{metr}(f, \mathcal{F}_n),$$

where the infimum is taken over all admissible sequences and we write $\mathtt{metr}(f, \mathcal{F}_n) = \inf_{s\in\mathcal{F}_n} \mathtt{metr}(f, s)$.

Lemma 12 (**Dudley's integral bound for γ_α functional, Talagrand (1996)**). *There exist a constant C_α depending only on α such that*

$$\gamma_\alpha(\mathcal{F}, \mathtt{metr}) \leq C_\alpha \int_0^{+\infty} (\log N(\varepsilon, \mathcal{F}, \mathtt{metr}))^{\frac{1}{\alpha}} d\varepsilon.$$

Lemma 13 (**generic chaining for a process with mixed tail increments, Dirksen (2015)**). *If $(X_f)_{f\in\mathcal{F}}$ has mixed sub-Gaussian-sub-exponential increments with respect to the pair $(\mathtt{metr}_1, \mathtt{metr}_2)$, then there are absolute constants $c, C > 0$ such that $\forall \delta \in (0, 1)$,*

$$\sup_{\theta\in\Theta} \|X_f - X_{f_0}\| \leq C(\gamma_2(\mathcal{F}, \mathtt{metr}_2) + \gamma_1(\mathcal{F}, \mathtt{metr}_1)) +$$
$$c\left(\sqrt{\log\frac{1}{\delta}} \sup_{f_1,f_2\in\mathcal{F}} [\mathtt{metr}_2(f_1, f_2)] + \log\frac{1}{\delta} \sup_{f_1,f_2\in\mathcal{F}} [\mathtt{metr}_1(f_1, f_2)]\right),$$

with probability at least $1 - \delta$.

D Proofs for Appendix A

D.1 Proof of Theorem 8

The proof consists of five parts. Among them, the main purpose of Part I and Part IV is to localized the strong convexity parameter. When there is no need to localized the strong convexity parameter (e.g., when one uses the square cost), the proof can be simplified—Part I and Part IV will be quite straightforward, and all the "upper-side" truncation analysis related to $\frac{2\|\xi\|_{L_2}}{\sqrt{c_\kappa}}$, $\frac{4\|\xi\|_{L_2}^2}{c_\kappa}$ or $\frac{4\|\xi\|_{L_2}^2}{\kappa^2 c_\kappa}$ will be unnecessary.

Part I: analysis of the concentrated functions.

Denote $T(h) = \|h - h^*\|_{L_2}^2$ and

$$v_h = \min\left\{\kappa\|h - h^*\|_{L_2}, \frac{2\|\xi\|_{L_2}}{\sqrt{c_\kappa}}\right\}.$$

For every $h \in \mathcal{H}$, define

$$f_h(x, y) = \frac{2}{\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)} \partial_1 \ell_{\mathtt{sv}}(h^*(x), y)(h(x) - h^*(x)),$$
$$g_h(x, y) = \min\left\{(h(x) - h^*(x))^2, v_h^2\right\} \cdot \mathbb{1}\left\{|\xi| \leq \frac{2\|\xi\|_{L_2}}{\sqrt{c_\kappa}}\right\}.$$

One can view g_h as a truncated version of the quadratic form $(h(x)-h^*(x))^2$. Later we will use concentration to control $(\mathbb{P}-\mathbb{P}_n)(f_h+g_h)$ uniformly.

From Lemma 14 (for which we defer to the end of Section D.1), we can show

$$\begin{aligned}\ell_{\mathtt{sv}}(h(x),y)-\ell_{\mathtt{sv}}(h^*(x),y)-\partial_1\ell_{\mathtt{sv}}(h^*(x),y)(h(x)-h^*(x)) &\geq \frac{\alpha(2v_h)}{2}\min\left\{(h(x)-h^*(x))^2, v_h^2\right\}\\ &\geq \frac{\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)}{2}g_h(x,y).\end{aligned}$$

The above inequality implies that

$$\mathbb{P}_n(f_h+g_h)\leq \frac{2}{\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)}\mathbb{P}_n[\ell_{\mathtt{sv}}(h(x),y)-\ell_{\mathtt{sv}}(h^*(x),y)]. \tag{D.1}$$

Recall that $\xi=h^*(x)-y$. By Markov's inequality,

$$\text{Prob}\left(|\xi|\geq\frac{2\|\xi\|_{L_2}}{\sqrt{c_\kappa}}\right)\leq\frac{c_\kappa}{4}, \tag{D.2}$$

From the definition of g_h and v_h, it is straightforward to show that

$$\begin{aligned}\mathbb{P}g_h &= \mathbb{P}\left[\min\left\{(h(x)-h^*(x))^2,v_h^2\right\}\cdot\mathbb{1}\left\{|\xi|\leq\frac{2\|\xi\|_{L_2}}{\sqrt{c_\kappa}}\right\}\right]\\ &\geq \mathbb{P}\left[\min\left\{(h(x)-h^*(x))^2,v_h^2\right\}\cdot\mathbb{1}\left\{|h(x)-h^*(x)|\geq\kappa\|h-h^*\|_{L_2}\right\}\cdot\mathbb{1}\left\{|\xi|\leq\frac{2\|\xi\|_{L_2}}{\sqrt{c_\kappa}}\right\}\right]\\ &\geq \mathbb{P}\left[v_h^2\cdot\mathbb{1}\left\{|h(x)-h^*(x)|\geq\kappa\|h-h^*\|_{L_2}\right\}\cdot\mathbb{1}\left\{|\xi|\leq\frac{2\|\xi\|_{L_2}}{\sqrt{c_\kappa}}\right\}\right]\\ &= v_h^2\cdot\text{Prob}\left(|h(x)-h^*(x)|\geq\kappa\|h-h^*\|_{L_2},\ |\xi|\leq\frac{2\|\xi\|_{L_2}}{\sqrt{c_\kappa}}\right)\\ &\geq v_h^2\cdot\left(\text{Prob}\left(|h(x)-h^*(x)|\geq\kappa\|h-h^*\|_{L_2}\right)-\text{Prob}\left(\|\xi|>\frac{2\|\xi\|_{L_2}}{\sqrt{c_\kappa}}\right)\right)\\ &\geq \frac{3c_\kappa}{4}v_h^2,\end{aligned}$$

where the first inequality is due to $1\geq\mathbb{1}\left\{|h(x)-h^*(x)|\geq\kappa\|h-h^*\|_{L_2}\right\}$; the second inequality is due to the definition of v_h; and the last inequality is due to Assumption 9 and (D.2). From Assumption 8, we have

$$\mathbb{P}(f_h+g_h)\geq\mathbb{P}g_h\geq\frac{3c_\kappa}{4}v_h^2. \tag{D.3}$$

Let us summarize the results from this part. We use the empirical average of the excess loss to upper bound $\mathbb{P}_n(f_h+g_h)$ in (D.1), and use the (truncated) quadratic form to lower bound $\mathbb{P}(f_h+g_h)$ in (D.3). The next steps are to prove concentration of f_h and g_h and establish a "uniform localized convergence" argument.

Part II: bound the localized empirical process.

Given $r>0$, we want to bound the localized empirical process

$$\sup_{\frac{r}{\lambda}\leq T(h)\leq r}(\mathbb{P}-\mathbb{P}_n)(f_h+g_h)$$

where $\lambda > 1$ is a fixed value that we will specify later. From the definition of $\varphi_{\text{noise}}(r;\delta)$ in Assumption 10, for any $\delta \in (0,1)$, with probability $1-\frac{\delta}{2}$, we have

$$\sup_{\frac{r}{\lambda}\le T(h)\le r}(\mathbb{P}-\mathbb{P}_n)(f_h+g_h) \le \sup_{\frac{r}{\lambda}\le T(h)\le r}(\mathbb{P}-\mathbb{P}_n)g_h + \frac{2}{\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)}\varphi_{\text{noise}}\left(r;\frac{\delta}{2}\right). \tag{D.4}$$

Given $r>0$, denote the hypothesis class $\mathcal{H}(\frac{r}{\lambda},r) = \{h\in\mathcal{H} : \frac{r}{\lambda}\le T(h)\le r\}$, and define the function $g_{h,r}$ as

$$g_{h,r}(x,y) = \min\left\{(h(x)-h^*(x))^2, \kappa^2 r, \frac{4\|\xi\|_{L_2}^2}{c_\kappa}\right\}\cdot \mathbb{1}\left\{|\xi| \le \frac{2\|\xi\|_{L_2}}{\sqrt{c_\kappa}}\right\}.$$

Recall that g_h is defined by

$$g_h(x,y) = \min\left\{(h(x)-h^*(x))^2, \kappa^2 T(h), \frac{4\|\xi\|_{L_2}^2}{c_\kappa}\right\}\cdot \mathbb{1}\left\{|\xi| \le \frac{2\|\xi\|_{L_2}}{\sqrt{c_\kappa}}\right\}.$$

For every $h\in\mathcal{H}(\frac{r}{\lambda},r)$ and any $(x,y)\in\mathcal{X}\times\mathcal{Y}$, we have

$$g_{h,\frac{r}{\lambda}}(x,y) \le g_h(x,y) \le g_{h,r}(x,y), \tag{D.5}$$

and

$$\begin{aligned}
&g_{h,r}(x,y) - g_{h,\frac{r}{\lambda}}(x,y)\\
&\le \min\left\{(h(x)-h^*(x))^2, \kappa^2 r, \frac{4\|\xi\|_{L_2}^2}{c_k}\right\} - \min\left\{(h(x)-h^*(x))^2, \frac{\kappa^2 r}{\lambda}, \frac{4\|\xi\|_{L_2}^2}{c_\kappa}\right\}\\
&\le \min\left\{\kappa^2 r, \frac{4\|\xi\|_{L_2}^2}{c_k}\right\} - \min\left\{\frac{\kappa^2 r}{\lambda}, \frac{4\|\xi\|_{L_2}^2}{c_\kappa}\right\}\\
&\le \left(1-\frac{1}{\lambda}\right)\min\left\{\kappa^2 r, \frac{4\|\xi\|_{L_2}^2}{c_\kappa}\right\}.
\end{aligned} \tag{D.6}$$

From (D.5) and (D.6), for every $h\in\mathcal{H}(\frac{r}{\lambda},r)$ and any $(x,y)\in\mathcal{X}\times\mathcal{Y}$,

$$-\left(1-\frac{1}{\lambda}\right)\min\left\{\kappa^2 r, \frac{4\|\xi\|_{L_2}^2}{c_\kappa}\right\} \le g_h(x,y) - g_{h,r}(x,y) \le 0,$$

which implies

$$(\mathbb{P}-\mathbb{P}_n)g_h \le (\mathbb{P}-\mathbb{P}_n)g_{h,r} + \left(1-\frac{1}{\lambda}\right)\min\left\{\kappa^2 r, \frac{4\|\xi\|_{L_2}^2}{c_\kappa}\right\}.$$

As a result, we have

$$\begin{aligned}
\sup_{\frac{r}{K}\le T(h)\le r}(\mathbb{P}-\mathbb{P}_n)g_h &\le \sup_{\frac{r}{K}\le T(h)\le r}(\mathbb{P}-\mathbb{P}_n)g_{h,r} + \left(1-\frac{1}{\lambda}\right)\min\left\{\kappa^2 r, \frac{4\|\xi\|_{L_2}^2}{c_\kappa}\right\}\\
&\le \sup_{T(h)\le r}(\mathbb{P}-\mathbb{P}_n)g_{h,r} + \left(1-\frac{1}{\lambda}\right)\min\left\{\kappa^2 r, \frac{4\|\xi\|_{L_2}^2}{c_\kappa}\right\}.
\end{aligned} \tag{D.7}$$

We know that $g_{h,r}$ is uniformly bounded by $\left[0, \min\left\{\kappa^2 r, \frac{4\|\xi\|_{L_2}^2}{c_\kappa}\right\}\right]$. Form the standard bound for global Rademacher complexity (Wainwright (2019)), $\forall \delta \in (0,1)$, with probability at least $1-\frac{\delta}{2}$,

$$\sup_{T(h)\le r} (\mathbb{P}-\mathbb{P}_n) g_{h,r} \le 2\mathfrak{R}\{g_{h,r} : T(h) \le r\} + \min\left\{\kappa^2 r, \frac{4\|\xi\|_{L_2}^2}{c_\kappa}\right\}\sqrt{\frac{\log\frac{2}{\delta}}{2n}}. \tag{D.8}$$

It is straightforward to verify that for all $h_1, h_2 \in \mathcal{H}$ and $(x,y) \in \mathcal{X}\times\mathcal{Y}$,

$$|g_{h_1,r}(x) - g_{h_2,r}(x)| \le 2\kappa\sqrt{r}|h_1(x) - h_2(x)|.$$

From the Lipchitz contraction property of Rademacher complexity (see, e.g., Theorem 7 in Meir and Zhang (2003)), we have

$$\mathfrak{R}\{g_{h,r}\} \le 2\kappa\sqrt{r}\mathfrak{R}\{h : T(h) \le r\} \le 2\kappa\sqrt{r}\varphi(r), \tag{D.9}$$

where $\varphi(r)$ is defined in Assumption 10. Define the ψ function as

$$\psi(r;\delta) = 4\kappa\sqrt{r}\varphi(r) + \left(\sqrt{\frac{\log\frac{2}{\delta}}{2n}} + 1 - \frac{1}{\lambda}\right)\min\left\{\kappa^2 r, \frac{4\|\xi\|_{L_2}^2}{c_\kappa}\right\} + \frac{2}{\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)}\varphi_{\text{noise}}\left(r; \frac{\delta}{2}\right). \tag{D.10}$$

Combining the definition (D.10) with (D.4) (D.7) (D.8) (D.9) , for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have

$$\sup_{\frac{r}{K}\le T(h)\le r} (\mathbb{P}-\mathbb{P}_n)(f_h + g_h) \le \psi(r;\delta). \tag{D.11}$$

Part III: the "uniform localized convergence" argument.

Applying Proposition 3, for any $\delta_1 \in (0,1)$ and $r_0 \in (0, 4\Delta^2)$, with probability at least $1-\delta_1$, for all $h \in \mathcal{H}$, either $T(h) \le r_0$ or

$$\begin{aligned}
(\mathbb{P}-\mathbb{P}_n)(f_h+g_h) &\le \psi\left(\lambda T(h); \delta_1(\log_K \frac{4K\Delta^2}{r_0})^{-1}\right) \\
&= 4\kappa\sqrt{\lambda T(h)}\varphi(\lambda T(h)) + \frac{2}{\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)}\varphi_{\text{noise}}\left(\lambda T(h); \frac{\delta_1}{2\log_\lambda \frac{4\lambda\Delta^2}{r_0}}\right) \\
&\quad + \underbrace{\min\left\{\lambda\kappa^2 T(h), \frac{4\|\xi\|_{L_2}^2}{c_\kappa}\right\}\left(\sqrt{\frac{\log\frac{2\log_\lambda\frac{4\lambda\Delta^2}{r_0}}{\delta_1}}{2n}} + 1 - \frac{1}{\lambda}\right)}_{\text{last term in (D.12)}}.
\end{aligned} \tag{D.12}$$

We specify

$$\lambda = \frac{8+2c_\kappa}{8+c_\kappa}.$$

Then when $n > \frac{32}{c_\kappa^2} \log \frac{2\log_\lambda \frac{4\lambda\Delta^2}{r_0}}{\delta_1}$, for all $h \in \mathcal{H}$ we have

$$\lambda \left(\sqrt{\frac{\log \frac{2\log_\lambda \frac{4\lambda\Delta^2}{r_0}}{\delta_1}}{2n}} + 1 - \frac{1}{\lambda} \right) < \frac{c_\kappa}{4},$$

which implies when $T(h) > 0$,

$$\text{last term in (D.12)} < \frac{c_\kappa}{4} \min \left\{ \kappa^2 r, \frac{4\|\xi\|_{L_2}^2}{\lambda c_\kappa} \right\} \le \frac{c_\kappa}{4} \min \left\{ \kappa^2 r, \frac{4\|\xi\|_{L_2}^2}{c_\kappa} \right\}. \tag{D.13}$$

Denote $C_{r_0} = 2 + \left(\frac{16}{c_\kappa} + 2 \right) \log \frac{4\Delta^2}{r_0}$, then

$$2\log_\lambda \frac{4\lambda\Delta^2}{r_0} = 2 + \frac{\log \frac{4\Delta^2}{r_0}}{\log \lambda} \le 2 + \left(\frac{16}{c_\kappa} + 2 \right) \log \frac{4\Delta^2}{r_0} = C_{r_0}.$$

For any $\delta \in (0,1)$, taking $\delta_1 = \frac{2\log_\lambda \frac{4\lambda\Delta^2}{r_0}}{C_{r_0}} \delta$, from (D.12) (D.13) and the fact $\lambda < 2$, we have the following conclusion: when $n > \frac{32}{c_\kappa^2} \log \frac{C_{r_0}}{\delta}$, with probability at least $1-\delta$, for all $h \in \mathcal{H}$, either $T(h) \le r_0$ or

$$\begin{aligned} &(\mathbb{P} - \mathbb{P}_n)(f_h + g_h) \\ &< 4\kappa \sqrt{2T(h)} \varphi(2T(h)) + \frac{2}{\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)} \varphi_{\text{noise}}\left(2T(h); \frac{\delta}{C_{r_0}}\right) + \frac{c_\kappa}{4} \min \left\{ \kappa^2 T(h), \frac{4\|\xi\|_{L_2}^2}{c_\kappa} \right\}. \end{aligned} \tag{D.14}$$

Let $\hat{h} \in \arg\min \mathbb{P}_n \ell_{\texttt{sv}}(h(x), y)$ be the empirical risk minimizer. From (D.1) and the property of $\hat{h}$, we have

$$\mathbb{P}_n(f_{\hat{h}} + g_{\hat{h}}) \le \frac{2}{\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)} \mathbb{P}_n[\ell_{\texttt{sv}}(\hat{h}(x) - y) - \ell_{\texttt{sv}}(h^*(x) - y)] \le 0. \tag{D.15}$$

Recall the result (D.3) proved in Part I,

$$\mathbb{P}(f_h + g_h) \ge \frac{3c_\kappa}{4} v_h^2 = \frac{3c_\kappa}{4} \min \left\{ \kappa^2 T(h), \frac{4\|\xi\|_{L_2}^2}{c_\kappa} \right\}. \tag{D.16}$$

From (D.14) (D.15) (D.16), when $n > \frac{32}{c_\kappa^2} \log \frac{C_{r_0}}{\delta}$, with probability at least $1-\delta$, either $T(\hat{h}) \le r_0$ or

$$\begin{aligned} &\frac{3c_\kappa}{4} \min \left\{ \kappa^2 T(\hat{h}), \frac{4\|\xi\|_{L_2}^2}{c_\kappa} \right\} \\ &< 4\kappa \sqrt{2T(\hat{h})} \varphi(2T(\hat{h})) + \frac{2}{\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)} \varphi_{\text{noise}}\left(2T(\hat{h}); \frac{\delta}{C_{r_0}}\right) + \frac{c_\kappa \kappa^2}{4} \min \left\{ T(\hat{h}), \frac{4\|\xi\|_{L_2}^2}{c_\kappa} \right\}, \end{aligned}$$

i.e.,

$$\begin{aligned}&\frac{c_\kappa}{2}\min\left\{\kappa^2 T(\hat{h}),\frac{4\|\xi\|_{L_2}^2}{c_\kappa}\right\}\\&<4\kappa\sqrt{2T(\hat{h})}\varphi(2T(\hat{h}))+\frac{2}{\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)}\varphi_{\text{noise}}\left(2T(\hat{h});\frac{\delta}{C_{r_0}}\right).\end{aligned}\tag{D.17}$$

In the theorem we have asked $n>\frac{72}{c_\kappa^2}\log\frac{C_{r_0}}{\delta}$. Denote the event

$$\mathscr{A}:=\{\text{either } T(\hat{h})\le r_0 \text{ or (D.17) is true}\}.$$

Then we have $\text{Prob}(\mathscr{A})\ge 1-\delta$.

Part IV: preliminary localization.

We first prove a preliminary localization result $T(\hat{h})\le\max\left\{\frac{4\|\xi\|_{L_2}^2}{\kappa^2 c_\kappa},r_0\right\}$ on the event $\mathscr{A}$. The essential purpose of this step is to localize the strong convexity parameter. If $T(\hat{h})\in\left(\max\left\{\frac{4\|\xi\|_{L_2}^2}{\kappa^2 c_\kappa},r_0\right\},4\Delta^2\right]$ is true, then on the event $\mathscr{A}$ one have

$$\text{RHS of (D.17)}>2\|\xi\|_{L_2}^2.\tag{D.18}$$

In the theorem we ask $n>\max\left\{\bar{N}_{\delta,r_0},\frac{72}{c_\kappa^2}\log\frac{C_{r_0}}{\delta}\right\}$. According to Assumption 10, this implies that

$$\varphi_{\text{noise}}\left(8\Delta^2;\frac{\delta}{C_{r_0}}\right)\le\frac{\alpha(4\|\xi\|_{L_2}/\sqrt{c_\kappa})\|\xi\|_{L_2}^2}{2},\tag{D.19}$$

$$\varphi\left(8\Delta^2\right)\le\frac{\sqrt{2c_\kappa}\|\xi\|_{L_2}^2}{16\Delta},\tag{D.20}$$

which further imply

$$\text{RHS of (D.17)}\le\|\xi\|_{L_2}^2+\|\xi\|^2=2\|\xi\|_{L_2}^2.\tag{D.21}$$

(D.18) and (D.21) result in a contradiction. Therefore, $T(\hat{h})$ must be bounded by $\max\left\{\frac{4\|\xi\|_{L_2}^2}{\kappa^2 c_\kappa},r_0\right\}$. Then on the event $\mathscr{A}$, either $T(\hat{h})\le r_0$ or

$$\frac{c_\kappa\kappa^2}{2}T(\hat{h})<\text{RHS of (D.17)}.\tag{D.22}$$

Part V: final steps.

Let r_{noise}^* be the fixed point of

$$\frac{4}{c_\kappa\kappa^2\cdot\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)}\varphi_{\text{noise}}\left(2r;\frac{\delta}{C_{r_0}}\right),$$

and r_{ver}^* be the fixed point of

$$\frac{8}{c_\kappa\kappa}\sqrt{2r}\varphi(2r).$$

From the definition of fixed points, when $T(\hat{h}) > \max\{r^*_{\text{ver}},\ r^*_{\text{noise}}\}$, we have

$$\frac{c_\kappa \kappa^2}{4} T(\hat{h}) > \frac{2}{\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)} \varphi_{\text{noise}}\left(2T(\hat{h}); \frac{\delta}{C_{r_0}}\right)$$

and

$$\frac{c_\kappa \kappa^2}{4} T(\hat{h}) > 4\kappa\sqrt{2T(\hat{h})}\varphi(2T(\hat{h})).$$

Contrasting the above two inequalities with our previous result (D.22), on the event $\mathscr{A}$ we have

$$T(\hat{h}) \le \max\{r^*_{\text{ver}},\ r^*_{\xi},\ r_0\}.$$

We conclude that when $n > \max\left\{\bar{N}_{\delta,r_0}, \frac{72}{c_\kappa^2}\log\frac{C_{r_0}}{\delta}\right\}$, with probability at least $1-\delta$,

$$\|\hat{h} - h^*\|_{L_2}^2 \le \max\{r^*_{\text{noise}},\ r^*_{\text{ver}},\ r_0\}. \tag{D.23}$$

Finally, from the optimality condition on h^* (Assumption 8), it is straightforward to prove that for all $h \in \mathcal{H}$,

$$\mathscr{E}(h) \le \frac{\beta_{\mathtt{sv}}}{2}\|h - h^*\|_{L_2}^2.$$

Combining the above inequality with (D.23), we have

$$\mathscr{E}(\hat{h}) \le \frac{\beta_{\mathtt{sv}}}{2}\max\left\{r^*_{\text{noise}},\ r^*_{\text{ver}},\ r_0\right\}.$$

This completes the proof. □

Lemma 14 (lower bound of the residual of the Taylor expansion). *Let $\ell_{\mathtt{sv}}$ be convex with respect to its first argument. Given $v > 0$, for all $u_1, u_2 \in \mathbb{R}$ and $y \in \mathcal{Y}$, we have*

$$\ell_{\mathtt{sv}}(u_1, y) - \ell_{\mathtt{sv}}(u_2, y) - \partial_1\ell_{\mathtt{sv}}(u_2, y)(u_1 - u_2) \ge \frac{\alpha(2v)}{2}\min\{|u_1 - u_2|^2, v^2\} \cdot \mathbb{1}\{|u_2 - y| \le v\}. \tag{D.24}$$

Proof of Lemma 14: we consider the following four cases: (1) $|u_2 - y| > v$; (2) $|u_2 - y| \le v$ and $|u_1 - u_2| \le v$; (3) $|u_2 - y| \le v$ and $u_1 - u_2 > v$; and (4) $|u_2 - y| \le v$ and $u_1 - u_2 < -v$. It is straightforward to prove (D.24) in case (1) and case (2). In case (3), because

$$\ell_{\mathtt{sv}}(u_1, y) - \ell_{\mathtt{sv}}(u_2, y) - \partial_1\ell_{\mathtt{sv}}(u_2, y)(u_1 - u_2) = \int_0^1 (\partial_1\ell_{\mathtt{sv}}(u_2 + t(u_1 - u_2)) - \partial_1\ell_{\mathtt{sv}}(u_2))(u_1 - u_2)dt,$$

and $(\partial_1\ell_{\mathtt{sv}}(u_2 + t(u_1 - u_2)) - \partial_1\ell_{\mathtt{sv}}(u_2))(u_1 - u_2) \ge 0$ for all $t \in [0, 1]$, we have

$$\begin{aligned}\ell_{\mathtt{sv}}(u_1, y) - \ell_{\mathtt{sv}}(u_2, y) - \partial_1\ell_{\mathtt{sv}}(u_2, y)(u_1 - u_2) &\ge \int_0^{\frac{v}{u_1 - u_2}} (\partial_1\ell_{\mathtt{sv}}(u_2 + t(u_1 - u_2)) - \partial_1\ell_{\mathtt{sv}}(u_2))(u_1 - u_2)dt \\ &= \ell_{\mathtt{sv}}(u_2 + v, y) - \ell_{\mathtt{sv}}(u_2, y) - \partial_1\ell_{\mathtt{sv}}(u_2, y)v \ge \frac{\alpha(2v)}{2}v^2.\end{aligned}$$

Similarly, we can prove (D.24) in case (4). This completes the proof of Lemma 14. □

D.2 Proof of Corollary 9

The proof is nearly identical to the proof of Theorem 8, but with the following modifications. First, we only need to consider the hypothesis set $\mathcal{H}_0$. Second, based on the definition of φ_{noise} in Corollary 9, we modify (D.4) to

$$\sup_{h\in\mathcal{H}_0,\frac{r}{\lambda}\le T(h)\le r}(\mathbb{P}-\mathbb{P}_n)(f_h+g_h)\le \sup_{\frac{r}{\lambda}\le T(h)\le r}(\mathbb{P}-\mathbb{P}_n)g_h+\frac{2}{\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)}\left(\varphi_{\text{noise}}\left(r;\frac{\delta}{2}\right)-\Phi(h^*)\right). \tag{D.25}$$

We also do similar modifications to (D.10) (D.11) (D.12) (D.14). Third, we modify (D.15) (note that this is the only place we use the property of empirical risk minimization) as follows:

$$\begin{aligned}\mathbb{P}_n(f_{\hat{h}}+g_{\hat{h}})&\le\frac{2}{\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)}\mathbb{P}_n[\ell_{\mathtt{sv}}(\hat{h}(x)-y)-\ell_{\mathtt{sv}}(h^*(x)-y)]\\&\le\frac{2}{\alpha\left(4\|\xi\|_{L_2}/\sqrt{c_\kappa}\right)}\Phi(h^*),\end{aligned} \tag{D.26}$$

where the first inequality is due to (D.1) and the second inequality is due to the definition (A.10) of the estimator $\hat{h}$. After all these modifications, the inequality (D.17) still hold true, and the remaining proof is identical to that of Theorem 8. □