On the Efficiency of ERM in Feature Learning

Ayoub El Hanchi
University of Toronto &
Vector Institute
aelhan@cs.toronto.edu

Chris J. Maddison
University of Toronto &
Vector Institute
cmaddis@cs.toronto.edu

Murat A. Erdogdu
University of Toronto &
Vector Institute
erdogdu@cs.toronto.edu

Abstract

Given a collection of feature maps indexed by a set $\mathcal{T}$, we study the performance of empirical risk minimization (ERM) on regression problems with square loss over the union of the linear classes induced by these feature maps. This setup aims at capturing feature learning, where the model is expected to jointly learn from the data an appropriate feature map as well as a linear predictor on top of it. We start by studying the asymptotic quantiles of the excess risk of ERM in this setting. Remarkably, we show that when there is a unique optimal feature map, these quantiles coincide, up to a factor of two, with those of the excess risk of the oracle procedure, which knows a priori this optimal feature map and deterministically outputs an empirical risk minimizer from the optimal linear class. We derive a non-asymptotic upper bound on the excess risk that captures a refined version of this phenomenon. Specifically, under a mild assumption, the upper bound we derive depends on the sample size n as C_n/n, where C_n is a sequence of monotonically decreasing distribution-dependent constants. Finally, we specialize our analysis to the case where the set $\mathcal{T}$ is finite, where we explicitly compute these constants in terms of moments of the feature maps and target as well as the size of $\mathcal{T}$.

1 Introduction

A central idea in modern machine learning is that of data-driven feature learning. Specifically, instead of performing linear prediction on top of hand-crafted features, the current dominant paradigm suggests to use models that select useful features for linear prediction in a data-dependent way [e.g. KSH12; LBH15; He+16; Vas+17]. Of course, by putting the burden of picking a feature map on the model and data, we should expect that the resulting learning problem will require more samples to be solved. But just how many more samples do we need to learn such feature-learning-based models?

In this paper, we investigate this question in a general setting. We study the performance of empirical risk minimization (ERM) on regression tasks with square loss and over model classes induced by arbitrary collections of features maps. More precisely, let X be the random input taking value in a set $\mathcal{X}$, and let $(\phi_t)_{t\in\mathcal{T}}$, $\phi_t : \mathcal{X} \to \mathbb{R}^d$, be a collection of feature maps indexed by a set $\mathcal{T}$. For a given regression task and i.i.d. samples, our aim is to understand the performance of ERM over the class of predictors $\cup_{t\in\mathcal{T}}\{x \mapsto \langle w, \phi_t(x)\rangle \mid w \in \mathbb{R}^d\}$ as a function of the sample size, the distribution of the data, and relevant properties of the collection of feature maps $(\phi_t)_{t\in\mathcal{T}}$.

Classical uniform-convergence-based analyses would suggest that the performance of ERM in this setting would be determined by an appropriately defined measure of size of the model class. The main message of this paper is that in this case, this is wrong in a strong sense. Specifically, we prove an upper bound on the excess risk of ERM on this problem whose dependence on the size of the model class decays monotonically with the sample size, and eventually depends only on the size of the model class induced by the collection of optimal feature maps, which is typically *much* smaller.

Preprint. Under review.

Formal setup. We briefly formalize our problem here. Let X be the random input taking value in a set $\mathcal{X}$, and let $(\phi_t)_{t \in \mathcal{T}}$, $\phi_t : \mathcal{X} \to \mathbb{R}^d$, be a collection of feature maps indexed by a set $\mathcal{T}$. Let $Y \in \mathbb{R}$ be the output random variable, jointly distributed with the input X. Our goal is to learn to predict the output Y given the input X as well as possible within the class of predictors $\{x \mapsto \langle w, \phi_t(x) \rangle \mid (t, w) \in \mathcal{T} \times \mathbb{R}^d\}$. We evaluate the quality of a single prediction $\hat{y}$ given the ground truth y through the loss function $\ell(\hat{y}, y) = (\hat{y} - y)^2/2$, and the overall quality of a predictor $(t, w) \in \mathcal{T} \times \mathbb{R}^d$ through its risk

$$R(t, w) := \mathrm{E}[\ell(\langle w, \phi_t(X) \rangle, Y)], \qquad R_* := \inf_{(t,w) \in \mathcal{T} \times \mathbb{R}^d} R(t, w).$$

We assume that we have access to n i.i.d. samples $(X_i, Y_i)_{i=1}^n$ with the same distribution as (X, Y), and that for each $t \in \mathcal{T}$, we have the ability to evaluate the feature map $\phi_t : \mathcal{X} \to \mathbb{R}^d$ pointwise. We perform ERM on the empirical risk

$$(\hat{t}_n, \hat{w}_n) \in \operatorname*{argmin}_{(t,w) \in \mathcal{T} \times \mathbb{R}^d} R_n(t, w) \quad \text{where} \quad R_n(t, w) := n^{-1} \sum_{i=1}^{n} \ell(\langle w, \phi_t(X_i) \rangle, Y_i).$$

Our goal in this paper is to derive in-probability upper bounds on the excess risk $R(\hat{t}_n, \hat{w}_n) - R_*$.

Related work. The study of upper bounds on the excess risk of ERM in a general setting is a classical topic. It was initiated by Vapnik and Chervonenkis [VC74] who established a link between the excess risk of ERM and the uniform convergence of the underlying empirical process. More recently, and fuelled by the development of Talagrand's concentration inequality [Tal96] and its refinements [e.g. BLM00; Bou02], a literature emerged that provided more fine-grained control of the excess risk of ERM [e.g. BBM05; Kol06; BM06]. A key idea emerging from this line of work is localization. This concept, and in particular the iterative localization method of Koltchinskii [Kol06], plays an important role in our development. We refer the reader to the books [Kol11; Wai19], as well as the recent articles [LRS15b; KRV22] for more on this idea.

Focusing on the task of regression with square loss, upper bounds on the excess risk of ERM are available for many classes of predictors, including finite [e.g. Aud07; JRT08; LM09], linear [e.g. LM16b; Oli16; Mou22], and convex classes [e.g. LM16a; Men14; LRS15a]. A key development in this area over the last decade has been the realization that such bounds can be obtained under much weaker assumptions than previously thought [Men14; Oli16], owing to the fact that only one-sided control of a certain empirical process is needed [Men14; KM15; Oli16], and which can be obtained under very weak assumptions. The line of work most closely related to ours is the one on random-design linear regression [AC11; HKZ12; Oli16; LM16c; Sau18; Mou22; EE23], and we view our work as an extension of this literature. We review these results in more detail in Section 2.

Finally, and on a more conceptual level, our work is related to the recent effort to understand the effect of feature learning on the performance of neural networks [e.g. COB19; Gho+19; Ba+22]. Beyond this conceptual connection however, our work is quite distinct from this literature. Among other things, our setting is more general since we consider arbitrary features maps. In the same vein, it is worth mentioning the line of work on multiple kernel learning [e.g. Lan+04; GA11; SD16], although we are not aware of results from this literature that are directly relevant to our setup.

Challenges. Our class of predictors is somewhat unstructured (e.g. it is in general non-convex), so that off-the-shelf results from the above literature are not directly applicable. Nevertheless, the analysis of the performance of ERM on linear classes provides a good starting point as we review in Section 2. Compared to that setting however, we are faced with two additional challenges. First, we need to control an additional source of error arising from the fact that ERM might select a suboptimal feature map. Second, we are lead to study the suprema of certain $\mathcal{T}$-indexed empirical processes, which in the linear setting reduce to single random variables that are easily dealt with.

Organization. The rest of the paper is organized as follows. In Section 2, we review known result on the excess risk of linear regression under square loss. In Section 3, we state our main results that hold for the excess risk of ERM for general index sets $\mathcal{T}$. In Section 4, we specialize our analysis to the case where the index set $\mathcal{T}$ is finite, and provide explicit estimates on certain distribution-dependent constants that arise in our main results. We conclude in Section 5 with a brief discussion.

2 Background

The goal of this section is to provide more context for our results. We review known results on the excess risk of ERM over linear classes, which corresponds in our setting to the special case where the set $\mathcal{T}$ indexing the feature maps is a singleton. As such, to avoid introducing further notation, we use the one from the previous section, while dropping the dependence on t whenever it occurs.

In the linear regression setup where the loss is quadratic, the excess risk of ERM has a closed form when the sample covariance matrix of the feature map is invertible. Specifically, define

$$\Sigma := \mathrm{E}\big[\phi(X)\phi(X)^T\big], \qquad \Sigma_n := \frac{1}{n}\sum_{i=1}^{n}\phi(X_i)\phi(X_i)^T, \tag{1}$$

and let w_* denote the unique minimizer of the risk $R(w)$. Then, an elementary calculation shows that when Σ_n is invertible, there is a unique empirical risk minimizer and it satisfies

$$\hat{w}_n = w_* - \Sigma_n^{-1}\nabla R_n(w_*). \tag{2}$$

Furthermore, since the risk is a quadratic function of w whose gradient at w_* vanishes, replacing $R(\hat{w}_n)$ by the equivalent exact second order Taylor expansion around w_* yields

$$R(\hat{w}_n) - R(w_*) = \frac{1}{2}\|\hat{w}_n - w_*\|_{\Sigma}^2 = \frac{1}{2}\big\|\Sigma_n^{-1}\nabla R_n(w_*)\big\|_{\Sigma}^2. \tag{3}$$

While exact, this expression is not readily interpretable. For example, how fast does this excess risk go to 0 as a function of the sample size? The following classical result from asymptotic statistics [e.g. Whi82; LC06; Van00] makes this more explicit. To state it, we define

$$g(X,Y) := \nabla_w \ell(\langle w_*, \phi(X)\rangle, Y), \qquad G := \mathrm{E}\big[g(X,Y)g(X,Y)^T\big]. \tag{4}$$

Theorem 1. *Assume that for all $j \in [d]$, $\mathrm{E}\big[\phi_j^2(X)\big] < \infty$, $\mathrm{E}\big[Y^2\big] < \infty$, and $\mathrm{E}[\|g(X,Y)\|_{\Sigma^{-1}}^2] < \infty$. Then, as $n \to \infty$,*

$$n \cdot [R(\hat{w}_n) - R_*] \xrightarrow{d} \frac{1}{2}\|Z\|_2^2,$$

where $Z \sim \mathcal{N}(0, \Sigma^{-1/2}G\Sigma^{-1/2})$. In particular, for any $\delta \in (0, 0.1)$,

$$\lim_{n\to\infty} n \cdot Q_{R(\hat{w}_n)-R_*}(1-\delta) \asymp \mathrm{E}[\|g(X,Y)\|_{\Sigma^{-1}}^2] + 2\lambda_{\max}(\Sigma^{-1/2}G\Sigma^{-1/2})\log(1/\delta),$$

where $Q_X(p) = \inf\{x \in \mathbb{R} \mid F_X(x) \geq p\}$ is the quantile function of a random variable X with $F_X(x) = \mathrm{P}(X \leq x)$, and the constants in the relation $\asymp$ can be taken as $C = 1$ and $c = 1/32$ for the upper and lower bounds.

We provide a proof in Appendix A for completeness. For our purposes, this theorem is most easily interpreted as follows: for large enough n and small enough δ, if the excess risk of ERM is bounded by some quantity with probability at least $1-\delta$, then this quantity is at least as large as the right-hand side of the second displayed equation divided by n. While our primary interest is in non-asymptotic upper bounds, this asymptotic result, by virtue of its exactness, provides us with a benchmark against which upper bounds can be compared. In particular, it identifies the quantity $\mathrm{E}[\|g(X,Y)\|_{\Sigma^{-1}}^2]$ as an intrinsic parameter determining the excess risk of ERM on this problem.

For large enough n, Theorem 1 gives an interpretable expression for the excess risk. However, it says nothing about how large n needs to be for this expression to be accurate. This motivates a non-asymptotic analysis of the excess risk of ERM, which has been carried out numerous times in recent years [e.g. Oli16; LM16c]. A goal of this literature has been to obtain upper bounds on the excess risk of ERM that hold in probability under weak moment assumptions, building on the observation that this is indeed possible [Men14]. The following theorem is comparable to the best known result in this area. We leave the proof to Appendix B. To state it, we define

$$\varphi(x) := \Sigma^{-1/2}\phi(x), \quad V := \mathrm{E}\Big[\big(\varphi(X)\varphi(X)^T - I\big)^2\Big], \quad L := \sup_{v\in S^{d-1}} \mathrm{E}\Big[\big(\langle v, \varphi(X)\rangle^2 - 1\big)^2\Big].$$

Theorem 2. *Assume that for all $j \in [d]$, $\mathrm{E}\big[\phi_j^4(X)\big] < \infty$, $\mathrm{E}\big[Y^2\big] < \infty$, and $\mathrm{E}[\|g(X,Y)\|_{\Sigma^{-1}}^2] < \infty$. Let $\delta \in (0,1)$. If*

$$n \geq (512\lambda_{\max}(V) + 6)\log(ed) + (128L + 11)\log(2/\delta),$$

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

$$R(\hat{w}_n) - R_* \leq \frac{4\,\mathrm{E}[\|g(X,Y)\|_{\Sigma^{-1}}^2]}{\delta n}.$$

At a high-level, this result says that above a certain explicit minimal sample size, the asymptotic expression of the excess risk of Theorem 1 is correct, up to a (significantly) worse dependence on δ. The restriction on the sample size is almost the best one can hope for. To see why, note that to get guarantees on the excess risk of *any* empirical risk minimizer, we need at least that Σ_n is invertible, otherwise there exists an empirical risk minimizer arbitrarily far away from w_*. To get quantitative guarantees, we need slightly more control in the form of a lower bound on $\lambda_{\min}(\Sigma^{-1/2}\Sigma_n\Sigma^{-1/2})$.

This result has two key qualities, which we aim to reproduce in our results. First, it is assumption-lean, requiring nothing more than a fourth moment assumption on the coordinates of the feature map compared to Theorem 1. Second, it recovers the right dependence on the intrinsic parameter $\mathrm{E}[\|g(X,Y)\|_{\Sigma^{-1}}^2]$ identified in Theorem 1. Of course, a downside of this generality is the bad dependence on δ. Without further assumptions, this cannot be improved; we refer the reader to the recent literature on robust linear regression for more on this topic [e.g. LM19a; LM19b; LL20].

3 Main Results

In this section we state our main results. They are most easily seen as extensions of Theorems 1 and 2 for general index sets $\mathcal{T}$. In particular, in Section 3.1, we study the asymptotics of the excess risk of ERM in our setting, and in Section 3.2, we state a non-asymptotic upper bound on the excess risk.

To state our results, we require multiple definitions and some additional notation. We start with the population and the sample covariance matrices

$$\Sigma(t) := \mathrm{E}\big[\phi_t(X)\phi_t(X)^T\big] \quad \text{and} \quad \Sigma_n(t) := n^{-1}\sum_{i=1}^{n}\phi_t(X_i)\phi_t(X_i)^T.$$

We define the following collection of minimizers,

$$w_*(t) := \operatorname*{argmin}_{w\in\mathbb{R}^d} R(t,w), \qquad \mathcal{T}_* := \operatorname*{argmin}_{t\in\mathcal{T}} R(t, w_*(t)),$$

the first is uniquely defined, while the second is set-valued in general. We define the gradient of the loss of these minimizers and their covariance matrices

$$g(t,(X,Y)) := \nabla_w \ell(\langle w_*(t), \phi_t(X)\rangle, Y), \qquad G(t) := \mathrm{E}\big[g(t,(X,Y))g(t,(X,Y))^T\big].$$

The following processes will play a key role in our development

$$\Lambda_n(t) := \sqrt{n}\cdot\lambda_{\max}(I - \Sigma^{-1/2}(t)\Sigma_n(t)\Sigma^{-1/2}(t)), \quad G_n(t) := \sqrt{n}\cdot\|\nabla_w R_n(t, w_*(t))\|_{\Sigma^{-1}(t)} \tag{5}$$

as well as, for $t_* \in \mathcal{T}$ and $t \in \mathcal{T}\setminus\mathcal{T}_*$,

$$\Delta_n(t,t_*) := \sqrt{n}\cdot\left(1 - \frac{R_n(t,w_*(t)) - R_n(t_*, w_*(t))}{R(t,w_*(t)) - R_*}\right). \tag{6}$$

We note that the process $(\Delta_n(t,t_*))_{t\in\mathcal{T}\setminus\mathcal{T}_*}$ is an empirical process, while $(\Lambda_n(t))_{t\in\mathcal{T}}$ and $(G_n(t))_{t\in\mathcal{T}}$ are partial suprema of empirical processes. In the sequel, we will slightly abuse this terminology, and call all of these empirical processes, with the understanding that they can be viewed as one with more indexing. We will further assume that these processes are separable; see [BLM13, p.305-306] for a definition. This covers a wide range of applications, while avoiding delicate measurability issues. The suprema of such separable processes, which is the only way they enter our results, can be studied by taking the supremum over a countable dense subset of the index set. Therefore, without loss of generality, we assume that $\mathcal{T}$ is countable.

Finally, in line with the literature on the theory of empirical processes [VW96], we say that a sequence of empirical processes is Glivenko–Cantelli if, when rescaled by $n^{-1/2}$, the supremum of their absolute value taken over their index set converges to zero in probability as $n \to \infty$. In other words, the weak law of large numbers holds uniformly over the index set of the process. Similarly, we say that a sequence of empirical processes is Donsker if it converges weakly to its limiting Gaussian process. In other words, the central limit theorem holds uniformly over the index set.

3.1 Asymptotic results

Our first main result is an asymptotic upper bound on the quantiles of the excess risk of ERM in our setting, which vastly generalizes that of Theorem 1.

Theorem 3. *Assume that $\mathcal{T}_* \neq \varnothing$ and for some $t_* \in \mathcal{T}_*$, assume that the empirical processes $(\Lambda_n(t))_{t\in\mathcal{T}}$, $(\Delta(t,t_*))_{t\in\mathcal{T}\setminus\mathcal{T}_*}$ and $(G_n(t))_{t\in\mathcal{T}}$ are Glivenko-Cantelli. Then, for all $\varepsilon > 0$,*

$$\lim_{n\to\infty} \mathrm{P}\big(R(\hat{t}_n, w_*(\hat{t}_n)) - R_* > \varepsilon\big) = 0.$$

Furthermore, if the sequence of processes $(G_n(t))_{t\in\mathcal{T}}$ is Donsker, then for $\delta \in (0,1)$,

$$\lim_{n\to\infty} n \cdot Q_{R(\hat{t}_n,\hat{w}_n)-R_*}(1-\delta) \leq \mathrm{E}\left[\sup_{s\in\mathcal{T}_*} Z^2(s)\right] + 2\log(1/\delta) \sup_{s\in\mathcal{T}_*} \lambda_{\max}(\Sigma^{-1/2}(s)G(s)\Sigma^{-1/2}(s))$$

where $(Z(t))_{t\in\mathcal{T}}$ is the limiting Gaussian process of the empirical process $(G_n(t))_{t\in\mathcal{T}}$.

We note that Theorem 3 reduces to the upper bound of Theorem 1 when $\mathcal{T}$ is a singleton, with the *exact* same assumptions. We are not aware of comparable results in the literature. The full proof can be found in Appendix D. Before discussing the content of the theorem, we note the following surprising and easier-to-parse corollary.

Corollary 1. *Assume that $\mathcal{T}_*$ is non-empty and finite, and that for some $t_* \in \mathcal{T}_*$, the empirical processes $(\Lambda_n(t))_{t\in\mathcal{T}}$, $(\Delta(t,t_*))_{t\in\mathcal{T}\setminus\mathcal{T}_*}$ and $(G_n(t))_{t\in\mathcal{T}}$ are Glivenko-Cantelli. Furthermore, assume that the sequence of processes $(G_n(t))_{t\in\mathcal{T}}$ is Donsker. Then*

$$\begin{aligned}\lim_{n\to\infty} n \cdot Q_{R(\hat{t}_n,\hat{w}_n)-R_*}(1-\delta) \leq 80 \cdot (1 + \log|\mathcal{T}_*|) \cdot \max_{s\in\mathcal{T}_*} \mathrm{E}[\|g(s,(X,Y)\|^2_{\Sigma^{-1}(s)}] \\ + 2 \cdot \max_{s\in\mathcal{T}_*} \lambda_{\max}(\Sigma^{-1/2}(s)G(s)\Sigma^{-1/2}(s)) \log(1/\delta).\end{aligned}$$

Furthermore, if $\mathcal{T}_*$ is a singleton with unique element t_*, then for all $\delta \in (0,1)$

$$\lim_{n\to\infty} n \cdot Q_{R(\hat{t}_n,\hat{w}_n)-R_*}(1-\delta) \asymp \frac{1}{2} \cdot Q_{\|Z\|_2^2}(1-\delta),$$

where $Z \sim \mathcal{N}(0, \Sigma^{-1/2}(t_*)G(t_*)\Sigma(t_*)^{-1/2})$, and the constants in the relation $\asymp$ are 2 and 1 for the upper and lower bounds, respectively.

To see why this result is surprising, let us focus on the second statement of Corollary 1, where the optimal feature map is unique. Consider the oracle procedure, which knows beforehand what the optimal feature map t_* is, and outputs t_* and a minimizer of $R_n(t_*, w)$. This statement says that, up to a factor of two, the asymptotic quantiles of the excess risk of ERM, which needs to learn over the large class $\cup_{t\in\mathcal{T}}\{x \mapsto \langle w, \phi_t(x)\rangle \mid w \in \mathbb{R}^d\}$, coincide with those of the oracle procedure (by Theorem 1), which only needs to learn over the *linear* class $\{x \mapsto \langle w, \phi_{t_*}(x)\rangle \mid w \in \mathbb{R}^d\}$!

In some cases, the uniqueness assumption might be too much to ask for, but the first statement of Corollary 1 shows that as long as the size of $\mathcal{T}_*$ is not unreasonably large, the asymptotic excess risk of ERM is not much worse than that of an oracle procedure which knows an optimal feature map beforehand. Finally, in the very unlikely case that the set of optimal feature maps is infinite, the second statement of Theorem 1 bounds the asymptotic quantiles of the excess risk by the quantiles of the supremum of a $\mathcal{T}_*$-indexed Gaussian chaos process. In all cases, the main takeaway is that

Asymptotically, only the complexity of the set of optimal feature maps $\mathcal{T}_*$ affects the excess risk of ERM. The global complexity of $\mathcal{T}$ is irrelevant for the excess risk.

3.2 Non-asymptotic results

The result in Theorem 3 hints at a dramatic localization phenomenon, whereby the influence of the size and complexity of the collection of feature maps $(\phi_t)_{t\in\mathcal{T}}$ on the excess risk of ERM vanishes as $n \to \infty$ under some assumptions. The root of this localization phenomenon is the first statement of Theorem 3: eventually, ERM picks near-optimal feature maps with probability approaching 1. For small enough sample sizes however, it is clear that ERM is likely to select suboptimal feature maps, so that this localization phenomenon cannot hold uniformly over n. This raises a host of questions: (i) How fast, as measured by the sample size, does ERM learn the optimal feature map? (ii) What is the effect of this localization on the rate of decay of the excess risk of ERM non-asymptotically? (iii) What properties of the feature maps $(\phi_t)_{t\in\mathcal{T}}$ influence these rates?

Our answers to these questions in this very general setting are summarized in Theorem 4 below. To state it, it will be useful to define the following normalized process and the parameter

$$\overline{G}_n(t) := \frac{G_n(t)}{\mathrm{E}[\lVert g(t,(X,Y))\rVert^2_{\Sigma^{-1}(t)}]^{1/2}} \quad \text{and} \quad L := \sup \mathrm{E}\left[\left(\sum_{t\in\mathcal{T}}\langle v_t, \Sigma^{-1/2}(t)\phi_t(X)\rangle^2 - 1\right)^2\right],$$

where the supremum is taken over vectors $(v_t)_{t\in\mathcal{T}}$ such that $\sum_{t\in\mathcal{T}}\lVert v_t\rVert_2^2 = 1$. This is simply the variance parameter in Bousquet's inequality [Bou02] applied to the supremum of the empirical process $\Lambda_n(t)$. We need one last definition before stating our result. For $n \in \mathbb{N}$ and $\delta \in (0,1)$, define the set function $F_{n,\delta}$, for any subset $\mathcal{S} \subset \mathcal{T}$, by

$$F_{n,\delta}(\mathcal{S}) := \left\{ t \in \mathcal{T} \;\middle|\; R(t, w_*(t)) - R_* \leq 2\,\mathrm{E}[\sup_{s\in\mathcal{S}} \overline{G}_n^2(s)] \cdot \frac{\mathrm{E}[\lVert g(t,(X,Y))\rVert^2_{\Sigma^{-1}(t)}]}{\delta n} \right\}. \tag{7}$$

This map acts as a contraction as shown in the next lemma, whose proof is deferred to Appendix F. For a function f, we use f^k to denote $f^k(x) := f(f^{k-1}(x))$ with $f^0(x) := x$.

Lemma 1. *Let $n \in \mathbb{N}$ and $\delta \in (0,1)$. Then for all $k \in \{0, 1, \dots\}$,*

- $F^{k+1}_{n,\delta}(\mathcal{T}) \subseteq F^k_{n,\delta}(\mathcal{T})$.
- *If $\exists\, n_0, B$ such that $\mathrm{E}[\sup_{t\in\mathcal{T}} \overline{G}_n^2(t)] \leq B$ for all $n \geq n_0$, then $\bigcap_{n\geq 1} F^k_{n,\delta}(\mathcal{T}) = \mathcal{T}_*$.*

With these definitions, we now state the second main result of the paper.

Theorem 4. *Assume that $\mathcal{T}_* \neq \varnothing$, $\mathrm{E}[Y^2] < \infty$, $\forall (t,j) \in \mathcal{T} \times [d]$, $\mathrm{E}[\phi^2_{t,j}(X)] < \infty$, and $\mathrm{E}[\lVert g(t,(X,Y))\rVert^2_{\Sigma^{-1}(t)}] < \infty$. Let $\delta \in (0,1)$ and $k \in \mathbb{N}$. If, for some $t_* \in \mathcal{T}_*$, n satisfies*

$$n \geq 64\,\mathrm{E}[\sup_{t\in\mathcal{T}} \Lambda_n(t)] + (128L + 11)\log(4/\delta) + 2 \cdot \delta^{-2} \cdot \mathrm{E}[\sup_{t\in\mathcal{T}\setminus\mathcal{T}_*} \Delta_n(t, t_*)],$$

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

$$\hat{t}_n \in F^k_{n,\delta/2k}(\mathcal{T}) =: \mathcal{S}_{n,\delta,k},$$

and

$$R(\hat{t}_n, \hat{w}_n) - R_* \leq 4 \cdot \mathrm{E}[\sup_{s\in\mathcal{S}_{n,\delta,k}} \overline{G}_n^2(s)] \cdot \frac{2k \cdot \mathrm{E}[\lVert g(\hat{t}_n,(X,Y))\rVert^2_{\Sigma^{-1}(\hat{t}_n)}]}{\delta n},$$

where the last expectation is with respect to (X,Y) only, independent of $(X_i,Y_i)_{i=1}^n$ and hence of $\hat{t}_n$, and where the processes Λ_n, Δ_n, and G_n are as in (5) and (6).

We make some general remarks before interpreting the content of the theorem. First, we note that when the index set $\mathcal{T}$ is a singleton, the last term in the sample size restriction vanishes, while the first matches the sample size restriction from Theorem 2 after an application of Lemma 3 below. Furthermore, taking the free parameter $k = 1$ in Theorem 3 recovers the upper bound on the excess risk of Theorem 2 up to a factor of 2. Theorem 3 may therefore be viewed as a broad generalization of Theorem 2. We mention that we are not aware of comparable results in the literature. Second, the statement of Theorem 4 is very general, and in fact, too general for us to be able to interpret it precisely. As such, we will discuss it in the context of the following mild assumption.

Assumption 1. There exists constants C_Λ, C_Δ and a set function C_G, independent of the sample size, but dependent on the remaining parameters of the problem, such that for all $\mathcal{S} \subseteq \mathcal{T}$ and $n \in \mathbb{N}$,

$$\mathrm{E}[\sup_{t\in\mathcal{T}} \Lambda_n(t)] \leq C_\Lambda, \quad \mathrm{E}\left[\sup_{t\in\mathcal{T}\setminus\mathcal{T}_*} \Delta_n(t, t_*)\right] \leq C_\Delta, \quad \mathrm{E}[\sup_{s\in\mathcal{S}} G_n(s)] \leq C_G(\mathcal{S}),$$

and C_G satisfies $C_G(\mathcal{S}') \leq C_G(\mathcal{S})$ for all $\mathcal{S}' \subseteq \mathcal{S}$, and where Λ_n, Δ_n, and G_n are as in (5) and (6).

In general, this assumption holds for reasonably sized index sets, as measured by appropriate entropy numbers [e.g. VW96; Wai19; GN21]. In particular, this assumption always holds for finite index sets, and we will derive in Section 4 explicit estimates of the constants and set function in Assumption 1 in terms of moments of the feature maps and the size of the set $\mathcal{T}$.

Let us now interpret the content of Theorem 4, which comes with a free parameter k, in the context of Assumption 1. We fix k here, and discuss its choice below. First, recalling the definition of $F_{n,\delta}$, this result says that above a certain sample size, both the suboptimality of the feature map picked by ERM and its excess risk decay at the fast rate n^{-1}. Second, it says that these bounds depend on the index set $\mathcal{T}$ and the associated feature maps $(\phi_t)_{t\in\mathcal{T}}$ *only* through the subset $\mathcal{S}_{n,\delta,k}$. Specifically, loosening the upper bound on the excess risk says that this dependence is at most

$$k \cdot C_G(\mathcal{S}_{n,\delta,k}) \sup_{s\in\mathcal{S}_{n,\delta,k}} \mathrm{E}\left[\|g(s,(X,Y))\|^2_{\Sigma^{-1}(s)}\right]. \tag{8}$$

Looking at the definition of $\mathcal{S}_{n,\delta,k}$, we see that the magnitude of this term depends on how easy it is to differentiate between good and bad feature maps for the given target. If they are roughly equally good, then many samples are needed to differentiate between them, so that $\mathcal{S}_{n,\delta,k}$ is only small for large n. On the other hand, if there is a strong separation between optimal and suboptimal feature maps, this subset is small even for moderate n. We mention that by the second item of Lemma 1, the upper bound on the excess risk of Theorem 4 eventually matches the main term in the asymptotic rate of Theorem 3, in the same way that Theorem 2 achieves this when compared with Theorem 1.

Finally, let us turn to the choice of k. Practically, we select the one that minimizes the bound on the excess risk. Looking at the first item of Lemma 1, this optimal k balances the following trade-off: on the one hand, for small k, applications of $F_{n,\delta/2k}$ constrain the input set more severely, but only a few iterations are performed; on the other hand, larger values of k allow multiple iterations, but at the cost of more weakly constraining the input set per application.

Stepping back, the value of Theorem 4 is twofold. First, and perhaps most importantly, it exhibits the localization phenomenon uncovered in Theorem 3 non-asymptotically, and clarifies the properties of the feature maps that affect the speed of this localization and its effect on the excess risk, answering the questions we raised at the beginning of this subsection. Secondly, it provides a template which can be used to derive more explicit excess risk bounds on ERM given estimates on the expected suprema of the relevant empirical processes. Deriving such accurate estimates for infinite $\mathcal{T}$ is in general a highly non-trivial task, and cannot be done at the level of generality we have been operating at. The case of finite $\mathcal{T}$ however is tractable in a general setting as we discuss in the next section.

4 Case study: Finite index sets

A desirable quality of our results of Section 3 is their vast generality. However, and in particular for the non-asymptotic bound of Theorem 4, this comes at the cost of abstract and difficult to interpret quantities appearing in the bounds. Ideally, one would like to relate the various expected suprema appearing in Theorem 4 to explicit functions of, say, moments of the feature maps and of the target, much like in the statement of Theorem 2. Unfortunately, in the very general setting of Section 3 with arbitrary index sets, not much more can be done without committing to a specific problem.

One setting which can be studied while maintaining some generality is the case where the index set is finite. Roughly speaking, this is because a worst-case analysis still yields non-trivial upper bounds on the expected suprema appearing in Theorem 4. This is decidedly not the case when $\mathcal{T}$ is infinite, in which case these expected suprema can be infinite in the worst case.

When the index set $\mathcal{T}$ is finite, the Glivenko-Cantelli and Donsker assumptions of Theorem 3 reduce to simple moment conditions. We record this in the following corollary.

Corollary 2. *Assume that for all* $(t,j) \in \mathcal{T} \times [d]$*,* $\mathrm{E}\big[\phi_{t,j}^2(X)\big] < \infty$*,* $\mathrm{E}\big[Y^2\big] < \infty$*, and for all* $t \in \mathcal{T}$*,* $\mathrm{E}[\|g(t,(X,Y))\|_{\Sigma^{-1}(t)}^2] < \infty$*, and that* $\mathcal{T}$ *is finite. Then the conclusions of Corollary 1 hold.*

Making Theorem 4 more explicit is a more laborious task. We recall here two known results that allow us to accomplish this. We start with the following bounds on the expectation of the supremum of a finitely-indexed empirical process, which we will later use to bound the suprema of the processes $(G_n(s))_{s\in\mathcal{S}}$ and $(\Delta_n(t,t_*))_{t\in\mathcal{T}\setminus\mathcal{T}_*}$ appearing in Theorem 4. A proof can be found in Appendix H.

Lemma 2. *Let* $n, d \in \mathbb{N}$*, and let* Z *be a random element taking value in a set* $\mathcal{Z}$*, and let* $(Z_i)_{i=1}^n$ *be i.i.d. samples with the same distribution as* Z*. Let* $\mathcal{F}$ *be a finite collection of* $\mathbb{R}^d$*-valued measurable functions. Define*

$$\sigma^2(\mathcal{F}) := \max_{f\in\mathcal{F}} \mathrm{E}\big[\|f(Z) - \mathrm{E}[f(Z)]\|_2^2\big], \qquad r(\mathcal{F}) := \mathrm{E}\left[\max_{(i,f)\in[n]\times\mathcal{F}} \|f(Z_i) - \mathrm{E}[f(Z)]\|_2^2\right]^{1/2},$$

and let $E_n(f) := \sqrt{n}\cdot(n^{-1}\sum_{i=1}^n f(Z_i) - \mathrm{E}[f(Z)])$*. Then, we have*

$$\frac{1}{2}\cdot\sigma(\mathcal{F}) + \frac{1}{4}\cdot\frac{r(\mathcal{F})}{\sqrt{n}} \le \mathrm{E}\left[\max_{f\in\mathcal{F}}\|E_n(f)\|_2^2\right]^{1/2} \le C(|\mathcal{F}|)\cdot\sigma(\mathcal{F}) + C^2(|\mathcal{F}|)\cdot\frac{r(\mathcal{F})}{\sqrt{n}},$$

where $C(x) := 5\sqrt{1+\log x}$*.*

Lemma 2 allows us to compute the expected supremum of a finitely-indexed empirical process, up to log factors in the size of the index set. It is known that these factors cannot be removed from the upper bound nor added to the lower bound without more assumptions, we refer the reader to a related discussion in [Tro16]. Finally, while the term $r(\mathcal{F})$ might grow with n, by bounding the maximum with the sum, it grows at most as $\sqrt{n}$. In many applications however, the random vectors $f(Z)$ are bounded almost surely, so that $r(\mathcal{F})$ is of order one, which justifies our presentation choice.

The second result we recall is the expectation version of a one sided Matrix Bernstein inequality due to Tropp [Tro15]. We use it below to bound the supremum of the process $(\Lambda_n(t))_{t\in\mathcal{T}}$ appearing in Theorem 4. We do not known of a matching non-asymptotic lower bound, but upper and lower bounds similar to those of Lemma 2 hold if instead one considers the expected operator norm instead of only the maximum eigenvalue [Tro16, Section 7].

Lemma 3 ([Tro15], Theorem 6.6.1.)**.** *Let* $n, d \in \mathbb{N}$ *and for each* $i \in [n]$*, let* $Z_i \in \mathbb{R}^{d\times d}$ *be i.i.d. positive semi-definite matrices with the same distribution as* Z*. Define*

$$V := \mathrm{E}\Big[(\mathrm{E}[Z] - Z)^2\Big], \quad W_n := \sqrt{n}\cdot\lambda_{\max}\Big(\mathrm{E}[Z] - \frac{1}{n}\sum_{i=1}^n Z_i\Big).$$

Then, we have

$$\mathrm{E}[\lambda_{\max}(W_n)] \le \sqrt{2\lambda_{\max}(V)\log(ed)} + \frac{\lambda_{\max}(\mathrm{E}[Z])\log(ed)}{3\sqrt{n}}.$$

Equipped with these estimates, we may now control the expected suprema of the empirical processes appearing in Theorem 4. To apply Lemma 2, define the following classes, for $\mathcal{S} \subset \mathcal{T}$ and $t_* \in \mathcal{T}_*$

$$\mathcal{G}(\mathcal{S}) := \left\{(x,y) \mapsto \frac{\Sigma^{-1/2}(s)g(s,(x,y))}{\mathrm{E}[\|g(s,(X,Y))\|_{\Sigma^{-1}(s)}^2]^{1/2}} \,\middle|\, s\in\mathcal{S}\right\},$$

$$\mathcal{D}(t_*) := \left\{(x,y) \mapsto \frac{\ell(\langle w_*(t),\phi_t(x)\rangle, y) - \ell(\langle w_*(t_*),\phi_{t_*}(x)\rangle, y)}{R(t,w_*(t)) - R_*} \,\middle|\, t\in\mathcal{T}\setminus\mathcal{T}_*\right\}.$$

Applying Lemma 2 on $\mathcal{G}(S)$ bounds the expected supremum of the empirical process $(G_n(s))_{s\in\mathcal{S}}$ while applying it on $\mathcal{D}(t_*)$ bounds that of $(\Delta_n(t,t_*))_{t\in\mathcal{T}\setminus\mathcal{T}_*}$. To control the supremum of $(\Lambda_n(t))_{t\in\mathcal{T}}$, the key idea is to notice that it can be expressed as the maximum eigenvalue of a block diagonal matrix whose blocks are $\sqrt{n}(I - \Sigma^{-1/2}(t)\Sigma_n(t)\Sigma^{-1/2}(t))$. Looking at Lemma 3, the relevant parameter is therefore a block diagonal matrix V with the following blocks

$$V(t) := \mathrm{E}\left[\Big(\Sigma^{-1/2}(t)\phi_t(X)\phi_t(X)^T\Sigma^{-1/2}(t) - I\Big)^2\right].$$

As the bound depends only on the maximum eigenvalue of V, the ordering of the blocks does not matter. Putting together these estimates, we arrive at our final result of the paper: a fully explicit version of Theorem 4 for the special case when $\mathcal{T}$ is finite.

Corollary 3. *Assume for all* $(t, j) \in \mathcal{T} \times [d]$, $\mathrm{E}\big[\phi^4_{t,j}(X)\big] < \infty$, $\mathrm{E}\big[Y^4\big] < \infty$, *and let* $\delta \in (0, 1)$. *Let* $k \in [1 + |\mathcal{T} \setminus \mathcal{T}_*|]$ *and* $C(\cdot)$ *as in Lemma 2. If*

$$n \geq (512\lambda_{\max}(V) + 6)\log(ed|\mathcal{T}|) + (128L + 11)\log(4/\delta) \\ + \min_{t_* \in \mathcal{T}_*}\left\{\frac{16C(|\mathcal{T}|)\sigma^2(\mathcal{D}(t_*))}{\delta} + \frac{8C^2(|\mathcal{T}|)r(\mathcal{D}(t_*))}{\delta^{1/2}}\right\},$$

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

$$\hat{t}_n \in \widetilde{F}^k_{n,\delta/2k}(\mathcal{T}),$$

and

$$R(\hat{t}_n, \hat{w}_n) - R_* \leq 4 \cdot A(\widetilde{F}^k_{n,\delta/2k}(\mathcal{T})) \cdot \frac{2k \cdot \mathrm{E}\Big[\|g(\hat{t}_n, (X, Y))\|^2_{\Sigma^{-1}(\hat{t}_n)}\Big]}{\delta n},$$

where the last expectation is with respect to (X, Y) *only, independent of* $(X_i, Y_i)^n_{i=1}$ *and hence of* $\hat{t}_n$*, and where, for* $\mathcal{S} \subset \mathcal{T}$,

$$A(\mathcal{S}) := C^2(\mathcal{S}) \cdot \left(1 + C(\mathcal{S}) \cdot \tfrac{r(\mathcal{G}(\mathcal{S}))}{\sqrt{n}}\right)^2.$$

Finally, $\widetilde{F}_{n,\delta}(\mathcal{S})$ *is the same as* $F_{n,\delta}(\mathcal{S})$ *defined in (7) but with* $A(\mathcal{S})$ *replacing* $\mathrm{E}[\sup_{s \in \mathcal{S}} \overline{G}^2_n(s)]$.

We close this section with a few remarks about Corollary 3. First, while the quantities $r(\mathcal{D}(t_*))$ and $r(\mathcal{G}(\mathcal{S}))$ depend on n, we reiterate that this is not problematic as was explained after Lemma 2. Second, the constant $A(\mathcal{S})$ controlling the contraction rate of the map $\widetilde{F}_{n,\delta}$, as well as the excess risk, has a pleasantly simple form: to first order, it is equal to $\log|\mathcal{S}|$. As the sets $\widetilde{F}^k_{n,\delta/2k}(\mathcal{T})$ are shrinking with n, this transparently shows the decaying effect of the global complexity of $\mathcal{T}$ on the excess risk. Finally, we note that the restriction on k in Corollary 3 is there only because after at most that many iterations, a fixed point is reached, so additional iterations are only harmful for the bound.

5 Conclusion

Broadly speaking, there are two main conclusions one can draw from this work. Firstly, in the large sample regime and under mild assumptions, asking a model to additionally pick a feature map on top of learning a linear predictor has a negligible effect on the excess risk of ERM on regression problems with square loss. Secondly, for moderate sample sizes, the magnitude of this effect depends on the size of the sublevel sets of the function $t \mapsto R(t, w_*(t))$. Plainly, learning feature maps is easy when only a small subset of them is good, as the bad ones can be quickly discarded.

The most tantalizing aspect of our results is their potential in explaining the experiments in [Zha+21]. It was shown there that complex neural networks trained by ERM were able to achieve good performance despite being expressive enough to fit random labels. This is paradoxical if one assumes that the performance of ERM is driven by the complexity of the model class. Our results refute this assumption for feature-learning-based models, of which neural networks are an example. While there are many works offering explanations for this (see e.g. [BMR21] for a survey), we are not aware of one that shows the vanishing influence of the size of the model class on the excess risk as Theorems 3 and 4 show. Formally connecting our statements to these experiments is beyond what we achieved here, yet, we believe that the new perspective we took might generate useful insights in this area.

We conclude by outlining a few limitations of our work. Firstly, we do not deal with the question of how to solve the ERM problem. Our focus is on understanding its statistical performance, and our setting is so general that such a question cannot be meaningfully tackled. Continuing on this last point, while the generality of our results is desirable in some aspects, it is detrimental in others. As an example, it would be desirable to specialize our results from Section 3 to specific collections of feature maps used in practice. Let us also mention that it is a priori unclear whether ERM is an optimal procedure, in a minimax sense, for the model classes we consider; we suspect that recently developed tools might be relevant [Mou22] to address this. Finally, while we focused on the case of regression with square loss, this was mostly done to simplify the presentation. Indeed, the only property of the loss used in the proofs is the exactness of its second order Taylor expansion. This is however not necessary if one can control the error term from above and below. It is known how to do this for many loss functions [e.g. OB21; EE23], and most importantly for logistic regression [Bac10; Bac14]. We have purposefully selected generic notation to make translating such arguments easier.

References

[AC11] J.-Y. Audibert and O. Catoni. “Robust linear least squares regression”. In: (2011).

[Aud07] J.-y. Audibert. “Progressive Mixture Rules Are Deviation Suboptimal”. In: *Advances in Neural Information Processing Systems*. 2007. URL: https://papers.nips.cc/paper_files/paper/2007/hash/ef575e8837d065a1683c022d2077d342-Abstract.html.

[Ba+22] J. Ba, M. A. Erdogdu, T. Suzuki, Z. Wang, D. Wu, and G. Yang. “High-dimensional asymptotics of feature learning: How one gradient step improves the representation”. In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 37932–37946.

[Bac10] F. Bach. “Self-concordant analysis for logistic regression”. In: (2010).

[Bac14] F. Bach. “Adaptivity of averaged stochastic gradient descent to local strong convexity for logistic regression”. In: *The Journal of Machine Learning Research* 15.1 (2014), pp. 595–627.

[BBM05] P. L. Bartlett, O. Bousquet, and S. Mendelson. “Local rademacher complexities”. In: (2005).

[BLM00] S. Boucheron, G. Lugosi, and P. Massart. “A sharp concentration inequality with applications”. In: *Random Structures & Algorithms* 16.3 (2000), pp. 277–292.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN: 9780199535255. URL: https://books.google.ca/books?id=5oo4YIz6tR0C.

[BM06] P. L. Bartlett and S. Mendelson. “Empirical minimization”. In: *Probability theory and related fields* 135.3 (2006), pp. 311–334.

[BMR21] P. L. Bartlett, A. Montanari, and A. Rakhlin. “Deep learning: a statistical viewpoint”. In: *Acta numerica* 30 (2021), pp. 87–201.

[Bou02] O. Bousquet. “A Bennett concentration inequality and its application to suprema of empirical processes”. In: *Comptes Rendus Mathematique* 334.6 (2002), pp. 495–500.

[COB19] L. Chizat, E. Oyallon, and F. Bach. “On lazy training in differentiable programming”. In: *Advances in neural information processing systems* 32 (2019).

[DG12] V. De la Pena and E. Giné. *Decoupling: from dependence to independence*. Springer Science & Business Media, 2012.

[EE23] A. El Hanchi and M. A. Erdogdu. “Optimal Excess Risk Bounds for Empirical Risk Minimization on p-Norm Linear Regression”. In: *Advances in Neural Information Processing Systems* 36 (2023).

[GA11] M. Gönen and E. Alpaydın. “Multiple kernel learning algorithms”. In: *The Journal of Machine Learning Research* 12 (2011), pp. 2211–2268.

[Gho+19] B. Ghorbani, S. Mei, T. Misiakiewicz, and A. Montanari. “Limitations of lazy training of two-layers neural network”. In: *Advances in Neural Information Processing Systems* 32 (2019).

[GN21] E. Giné and R. Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

[He+16] K. He, X. Zhang, S. Ren, and J. Sun. “Deep residual learning for image recognition”. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2016, pp. 770–778.

[HKZ12] D. Hsu, S. M. Kakade, and T. Zhang. “Random design analysis of ridge regression”. In: *Conference on learning theory*. JMLR Workshop and Conference Proceedings. 2012, pp. 9–1.

[JRT08] A. Juditsky, P. Rigollet, and A. B. Tsybakov. “Learning by Mirror Averaging”. In: *The Annals of Statistics* (Oct. 2008). DOI: 10.1214/07-AOS546.

[KM15] V. Koltchinskii and S. Mendelson. “Bounding the smallest singular value of a random matrix without concentration”. In: *International Mathematics Research Notices* 2015.23 (2015), pp. 12991–13008.

[Kol06] V. Koltchinskii. “Local Rademacher complexities and oracle inequalities in risk minimization”. In: (2006).

[Kol11] V. Koltchinskii. *Oracle inequalities in empirical risk minimization and sparse recovery problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*. Vol. 2033. Springer Science & Business Media, 2011.

[KRV22] V. Kanade, P. Rebeschini, and T. Vaskevicius. "Exponential tail local rademacher complexity risk bounds without the bernstein condition". In: *arXiv preprint arXiv:2202.11461* (2022).

[KSH12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25 (2012).

[Lan+04] G. R. Lanckriet, N. Cristianini, P. Bartlett, L. E. Ghaoui, and M. I. Jordan. "Learning the kernel matrix with semidefinite programming". In: *Journal of Machine learning research* 5.Jan (2004), pp. 27–72.

[LBH15] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". In: *nature* 521.7553 (2015), pp. 436–444.

[LC06] E. L. Lehmann and G. Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

[LL20] G. Lecué and M. Lerasle. "Robust machine learning by median-of-means: theory and practice". In: (2020).

[LM09] G. Lecué and S. Mendelson. "Aggregation via Empirical Risk Minimization". en. In: *Probability Theory and Related Fields* (Nov. 1, 2009). DOI: 10.1007/s00440-008-0180-8.

[LM16a] G. Lecué and S. Mendelson. *Learning Subgaussian Classes : Upper and Minimax Bounds*. Sept. 17, 2016. DOI: 10.48550/arXiv.1305.4825.

[LM16b] G. Lecué and S. Mendelson. "Performance of Empirical Risk Minimization in Linear Aggregation". In: *Bernoulli* (Aug. 2016). DOI: 10.3150/15-BEJ701.

[LM16c] G. Lecué and S. Mendelson. "Performance of empirical risk minimization in linear aggregation". In: (2016).

[LM19a] G. Lugosi and S. Mendelson. "Risk minimization by median-of-means tournaments". In: *Journal of the European Mathematical Society* 22.3 (2019), pp. 925–965.

[LM19b] G. Lugosi and S. Mendelson. "Mean estimation and regression under heavy-tailed distributions: A survey". In: *Foundations of Computational Mathematics* 19.5 (2019), pp. 1145–1190.

[LRS15a] T. Liang, A. Rakhlin, and K. Sridharan. "Learning with Square Loss: Localization through Offset Rademacher Complexity". en. In: *Proceedings of The 28th Conference on Learning Theory*. June 26, 2015. URL: https://proceedings.mlr.press/v40/Liang15.html.

[LRS15b] T. Liang, A. Rakhlin, and K. Sridharan. "Learning with square loss: Localization through offset rademacher complexity". In: *Conference on Learning Theory*. PMLR. 2015, pp. 1260–1285.

[Men14] S. Mendelson. "Learning without Concentration". en. In: *Proceedings of The 27th Conference on Learning Theory*. May 29, 2014. URL: https://proceedings.mlr.press/v35/mendelson14.html.

[Mou22] J. Mourtada. "Exact Minimax Risk for Linear Least Squares, and the Lower Tail of Sample Covariance Matrices". In: *The Annals of Statistics* (Aug. 2022). DOI: 10.1214/22-AOS2181.

[OB21] D. M. Ostrovskii and F. Bach. "Finite-sample analysis of m-estimators using self-concordance". In: (2021).

[Oli16] R. I. Oliveira. "The lower tail of random quadratic forms with applications to ordinary least squares". In: *Probability Theory and Related Fields* 166 (2016), pp. 1175–1194.

[Sau18] A. Saumard. "On optimality of empirical risk minimization in linear aggregation". In: (2018).

[SD16] A. Sinha and J. C. Duchi. "Learning kernels with random features". In: *Advances in neural information processing systems* 29 (2016).

[Tal96] M. Talagrand. "New concentration inequalities in product spaces". In: *Inventiones mathematicae* 126.3 (1996), pp. 505–563.

[Tro15] J. A. Tropp. “An introduction to matrix concentration inequalities”. In: *Foundations and Trends in Machine Learning* 8.1-2 (2015), pp. 1–230.

[Tro16] J. A. Tropp. “The expected norm of a sum of independent random matrices: An elementary approach”. In: *High Dimensional Probability VII: The Cargese Volume*. Springer. 2016, pp. 173–202.

[Van00] A. W. Van der Vaart. *Asymptotic statistics*. Vol. 3. Cambridge university press, 2000.

[Van14] R. Van Handel. “Probability in high dimension”. In: *Lecture Notes (Princeton University)* (2014).

[Vas+17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. “Attention is all you need”. In: *Advances in neural information processing systems* 30 (2017).

[VC74] V. Vapnik and A. Chervonenkis. “Theory of pattern recognition”. In: (1974).

[VW96] A. W. Van Der Vaart and J. A. Wellner. *Weak convergence and empirical processes*. Springer, 1996.

[Wai19] M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*. Vol. 48. Cambridge university press, 2019.

[Whi82] H. White. “Maximum likelihood estimation of misspecified models”. In: *Econometrica: Journal of the econometric society* (1982), pp. 1–25.

[Zha+21] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. “Understanding deep learning (still) requires rethinking generalization”. In: *Communications of the ACM* 64.3 (2021), pp. 107–115.

A Proof of Theorem 1

Let A_n denote the event that Σ_n is invertible. By the weak law of large numbers, Σ_n converges to Σ in probability so that $\lim_{n\to\infty} \mathrm{P}(A_n^c) = 0$. Now on the event A_n, we have by (2)

$$\sqrt{n} \cdot (\hat{w}_n - w_*) = \Sigma_n^{-1} \cdot (\sqrt{n} \cdot \nabla R_n(w_*)).$$

By the continuous mapping theorem, Σ_n^{-1} converges to Σ^{-1} in probability and by the central limit theorem

$$\sqrt{n} \cdot \nabla R_n(w_*) \xrightarrow{d} \mathcal{N}(0, G)$$

Therefore, by Slutsky's theorem

$$\sqrt{n} \cdot (\hat{w}_n - w_*) \xrightarrow{d} \mathcal{N}(0, \Sigma^{-1} G \Sigma^{-1})$$

Now since the risk is quadratic and the gradient vanishes at w_*,

$$n \cdot [R(\hat{w}) - R(w_*)] = \frac{1}{2} \cdot \|\sqrt{n} \cdot (\hat{w} - w_*)\|_\Sigma^2 \xrightarrow{d} \frac{1}{2}\|Z\|_2^2$$

where Z is as in the theorem, and where the last statement follows by the continuous mapping theorem. This proves the first statement in the theorem. The bounds on the quantiles are a simple consequence of concentration bounds for the norm of Gaussian vectors, we omit the proof here.

B Proof of Theorem 2

Denote by A_n the event that

$$\lambda_{\text{min}}\left(\Sigma^{-1/2}\Sigma_n\Sigma^{-1/2}\right) \geq \frac{1}{2}.$$

We show that under the sample size restriction, $\mathrm{P}(A_n) \geq 1 - \delta/2$. Indeed we have the variational representation

$$\lambda_{\text{max}}(I - \Sigma^{-1/2}\Sigma_n\Sigma^{-1/2}) = \sup_{v \in S^{d-1}} \frac{1}{n}\sum_{i=1}^n 1 - \langle v, \Sigma^{-1/2}\phi(X_i)\rangle^2$$

Each element in the sum is upper bounded by 1, and the variance parameter in Bousquet's inequality [Bou02] is given by the parameter L in the statement of the theorem. Applying this inequality yields that with probability at least $1 - \delta/2$

$$\lambda_{\text{max}}(I - \Sigma^{-1/2}\Sigma_n\Sigma^{-1/2}) \leq 2\,\mathrm{E}\Big[\lambda_{\text{max}}(I - \Sigma^{-1/2}\Sigma_n\Sigma^{-1/2})\Big] + \sqrt{\frac{2L\log(2/\delta)}{n}} + \frac{4\log(2/\delta)}{3n}$$

Using Lemma 3 to upper bound the above expectation, and replacing the sample size n in the resulting inequality with the minimal allowed by the theorem proves that $\mathrm{P}(A_n) \geq 1 - \delta/2$ for all sample sizes allowable by the theorem. Now on this event we have, using (3),

$$R(\hat{w}_n) - R(w_*) = \frac{1}{2} \cdot \|\Sigma_n^{-1}\nabla R_n(w_*)\|_\Sigma^2 \leq 2 \cdot \|\nabla R_n(w_*)\|_{\Sigma^{-1}}^2$$

Now by an elementary calculation

$$\mathrm{E}\big[\|\nabla R_n(w_*)\|_{\Sigma^{-1}}^2\big] = n^{-1}\,\mathrm{E}\big[\|g(X,Y)\|_{\Sigma^{-1}}^2\big]$$

so that an application of Markov's inequality yields that there is an event B_n that holds with probability $1 - \delta/2$ and on which

$$\|\nabla R_n(w_*)\|_{\Sigma^{-1}}^2 \leq 2 \cdot (n\delta)^{-1} \cdot \mathrm{E}\big[\|g(X,Y)\|_{\Sigma^{-1}}^2\big]$$

The union bound $\mathrm{P}(A_n \cap B_n) = 1 - \mathrm{P}(A_n \cup B_n) \geq 1 - \delta$ finishes the proof.

C Main Lemma

We state here a core lemma, which we use in many of our proofs. To state it, we define, for a function $F : \mathcal{S} \to \mathbb{R}$ on a subset $\mathcal{S} \subseteq \mathcal{T}$,

$$\|F\|_\infty := \sup_{s \in \mathcal{S}} |F(s)|, \quad \|F\|_{\infty,-} := \sup_{s \in \mathcal{S}} \{-F(s)\}, \quad \|F\|_{\infty,+} := \sup_{s \in \mathcal{S}} F(s),$$

where the first quantity is the ℓ^∞ norm of the function F, and the remaining are one-sided versions of it. The processes appearing in the next statement are defined in (5) and (6).

Lemma 4. *Assume that $\mathcal{T}_* \neq \varnothing$ and let $t_* \in \mathcal{T}_*$. On the event that $\|n^{-1/2}\Delta_n(\cdot, t_*)\|_{\infty,+} < 1$ and $\|n^{-1/2}\Lambda_n\|_{\infty,+} < 1$, we have*

$$R(\hat{t}_n, w_*(\hat{t}_n)) - R_* \leq \frac{1}{2} \cdot \frac{1}{1 - \|n^{-1/2}\Delta_n(\cdot, t_*)\|_{\infty,+}} \cdot \frac{1}{1 - \|n^{-1/2}\Lambda_n\|_{\infty,+}} \cdot (n^{-1}G_n^2(\hat{t}_n)),$$

and

$$R(\hat{t}_n, \hat{w}_n) - R(\hat{t}_n, w_*(\hat{t}_n)) \leq \frac{1}{2} \cdot \frac{1}{(1 - \|n^{-1/2}\Lambda_n\|_{\infty,+})^2}(n^{-1}G_n^2(\hat{t}_n)).$$

Proof. To lighten the notation, we drop the dependence on n, and write $\hat{t}$ instead of $\hat{t}_n$. We start with the first statement. First, we note that if $\hat{t} \in \mathcal{T}_*$, then the statement holds trivially as the left-hand side is zero, so we only consider the other case in what follows. For any $t \in \mathcal{T}$, define

$$\hat{w}(t) \in \operatorname*{argmin}_{w \in \mathbb{R}^d} R_n(t, w),$$

where the choice of minimizer is arbitrary. With this definition, we have $\hat{w}_n = \hat{w}(\hat{t})$. Now, by definition of ERM,

$$R_n(\hat{t}, \hat{w}(\hat{t})) - R_n(t_*, w_*(t_*)) \leq 0. \tag{9}$$

On the other hand, for any $t \in \mathcal{T} \setminus \mathcal{T}_*$, we have the decomposition

$$R_n(t, \hat{w}(t)) - R_n(t_*, w_*) = [R_n(t, \hat{w}(t)) - R_n(t, w_*(t))] + [R_n(t, w_*(t)) - R_n(t_*, w_*(t_*))]. \tag{10}$$

We study each of the terms of (10) separately, and we start with the first. Note that since we are in the event

$$\inf_{t \in \mathcal{T}} \lambda_{\text{min}}(\Sigma^{-1/2}(t)\Sigma_n(t)\Sigma^{-1/2}(t)) = 1 - \|n^{-1/2}\Lambda_n\|_{\infty,+} > 0,$$

the sample covariance matrices $\Sigma_n(t)$ are invertible for all $t \in \mathcal{T}$, so that $\hat{w}(t)$ is uniquely defined and satisfies

$$\hat{w}(t) = w_*(t) - \Sigma_n^{-1}(t)\nabla_w R_n(t, w_*(t)). \tag{11}$$

Furthermore, since the function $w \mapsto R_n(t, w)$ is quadratic in w and its gradient vanishes at its minimizer $\hat{w}(t)$, we have

$$R_n(t, \hat{w}(t)) - R_n(t, w_*(t)) = -\frac{1}{2}\|\hat{w}(t) - w_*(t)\|^2_{\Sigma_n(t)} = -\frac{1}{2}\|\nabla_w R_n(t, w_*(t))\|^2_{\Sigma_n^{-1}(t)}, \tag{12}$$

where the last equality follows from (11). To bound this last term, define

$$\widetilde{\Sigma}_n(t) := \Sigma^{-1/2}(t)\Sigma_n(t)\Sigma^{-1/2}(t).$$

Then we have,

$$\begin{aligned}
\|\nabla_w R_n(t, w_*(t))\|^2_{\Sigma_n^{-1}(t)} &= \left\{\Sigma^{-1/2}(t)\nabla_w R_n(t, w_*(t))\right\}^T \widetilde{\Sigma}_n^{-1}(t)\left\{\Sigma^{-1/2}(t)\nabla_w R_n(t, w_*(t))\right\} \\
&\leq \lambda_{\text{max}}(\widetilde{\Sigma}_n^{-1}(t)) \cdot \|\nabla_w R_n(t, w_*(t))\|^2_{\Sigma^{-1}(t)} \\
&= \frac{1}{1 - \lambda_{\text{max}}(I - \widetilde{\Sigma}_n(t))} \cdot (n^{-1}G_n^2(t)) \\
&\leq \frac{1}{1 - \|n^{-1/2}\Lambda_n\|_{\infty,+}} \cdot (n^{-1}G_n^2(t)).
\end{aligned} \tag{13}$$

Finally, the second term of (10) is lower bounded by

$$\begin{aligned}R_n(t, w_*(t)) - R_n(t_*, w_*(t_*))) &= (1 - n^{-1/2}\Delta_n(t, t_*))[R(t, w_*(t)) - R_*]\\ &\geq (1 - \|n^{-1/2}\Delta_n(\cdot, t_*)\|_{\infty,+})[R(t, w_*(t)) - R_*] \qquad (14)\end{aligned}$$

Combining (13) and (12) lower bounds the first term of (10), while (14) lower bounds the second. Combining the resulting lower bound on (10) with (9) and rearranging yields the first statement.

For the second statement, not that for all $t \in \mathcal{T}$,

$$\begin{aligned}R(t, \hat{w}(t)) - R(t, w_*(t)) &= \frac{1}{2} \cdot \|\hat{w}(t) - w_*(t)\|^2_{\Sigma(t)}\\ &= \frac{1}{2} \cdot \|\Sigma_n^{-1}(t)\nabla_w R_n(t, w_*(t))\|^2_{\Sigma(t)}\\ &\leq \frac{1}{2} \cdot \lambda_{\max}(\widetilde{\Sigma}_n^{-2}(t)) \cdot \|\nabla_w R_n(t, w_*(t))\|^2_{\Sigma^{-1}(t)}\\ &= \frac{1}{2} \cdot \frac{1}{(1 - \|n^{-1/2}\Lambda_n\|_{\infty,+})^2} \cdot (n^{-1}G_n^2(t)).\end{aligned}$$

where the second line follows from (11), and in particular the inequality holds for $\hat{t}$. □

D Proof of Theorem 3

Consistency of $\hat{t}_n$. We want to show that, as $n \to \infty$,

$$R(\hat{t}_n, w_*(\hat{t}_n)) - R_* \xrightarrow{p} 0. \qquad (15)$$

Using the notation introduced in Appendix C, the Glivenko-Cantelli assumptions in Theorem 3 amount to the statements that, for some $t_* \in \mathcal{T}_*$, and as $n \to \infty$,

$$\|n^{-1/2}\Lambda_n\|_\infty \xrightarrow{p} 0, \quad \|n^{-1/2}\Delta_n(\cdot, t_*)\|_\infty \xrightarrow{p} 0, \quad \|n^{-1/2}G_n\|_\infty \xrightarrow{p} 0. \qquad (16)$$

Let A_n denote the event that both $\|n^{-1/2}\Lambda_n\|_\infty < 1$ and $\|n^{-1/2}\Delta_n(\cdot, t_*)\|_\infty < 1$. The union bound and (16) show that

$$\lim_{n\to\infty} \mathrm{P}(A_n^c) = 0 \qquad (17)$$

Furthermore, on the event A_n, the first bound of Lemma 4 holds, and bounding $n^{-1}G_n^2(\hat{t}_n)$ by $\|n^{-1/2}G_n\|^2_\infty$ yields that on A_n

$$R(\hat{t}_n, w_*(\hat{t}_n)) - R_* \leq \frac{1}{2} \cdot \frac{1}{1 - \|n^{-1/2}\Delta_n(\cdot, t_*)\|_\infty} \cdot \frac{1}{1 - \|n^{-1/2}\Lambda_n\|_\infty} \cdot \|n^{-1/2}G_n\|^2_\infty \qquad (18)$$

Now let $\varepsilon > 0$, and denote by $B_n(\varepsilon)$ the event that the right hand side of (18) is strictly larger than ε. Then the statements (16) together with the continuous mapping theorem show that $\lim_{n\to\infty} \mathrm{P}(B_n(\varepsilon)) = 0$. Therefore, again by (18), we have

$$\mathrm{P}\big(R(\hat{t}_n, w_*(\hat{t}_n)) - R_* > \varepsilon\big) \leq \mathrm{P}(B_n(\varepsilon)) + \mathrm{P}(A_n^c),$$

and taking $n \to \infty$ proves (15).

Asymptotic quantiles. We derive a few intermediate statements from which we deduce the bound on the quantiles of the excess risk appearing in the statement of the theorem.

We start with the simple decomposition

$$n \cdot \big[R(\hat{t}_n, \hat{w}_n) - R_*\big] = n\big[R(\hat{t}_n, \hat{w}_n) - R(\hat{t}_n, w_*(\hat{t}_n))\big] + n\big[R(\hat{t}_n, w_*(\hat{t}_n)) - R_*\big]. \qquad (19)$$

Now on the event A_n defined above, we have, by an application of Lemma 4, combining the two bounds in the lemma along with (19), that the rescaled excess risk (19) is upper bounded by

$$\frac{1}{2} \cdot \frac{1}{1 - \|n^{-1/2}\Lambda_n\|_\infty} \cdot \left(\frac{1}{1 - \|n^{-1/2}\Delta_n(\cdot, t_*)\|_\infty} + \frac{1}{1 - \|n^{-1/2}\Lambda_n\|_\infty}\right) \cdot G_n^2(\hat{t}_n). \qquad (20)$$

From the Glivenko-Cantelli assumptions (16), the first three factors converge in probability to 1. Our aim will be to derive a bound on the upper tail of the last factor, which will imply a bound on the upper tail of the rescaled excess risk.

We briefly make explicit the Donsker assumption before deriving this bound. Both define and note that

$$G_n(t,v) := \langle v, \Sigma^{-1/2}(t) \nabla_w R_n(t, w_*(t)) \rangle, \qquad G_n(t) = \sup_{v \in S^{d-1}} G_n(t,v),$$

where S^{d-1} is the Euclidean unit sphere in $\mathbb{R}^d$. As pointed out in Section 3, the processes $G_n(t)$ are partial suprema of the empirical processes $G_n(t,v)$. The Donsker assumption of the theorem states that the empirical processes $G_n(t,v)$ take value in the space of bounded functions on $\mathcal{T} \times S^{d-1}$, equipped with the $\ell^\infty(\mathcal{T} \times S^{d-1})$ norm and the metric it induces, and converge weakly to their unique Gaussian limit $(Z(t,v))_{(t,v)\in \mathcal{T}\times S^{d-1}}$ as $n \to \infty$. We define the partial supremum with respect to $v \in S^{d-1}$ of the latter by $Z(t) := \sup_{v \in S^{d-1}} Z(t,v)$ in analogy with the definition of $G_n(t)$.

We now bound the upper tail of $G_n^2(\hat{t}_n)$ in (20). Let $(\varepsilon_k)_{k=1}^\infty$ be a decreasing sequence of positive numbers such that $\varepsilon_k \to 0$ as $k \to \infty$, and define the sets

$$\mathcal{T}_*(\varepsilon) := \{ t \in \mathcal{T} \mid R(t, w_*(t)) - R_* \le \varepsilon \} \tag{21}$$

as well as the function $F_k : \ell^\infty(\mathcal{T} \times S^{d-1}) \to \mathbb{R}$ by

$$F_k(z) := \sup_{s \in \mathcal{T}_*(\varepsilon_k)} \sup_{v \in S^{d-1}} z(s,v).$$

Note on the one hand that $\cap_{k\ge 1} \mathcal{T}_*(\varepsilon_k) = \mathcal{T}_*$, and on the other that F_k is continuous for all $k \in \mathbb{N}$, and in fact Lipschitz. Indeed, let $z, z' \in \ell^\infty(\mathcal{T} \times S^{d-1})$. Then

$$|F_k(z) - F_k(z')| = \left| \sup_{s \in \mathcal{T}_*(\varepsilon_k)} \sup_{v \in \mathbb{S}^{d-1}} z(s,v) - \sup_{s \in \mathcal{T}_*(\varepsilon_k)} \sup_{v \in \mathbb{S}^{d-1}} z'(s,v) \right| \le \| z - z' \|_\infty$$

Now let $k \in \mathbb{N}$ and $x \in [0, \infty)$. Then

$$\begin{aligned}
\mathrm{P}\big(G_n^2(\hat{t}_n) > x\big) &= \mathrm{P}\big(\big\{G_n^2(\hat{t}_n) > x\big\} \cap \big\{\hat{t}_n \in \mathcal{T}_*(\varepsilon_k)\big\}\big) + \mathrm{P}\big(\big\{G_n^2(\hat{t}_n) > x\big\} \cap \big\{\hat{t}_n \notin \mathcal{T}_*(\varepsilon_k)\big\}\big) \\
&\le \mathrm{P}\big(\big\{G_n^2(\hat{t}_n) > x\big\} \cap \big\{\hat{t}_n \in \mathcal{T}_*(\varepsilon_k)\big\}\big) + \mathrm{P}\big(\big\{\hat{t}_n \notin \mathcal{T}_*(\varepsilon_k)\big\}\big) \\
&\le \mathrm{P}\left(\sup_{s \in \mathcal{T}_*(\varepsilon_k)} G_n^2(s) > x \right) + \mathrm{P}\big(R(\hat{t}_n, w_*(\hat{t}_n)) - R_* > \varepsilon_k\big) \\
&= \mathrm{P}\big(F_k^2(G_n) > x\big) + \mathrm{P}\big(R(\hat{t}_n, w_*(\hat{t}_n)) - R_* > \varepsilon_k\big)
\end{aligned}$$

taking the limit as $n \to \infty$, the first term converges, by the continuous mapping theorem, to the probability of the event $\big\{F_k^2(Z) > x\big\}$, where Z is the limiting Gaussian process discussed above, while the second term vanishes by the first part of Theorem 3. Therefore, for all $k \in \mathbb{N}$,

$$\lim_{n\to\infty} \mathrm{P}\big(G_n^2(\hat{t}_n) > x\big) \le \mathrm{P}\left(\sup_{s \in \mathcal{T}_*(\varepsilon_k)} Z^2(s) > x \right)$$

Taking the limit as $k \to \infty$, noticing that the events

$$\left\{ \sup_{s \in \mathcal{T}_*(\varepsilon_k)} Z^2(s) > x \right\}$$

are nested, using the continuity of probability from above, and recalling that $\cap_{k \ge 1} \mathcal{T}_*(\varepsilon_k) = \mathcal{T}_*$ gives

$$\lim_{n\to\infty} \mathrm{P}\big(G_n^2(\hat{t}_n) > x\big) \le \mathrm{P}\left(\sup_{s \in \mathcal{T}_*} Z^2(s) > x \right). \tag{22}$$

Using the fact that the Gaussian process $\big\{Z(s,v) \mid (s,v) \in \mathcal{T}_* \times S^{d-1}\big\}$ is $\sup\big\{\mathrm{Var}[Z(s,v)] \mid (s,v) \in \mathcal{T}_* \times S^{d-1}\big\}$ subgaussian (e.g. [Van14], Lemma 6.12) to bound the tail probability finishes the proof.

E Proof of Corollary 1

For the first statement, by Theorem 3, it is enough to show that

$$\mathrm{E}\Big[\max_{s\in\mathcal{T}_*} Z^2(s)\Big] \le 80\cdot(1+\log|\mathcal{T}_*|)\cdot\max_{s\in\mathcal{T}_*}\mathrm{E}\Big[\|g(s,(X,Y))\|^2_{\Sigma^{-1}(s)}\Big]$$

This follows from the following observation. Since $(Z(s))_{s\in\mathcal{T}_*}$ is a finite-dimensional Gaussian vector, we have

$$\begin{aligned}\mathrm{E}\Big[\max_{s\in\mathcal{T}_*} Z^2(s)\Big] &\le \mathrm{E}\left[\left(\sum_{s\in\mathcal{T}_*} Z^{2p}(s)\right)^{1/p}\right]\\ &\le \left(\sum_{s\in\mathcal{T}_*}\mathrm{E}\big[Z^{2p}(s)\big]\right)^{1/p}\\ &\le 32\cdot p\cdot\left(\sum_{s\in\mathcal{T}_*}\mathrm{E}\big[Z^{2}(s)\big]^p\right)^{1/p}\end{aligned}$$

where the last estimate follows from Gaussian concentration. Taking $p = 1+\log|\mathcal{T}_*|$, and recalling that for $x\in\mathbb{R}^d$, $\|x\|_p \le d^{1/p}\|x\|_\infty$ yields the result.

For the second statement, we first note that the upper bound follows from the proof of Theorem 3. Indeed, the statement (22) reads in our case, where $\mathcal{T}_* = \{t_*\}$.

$$\lim_{n\to\infty}\mathrm{P}\big(n\cdot[R(\hat{w}_n,\hat{t}_n)-R_*]>x\big)\le\mathrm{P}(Z(t_*)>x),$$

where $Z(t_*)\sim\mathcal{N}(0,\Sigma^{-1/2}(t_*)G(t_*)\Sigma^{-1/2}(t_*))$. This proves the upper bound on the quantile. For the lower bound, we develop a new but similar argument to the one in Theorem 3. We have the following lower bound on the rescaled excess risk

$$\begin{aligned}n\cdot[R(\hat{t}_n,\hat{w}_n)-R_*] &\ge n\cdot[R(\hat{t}_n,\hat{w}_n)-R(\hat{t}_n,w_*(\hat{t}_n))]\\ &= n\cdot\frac{1}{2}\cdot\|\Sigma_n^{-1}(\hat{t}_n)\nabla_w R_n(\hat{t}_n,w_*(\hat{t}_n))\|^2_{\Sigma(t)}\\ &\ge n\cdot\frac{1}{2}\cdot\lambda_{\text{min}}(\widetilde{\Sigma}_n^2(\hat{t}_n))\cdot\|\nabla_w R(\hat{t}_n,w_*(\hat{t}_n))\|^2_{\Sigma^{-1}(t)}\\ &\ge\frac{1}{2}\frac{1}{(1+\|n^{-1/2}\Lambda_n\|_{\infty,-})^2}G_n^2(\hat{t}_n).\end{aligned}$$

By the Glivenko-Cantelli assumption on Λ_n, the first factor converges to $1/2$. For the second, we will lower bound its upper tails. Similar to the proof of Theorem 3, we let ε_k be a sequence of positive decreasing constants converging to 0, and define $F_k:\ell^\infty(\mathcal{T}\times S^{d-1})\to\mathbb{R}$ by

$$F_k(z) := \inf_{t\in\mathcal{T}_*(\varepsilon_k)}\sup_{v\in S^{d-1}} z(t,v)$$

where the subsets $\mathcal{T}_*(\varepsilon_k)$ are as defined in (21). Clearly, for $z,z'\in\ell^\infty(\mathcal{T}\times S^{d-1})$,

$$|F_k(z)-F_k(z')|\le\|z-z'\|_\infty$$

so F_k is continuous. Now

$$\begin{aligned}\mathrm{P}\big(G_n^2(\hat{t}_n)>x\big) &\ge \mathrm{P}\big(\{G_n^2(\hat{t}_n)>x\}\cap\{\hat{t}_n\in\mathcal{T}_*(\varepsilon_k)\}\big)\\ &\ge \mathrm{P}\left(\inf_{s\in\mathcal{T}_*(\varepsilon_k)} G_n^2(s)>x\right)-\mathrm{P}\big(\{\hat{t}_n\notin\mathcal{T}_*(\varepsilon_k)\}\big)\\ &= \mathrm{P}\big(F_k^2(G_n)>x\big)-\mathrm{P}\big(R(\hat{t}_n,w_*(\hat{t}_n))-R_*>\varepsilon_k\big)\end{aligned}$$

By the same argument as in Theorem 3, we obtain, as $n\to\infty$, and for all $k\in\mathbb{N}$,

$$\mathrm{P}\big(G_n^2(\hat{t}_n)>x\big)\ge\mathrm{P}\left(\inf_{s\in\mathcal{T}_*(\varepsilon_k)} Z^2(s)>x\right)$$

Taking the limit as $k\to\infty$, and noticing that

$$\bigcup_{k\ge1}\left\{\inf_{s\in\mathcal{T}_*(\varepsilon_k)} Z^2(s)>x\right\}=\{Z^2(t_*)>x\}$$

completes the proof.

F Proof of Lemma 1

We prove the first statement by induction. For $k = 0$, this follows directly from the fact that by definition $F^0_{n,\delta}(\mathcal{T}) = \mathcal{T}$ and $F_{n,\delta}(\mathcal{T}) \subseteq \mathcal{T}$. Now let $k \in \mathbb{N}$ and assume that the statement holds for $k-1$. Let $s \in F^{k+1}_{n,\delta}(\mathcal{T})$. Then by definition

$$\begin{aligned} R(s, w_*(s)) - R_* &\leq 2\,\mathrm{E}[\sup_{s \in F^k_{n,\delta}(\mathcal{T})} \overline{G}^2_n(s)] \cdot \frac{\mathrm{E}\left[\|g(s,(X,Y))\|^2_{\Sigma^{-1}(s)}\right]}{n\delta} \\ &\leq 2\,\mathrm{E}[\sup_{s \in F^{k-1}_{n,\delta}(\mathcal{T})} \overline{G}^2_n(s)] \cdot \frac{\mathrm{E}\left[\|g(s,(X,Y))\|^2_{\Sigma^{-1}(s)}\right]}{n\delta}. \end{aligned}$$

where the second inequality follows from the fact that by the induction hypothesis, $F^k_{n,\delta}(\mathcal{T}) \subset F^{k-1}_{n,\delta}(\mathcal{T})$, and that the supremum is increasing. Therefore $s \in F^k_{n,\delta}(\mathcal{T})$ since the last inequality is the defining inequality for $F^k_{n,\delta}(\mathcal{T})$. We now turn to the second statement. Fix k and δ. On the one hand, $\mathcal{T}_* \subseteq \bigcap_{n \geq 1} F^k_{n,\delta}(\mathcal{T})$. On the other, for any $t \in \bigcap_{n\geq 1} F^k_{n,\delta}(\mathcal{T})$, we have for all $n \geq n_0$, $R(t, w_*(t)) - R_* \leq C(t,\delta) \cdot n^{-1}$. Therefore $R(t, w_*(t)) - R_* = 0$, and $t \in \mathcal{T}_*$.

G Proof of Theorem 4

Recall the notation introduced in Appendix C. Let $A_n(t_*)$ be the event that:

$$\|n^{-1/2}\Lambda_n\|_{\infty,+} \leq 1/2, \qquad \text{and} \qquad \|n^{-1/2}\Delta_n(\cdot, t_*)\| \leq \frac{1}{2}.$$

We start by showing that under the sample size inequality stated in the theorem, there exists a $t_* \in \mathcal{T}_*$ such that $\mathrm{P}(A_n(t_*)) \geq 1 - \delta/2$. Indeed, we have

$$\|n^{-1/2}\Lambda_n\|_{\infty,+} = \sup_{(t,v)\in(\mathcal{T}, S^{d-1})} \frac{1}{n}\sum_{i=1}^n 1 - \langle v, \Sigma^{-1/2}(t)\phi_t(X_i)\rangle^2$$

The elements of this sum are bounded by 1, and the variance parameter of Bousquet's inequality [Bou02] is given by L as defined in Section 3.2. Applying this inequality yields that with probability at least $1 - \delta/4$

$$\|n^{-1/2}\Lambda_n\|_{\infty,+} \leq \frac{2}{n^{1/2}} \cdot \mathrm{E}\left[\sup_{t\in\mathcal{T}} \Lambda_n(t)\right] + \sqrt{\frac{2L\log(4/\delta)}{n}} + \frac{4\log(4/\delta)}{3n}$$

Furthermore, by Markov's inequality, with probability at least $1 - \delta/4$

$$\|n^{-1/2}\Delta(\cdot, t_*)\|_{\infty,+} \leq \frac{4 \cdot \mathrm{E}[\sup_{t \in \mathcal{T}\setminus\mathcal{T}_*} \Delta(t, t_*)]}{n^{1/2}\cdot\delta}$$

Hence, when the inequality on the sample size stated in the theorem holds for some t_*, the event $A_n(t_*)$ holds with probability at least $1-\delta/2$. Now on this event, the first bound of Lemma 4 applies, and we have, replacing G_n by $\overline{G}_n$

$$R(\hat{t}_n, \hat{w}_*(\hat{t}_n)) - R_* \leq 2 \cdot \overline{G}^2_n(\hat{t}_n) \cdot n^{-1} \cdot \mathrm{E}\left[\|g(\hat{t}_n, (X,Y))\|^2_{\Sigma^{-1}(\hat{t}_n)}\right] \tag{23}$$

Now we use the iterative localization method of Koltchinskii [Kol06]. Initially, we have no information about where $\hat{t}_n$ is located aside from belonging to $\mathcal{T}$, so we start with the bound

$$R(\hat{t}_n, \hat{w}_*(\hat{t}_n)) - R_* \leq 2 \cdot \left(\sup_{t\in\mathcal{T}} \overline{G}^2_n(t)\right) \cdot n^{-1} \cdot \mathrm{E}\left[\|g(\hat{t}_n, (X,Y))\|^2_{\Sigma^{-1}(\hat{t}_n)}\right].$$

Using Markov's inequality, we have on an event $B_{n,1}$ which holds with probability at least $1 - \delta/2k$

$$\sup_{t\in\mathcal{T}} \overline{G}^2_n(t) \leq \frac{2 \cdot k \cdot \mathrm{E}[\sup_{t\in\mathcal{T}} \overline{G}^2_n(t)]}{\delta}.$$

Replacing in (23) yields that on the event $A_n(t_*) \cap B_{n,1}$,

$$R(\hat{t}_n, w_*(\hat{t}_n)) - R_* \leq 2 \cdot \mathrm{E}\left[\sup_{t \in \mathcal{T}} \overline{G}_n^2(t)\right] \cdot k \cdot (n\delta)^{-1} \cdot \mathrm{E}\left[\left\|g(\hat{t}_n, (X, Y))\right\|^2_{\Sigma^{-1}(\hat{t}_n)}\right],$$

which shows that on this event, $\hat{t}_n \in F_{n,\delta/2k}(\mathcal{T})$, by definition of the map $F_{n,\delta/2k}$. With this knowledge, we now reuse the bound (23) to obtain that on $A_n(t_*) \cap B_{n,1}$

$$R(\hat{t}_n, w_*(\hat{t}_n)) - R_* \leq 2 \cdot \left(\sup_{t \in F_{n,\delta/2k}(\mathcal{T})} \overline{G}_n^2(t)\right) \cdot n^{-1} \cdot \mathrm{E}\left[\left\|g(\hat{t}_n, (X, Y))\right\|^2_{\Sigma^{-1}(\hat{t}_n)}\right].$$

Iterating the procedure we just described k times, we obtain that on an event $A_n(t_*) \cap (\cap_{j=1}^k B_{n,j})$, where $\mathrm{P}(B_{n,j}) \geq 1 - \delta/2k$ for all $j \in [k]$, we obtain

$$\hat{t}_n \in F^k_{n,\delta/2k} = \mathcal{S}_{n,\delta,k} \tag{24}$$

where $\mathcal{S}_{n,\delta,k}$ was defined in the statement of the theorem, as well as

$$\sup_{t \in \mathcal{S}_{n,\delta,k}} \overline{G}_n^2(t) \leq \frac{2 \cdot k \cdot \mathrm{E}[\sup_{t \in \mathcal{S}_{n,\delta,k}} \overline{G}_n^2(t)]}{\delta}. \tag{25}$$

Since

$$\mathrm{P}(A_n(t_*) \cap (\cap_{j=1}^k B_{n,k})) \geq 1 - \delta/2 - \sum_{j=1}^k \delta/2k = 1 - \delta$$

equation (24) proves the first statement of the theorem. For the second statement, we have on the same event, and combining the two bounds from Lemma 4,

$$R(\hat{t}_n, \hat{w}_n) - R_* \leq 4 \cdot \overline{G}_n^2(\hat{t}_n) \cdot n^{-1} \cdot \mathrm{E}\left[\left\|g(\hat{t}_n, (X, Y))\right\|^2_{\Sigma^{-1}(\hat{t}_n)}\right].$$

Using (24) and (25) proves the second statement.

H Proof of Lemma 2

We prove a slightly more general result, from which Lemma 2 can be immediately deduced.

Lemma 5. *Let $n, d \in \mathbb{N}$ and let $\mathcal{T}$ be a finite set. For each $(i, t) \in [n] \times \mathcal{T}$, let $Z_{i,t} \in \mathbb{R}^d$ be random vectors such that for each $t \in \mathcal{T}$, $(Z_{i,t})_{i=1}^n$ are i.i.d. with the same distribution as Z_t. For all $t \in \mathcal{T}$, assume that $\mathrm{E}[Z_t] = 0$, and define*

$$\sigma^2(\mathcal{T}) := \sup_{t \in \mathcal{T}} \mathrm{E}\left[\|Z_t\|_2^2\right], \qquad r(\mathcal{T}) := \mathrm{E}\left[\sup_{(i,t) \in [n] \times \mathcal{T}} \|Z_{i,t}\|_2^2\right]^{1/2}.$$

Then

$$\frac{1}{2} \cdot \frac{\sigma(\mathcal{T})}{n^{1/2}} + \frac{1}{4} \cdot \frac{r(\mathcal{T})}{n} \leq \mathrm{E}\left[\sup_{t \in \mathcal{T}} \left\|\frac{1}{n}\sum_{i=1}^n Z_{i,t}\right\|_2^2\right]^{1/2} \leq C(\mathcal{T}) \cdot \frac{\sigma(\mathcal{T})}{n^{1/2}} + C^2(\mathcal{T}) \cdot \frac{r(\mathcal{T})}{n},$$

where $C(\mathcal{T}) := 5\sqrt{1 + \log|\mathcal{T}|}$.

To prove Lemma 5, we need to recall a few preliminary results. The first is a classical symmetrization inequality, see e.g. [BLM13, Lemma 11.4] or [Wai19, Proposition 4.11] for a proof.

Lemma 6. *For each $(i, t) \in [n] \times \mathcal{T}$, let $W_{i,t} \in \mathbb{R}^d$ be random vectors such that for each $t \in \mathcal{T}$, $(W_{i,t})_{i=1}^n$ are i.i.d. with the same distribution as W_t. Let $(\varepsilon_i)_{i=1}^n$ be independent Rademacher random variables, independent of the collection of random vectors $W_{i,t}$. Define $\overline{W}_{i,t} := W_{i,t} - \mathrm{E}[W_t]$. Then*

$$\frac{1}{2}\mathrm{E}\left[\sup_{t \in \mathcal{T}} \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \overline{W}_{i,t}\right\|_2^2\right]^{1/2} \leq \mathrm{E}\left[\sup_{t \in \mathcal{T}} \left\|\frac{1}{n}\sum_{i=1}^n \overline{W}_{i,t}\right\|_2^2\right]^{1/2} \leq 2\,\mathrm{E}\left[\sup_{t \in \mathcal{T}} \left\|\frac{1}{n}\sum_{i=1}^n \varepsilon_i W_{i,t}\right\|_2^2\right]^{1/2}.$$

The second result we recall is the Khinchin-Kahane inequality, the specific form we require is obtained from De la Pena and Giné [DG12, Theorem 1.3.1] by setting $q = p$ and $p = 2$ in that theorem, see also Boucheron, Lugosi, and Massart [BLM13, page 141].

Lemma 7. *For $i \in [n]$, let $z_i \in \mathbb{R}^d$ be fixed vectors. Let $(\varepsilon_i)_{i=1}^n$ be independent Rademacher random variables. Then for all $p \geq 2$,*

$$\mathrm{E}\left[\left\|\sum_{i=1}^n \varepsilon_i z_i\right\|_2^p\right]^{1/p} \leq \sqrt{p-1}\cdot\left(\sum_{i=1}^n \|z_i\|_2^2\right)^{1/2}.$$

A straightforward consequence of Lemma 7 is the following result, which follows from the elementary observation that for a vector $x \in \mathbb{R}^d$, $\|x\|_\infty \leq \|x\|_p \leq d^{1/p}\|x\|_\infty$.

Lemma 8. *For $(i,t) \in [n]\times\mathcal{T}$, let $z_{i,t} \in \mathbb{R}^d$ be fixed vectors. Let $(\varepsilon_i)_{i=1}^n$ be independent Rademacher random variables. Then*

$$\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left\|\sum_{i=1}^n \varepsilon_i z_{i,t}\right\|_2^2\right]^{1/2} \leq \frac{5}{2}\sqrt{1+\log|\mathcal{T}|}\cdot\left(\sup_{t\in\mathcal{T}}\sum_{i=1}^n \|z_{i,t}\|_2^2\right)^{1/2}.$$

Proof. Let $p \geq 1$. Then, by Jensen's inequality and Lemma 7

$$\begin{aligned}
\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left\|\sum_{i=1}^n \varepsilon_i z_{i,t}\right\|_2^2\right] &\leq \mathrm{E}\left[\left(\sum_{t\in\mathcal{T}}\left\|\sum_{i=1}^n \varepsilon_i z_{i,t}\right\|_2^{2p}\right)^{1/p}\right] \\
&\leq \left(\sum_{t\in\mathcal{T}}\mathrm{E}\left[\left\|\sum_{i=1}^n \varepsilon_i z_{i,t}\right\|_2^{2p}\right]\right)^{1/p} \\
&\leq (2p-1)\cdot\left(\sum_{t\in\mathcal{T}}\left\{\sum_{i=1}^n \|z_{i,t}\|_2^2\right\}^p\right)^{1/p}.
\end{aligned}$$

Recalling that $\|x\|_p \leq d^{1/p}\|x\|_\infty$ for all $x \in \mathbb{R}^d$ and taking $p := 1 + \log|\mathcal{T}|$ yields the result. □

Finally, we need the following consequence of Lemmas 6 and 8. The proof idea is taken from [Tro16].

Lemma 9. *For each $(i,t) \in [n]\times\mathcal{T}$, let $W_{i,t} \in \mathbb{R}$ be random variables such that for each $t \in \mathcal{T}$, $(W_{i,t})_{i=1}^n$ are i.i.d. with the same distribution as W_t, with $W_t \geq 0$ almost surely. Then*

$$\mathrm{E}\left[\sup_{t\in\mathcal{T}}\sum_{i=1}^n W_{i,t}\right]^{1/2} \leq \left(\sup_{t\in\mathcal{T}}\sum_{i=1}^n \mathrm{E}[W_{i,t}]\right)^{1/2} + 5\sqrt{1+\log|\mathcal{T}|}\cdot\mathrm{E}\left[\sup_{(i,t)\in[n]\times\mathcal{T}} W_{i,t}\right]^{1/2}.$$

Proof. We have by Jensen's inequality and Lemma 6,

$$\begin{aligned}
\mathrm{E}\left[\sup_{t\in\mathcal{T}}\sum_{i=1}^n W_{i,t}\right] &\leq \mathrm{E}\left[\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^n W_{i,t} - \mathrm{E}[W_{i,t}]\right|\right] + \sup_{t\in\mathcal{T}}\sum_{i=1}^n \mathrm{E}[W_{i,t}], \\
&\leq 2\,\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^n \varepsilon_i W_{i,t}\right|^2\right]^{1/2} + \sup_{t\in\mathcal{T}}\sum_{i=1}^n \mathrm{E}[W_{i,t}]. \qquad (26)
\end{aligned}$$

Conditioning on the random vectors $W_{i,t}$, we have by Lemma 8 and the assumption $W_{i,t} \geq 0$ a.s.

$$\begin{aligned}
2\,\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^n \varepsilon_i W_{i,t}\right|^2\right]^{1/2} &\leq 5\sqrt{(1+\log|\mathcal{T}|)}\cdot\left(\sup_{t\in\mathcal{T}}\sum_{i=1}^n W_{i,t}^2\right)^{1/2}, \\
&\leq 5\sqrt{1+\log|\mathcal{T}|}\cdot\left(\sup_{(i,t)\in[n]\times\mathcal{T}} W_{i,t}\right)^{1/2}\cdot\left(\sup_{t\in\mathcal{T}}\sum_{i=1}^n W_{i,t}\right)^{1/2}.
\end{aligned}$$

Taking expectation with respect to $W_{i,t}$, and using the Cauchy-Schwartz inequality yields

$$2\,\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left|\sum_{i=1}^{n}\varepsilon_i W_{i,t}\right|^2\right]^{1/2} \leq \sqrt{6(1+\log|\mathcal{T}|)}\cdot\mathrm{E}\left[\sup_{(i,t)\in[n]\times\mathcal{T}} Z_{i,t}\right]^{1/2}\cdot\mathrm{E}\left[\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}W_{i,t}\right]^{1/2}.$$

Replacing in (26) and solving the resulting quadratic inequality yields the result. □

Equipped with these results, we now prove Lemma 1. The proof idea is taken from [Tro16].

Proof of Lemma 1. We start with the lower bound. We have on the one hand

$$\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left\|\frac{1}{n}\sum_{i=1}^{n}Z_{i,t}\right\|_2^2\right] \geq \sup_{t\in\mathcal{T}}\mathrm{E}\left[\left\|\frac{1}{n}\sum_{i=1}^{n}Z_{i,t}\right\|_2^2\right] = \sigma^2(\mathcal{T}). \tag{27}$$

On the on other hand, by Lemma 6, we have

$$\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left\|\frac{1}{n}\sum_{i=1}^{n}Z_{i,t}\right\|_2^2\right]^{1/2} \geq \frac{1}{2}\,\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i Z_{i,t}\right\|_2^2\right]^{1/2}.$$

Define the random index

$$I \in \operatorname*{argmax}_{i\in[n]}\max_{t\in\mathcal{T}}\|Z_{i,t}\|_2^2.$$

Conditioning on $Z_{i,t}$, we have by Jensen's inequality

$$\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i Z_{i,t}\right\|_2^2\right] \geq \sup_{t\in\mathcal{T}}\mathrm{E}\left[\left\|\mathrm{E}\left[\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i Z_{i,t}\right]\right\|_2^2\right] = \sup_{t\in\mathcal{T}}\frac{\|Z_{I,t}\|_2^2}{n^2} = \sup_{(i,t)\in[n]\times\mathcal{T}}\frac{\|Z_{i,t}\|_2^2}{n^2},$$

where in the inequality, the outer expectation is with respect to ε_I, and the inner one is with respect to $(\varepsilon_i)_{i\neq I}$. Taking expectation with respect to $Z_{i,t}$ gives

$$\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left\|\frac{1}{n}\sum_{i=1}^{n}Z_{i,t}\right\|_2^2\right]^{1/2} \geq \frac{1}{2}\cdot\frac{r(\mathcal{T})}{n} \tag{28}$$

Averaging the lower bounds (27) and (28) yields the desired lower bound. We now turn to the upper bound. We have by Lemmas 6 and 8.

$$\begin{aligned}\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left\|\frac{1}{n}\sum_{i=1}^{n}Z_{i,t}\right\|_2^2\right]^{1/2} &\leq 2\,\mathrm{E}\left[\sup_{t\in\mathcal{T}}\left\|\frac{1}{n}\sum_{i=1}^{n}\varepsilon_i Z_{i,t}\right\|_2^2\right]^{1/2}\\ &\leq 5\sqrt{1+\log|\mathcal{T}|}\cdot\mathrm{E}\left[\sup_{t\in\mathcal{T}}\sum_{i=1}^{n}\left\|\frac{1}{n}Z_{i,t}\right\|_2^2\right]^{1/2}\end{aligned}$$

Applying Lemma 9 on the last term yields the desired upper bound. □

I Proof of Corollary 3

The statement follows from the same argument as Theorem 4 with only a few simple modifications. As explained in the main text, we use Lemma 3 to bound the quantity $\mathrm{E}[\max_{t\in\mathcal{T}}\Lambda_n(t)]$ by constructing a block diagonal matrix. We use Lemma 2 to control, for any subset $\mathcal{S}$, $\mathrm{E}[\max_{s\in\mathcal{S}}G_n(s)]$. The only minor deviation from Theorem 4 is that we bound the second moment

$$\mathrm{E}\left[\sup_{t\in\mathcal{T}\setminus\mathcal{T}_*}\Delta_n^2(t,t_*)\right]$$

instead of the first. This explains the slightly better dependence on δ in the sample size restriction of Theorem 3 compared to Theorem 4.