Optimistic Rates: A Unifying Theory for Interpolation Learning and Regularization in Linear Regression

LIJIA ZHOU[*], Department of Statistics, University of Chicago, USA
FREDERIC KOEHLER[*], Department of Computer Science, Stanford University, USA
DANICA J. SUTHERLAND, University of British Columbia; Alberta Machine Intelligence Institute, Canada
NATHAN SREBRO, Toyota Technological Institute at Chicago, USA

We study a localized notion of uniform convergence known as an "optimistic rate" [34, 39] for linear regression with Gaussian data. Our refined analysis avoids the hidden constant and logarithmic factor in existing results, which are known to be crucial in high-dimensional settings, especially for understanding interpolation learning. As a special case, our analysis recovers the guarantee from Koehler et al. [21], which tightly characterizes the population risk of low-norm interpolators under the benign overfitting conditions. Our optimistic rate bound, though, also analyzes predictors with arbitrary training error. This allows us to recover some classical statistical guarantees for ridge and LASSO regression under random designs, and helps us obtain a precise understanding of the excess risk of near-interpolators in the over-parameterized regime.

CCS Concepts: • **Theory of computation → Sample complexity and generalization bounds**; • **Mathematics of computing → Multivariate statistics**; Markov processes.

Additional Key Words and Phrases: Generalization theory, High-dimensional Statistics, Interpolation, Regularization, Uniform Convergence

ACM Reference Format:
Lijia Zhou, Frederic Koehler, Danica J. Sutherland, and Nathan Srebro. 2023. Optimistic Rates: A Unifying Theory for Interpolation Learning and Regularization in Linear Regression. *ACM/IMS J. Data Sci.* 1, 1, Article 1 (January 2023), 53 pages. https://doi.org/10.1145/3594234

1 INTRODUCTION

One of the core mysteries behind the success of deep learning is that a neural network with a huge number of parameters can be trained with little to no regularization to fit noisy observations, and yet can still achieve good generalization on unseen data points. Even more mind-boggling is the observation that models with larger parameter counts actually tend to generalize *better* [7, 31, 52]. This turns out to be a quite universal phenomenon, not unique to deep learning [8, 9, 18]. As over-parameterized models become more and more important in applications, it seem imperative to understand the mathematical reasons behind their success.

[*] These authors contributed equally.

Authors' addresses: Lijia Zhou, Department of Statistics, University of Chicago, Illinois, USA, zlj@uchicago.edu; Frederic Koehler, Department of Computer Science, Stanford University, California, USA, fkoehler@stanford.edu; Danica J. Sutherland, University of British Columbia; Alberta Machine Intelligence Institute, Canada, dsuth@cs.ubc.ca; Nathan Srebro, Toyota Technological Institute at Chicago, Illinois, USA, nati@ttic.edu.

Manuscript submitted to ACM

In high-dimensional settings, there are usually possible solutions with low training error but very high population risk, and so any analysis based only on the number of parameters will be extremely loose. To explain over-parameterized learning, we need some alternative measure of complexity. Finding the relevant complexity measure of a neural network remains an open question, but we have come to understand that the appropriate complexity measure for linear regression is the norm of the coefficients. Much recent work [e.g. 4, 8, 13, 18, 19, 23, 28, 30, 44, 54] has considered linear regression as a testbed problem which also exhibits some of the surprising behaviors found in deep learning. In particular, Bartlett et al. [4] show that it is possible for the minimal norm interpolator $\hat{w}$ to be consistent even when the number of dimensions grows much faster than the sample size.

A very natural idea to recover this fact is the following: we can consider the set of predictors with norm smaller than $\|\hat{w}\|$, and argue that the difference between training error and population error is small uniformly for all predictors in this set. Because this set is simple in the sense that all predictors have small norm, we can hope for a uniform law of large numbers to show that the population risk of the minimal norm interpolator is also small. This idea, known as uniform convergence, has been the core workhorse of learning theory for decades. Unfortunately, there are lower bounds that show this approach cannot explain consistency in many natural high-dimensional problems [3, 29, 30, 54]. At a high level, this is because the norm required to perfectly fit the noisy labels need to scale with the sample size, and so the set of predictors with norm smaller than $\|\hat{w}\|$ can actually be quite large, and in particular will include predictors with high training error. To sidestep these negative results, Zhou et al. [54] argue that we should focus on upper bounds only for predictors with low training error. Koehler et al. [21] subsequently show that if we only consider the low-norm predictors with *exactly zero* training error, then a uniform convergence argument can actually tightly control the population risk of low-norm interpolators in Gaussian linear regression. Uniform convergence of interpolators has shown itself to be a powerful tool for analyzing interpolation learning, especially when closed form solutions are not available: Koehler et al. [21] proved the first consistency result for basis pursuit (minimum ℓ_1-norm interpolation), and Wang et al. [51] sharpened the analysis to show that basis pursuit can be consistent when the covariates are isotropic, in strong contrast to the ℓ_2 setting.

Though their works highlight the importance of localized uniform convergence and very clearly demonstrates that it is sufficient for interpolation learning, in practice we do not only care about *exact* interpolators. For example, there can be interesting high dimensional settings where interpolation is not possible. When interpolation is possible, we can also obtain good non-interpolating predictors by early stopping or some amount of regularization. Even if we intend to perfectly memorize the labels, numerical precision issues will likely prevent us from fitting them to literally zero error. Thus, we want a more general notion of risk-dependent uniform convergence that is robust to non-interpolation. In the context of linear regression, we want to understand the population risk of any low-norm predictor with small, but not exactly zero training error.

In this paper, we revisit the "optimistic rate" bound of Srebro et al. [39], and perform a tighter analysis based on Gordon's comparison inequality for Gaussian processes [17, 43]. Our new analysis is tight enough to recover the consistency result of the minimal-norm interpolator from Bartlett et al. [4] and Koehler et al. [21] for Gaussian linear regression, which previous work on optimistic rates cannot achieve due to hidden constants and logarithmic factors.[1] At the same time, our result allows us to have a very precise and accurate understanding of the finite-sample risk of non-interpolating estimators. For example, our upper bound for the ordinary least square estimator matches the exact expectation formula given by Hastie et al. [18] in the proportional scaling limit, even though the estimator is not

[1] A more detailed discussion can be found at the beginning of Section 3.

consistent. In Section 4, we also apply our generalization framework to analyze ridge and LASSO regression. We show that it is possible to understand classical statistical theory as well as recent progress in interpolation learning under the same unified framework of optimistic rates.

2 PROBLEM SETTING

Notation. We use $\|\cdot\|_p$ for the ℓ_p norm, $\|x\|_p = (\sum_i |x_i|^p)^{1/p}$. For a positive semidefinite matrix A, the *Mahalanobis (semi-)norm* is $\|x\|_A^2 := \langle x, Ax\rangle$. For a matrix A and set S, AS denotes the set $\{Ax : x \in S\}$. We always use $\max_{x\in S} f(x)$ to be $-\infty$ when S is empty, and similarly $\min_{x\in S} f(x)$ to be ∞. We use $a \vee b$ to denote the maximum between a and b and $a \wedge b$ to denote the minimum. We use standard $O(\cdot)$ notation, and $a \lesssim b$ for inequality up to an absolute constant.

Data model. We assume that the data (X, Y) is generated as

$$Y = Xw^* + \xi, \qquad X_i \overset{iid}{\sim} N(0, \Sigma), \qquad \xi \sim N(0, \sigma^2 I_n), \tag{1}$$

where $X \in \mathbb{R}^{n\times d}$ has i.i.d. Gaussian rows $X_1, \dots, X_n \in \mathbb{R}^d$, w^* is arbitrary, and ξ is Gaussian and independent of X. The *empirical* and *population loss* are defined as, respectively,

$$\hat{L}(w) = \frac{1}{n}\|Y - Xw\|_2^2, \qquad L(w) = \mathop{\mathbb{E}}_{(x,y)} (y - \langle w, x\rangle)^2 = \sigma^2 + \|w - w^*\|_\Sigma^2,$$

where in the expectation $y = \langle x, w^*\rangle + \xi_0$ with $x \sim N(0, \Sigma)$ independent of $\xi_0 \sim N(0, \sigma^2)$. When $d < n$, there is an unique minimizer of $\hat{L}$ which is the ordinary least square estimator $\hat{w}_{\text{OLS}} = (X^T X)^{-1} X^T Y$. When $d \geq n$, for an arbitrary norm $\|\cdot\|$, the minimal norm interpolator is $\hat{w} = \arg\min_{\hat{L}(w)=0} \|w\|$.

3 OPTIMISTIC RATES THEORY

As discussed by Zhou et al. [54], a promising version of localized uniform convergence is to use bounds with “optimistic rates” [34, 39], which establish different generalization guarantees depending on the size of the training error. (This broad concept has been studied at least since the work of Vapnik [47, Theorem 6.3].) In particular, Srebro et al. [39] show that with high probability, it holds uniformly over all $w \in \mathcal{H}$ that

$$L(w) - \hat{L}(w) \leq \tilde{O}\left(\sqrt{\hat{L}(w) \cdot \mathcal{R}_n^2(\mathcal{H})} + \mathcal{R}_n^2(\mathcal{H})\right) \tag{2}$$

where $\mathcal{R}_n(\mathcal{H})$ is the Rademacher complexity[2] of $\mathcal{H}$ for any $n \in \mathbb{N}$. Considering only interpolators in $\mathcal{H}$, the points for which $\hat{L}(w) = 0$), this bound becomes

$$L(w) \leq \tilde{O}\left(\mathcal{R}_n^2(\mathcal{H})\right). \tag{3}$$

In classical settings, it is typically the case that $\mathcal{R}_n(\mathcal{H}) \leq \sqrt{R/n}$ for some constant $R > 0$, and so (2) implies a graceful degradation from a learning rate of $\tilde{O}(1/n)$ in realizable settings to a learning rate of $\tilde{O}(1/\sqrt{n})$ in the more general non-realizable settings. The hidden constant and log factor in the $\tilde{O}$ notation are not so problematic in this regime, because the quantity inside is vanishing.

In interpolation learning, however, we no longer have the scaling of $\mathcal{R}_n(\mathcal{H}) \leq \sqrt{R/n}$: the complexity required to perfectly fit the noisy observations needs to scale with the sample size, and Zhou et al. [54] show in some cases that we can expect $\mathcal{R}_n^2(\mathcal{H})$ to be approximately as large as the Bayes risk σ^2. Therefore, any hidden factor greater than 1 inside

[2] Srebro et al. [39] consider the worst-case Rademacher complexity. In our results, we use a smaller quantity known as the average Rademacher complexity. The formal definition is given in Section 3.2.

the $\tilde{O}$ notation of (3) will not be tight enough to establish consistency. In this work, we improve the hidden factor of $200\,000 \log^3 n$ from Srebro et al. [39] to exactly 1, in the particular setting of Gaussian linear regression. Ignoring lower-order terms, we show that with high probability, the following inequality is approximately true for all $w \in \mathcal{H}$:

$$L(w) - \hat{L}(w) \leq 2\sqrt{\hat{L}(w) \cdot \mathcal{R}_n^2(\mathcal{H})} + \mathcal{R}_n^2(\mathcal{H}),$$

which can be more elegantly written as

$$L(w) \leq \left(\sqrt{\hat{L}(w)} + \mathcal{R}_n(\mathcal{H})\right)^2. \tag{4}$$

The formal statement is given in Theorem 2. It will be clear from our applications in Section 4 that the constants in (4) are in fact tight, and that the bound allows us to get precise generalization bounds for minimal-norm interpolation as well as ridge and LASSO regression.

3.1 Main Bound

We now give our main result, which will be used in Section 3.2 to obtain (4).

Theorem 1. *Under the model assumption in* (1)*, let* $F : \mathbb{R}^d \to [0, \infty]$ *be a function such that for* $x \sim N(0, \Sigma)$*, with probability at least* $1 - \delta'$*, it holds uniformly over all* $w \in \mathbb{R}^d$ *that*

$$\langle w - w^*, x\rangle \leq F(w). \tag{5}$$

For any $\delta > 0$*, assume* $n \geq 196 \log(12/\delta)$*. Then there exists* $\beta_1 \leq 14\sqrt{\frac{\log(12/\delta)}{n}}$ *such that with probability at least* $1 - 2(\delta' + \delta)$*, it holds uniformly over all* $w \in \mathbb{R}^d$ *that*

$$L(w) \leq (1 + \beta_1)\left(\sqrt{\hat{L}(w)} + \frac{F(w)}{\sqrt{n}}\right)^2. \tag{6}$$

The full proof can be found in Appendix B; we briefly sketch the proof here.

Proof sketch of Theorem 1. We do this via Gordon's Theorem (also known as the Gaussian Minmax Theorem; see Theorem 16). It suffices to prove that

$$\sup_w \sqrt{\frac{L(w)}{1 + \beta_1}} - \left(\sqrt{\hat{L}(w)} + \frac{F(w)}{\sqrt{n}}\right) \leq 0.$$

Write $X = Z\Sigma^{1/2}$, where Z is a matrix of standard Gaussian entries. By the definitions of $\hat{L}(w)$ and Y, we have

$$\sup_w \sqrt{\frac{L(w)}{1 + \beta_1}} - \frac{1}{\sqrt{n}}\left(\|Y - Xw\|_2 + F(w)\right) = \sup_w \inf_{\|\lambda\|_2 = 1} \sqrt{\frac{L(w)}{1 + \beta_1}} + \frac{1}{\sqrt{n}}\left(\langle \lambda, Z\Sigma^{1/2}(w - w^*) - \xi\rangle - F(w)\right).$$

The last expression is a max-min optimization with a random Gaussian matrix Z, so by Gordon's Theorem we can prove a high-probability upper bound on this quantity (the "Primary Optimization") by upper-bounding the following

“Auxiliary Optimization” problem with standard Gaussian vectors $H \sim N(0, I_d)$ and $G \sim N(0, I_n)$:

$$\begin{aligned}
&\sup_{w} \inf_{\|\lambda\|_2=1} \sqrt{\frac{L(w)}{1+\beta_1}} + \frac{1}{\sqrt{n}} \Big(\|\lambda\|_2 \langle H, \Sigma^{1/2}(w - w^*) \rangle + \|\Sigma^{1/2}(w - w^*)\|_2 \langle G, \lambda \rangle - \langle \lambda, \xi \rangle - F(w) \Big) \\
&= \sup_{w} \inf_{\|\lambda\|_2=1} \sqrt{\frac{L(w)}{1+\beta_1}} + \frac{1}{\sqrt{n}} \Big(\langle H, \Sigma^{1/2}(w - w^*) \rangle + \langle G\|\Sigma^{1/2}(w - w^*)\|_2 - \xi, \lambda \rangle - F(w) \Big) \\
&= \sup_{w} \left[\sqrt{\frac{L(w)}{1+\beta_1}} - \frac{1}{\sqrt{n}} \|G\|\Sigma^{1/2}(w - w^*)\|_2 - \xi\|_2 \right] + \frac{1}{\sqrt{n}} \Big[\langle \Sigma^{1/2} H, w - w^* \rangle - F(w) \Big].
\end{aligned}$$

The first term is negative with high probability, because

$$L(w) = \|\Sigma^{1/2}(w - w^*)\|_2^2 + \sigma^2;$$

since G, ξ are approximately orthogonal, we have

$$\|G\|\Sigma^{1/2}(w^* - w)\|_2 - \xi\|_2^2 \approx \|G\|_2^2 \|\Sigma^{1/2}(w^* - w)\|_2^2 + \|\xi\|_2^2 \approx n(\|\Sigma^{1/2}(w - w^*)\|_2^2 + \sigma^2).$$

The $1 + \beta_1$ terms accounts for the variations in G and ξ. The second term is also negative with high probability by the fact that $\Sigma^{1/2} H \sim \mathcal{N}(0, \Sigma)$ and our definition of F. □

3.2 Gaussian width/Rademacher Bound

Now we discuss how to recover the Rademacher bound (4) by choosing an F to satisfy the criterion (5). In the context of our model assumption (1), the average Rademacher complexity is given by the following:

Definition 1. Given a positive semi-definite matrix Σ and sample size $n \in \mathbb{N}$, the *Rademacher complexity* of a hypothesis class $\mathcal{H}$ is given by

$$\mathcal{R}_n(\mathcal{H}) = \mathop{\mathbb{E}}_{\substack{x_1,\ldots,x_n \sim \mathcal{N}(0,\Sigma) \\ s \sim \mathrm{Unif}(\{\pm 1\}^n)}} \left[\sup_{h \in \mathcal{H}} \left| \frac{1}{n} \sum_{i=1}^{n} s_i h(x_i) \right| \right].$$

Rademacher complexity measures the ability of $\mathcal{H}$ to fit random Rademacher noise (±1) on an average training set sampled from the ground truth distribution. For more background, see for example the work of Bartlett et al. [2], Bartlett and Mendelson [5], Srebro et al. [39], Wainwright [50].

A closely related geometric complexity measure is the Gaussian width [see, e.g., 5, 48]. The following definitions match the notation of Koehler et al. [21].

Definition 2. The *Gaussian width* and the *radius* of a set $S \subset \mathbb{R}^d$ are

$$W(S) := \mathop{\mathbb{E}}_{H \sim \mathcal{N}(0, I_d)} \sup_{s \in S} |\langle s, H \rangle| \quad \text{and} \quad \mathrm{rad}(S) := \sup_{s \in S} \|s\|_2.$$

We also define the notation

$$W_\Sigma(S) := W(\Sigma^{1/2} S)$$

to represent the Gaussian width with respect to covariance matrix Σ.

As it turns out, when the hypothesis class $\mathcal{H}$ is linear, the Rademacher complexity is actually equivalent to Gaussian width (up to a scaling of $1/\sqrt{n}$).

Proposition 1. *Let $\mathcal{K}$ be an arbitrary subset of $\mathbb{R}^d$ and consider $\mathcal{H} = \{x \mapsto \langle w, x\rangle : w \in \mathcal{K}\}$. Then, for any positive semi-definite matrix Σ, it holds that*

$$\mathcal{R}_n(\mathcal{H}) = \frac{W_\Sigma(\mathcal{K})}{\sqrt{n}}. \tag{7}$$

Proof. Observe that for $x_1, ..., x_n \sim \mathcal{N}(0, \Sigma)$ independent of $s \sim \text{Unif}(\{\pm 1\}^n)$, we have $\frac{1}{n}\sum_{i=1}^n s_i x_i \sim \mathcal{N}\left(0, \frac{1}{n}\Sigma\right)$. The rest just follows from definitions:

$$\begin{aligned}
\mathcal{R}_n(\mathcal{H}) &= \mathop{\mathbb{E}}_{\substack{x_1,\ldots,x_n\sim\mathcal{N}(0,\Sigma)\\ s\sim\text{Unif}(\{\pm1\}^n)}}\left[\sup_{w\in\mathcal{K}}\left|\frac{1}{n}\sum_{i=1}^n s_i\langle w, x_i\rangle\right|\right] \\
&= \mathop{\mathbb{E}}_{\substack{x_1,\ldots,x_n\sim\mathcal{N}(0,\Sigma)\\ s\sim\text{Unif}(\{\pm1\}^n)}}\left[\sup_{w\in\mathcal{K}}\left|\left\langle w, \frac{1}{n}\sum_{i=1}^n s_i x_i\right\rangle\right|\right] = \mathop{\mathbb{E}}_{H\sim\mathcal{N}(0,I_d)}\left[\sup_{w\in\mathcal{K}}\left|\left\langle w, \tfrac{1}{\sqrt{n}}\Sigma^{\frac{1}{2}}H\right\rangle\right|\right] \\
&= n^{-1/2}W_\Sigma(\mathcal{K}).
\end{aligned}$$

□

Consequently, to prove (4), we can replace Rademacher complexity with Gaussian width, and we can see that the definition of F in Theorem 1 is very related to Gaussian width. To get tighter upper bounds, we recall the definition of covariance splitting [21], which is also used by Bartlett et al. [4]:

Definition 3 (Covariance splitting)**.** Given a positive semidefinite matrix $\Sigma \in \mathbb{R}^{d\times d}$, we write $\Sigma = \Sigma_1 \oplus \Sigma_2$ if $\Sigma = \Sigma_1 + \Sigma_2$, each matrix is positive semidefinite, and their spans are orthogonal.

To satisfy the definition of F in condition (5), we can write $x = \Sigma^{1/2}H$, where $H \sim N(0, I_d)$. For any splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$, let H_1 be the orthogonal projection of H onto the span of Σ_1, and H_2 that onto the span of Σ_2.

Example 1 (Gaussian width and Theorem 1)**.** If we are only interested in predictors from a fixed hypothesis class $\mathcal{K}$, then by orthogonality, it holds that for all $w \in \mathcal{K}$,

$$\begin{aligned}
\langle w^* - w,\, x\rangle &= \langle w^* - w,\, \Sigma_1^{1/2}H\rangle + \langle w^* - w,\, \Sigma_2^{1/2}H\rangle \\
&= \langle w^* - w,\, \Sigma_1^{1/2}H_1\rangle + \langle w^* - w,\, \Sigma_2^{1/2}H_2\rangle \\
&\le \|\Sigma^{1/2}(w - w^*)\|_2 \cdot \|H_1\|_2 + |\langle \Sigma_2^{1/2}w^*, H_2\rangle| + \sup_{w\in\Sigma_2^{1/2}\mathcal{K}} |\langle w, H_2\rangle|.
\end{aligned}$$

Hence, by standard concentration results and the fact that $\|\Sigma^{1/2}(w - w^*)\|_2 = \sqrt{L(w) - \sigma^2}$, we can choose

$$F(w) = \left(\sqrt{\operatorname{rank}\Sigma_1} + 2\sqrt{\log(16/\delta')}\right)\sqrt{L(w) - \sigma^2} + W_{\Sigma_2}(\mathcal{K}) + \left(\operatorname{rad}(\Sigma_2^{1/2}\mathcal{K}) + \|w^*\|_{\Sigma_2}\right)\sqrt{2\log(16/\delta')}$$

for $w \in \mathcal{K}$, and let $F(w) = \infty$ for $w \notin \mathcal{K}$.

Plugging into Theorem 1 and rearranging the $\sqrt{L(w) - \sigma^2}$ term, we obtain the following:

Theorem 2. *Under the model assumptions in* (1)*, let $\mathcal{K}$ be an arbitrary compact set, and take any covariance splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$. Fixing $\delta \le 1/4$, let $\beta_2 = 32\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\operatorname{rank}(\Sigma_1)}{n}}\right)$. If n is large enough that $\beta_2 \le 1$, then the following holds with probability at least $1 - \delta$ for all $w \in \mathcal{K}$:*

$$L(w) \le (1 + \beta_2)\left(\sqrt{\hat{L}(w)} + \frac{W_{\Sigma_2}(\mathcal{K})}{\sqrt{n}} + \left[\|w^*\|_{\Sigma_2} + \operatorname{rad}(\Sigma_2^{1/2}\mathcal{K})\right]\sqrt{\frac{2\log(32/\delta)}{n}}\right)^2. \tag{8}$$

Moreover, a stronger version of the above is also true: it holds that uniformly over all dilation factors $\alpha \geq 0$ *and* $w \in \alpha\mathcal{K}$, *we have*

$$L(w) \leq (1+\beta_2)\left(\sqrt{\hat{L}(w)} + \frac{\alpha W_{\Sigma_2}(\mathcal{K})}{\sqrt{n}} + \left[\|w^*\|_{\Sigma_2} + \alpha\,\mathrm{rad}(\Sigma_2^{1/2}\mathcal{K})\right]\sqrt{\frac{2\log(32/\delta)}{n}}\right)^2. \tag{9}$$

The full proof can be found in Appendix B. As discussed by Koehler et al. [21], we can usually find a split such that the $\sqrt{\log(32/\delta)/n}$ term is negligible compared to the Gaussian width term, and so ignoring lower-order terms, our Equation (8) basically shows that

$$L(w) \leq \left(\sqrt{\hat{L}(w)} + \frac{W_{\Sigma_2}(\mathcal{K})}{\sqrt{n}}\right)^2,$$

which, in light of Proposition 1, is the same as (4). In addition, our stronger bound (9) shows that for any predictor w outside $\mathcal{K}$, we can always dilate $\mathcal{K}$ by α and the Gaussian width term inside the corresponding upper bound will also be scaled by α. Since our guarantee is uniform over α, we are able to adapt our upper bounds to predictors with different norms and training errors at the same time. This will be useful for our applications in Section 3.4, where we prove uniform generalization guarantees for all predictors along the regularization path.

3.3 Special Case: Uniform Convergence of Interpolators

If we only look at interpolators in the set $\mathcal{K}$, we immediately recover the uniform convergence of interpolators guarantee from (8):

Corollary 1 (Theorem 1 of [21]). *Under the assumptions of Theorem 2, we have with probability at least* $1-\delta$ *that*

$$\sup_{w\in\mathcal{K},\hat{L}(w)=0} L(w) \leq \frac{1+\beta_2}{n}\left[W_{\Sigma_2}(\mathcal{K}) + \left[\|w^*\|_{\Sigma_2} + \mathrm{rad}(\Sigma_2^{1/2}\mathcal{K})\right]\sqrt{2\log\left(\frac{32}{\delta}\right)}\right]^2. \tag{10}$$

It was shown that the above result can be used to tightly characterize the population risk of interpolating predictors. In particular, when the set $\mathcal{K} = \{w \in \mathbb{R}^d : \|w\| \leq B\}$ is a norm ball for some arbitrary choice of norm $\|\cdot\|$ and $B > 0$, then the Gaussian width is

$$W_\Sigma(\mathcal{K}) = B \cdot \mathbb{E}\,\|x\|_*$$

where $\|\cdot\|_*$ is the dual norm and $x \sim \mathcal{N}(0,\Sigma)$. For example, if we consider the minimal-norm interpolator $\hat{w} = \arg\min_{w:\hat{L}(w)=0}\|w\|$ and choose B to be a high probability upper bound of $\|\hat{w}\|$, then we approximately have

$$L(\hat{w}) \leq (1+o(1)) \cdot \frac{B^2\,(\mathbb{E}\,\|x\|_*)^2}{n}. \tag{11}$$

Combined with a norm analysis, Koehler et al. [21] show that Corollary 1 can recover the nearly-matching necessary and sufficient conditions from Bartlett et al. [4] for the consistency of the minimal ℓ_2 norm interpolator. In particular, they show that

$$B^2 \approx \sigma^2 \frac{n}{(\mathbb{E}\,\|x\|_*)^2}$$

with lower-order terms depending on the effective ranks. In the context of ℓ_2 penalty, recall the following definition of effective ranks:

Definition 4 ([4]). The *effective ranks* of a covariance matrix Σ are

$$r(\Sigma) = \frac{\mathrm{Tr}(\Sigma)}{\|\Sigma\|_{op}} \quad \text{and} \quad R(\Sigma) = \frac{\mathrm{Tr}(\Sigma)^2}{\mathrm{Tr}(\Sigma^2)}.$$

The lower-order terms will vanish when the ℓ_2 benign overfitting conditions hold: there exists a sequence of covariance splits $\Sigma = \Sigma_1 \oplus \Sigma_2$ such that

$$\frac{\operatorname{rank}(\Sigma_1)}{n} \to 0, \qquad \left\|w^*\right\|_2 \sqrt{\frac{\operatorname{Tr}(\Sigma_2)}{n}} \to 0, \qquad \frac{n}{R(\Sigma_2)} \to 0. \tag{12}$$

In this case, we have $L(\hat{w}) \to \sigma^2$ in probability with $\hat{w} = X^T(XX^T)^{-1}Y$ when $\|\cdot\|$ is the Euclidean norm, recovering in the Gaussian case the consistency result of Bartlett et al. [4], Tsigler and Bartlett [44]. Koehler et al. [21] also demonstrated that Corollary 1 can establish the consistency of minimal-ℓ_1 norm interpolators in certain settings. This was established by generalizing the sufficient benign overfitting conditions (12) to general norms. Wang et al. [51] observed that these conditions are too pessimistic in the case of basis pursuit with isotropic covariance — this happens because the analysis of the auxiliary optimization problems arising in the proof of the general result is loose. Sharpening that step, they showed uniform convergence of interpolators can establish the optimal consistency result as well as a matching lower bound, c.f. [12, 28].

Remark 1 (Comparison to the proof of [21])**.** The proof technique introduced in Koehler et al. [21] proceeds by an application of the Gaussian minmax theorem directly to the left hand side of (10), the uniform generalization gap for interpolators. In order to accommodate covariance splitting, and also because the auxiliary problem arising this way may concentrate poorly, the analysis was performed conditional on high probability events over the span of the low rank part Σ_1. In the new analysis, the splitting is performed by choosing the complexity functional $F(w)$ appropriately; this roughly mirrors and simplifies the aforementioned step in the old analysis, as well as generalizing the result beyond exact interpolators.

3.4 General Consequence: Flatness of Loss under Benign Overfitting Conditions

In this section, we illustrate another consequence of Theorem 2 in the context of benign overfitting. As just discussed, even in situations where the labels have noise, there can be low-norm predictors that exactly interpolate the data and nevertheless generalize well. We see that our bounds from Theorem 2 and its special case Corollary 1 are sufficient to explain this phenomenon. In fact, they can tell us something more: the curve of the population loss along the regularization path will become flat in these settings, as long as the regularization parameter is small enough for us to obtain a predictor with norm larger than $\|w^*\|$. In other words, once we fit all of the signals, it does not matter how much noise is fitted, and all low norm near-interpolators can achieve consistency at the same time.

In particular, if we take $\mathcal{K} = \{w : \|w\| \leq 1\}$, then it is clear that for any $w \in \mathbb{R}^d$, we have $w \in \|w\| \cdot \mathcal{K}$. To apply (9) of Theorem 2, we define

$$C_\Sigma(\|w\|) := \frac{\|w\| W_\Sigma(\mathcal{K})}{\sqrt{n}} + \left[\|w^*\|_\Sigma + \|w\| \operatorname{rad}(\Sigma^{1/2}\mathcal{K})\right] \sqrt{\frac{2\log(32/\delta)}{n}}. \tag{13}$$

By virtue of (9), if $w' \in \mathbb{R}^d$ (e.g., the minimal-norm interpolator) satisfies

$$\hat{L}(w') = 0 \quad \text{and} \quad C_{\Sigma_2}(\|w'\|) = \sigma + o(1),$$

then w' is a benign interpolator: $L(w') = \sigma^2 + o(1)$. Moreover, when the above holds, we can also establish consistency for any constrained empirical risk minimizer $\hat{w}_R$ of the form:

$$\hat{w}_R := \operatorname*{arg\,min}_{\|w\| \leq R} \hat{L}(w) \tag{14}$$

as long as R is larger than $\|w^*\|$, and with the convention that if there are multiple minimizers then the minimum-norm minimizer is chosen.

Theorem 3. *Under the model assumptions in* (1), *let* $\|\cdot\|$ *be an arbitrary norm on* $\mathbb{R}^d$ *and consider the complexity functional* C_Σ *and the constrained ERM* $\hat{w}_R$ *given by* (13) *and* (14). *Suppose there is a split* $\Sigma = \Sigma_1 \oplus \Sigma_2$ *and* $\epsilon > 0$ *such that with probability at least* $1-\delta$, *it holds that*

$$\sqrt{\hat{L}(w^*)} \leq (1+\epsilon)\sigma \quad \text{and} \quad C_{\Sigma_2}(\|w^*\|) \leq \epsilon \tag{15}$$

and there exists $w' \in \mathbb{R}^d$ *such that*

$$\hat{L}(w') = 0 \quad \text{and} \quad C_{\Sigma_2}(\|w'\|) \leq (1+\epsilon)\sigma + \epsilon. \tag{16}$$

Then, with probability at least $1-2\delta$, *it holds uniformly over any* $R \geq \|w^*\|$ *that*

$$L(\hat{w}_R) \leq (\sigma + 5(\epsilon + \beta_2)(\sigma \vee 1))^2 . \tag{17}$$

for the same choice of β_2 *as in Theorem 2.*

The full proof, in Appendix B, follows based on a simple argument (Lemma 8) which can be applied even more generally. The condition (15) can easily be satisfied using standard concentration results, whereas (16) requires some benign overfitting conditions. When there exists a benign interpolator, we can expect $\epsilon \to 0$ for a sufficiently large sample size, and so $L(\hat{w}_R)$ will converge to σ^2 uniformly. In the context of ridge regression (ℓ_2 penalty), we want the condition (12) to hold.

Corollary 2. *Let* $\sigma > 0$ *be fixed. Under the assumptions of Theorem 3 with* $\|\cdot\|$ *as the Euclidean norm, suppose that* $\Sigma = \Sigma(n)$ *is a sequence of covariance matrices with splits* $\Sigma = \Sigma_1 \oplus \Sigma_2$ *satisfying the benign overfitting conditions* (12). *Then it holds that*

$$\sup_{R \geq \|w^*\|_2} L(\hat{w}_R) \to \sigma^2 \quad \text{in probability.} \tag{18}$$

In other words, we get a uniform convergence result along this entire component of the regularization path. It is straightforward to make this into a finite-sample bound by using the non-asymptotic bounds on the norm of the minimum-norm interpolator from Koehler et al. [21], as well as to generalize the result to other norms under the appropriate benign overfitting conditions from that work. We omit the details here.

4 APPLICATIONS

In this section, we show how to apply our generalization bound to a variety of settings by choosing the appropriate complexity functional F in Theorem 1, and by doing so we recover versions of classical results from compressed sensing, high-dimensional statistics, and statistical learning theory. Some aspects of our results are new: in particular, applying our theory always recovers finite-sample bounds and generally gives guarantees which apply to *all predictors* in a class, not just the particular empirical risk minimizer. As further explained by Koehler et al. [21], Zhou et al. [54], this is a crucial advantage of uniform-convergence based generalization bounds compared to other methods of analysis. For example, analyses based on random matrix theory methods or the asymptotic framework for applying the Convex Gaussian Minmax Theorem (CGMT) developed by Thrampoulidis et al. [41, 43] usually only give guarantees for the empirical risk minimizer and only apply in certain asymptotic limits ("proportional scaling"). Gordon's Theorem/GMT itself has long been used in the analysis of M-estimation, both in regularization and interpolation settings [e.g. 1, 11,

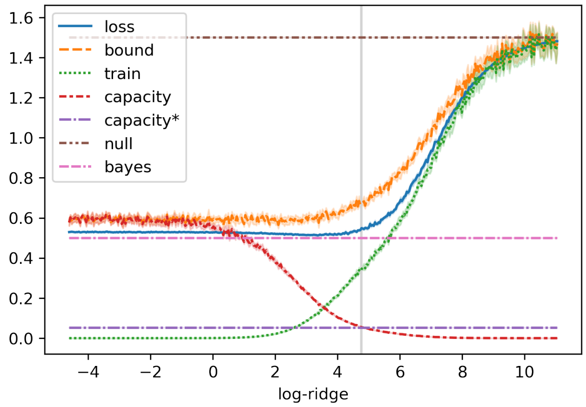

Fig. 1. Loss along regularization path for ridge regression under benign overfitting conditions. Curve and error bars are computed from 10 trials with covariance matrix $\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & \alpha^2 I_d \end{bmatrix}$, $\sigma^2 = 0.5$, $\alpha = 0.05$, and ground truth $w^* = (1, 0, \dots, 0)$ from $n = 600$ samples with aspect ratio $d/n = 20$; the x-axis corresponds to the log of the ridge parameter. The curve "bound" corresponds to the generalization guarantee of Theorem 2; it is close to the population loss ("loss") along the whole regularization path. "Null" and "bayes" are $L(0)$ and $L(w^*)$. "Capacity" corresponds to the term $W(\mathcal{K})^2/n \approx \|w\|^2 \operatorname{Tr}(\Sigma)/n$ for the ridge output w, and "capacity*" is the same term with $\|w\|$ replaced by $\|w^*\|$. As predicted by Theorem 3, the population loss of the ridge regression is roughly flat once $\|w\| > \|w^*\|$ (threshold indicated by grey vertical line), and this is matched by the generalization bound, even though it is determined by the training error $\hat{L}(w)$ (curve "train") and capacity/norm $\|w\|$ which vary significantly.

14, 24, 27, 32, 33, 35, 38, 40] — what is new is how we do this by controlling the generalization gap. As the examples will illustrate, the optimistic rates theory developed in the previous section explains many different phenomena with a simple and natural generalization theory approach.

4.1 Consistency of Optimally-tuned Regularized Regression

To demonstrate the applicability of our Theorem 2 outside of the interpolation setting, we show how to apply it to derive consistency of optimally-tuned regularized least squares estimators such as the LASSO and Ridge regression. In particular, we will show the ridge estimator is consistent under a low effective dimension assumption on Σ; this kind of effective dimension condition was used, for example, by Mendelson [25], Tsigler and Bartlett [44], Zhang [53].

Given any predictor w, by the same reasoning in Section 3.3, we obtain

$$L(w) \leq (1 + o(1)) \cdot \left(\sqrt{\hat{L}(w)} + \frac{\|w\| \cdot \mathbb{E}\|x\|_*}{\sqrt{n}} \right)^2. \tag{19}$$

For any $\lambda > 0$, consider the regularized linear regression problem

$$\hat{w}_\lambda = \arg\min_w \hat{L}(w) + \lambda\|w\|. \tag{20}$$

By comparing the KKT conditions, it is easy to see that there is some choice of λ^* such that

$$\hat{w}_{\lambda^*} = \mathop{\arg\min}_{\hat{L}(w) \leq \|\xi\|_2^2/n} \|w\|.$$

Since $\hat{L}(w^*) = \|\xi\|_2^2/n \approx \sigma$, it naturally follows that $\|\hat{w}_{\lambda^*}\| \leq \|w^*\|$. Plugging in the estimates into (19), we obtain the following:

Corollary 3. *Under the assumptions of Theorem 2, consider the regularized regression estimators $\hat{w}_\lambda$ as in* (20) *with an arbitrary norm $\|\cdot\|$. With probability at least $1-\delta$, there exists a $\lambda^* \geq 0$ such that*

$$L(\hat{w}_{\lambda^*}) \leq (1+3\beta_2)\left(\sigma + \frac{\|w^*\|}{\sqrt{n}}\left(\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,\Sigma_2)}\|x\|_* + \sup_{\|u\|\leq 1}\|u\|_{\Sigma_2}\cdot\sqrt{8\log(36/\delta)}\right)\right)^2. \tag{21}$$

Hence, we have $L(\hat{w}_{\lambda^*}) \to \sigma^2$ in probability if

$$\frac{\operatorname{rank}(\Sigma_1)}{n} \to 0, \quad \frac{\|w^*\|\cdot\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_2)}\|x\|_*}{\sqrt{n}} \to 0, \quad \textit{and} \quad \frac{\|w^*\|\cdot\sup_{\|u\|\leq 1}\|u\|_{\Sigma_2}}{\sqrt{n}} \to 0. \tag{22}$$

In the context of ridge regression, (21) can be simplified to

$$L(\hat{w}_{\lambda^*}) \leq (1+3\beta_2)\left(\sigma + \sqrt{32\log(36/\delta)\cdot\frac{\|w^*\|_2^2\operatorname{Tr}(\Sigma_2)}{n}}\right)^2 \tag{23}$$

because both $\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_2)}\|x\|_2$ and $\sup_{\|u\|_2\leq 1}\|u\|_{\Sigma_2} = \|\Sigma_2\|_{op}^{1/2}$ can be upper bounded by $\sqrt{\operatorname{Tr}(\Sigma_2)}$. Therefore, a sufficient condition for the consistency of optimally-tuned ridge regression is

$$\frac{\operatorname{rank}(\Sigma_1)}{n} \to 0 \quad \text{and} \quad \|w^*\|_2\sqrt{\frac{\operatorname{Tr}(\Sigma_2)}{n}} \to 0. \tag{24}$$

We see that the above is weaker than the benign overfitting condition (12) because we don't need the last condition $\frac{n}{R(\Sigma_2)} \to 0$. However, from Section 3.4, having that condition means we no longer need to tune the ridge parameter λ: any sufficiently small λ will lead to consistency.

4.2 LASSO

Slow Rate under Bounded ℓ_1 Norm. In the context of LASSO regression, assume without loss of generality that the maximum diagonal entry of Σ is 1. Then we have

$$\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,\Sigma_2)}\|x\|_\infty + \sup_{\|u\|_1\leq 1}\|u\|_{\Sigma_2}\cdot\sqrt{8\log(36/\delta)} \lesssim \sqrt{\log(d)},$$

and (21) translates to the convergence rate of $\sigma\|w^*\|_1\sqrt{\frac{\log(d)}{n}} + \|w^*\|_1^2\cdot\frac{\log(d)}{n}$ to σ^2, which is also known as the "slow" rate of LASSO. Moreover, if w^* is k-sparse, then we can bound

$$\|w^*\|_1 \leq k\|w^*\|_\infty$$

and so under these assumptions, the LASSO slow rate guarantee becomes $\sigma k\|w^*\|_\infty\sqrt{\frac{\log(d)}{n}} + k^2\|w^*\|_\infty^2\cdot\frac{\log(d)}{n}$. This analysis works for all predictors w^* of bounded ℓ_1-norm, and it is minimax optimal over this class, but when we assume that w^* is k-sparse it is generally suboptimal and in particular does not give exact recovery when $\sigma = 0$. We now explain how our theory recovers the correct behavior in the sparse and well-conditioned setting commonly studied in the sparse linear regression literature.

Performance under Sparsity and Compatability/Restricted Eigenvalue Condition. We show how to recover well-known results from compressed sensing and high-dimensional statistics about sparse linear regression with Gaussian designs. In particular, we prove a performance guarantee for the LASSO when the covariance matrix is well-conditioned, as previously analyzed by Raskutti et al. [35], or more generally satisfies a version of the *compatability condition* [45]. We start with the following well-known lemma commonly used in the analysis of the LASSO [see, e.g., 48].

Lemma 1. *Suppose w^* is k-sparse, i.e. supported on coordinate set $S \subset [d]$ with $|S| \le k$. Every w with $\|w\|_1 \le \|w^*\|_1$ satisfies*

$$\|(w - w^*)_{S^C}\|_1 \le \|(w^* - w_S)\|_1. \tag{25}$$

The above lemma shows that the vector $w - w^*$ lies in the covex cone

$$C(S) := \{u : \|u_{S^C}\|_1 \le \|u_S\|_1\},$$

where S is the support of w^*. Now we can state the version of the *compatibility condition* [45] we use; the compatibility condition is a weakening of the *restricted eigenvalue condition* [10, 35], and the compatibility condition is known to be a sufficient and almost necessary condition for the LASSO to perform exact recovery from $O(k \log d)$ samples in the Gaussian random design setting [20].

Definition 5 (Compatibility Condition; see [45])**.** For a positive semidefinite matrix $\Sigma : n \times n$, $L \ge 1$, and set $S \subset [n]$, we say Σ has *S-restricted ℓ_1-eigenvalue*

$$\phi^2(\Sigma, S) = \min_{u \in C(S)} \frac{|S| \cdot \langle u, \Sigma u \rangle}{\|u_S\|_1^2}.$$

We say the *S-compatibility condition* holds if the S-restricted ℓ_1-eigenvalue is nonzero.

Example 2 (Application of Theorem 1 to LASSO with sparsity)**.** Observe that for $x \sim N(0, \Sigma)$, we have by Holder's inequality, the standard Gaussian tail bound, and the union bound that with probability at least $1 - \delta'$,

$$\langle w - w^*, x \rangle \le \|w - w^*\|_1 \|x\|_\infty \le \|w - w^*\|_1 \max_i \sqrt{2\Sigma_{ii} \log(2d/\delta')}. \tag{26}$$

Thus, we can take $F(w)$ to be the right hand side of this inequality when applying Theorem 1.

Combining (26) with Lemma 1 and the compatibility condition, we obtain the following:

Theorem 4. *Under the model assumptions in* (1)*, additionally assume that:*

(1) w^* is a k-sparse vector.
(2) For $S \subset [d]$ the support of w^*, the covariance matrix Σ satisfies the S-compatibility condition.
(3) The number of samples n satisfies

$$n > \frac{32 \max_i \Sigma_{ii}}{\phi^2(\Sigma, S)} \cdot k \log\left(\frac{32d}{\delta}\right).$$

Then, for all w satisfying $\|w\|_1 \le \|w^*\|_1$ and $\hat{L}(w) \le (1 + \epsilon)\sigma^2$ for an arbitrary ϵ, we have

$$L(w) - \sigma^2 \lesssim (\beta_1 + \epsilon)\sigma^2 + (1 + \epsilon) \frac{\max_i \Sigma_{ii}}{\phi(\Sigma, S)^2} \cdot \frac{\sigma^2 k \log(32d/\delta)}{n}, \tag{27}$$

where $\beta_1 = O(\sqrt{\log(1/\delta)/n})$ is as defined in Theorem 1. In particular, when $\sigma = 0$ we have that $\|w - w^*\|_\Sigma = 0$, and so if Σ is positive definite then we have $w = w^*$ (exact recovery).

To interpret the above bound, observe that when we consider the ERM, we know that $\epsilon = O(1/\sqrt{n})$ based on concentration of the norm of the noise (Lemma 2) and so the first term is $\sigma^2/\sqrt{n}$ and the second term, assuming Σ is well-conditioned, is $O(\sigma^2 k \log(d/\delta)/n)$, which is the well-known minimax rate for sparse linear regression [see, e.g., 36]. The above analysis is not very careful in terms of constant factors; in Section 4.5 we show how to get sharp constants in the isotropic setting. Also, in Section 6.1 we show how to get rid of the first term on the right hand side of the bound above, when we are specially considering the constrained ERM $\hat{w}$ minimizing the squared loss over all $\|w\|_1 \leq \|w^*\|_1$, i.e. the LASSO solution: see Corollary 5.

4.3 Ordinary Least Squares

Next, we consider a high-dimensional setting when d is smaller than n. For example, when $d = n/2$, the ordinary least squares estimator $\hat{w}_{\text{OLS}}$ is the unique minimizer of the training error, but it does not interpolate the training data and so the uniform convergence analysis of Koehler et al. [21] cannot be applied. As it turns out, our Theorem 1 is enough to tightly characterize the excess risk of $\hat{w}_{\text{OLS}}$.

Example 3 (Application of Theorem 1 to OLS)**.** By the Cauchy-Schwarz inequality, it holds that

$$\langle w^* - w, x \rangle \leq \|H\|_2 \, \|w^* - w\|_\Sigma.$$

Using standard concentration inequalities and $L(w) - \sigma^2 = \|w - w^*\|_\Sigma^2$, we can choose

$$F(w) = \left(\sqrt{d} + 2\sqrt{\log(4/\delta')}\right) \sqrt{L(w) - \sigma^2}. \tag{28}$$

Theorem 5. *Under the model assumptions in* (1)*, let* $\gamma = d/n < 1$*. There exists some* $\epsilon \lesssim \left(\frac{\log(36/\delta)}{n}\right)^{1/2}$ *such that for all sufficiently large n, with probability* $1 - \delta$ *it holds uniformly for all* $w \in \mathbb{R}^d$ *that*

$$\left| \sqrt{L(w) - \sigma^2} - \sqrt{\frac{\gamma \hat{L}(w)}{(1-\gamma)^2}} \right| \leq \epsilon \sqrt{\hat{L}(w)} + \sqrt{\frac{1}{1-\gamma}\left(\frac{\hat{L}(w)}{1-\gamma} - \sigma^2\right) + \epsilon \hat{L}(w)}. \tag{29}$$

For the empirical risk minimizer $\hat{w}_{\text{OLS}} = (X^T X)^{-1} X^T Y$*, the right hand side of* (29) *is approximately zero because we also have*

$$\hat{L}(\hat{w}_{\text{OLS}}) \leq \sigma^2(1-\gamma) + \sigma^2 \epsilon \sqrt{1-\gamma}. \tag{30}$$

Therefore, we obtain the following generalization bound:

$$L(\hat{w}_{\text{OLS}}) - \frac{\sigma^2}{1-\gamma} \lesssim \sigma^2 \left(\frac{\log(36/\delta)}{n}\right)^{1/4}. \tag{31}$$

We have a relatively complicated expression in (29) because our choice of F according to (28) depends on the excess risk $L(w) - \sigma^2$, and so after applying (6) we need to solve a quadratic equation. All quantities in (29) are well-defined because $\hat{L} \geq 0$ and the $\epsilon \hat{L}(w)$ term inside the last square root ensures that with high probability it is positive. If we think of ϵ as zero for simplicity, then our uniform convergence guarantee (29) predicts that the excess risk $L(w) - \sigma^2$ of a predictor with training error $\hat{L}(w)$ cannot be larger than

$$\frac{1}{1-\gamma}\left(\sqrt{\frac{\gamma \hat{L}(w)}{1-\gamma}} + \sqrt{\frac{\hat{L}(w)}{1-\gamma} - \sigma^2}\right)^2.$$

The minimal error is approximately $\sigma^2(1-\gamma)$ and so all near empirical risk minimizer should enjoy an excess risk of $\sigma^2 \frac{\gamma}{1-\gamma}$, which agrees with the exact expectation formula in Hastie et al. [18]; see their discussion for additional references. Since our approach also gives us a lower bound for free (by solving the quadratic equation), Theorem 5 is enough to show that $L(\hat{w}_{\text{OLS}})$ converges to $\sigma^2 \frac{1}{1-\gamma}$ in probability. We see that even though the empirical risk minimizer is not consistent, our localized uniform convergence approach can still provide an accurate understanding of the excess risk, and our bound for OLS is tight at least for the leading term.

Remark 2. The $O(n^{-1/4})$ rate of (31) comes from the fact that we need to take the square root of ϵ in the last term of (29); it is not too difficult to see that this is sub-optimal for OLS. In fact, in Theorem 13, we explicitly calculate the variance of $L(\hat{w}_{\mathrm{OLS}})$ and show that in the proportional scaling regime (e.g., $\gamma = 0.5$), the right amount of deviation is of order $O(n^{-1/2})$. In the fixed-d regime, the convergence rate can be accelerated to the more familiar rate of $O(n^{-1})$. In Theorem 14, we show how to use a more direct approach to obtain high probability bounds that match these variance calculations. Surprisingly, we can also show that the $O(n^{-1/4})$ rate is generally unavoidable for any uniform convergence analysis that only considers the size of $\hat{L}(w)$. Our analysis is tight in the sense that there are estimators whose training error is indistinguishable from $\hat{w}_{\mathrm{OLS}}$, but whose convergence rate is provably slower than $\Omega(n^{-1/4})$. For readers interested in the tightest rate of convergence, more details can be found in Section 6.2.

4.4 Minimum-Euclidean Norm Interpolation with Isotropic Data and Proportional Scaling

In the previous section, we saw that for OLS in the proportional scaling regime a simple application of our optimistic-rate bound recovers the limiting asymptotic population loss as a function of $d/n < 1$. For $d/n > 1$, the OLS estimator is no longer defined, and instead we study the performance of the minimum-norm interpolator of the data. In Theorem 6 below, we show that with a slightly more careful[3] application of Theorem 1, we can recover the loss curve at any aspect ratio (see Figure 2). Together with the previous result, we show that the optimistic-rate bound can capture the behavior of the pseudoinverse estimator $\hat{w} = X^+ Y$ on both sides of the double descent curve.

Theorem 6. *Under the model assumptions in* (1) *with* $\gamma = d/n > 1$ *and* $\Sigma = I_d$*, there exists* $\epsilon \lesssim \left(\frac{\log(18/\delta)}{n}\right)^{1/2}$ *such that with probability at least* $1-\delta$*, the following holds uniformly over all* w *such that* $\hat{L}(w) = 0$*:*

$$\left| L(w) - \left[\sigma^2 + \|w\|_2^2 + \left(1 - \frac{2}{(1+\epsilon)\gamma}\right)\|w^*\|_2^2\right]\right| \leq 2\|w^*\|_2 \sqrt{\left(1-\frac{1}{\gamma}\right)\left(\|w\|_2^2 - \frac{\|w^*\|_2^2}{\gamma}\right) - \frac{\sigma^2}{\gamma}} + 3\epsilon\|w\|_2^2. \tag{32}$$

It is clear from Figure 2 below that Theorem 6 is capturing the asymptotic behavior of the minimum-norm interpolator; we prove this formally in Theorem 7 below by combining the generalization bound with a norm calculation, recovering the asymptotic formula for this setting computed by Hastie et al. [18] using random matrix theory techniques.

Theorem 7. *Under the model assumptions in* (1) *with* $\gamma = d/n > 1$ *and* $\Sigma = I_d$*, there exists* $\epsilon \lesssim \left(\frac{\log(40/\delta)}{n}\right)^{1/2}$ *such that with probability at least* $1-\delta$*, it holds that*

$$\min_{w: Xw = Y} \|w\|_2^2 \leq (1+\epsilon)\left(\frac{\|w^*\|_2^2}{\gamma} + \frac{\sigma^2}{\gamma - 1}\right). \tag{33}$$

Thus, by Theorem 6, we have

$$L(\hat{w}) - \left[\left(1 - \frac{1}{\gamma}\right)\|w^*\|_2^2 + \sigma^2 \frac{\gamma}{\gamma-1}\right] \leq \epsilon\left(\frac{\|w^*\|_2^2}{\gamma} + \frac{\sigma^2}{\gamma-1}\right) + \|w^*\|_2 \sqrt{\epsilon\left(\frac{\|w^*\|_2^2}{\gamma} + \frac{\sigma^2}{\gamma-1}\right)} \tag{34}$$

where $\hat{w}$ *is the minimal-*ℓ_2 *norm interpolator. If we fix* σ^2, γ *and* $\|w^*\|_2$*, then as* $n \to \infty$

$$L(\hat{w}) \to \left(1 - \frac{1}{\gamma}\right)\|w^*\|_2^2 + \sigma^2 \frac{\gamma}{\gamma - 1} \quad \textit{in probability.} \tag{35}$$

[3] The specific choice of the complexity function F follows from our Lemma 10 in the appendix.

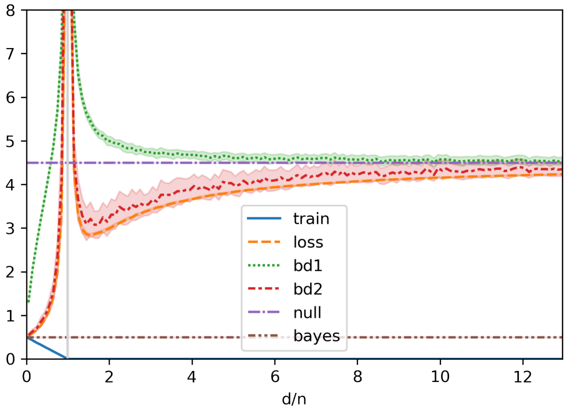

Fig. 2. Generalization bounds for OLS/minimum-ℓ_2 norm interpolation with isotropic covariance $\Sigma = I$, $\|w^*\| = 2$, $\sigma^2 = 0.5$, $n = 4096$, and varying aspect ratio d/n. The vertical line at $d/n = 1$ represents the double descent peak: on the left ($d/n < 1$) the predictor w considered is the Ordinary Least Squares solution and on the right the minimum ℓ_2-norm interpolator. The line "train" is the empirical loss $\hat{L}(w)$, the line "loss" is the test/population loss $L(w)$, "bayes" is the minimal population loss $L(w^*)$, and "null' is $L(0)$. Each curve correspond to the means from 30 trials at each value of d/n, and the error bars correspond to standard deviations. The line "bd1" corresponds to the bound $\left(\sqrt{\hat{L}(w)} + \|w\|\sqrt{d/n}\right)^2$ from Theorem 2, and "bd2" is the upper bound from Theorem 5 for $d/n < 1$ and Theorem 6 for $d/n > 1$. As we see, bd2 is much closer to the true loss around the double descent peak. As explained in Section 5, bd2 can be recovered by looking at a localized version of Gaussian width. Both bd1 and bd2 are derived from our main optimistic rates bound Theorem 1.

Remark 3. Similar to the application in the last section, we also have a lower order $O(n^{-1/4})$ term. It is suboptimal, and we suspect that this is unavoidable for any uniform convergence analysis that only considers the typical size of $\|\hat{w}\|$. Nonetheless, this bound recovers the leading term, and the lower-order term is negligible if we only care about the difference with σ^2.

4.5 Sharp analysis of LASSO in the Isotropic Setting

A well-known application of the Gaussian Minmax Theorem is to the sharp analysis of the LASSO in the setting where the covariates are isotropic and Gaussian [see, e.g., 1, 40]. Our optimistic rates bound Theorem 1 recovers a corresponding generalization bound for all predictors w with $\|w\|_1 \le \|w^*\|_1$, which when specialized to the constrained ERM (i.e. the LASSO solution) recovers these results.

Theorem 8. *Using the notation of Theorem 5, we have with probability at least $1-\delta$ that for all w with $\|w\|_1 \le \|w^*\|_1$,*

$$\left| \sqrt{L(w) - \sigma^2} - \sqrt{\frac{\gamma \hat{L}(w)}{(1-\gamma)^2}} \right| \le \epsilon\sqrt{\hat{L}(w)} + \sqrt{\frac{1}{1-\gamma}\left(\frac{\hat{L}(w)}{1-\gamma} - \sigma^2\right) + \epsilon \hat{L}(w)} \tag{36}$$

provided $\gamma + 2\epsilon/\sqrt{n} < 1$, where

$$\mathcal{K}' := \{u : \exists \delta > 0, \|w^* + \delta u\|_1 \le \|w^*\|_1\} \quad \text{and} \quad \gamma := \frac{1}{n} \cdot W(\mathcal{K}' \cap S^{n-1})^2.$$

Observe that if $\sigma = 0$ and $\hat{L}(w) = 0$ then we get exact recovery provided $\gamma + 2\epsilon/\sqrt{n} < 1$ which is sharp up to the constant in the confidence term [see, e.g., 1, 11]. Informally, exact recovery occurs when $n > \omega^2$, i.e. the number of observations exceeds the statistical dimension. Moreover, we can consider the asymptotic setting where $\sigma = o(1)$ and the proportional scaling limit where γ converges to constant. In this case, it is known [42, Equation 40(a)] that we have $\hat{L}(\hat{w}_{LASSO})/\sigma^2 \to 1 - \gamma$, so the right hand side of (84) converges to zero and we have

$$\frac{1}{\sigma^2} L(\hat{w}_{LASSO}) - 1 \to \frac{\gamma}{1-\gamma}.$$

Thus we recover the characterization of the performance of LASSO in this regime [40, 42]. It is possible, as in the OLS setting, to also derive non-asymptotic bounds on $\hat{L}(\hat{w}_{LASSO})$ and therefore obtain non-asymptotic bounds on the performance of the LASSO; we omit the details.

Remark 4. The Gaussian width of the tangent cone $\mathcal{K}'$ has been sharply characterized in previous work [e.g. 1, 11]. In particular, from the work of Amelunxen et al. [1] we know that if w^* is k-sparse,

$$\omega = W(\mathcal{K}' \cap S^{n-1}) \le W(\text{cone}(\mathcal{K}') \cap S^{n-1}) \le \sqrt{d\psi(s/d)}$$

where

$$\psi(\rho) := \inf_{\tau \ge 0} \left\{ \rho(1+\tau^2) + (1-\rho)\sqrt{2/\pi} \int_\tau^\infty (u-\tau)^2 e^{-u^2/2} du \right\},$$

as well as a corresponding lower bound which characterizes ω.

5 LOCALIZED UNIFORM CONVERGENCE MEETS LOCALIZED COMPLEXITY MEASURE: THE OPTIMALITY OF LOCAL GAUSSIAN WIDTH

Although our choice of the complexity function F in the applications so far can seem quite mysterious, we show how it can be chosen systematically based on the regularizer or the geometry of the constraint set in this section. As we will see, the fact that we can obtain the sharp constants our analysis is not coincidental: the local Gaussian width theory can explain it and elucidate the connection to the previous asymptotic statistics literature (see Remark 5). Consider the following localized version of a convex set $\mathcal{K}$:

$$\mathcal{K}_r := \{w \in \mathcal{K} : \|w^* - w\|_\Sigma \le r\}$$

Based on Proposition 1, the corresponding Gaussian width $W_\Sigma(\mathcal{K}_r)$ can be interpreted as a localized version of the Rademacher Complexity of the function class [see, e.g., 2, 26].

The optimal complexity functional. Ignoring relatively minor technical issues involving the uniform concentration of Gaussian width, we can take $F(w) = W_\Sigma(\mathcal{K}_{\|w-w^*\|_\Sigma})$ in the optimistic rates bound (Theorem 1). This choice of F will lead to an optimal asymptotic guarantee in certain limits, particularly the proportional scaling limit. To see why, first note that if $r = \|w - w^*\|_\Sigma$, then we have from the optimistic rates bound that

$$\sqrt{\sigma^2 + r^2} \le (1+\beta_1)\left(\sqrt{\hat{L}(\hat{w})} + W_\Sigma(\mathcal{K}_r)/\sqrt{n}\right).$$

Rearranging and using $1/(1+\beta_1) \ge 1 - \beta_1$ gives

$$(1-\beta_1)\sqrt{\sigma^2 + r^2} - W_\Sigma(\mathcal{K}_r)/\sqrt{n} \le \sqrt{\hat{L}(\hat{w})}. \tag{37}$$

For simplicity, denote the left hand side of (37) as a function of r called ψ. To obtain a learning guarantee in terms of r, we can find the sublevel set of ψ based on the empirical loss. As the empirical loss becomes smaller, we will pick out a smaller and smaller sublevel set. When $\mathcal{K}$ is convex, it is just an interval because ψ will be convex [4]. On the other hand, we can use CGMT to analyze the minimal training error in $\mathcal{K}$ and show that it nearly match the minimal value of ψ, see Theorem 9 below. This means that (37) is nearly an equality for the ERM in $\mathcal{K}$ and its excess risk r is precisely determined by the minimizer of ψ. In applications, ψ usually admits an unique minimizer, which confirms the approximate optimality of our generalization bound. We note that most of this discussion can also be generalized to non-convex sets $\mathcal{K}$, but the minimal error in $\mathcal{K}$ may no longer be determined by CGMT when $\mathcal{K}$ is not convex.

We can now formalize this argument. First, we define two summary functionals similar to the left hand side of (37). For some absolute constant $C > 0$ and β_1 as defined in Theorem 1, we let the upper summary functional $\psi_\delta^+(x)$ at confidence level $\delta \in (0,1)$ to be

$$\psi_\delta^+(r) := \max\left\{0, (1+\beta_1)\sqrt{\sigma^2 + r^2} - W_\Sigma(\mathcal{K}_r)/\sqrt{n} + Cr\sqrt{\log(2/\delta)/n}\right\} \tag{38}$$

and the lower summary functional $\psi_\delta^-(x)$ at confidence level $\delta \in (0,1)$ to be

$$\psi_\delta^-(r) := \max\left\{0, (1-\beta_1)\sqrt{\sigma^2 + r^2} - W_\Sigma(\mathcal{K}_r)/\sqrt{n} - Cr\sqrt{\log(2/\delta)/n}\right\}. \tag{39}$$

The upper functional comes from the CGMT analysis of the minimal error while the lower functional comes from the application of Theorem 1. As discussed, they match except for a lower order term.

Theorem 9. *Suppose that $\mathcal{K}$ is a convex set and consider the upper summary function ψ_δ^+ as defined in* (38). *It holds with probability at least* $1-\delta$*,*

$$\min_{w\in\mathcal{K}} \sqrt{\hat{L}(w)} \le \min_{r\ge 0} \psi_\delta^+(r) \tag{40}$$

The following result, which is a formalization of (37), informally states that when a training error of μ^2 is approximately achievable by any predictor in $\mathcal{K}$, then only predictors w with $\psi_\delta^-(\|w-w^*\|_\Sigma) \le \mu$ can achieve it — note that by convexity, the set $\{r : \psi_\delta^-(r) \le \mu\}$ will always be an interval which shrinks as we decrease μ. For the lower bound direction, the argument requires a union bound so we adjust the value of δ slightly to τ; the difference is generally negligible since these confidence parameters only appear inside of logarithms.

Theorem 10. *Suppose that $\mathcal{K}$ is a convex set and consider the summary functional $\psi_\delta^+, \psi_\delta^-$ as defined in* (38) *and* (39)*. Let $\delta > 0$ and μ be arbitrary such that $\mu > \mu^* := \min_{r\ge0}\psi_\delta^+(r)$ and define $r^* := \inf\{r : \psi_\delta^+(r) = \mu^*\}$. Then with probability at least $1-4\delta$, it holds that uniformly over all $w \in \mathcal{K}$ such that $\sqrt{\hat{L}(w)} \le \mu$ that:*

$$\|w-w^*\|_\Sigma \le r_+ := \sup\{r \ge 0 : \psi_\delta^-(r) \le \mu\} \tag{41}$$

and also

$$\|w-w^*\|_\Sigma \ge r_- := \inf\left\{r \ge 0 : \psi_\tau^-(r) \le \mu\right\} \tag{42}$$

where $\tau := \delta/\lceil \frac{\mu-\mu^*}{r^*} \rceil$.

If we want to specifically analyze near-empirical risk minimizers, we can apply Theorem 10 with μ of the form $\mu^* + \epsilon$ with a small $\epsilon > 0$, and the conclusion is that their generalization error $\|w-w^*\|_\Sigma$ will be an approximate minimizer of the summary functional ψ_δ^-.

Example 4. To illustrate Theorem 10, we briefly explain how to apply this result in the settings of OLS and minimum norm interpolation with isotropic data. Since we already have given precise nonasymptotic results for these settings in

[4] For a proof, see Lemma 12 and Lemma 13 in the appendix.

the previous sections, we only give a high-level summary of how to apply Theorem 10 in these examples and ignore, for example, the small difference between ψ^-,ψ^+ which is relevant for finite sample bounds. For OLS, we take $\mathcal{K} = \mathbb{R}^d$ so $W_\Sigma(\mathcal{K}_r) \approx r\sqrt{d/n}$ so the limiting summary functional is

$$\psi(r) \approx \sqrt{\sigma^2 + r^2} - r\sqrt{d/n}$$

which is minimized at

$$r^2 = \sigma^2(d/n)/(1 - d/n),$$

so taking $\mu \longrightarrow \psi(r)$ from above, we see by Theorem 10 that the OLS solution $\hat{w}$ satisfies $\|\hat{w} - w^*\|_\Sigma \in \psi^{-1}([0, \mu]) = \{\psi^{-1}(\mu)\} = \{r\}$ informally recovering the conclusion of Theorem 5. For ridge regression (and in particular minimun norm interpolation) in the isotropic setting, we can reduce without loss of generality to the case where $\mathcal{K}$ is the unit ball in which case $\mathcal{K}_r$ is the intersection of the unit ball with a ball of radius r about w^*: the Gaussian width of this intersection can be explicitly computed by solving a two-dimensional Euclidean geometry problem, and this essentially corresponds to the key Lemma 10 in the proof of Theorem 6.

Remark 5 (Comparison to Moreau Envelope Theory [41])**.** In asymptotic settings where the two two summary functionals ψ_δ^- and ψ_δ^+ both converge to a single limit ψ with a unique minimizer, Theorem 10 implies that the asymptotic error of the constrained empirical risk minimizer is given by the equation

$$\|\hat{w} - w^*\|_\Sigma = \arg\min_{r \geq 0} \psi(r).$$

In particular, the functional $\psi(r)$ serves as a "summary functional" which encapsulates all of the relevant information about the geometry of w^* and $\mathcal{K}$. In such an asymptotic setting, Theorem 3.1 of Thrampoulidis et al. [41] gives an asymptotic characterization of the performance of the constrained ERM (without any finite sample bounds) in terms of a summary functional called the "expected Moreau envelope": this can be understood as encoding almost the same information as $\psi(r)$. Some of the main advantages of Theorem 10 are that (1) it is nonasymptotic (in particular, it applies outside of the proportional scaling regime), (2) arguably easier to use and interpret, with a simple and direct connection to established notions of local complexity used in generalization theory [see, e.g., 2, 26], and (3) it describes the generalization behavior of predictors w besides the Empirical Risk Minimizer. Their result, while only applying in the proportional scaling limit, has the advantage of being applicable to other loss functions such as the Huber loss, being stated for more general noise models, and giving formulas directly in terms of regularization parameters without rewriting the optimization as a constrained optimization.

6 IMPROVED FINITE-SAMPLE RATE

In this section, we discuss how to obtain improved finite sample rates and explain why the precise rates will depend on the particular information we have about the predictor.

6.1 Faster rates for low-complexity classes

When the set $\mathcal{K}$ is low complexity, as in the case of ordinary least squares when d is fairly small compared to n, the optimal rate for the empirical risk minimizer in $\mathcal{K}$ goes at a "parametric rate" of $1/n$, faster than a $1/\sqrt{n}$ rate. At first glance, it may appear impossible to get faster than a $1/\sqrt{n}$ rate from the main optimistic rates bound Theorem 1 because of the presence of the $\beta_1 = O(\sqrt{\log(2/\delta)/n})$ term. As we will show, one can actually get fast/optimal rates from this theorem, but there is a different sense in which the $1/\sqrt{n}$ is unavoidable: this rate is actually the best we can hope for if we are only allowed to use certain summary statistics of the predictor (for example, see Remark 2). Nevertheless, it is still possible to obtain fast/optimal rates for the empirical risk minimizer by a black-box application of Theorem 1. The strategy we use is to bound the error $\|w - w^*\|_{\hat{\Sigma}}$ in the empirical metric by using a direct and very simple argument based on the KKT condition, and then apply Theorem 1 to bound the error in the population metric. The general idea of analyzing the population loss by going through the empirical metric is very common in statistics and learning theory [e.g. 6, 22, 26].

Theorem 11. *Let $\mathcal{K}$ be a closed convex set in $\mathbb{R}^d$ containing w^* and suppose $\delta' \geq 0, p \geq 0$ are such that with probability at least $1 - \delta'$ over the randomness of $x \sim N(0, \Sigma)$, uniformly over all $w \in \mathcal{K}$ we have*

$$\langle w - w^*, x\rangle \leq \|w - w^*\|_\Sigma \sqrt{p}. \tag{43}$$

Suppose that $\hat{w} = \arg\min_{w \in \mathcal{K}} \hat{L}(w)$ and $p/n \leq 0.999$, then for all $n \geq C \log(2/\delta)$ for some absolute constant $C > 0$, it holds with probability at least $1 - (\delta + \delta')$ that

$$L(\hat{w}) - \sigma^2 \leq (1 + \tau)\sigma^2 \cdot \frac{p}{n}. \tag{44}$$

where $\tau = \tau(p, n, \delta)$ is upper bounded by an absolute constant and satisfies $\tau(p, n, \delta) \to 1$ in any joint limit $[p + \log(2/\delta)]/n \to 0, n \to \infty$.

The details of the proof can be found in Appendix E, where it is obtained as a special case of a more general result (Lemma 15). To illustrate the application of this result, we show how it is used in the analysis of OLS.

Corollary 4. *Under the model assumptions* (1) *with $d < n$ and assuming a sufficiently large n, it holds with probability at least $1 - \delta$ that*

$$L(\hat{w}_{\text{OLS}}) - \sigma^2 \lesssim \sigma^2 \left(\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(36/\delta)}{n}}\right)^2 \tag{45}$$

Theorem 11 can be applied in a very similar way to analyze other models in the low complexity regime, for example the LASSO when the sparsity level is small, which we illustrate below. Provided the ℓ_1-eigenvalue φ and maximum diagonal entry of Σ are constants, we recover the sharp $\Theta(\sigma^2 k \log(d)/n)$ minimax rate for sparse linear regression (which is sharp provided $k \ll d$; see, e.g., [36]). This recovers the guarantee for the LASSO in the Gaussian random design setting given by combining the result of Raskutti et al. [35] with the appropriate analysis of LASSO in the fixed design setting [e.g. 10, 45].

Corollary 5. *Applying Theorem 11 with $\mathcal{K} = \{\|w\|_1 \leq \|w^*\|_1\}$ the rescaled ℓ_1-ball and under the sparsity and compatability condition assumptions of Theorem 4, we have with probability at least $1 - \delta$ that the LASSO solution*

$$\hat{w}_{LASSO} = \operatorname*{arg\,min}_{w : \|w\|_1 \leq \|w^*\|_1} \hat{L}(w)$$

satisfies

$$L(\hat{w}_{LASSO}) - \sigma^2 \lesssim \frac{\max_i \Sigma_{ii}}{\phi(\Sigma, S)^2} \cdot \frac{\sigma^2 k \log(16d/\delta)}{n} \tag{46}$$

provided n is sufficiently large that

$$\sqrt{\frac{\max_i \Sigma_{ii}}{\phi(\Sigma, S)^2} \cdot \frac{8k \log(16d/\delta)}{n}} \leq 0.999.$$

6.2 Sharp Rate for OLS

We now zero in on the question of sharp rates for Ordinary Least Squares, returning to the discussion from Remark 2. Unlike all of the previous sections, in this section we will use tools beyond Theorem 1 in order to precisely compute second order terms in the generalization gap. Surprisingly, even though we can match the high probability bound with an exact calculation up to first order term (see Theorem 5), the existence of certain near-ERM can prevent us from recovering the correct variance term:

Theorem 12. *Under the model assumptions in* (1)*, fix $\gamma = d/n$ to be some value in $(0, 1)$ and pick any $c > 0$. Then there exists another absolute constant $c' > 0$ such that for all sufficiently large n, with probability at least $1 - \delta$, there exists a $w \in \mathbb{R}^d$ such that*

$$\hat{L}(w) - \hat{L}(\hat{w}_{\text{OLS}}) \leq c \cdot \frac{\sigma^2}{n^{1/2}}, \tag{47}$$

but the population error satisfies

$$L(w) - L(\hat{w}_{\text{OLS}}) \geq c' \cdot \frac{\sigma^2}{n^{1/4}}. \tag{48}$$

If we know that $\hat{L}(w) = \hat{L}(\hat{w}_{\text{OLS}})$, then it is necessarily the case that $w = \hat{w}_{\text{OLS}}$ and as we will see, we can use Theorem 14 to get the tightest possible convergence rates. On the other hand, it is not difficult to see that $n\hat{L}(\hat{w}_{\text{OLS}})/\sigma^2$ follows a chi-squared distribution with $n-d$ degrees of freedom, and by the variance formula of chi-squared distributions, we have

$$\operatorname{Var}(\hat{L}(\hat{w}_{\text{OLS}})) = \frac{2\sigma^4(1-\gamma)}{n}.$$

Consequently, $\hat{L}(\hat{w}_{\text{OLS}})$ can in fact deviate from $\mathbb{E}\,\hat{L}(\hat{w}_{\text{OLS}}) = \sigma^2(1-\gamma)$ by the order of $\sigma^2/\sqrt{n}$. If we only know that $\hat{L}(w)$ is within the normal range of $\hat{L}(\hat{w}_{\text{OLS}})$, then the above theorem says that the sub-optimal rate of $O(n^{-1/4})$ that we show from Theorem 5 is actually tight and unavoidable. We can show a similar negative result for the fixed d regime that the convergence cannot be faster than $O(n^{-1/2})$, but as we can see from the last section, using $\|w - w^*\|_{\hat{\Sigma}}^2 \approx \sigma^2\gamma$ as the empirical metric instead is enough to recover the parameteric rate $O(1/n)$. This argument fails for the proportional limit regime because the smallest eigenvalue of $\hat{\Sigma}$ is $(1-\sqrt{\gamma})^2$ and so we can only get the larger quantity $\sigma^2 \frac{\gamma}{(1-\sqrt{\gamma})^2}$ which fails to capture the first order behavior of $\sigma^2 \frac{\gamma}{1-\gamma}$.

Finally, we show how to prove the tight finite sample rate using more direct methods. In fact, we can use the higher order moments of the inverse Wishart distribution [49] to obtain the exact closed-form expressions for both the mean and variance of $L(\hat{w}_{\text{OLS}})$ with any finite value of n and d.

Theorem 13. *Under the model assumptions in* (1) *with $d \leq n$, consider the ordinary least square estimator $\hat{w}_{\text{OLS}} = (X^T X)^{-1} X^T Y$. It holds that*

$$\begin{aligned} \mathbb{E}\,L(\hat{w}_{\text{OLS}}) &= \sigma^2 \frac{n-1}{n-d-1} \\ \operatorname{Var}(L(\hat{w}_{\text{OLS}})) &= 2\sigma^4 \frac{d(n-1)}{(n-d-1)^2(n-d-3)} \end{aligned} \tag{49}$$

Hence as $d/n \to \gamma$, it holds that

$$\mathbb{E}\,L(\hat{w}_{\text{OLS}}) \to \frac{\sigma^2}{1-\gamma} \quad \textit{and} \quad \frac{n}{\sigma^4}\operatorname{Var}(L(\hat{w}_{\text{OLS}})) \to \frac{2\gamma}{(1-\gamma)^3}. \tag{50}$$

If d is held constant, as $n \to \infty$, we have

$$n\,\mathbb{E}[L(\hat{w}_{\mathrm{OLS}}) - \sigma^2] \to \sigma^2 d \quad \textit{and} \quad \frac{n^2}{\sigma^4}\operatorname{Var}(L(\hat{w}_{\mathrm{OLS}})) \to 2d. \tag{51}$$

We can also show a matching high probability version of Theorem 13 based on the Gaussian minimax theorem:

Theorem 14. *Under the model assumptions in* (1) *with $d \le n$, consider the ordinary least square estimator $\hat{w}_{\mathrm{OLS}} = (X^T X)^{-1} X^T Y$ and denote $\gamma = d/n$. Assume that $\gamma \le 0.999$, then with probability at least $1 - \delta$, it holds that*

$$L(\hat{w}_{\mathrm{OLS}}) - \frac{\sigma^2}{1-\gamma} \lesssim \sigma^2 \sqrt{\frac{\gamma \log(36/\delta)}{n}}.$$

The full proof can be found in Appendix E. As we can see from Theorem 13, the variance of $L(\hat{w}_{\mathrm{OLS}})$ is of order $O(1/\sqrt{n})$ when d is proportional to n, and of order $O(1/n)$ when d is fixed. In both cases, the expectation is close to $\sigma^2/(1-\gamma)$. Theorem 14 shows exactly this and interpolates the two regimes: when γ is of constant order, then we recover the $O(1/\sqrt{n})$ rate, but when d is fixed, $\gamma = O(1/n)$ and so we can accelerate the convergence rate to $O(1/n)$.

Remark 6. Hastie et al. [18] provide a similar expectation calculation. On one hand, their results are more general in the sense that they do not assume the data is Gaussian, although the data is "almost Gaussian" because they require the existence of high-order moments. On the other hand, their results are asymptotic because their proof relies on the Marchenko-Pastur law and requires proportional scaling. In contrast, we obtain finite-sample bounds. After posting an initial preprint of this work, we learned that [15] has independently obtained a result equivalent to Theorem 13, also using the moments of the inverse Wishart distribution.

7 DISCUSSION

In this work, we push the limit of what bounds with an optimistic rate can do. At least for well-specified linear regression with Gaussian data, we see that they are flexible enough to simultaneously understand interpolation learning and recover many classical results from compressed sensing, high dimensional statistics and learning theory. In the context of benign overfitting, not only can we establish the consistency of the minimal norm interpolator, we actually show that any predictor with a sufficiently low norm and training error can achieve consistency. In a variety of applications, we use our main theorem to obtain bounds with very sharp constants and our general theory suggests that we can always get a nearly optimal analysis for ERM in any convex set by choosing the complexity functional F in Theorem 1 based on local Gaussian width.

A natural next step will be to relax the Gaussian assumption in our model (1) and also to consider situations where our linear model is misspecified in the sense that the Bayes optimal predictor is not linear. One of the key advantages of past works on uniform convergence, including the optimistic rate bound of Srebro et al. [39], is that they do not need to make strong parameteric assumptions on the data distribution. Though the Gaussian width formulation of optimistic rate bounds, as in (8), seems to crucially depend on the data being Gaussian, the connection to Rademacher complexity gives us hope that a version of our theory might apply to non-Gaussian data. (Some care must be taken in precisely formulating such a bound, due to the negative results discussed by Foygel and Srebro [16], Srebro et al. [39].) We also think that extending our results to generalized linear models, such as analyzing benign overfitting in linear classification, is an interesting direction. At least when the features are Gaussian, our techniques should be applicable; we leave this to future work.

REFERENCES

[1] Dennis Amelunxen, Martin Lotz, Michael B McCoy, and Joel A Tropp. 2014. Living on the edge: Phase transitions in convex programs with random data. *Information and Inference: A Journal of the IMA* 3, 3 (2014), 224–294.

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

[3] Peter L. Bartlett and Philip M. Long. 2020. Failures of model-dependent generalization bounds for least-norm interpolation. arXiv:2010.08479

[4] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. 2020. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences* 117, 48 (2020), 30063–30070. arXiv:1906.11300

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

[6] Peter L. Bartlett and Shahar Mendelson. 2006. Empirical minimization. *Probability theory and related fields* 135, 3 (2006), 311–334.

[7] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. 2019. Reconciling modern machine learning practice and the bias-variance trade-off. *Proceedings of the National Academy of Sciences* 116, 32 (2019), 15849–15854. arXiv:1812.11118

[8] Mikhail Belkin, Daniel Hsu, and Ji Xu. 2020. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science* 2, 4 (2020), 1167–1180. arXiv:1903.07571

[9] Mikhail Belkin, Daniel J. Hsu, and Partha Mitra. 2018. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems*. arXiv:1806.05161

[10] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. 2009. Simultaneous analysis of Lasso and Dantzig selector. *The Annals of statistics* 37, 4 (2009), 1705–1732.

[11] Venkat Chandrasekaran, Benjamin Recht, Pablo A Parrilo, and Alan S Willsky. 2012. The convex geometry of linear inverse problems. *Foundations of Computational mathematics* 12, 6 (2012), 805–849.

[12] Niladri S. Chatterji and Philip M. Long. 2021. Foolish Crowds Support Benign Overfitting. arXiv:2110.02914

[13] Geoffrey Chinot, Matthias Löffler, and Sara van de Geer. 2020. On the robustness of minimum-norm interpolators. *arXiv preprint arXiv:2012.00807* (2020).

[14] Zeyu Deng, Abla Kammoun, and Christos Thrampoulidis. 2021. A model of double descent for high-dimensional binary linear classification. *Information and Inference: A Journal of the IMA* (2021). arXiv:1911.05822

[15] Karthik Duraisamy. 2021. Basic'Generalization Error Bounds for Least Squares Regression with Well-specified Models. *arXiv preprint arXiv:2109.09647* (2021).

[16] Rina Foygel and Nathan Srebro. 2011. Concentration-based guarantees for low-rank matrix reconstruction. In *Proceedings of the 24th Annual Conference on Learning Theory*. JMLR Workshop and Conference Proceedings, 315–340.

[17] Yehoram Gordon. 1985. Some inequalities for Gaussian processes and applications. *Israel Journal of Mathematics* 50, 4 (1985), 265–289.

[18] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. 2019. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics* (2019). arXiv:1903.08560

[19] Peizhong Ju, Xiaojun Lin, and Jia Liu. 2020. Overfitting Can Be Harmless for Basis Pursuit: Only to a Degree. In *Advances in Neural Information Processing Systems*. arXiv:2002.00492

[20] Jonathan Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. 2021. On the Power of Preconditioning in Sparse Linear Regression. arXiv:2106.09207

[21] Frederic Koehler, Lijia Zhou, Danica J. Sutherland, and Nathan Srebro. 2021. Uniform Convergence of Interpolators: Gaussian Width, Norm Bounds and Benign Overfitting. In *Advances in Neural Information Processing Systems*. arXiv:2106.09276

[22] Guillaume Lecué and Shahar Mendelson. 2013. Learning subgaussian classes: Upper and minimax bounds. arXiv:1305.4825

[23] Yue Li and Yuting Wei. 2021. Minimum \ell_ {1}-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502* (2021).

[24] Tengyuan Liang and Pragya Sur. 2020. A Precise High-Dimensional Asymptotic Theory for Boosting and Minimum-L1-Norm Interpolated Classifiers. arXiv:2002.01586

[25] Shahar Mendelson. 2003. On the performance of kernel classes. *Journal of Machine Learning Research* (2003).

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

[27] Andrea Montanari, Feng Ruan, Youngtak Sohn, and Jun Yan. 2019. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. arXiv:1911.01544

[28] Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. 2020. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory* (2020). arXiv:1903.09139

[29] Vaishnavh Nagarajan and J. Zico Kolter. 2019. Uniform convergence may be unable to explain generalization in deep learning. In *Advances in Neural Information Processing Systems*. arXiv:1902.04742

[30] Jeffrey Negrea, Gintare Karolina Dziugaite, and Daniel M. Roy. 2020. In Defense of Uniform Convergence: Generalization via derandomization with an application to interpolating predictors. In *International Conference on Machine Learning*. arXiv:1912.04265

[31] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. 2015. In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. In *International Conference on Learning Representations – Workshop*. arXiv:1412.6614

[32] Samet Oymak and Babak Hassibi. 2010. New null space results and recovery thresholds for matrix rank minimization. arXiv:1011.6326

[33] Samet Oymak and Joel A Tropp. 2018. Universality laws for randomized dimension reduction, with applications. *Information and Inference: A Journal of the IMA* 7, 3 (2018), 337–446.

[34] Dmitriy Panchenko. 2002. Some Extensions of an Inequality of Vapnik and Chervonenkis. *Electronic Communications in Probability* 7 (2002), 55–65. arXiv:0405342

[35] Garvesh Raskutti, Martin J Wainwright, and Bin Yu. 2010. Restricted eigenvalue properties for correlated Gaussian designs. *The Journal of Machine Learning Research* 11 (2010), 2241–2259.

[36] Phillippe Rigollet and Jan-Christian Hütter. 2015. High dimensional statistics. *Lecture notes for course 18S997* 813 (2015), 814.

[37] R Tyrrell Rockafellar. 1997. *Convex analysis*. Vol. 11. Princeton university press.

[38] Mark Rudelson and Roman Vershynin. 2008. On sparse reconstruction from Fourier and Gaussian measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences* 61, 8 (2008), 1025–1045.

[39] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. 2010. Optimistic Rates for Learning with a Smooth Loss. arXiv:1009.3896

[40] Mihailo Stojnic. 2013. A framework to characterize performance of LASSO algorithms. arXiv:1303.7291

[41] Christos Thrampoulidis, Ehsan Abbasi, and Babak Hassibi. 2018. Precise error analysis of regularized M-estimators in high dimensions. *IEEE Transactions on Information Theory* 64, 8 (2018), 5592–5628.

[42] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. 2014. The Gaussian min-max theorem in the presence of convexity. arXiv:1408.4837

[43] Christos Thrampoulidis, Samet Oymak, and Babak Hassibi. 2015. Regularized linear regression: A precise analysis of the estimation error. In *Conference on Learning Theory*.

[44] Alexander Tsigler and Peter L. Bartlett. 2020. Benign overfitting in ridge regression. (2020). arXiv:2009.14286

[45] Sara A Van De Geer and Peter Bühlmann. 2009. On the conditions used to prove oracle results for the Lasso. *Electronic Journal of Statistics* 3 (2009), 1360–1392.

[46] Ramon van Handel. 2014. Probability in High Dimension. Lecture notes, Princeton University. https://web.math.princeton.edu/~rvan/APC550.pdf

[47] Vladimir Vapnik. 1982. *Estimation of dependences based on empirical data*. Springer Science & Business Media.

[48] Roman Vershynin. 2018. *High-dimensional probability: An introduction with applications in data science*. Vol. 47. Cambridge university press.

[49] Dietrich von Rosen. 1988. Moments for the Inverted Wishart Distribution. *Scandinavian Journal of Statistics* 15, 2 (1988), 97–109.

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

[51] Guillaume Wang, Konstantin Donhauser, and Fanny Yang. 2021. Tight bounds for minimum l1-norm interpolation of noisy data. *arXiv preprint arXiv:2111.05987* (2021).

[52] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. 2017. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*. arXiv:1611.03530

[53] Tong Zhang. 2002. Effective dimension and generalization of kernel learning. In *Advances in Neural Information Processing Systems*, Vol. 4. 454–461.

[54] Lijia Zhou, Danica J. Sutherland, and Nathan Srebro. 2020. On Uniform Convergence and Low-Norm Interpolation Learning. In *Advances in Neural Information Processing Systems*. arXiv:2006.05942

A PRELIMINARIES

Concentration of Lipschitz functions. Recall that a function $f : \mathbb{R}^n \to \mathbb{R}$ is L-Lipschitz with respect to the norm $\|\cdot\|$ if it holds for all $x, y \in \mathbb{R}^n$ that $|f(x) - f(y)| \le L\|x - y\|$. We use the concentration of Lipschitz functions of a Gaussian.

Theorem 15 ([46], Theorem 3.25). *If f is L-Lipschitz with respect to the Euclidean norm and $Z \sim N(0, I_n)$, then*

$$\Pr(|f(Z) - \mathbb{E} f(Z)| \ge t) \le 2e^{-t^2/2L^2}. \tag{52}$$

The proof of the following results can be found in Koehler et al. [21].

Lemma 2. *Suppose that $Z \sim N(0, I_n)$. Then*

$$\Pr(\big|\|Z\|_2 - \sqrt{n}\big| \ge t) \le 4e^{-t^2/4}. \tag{53}$$

Lemma 3. *Suppose that S is a fixed subspace of dimension d in $\mathbb{R}^n$ with $n \ge 4$, P_S is the orthogonal projection onto S, and V is a spherically symmetric random vector (i.e. $V/\|V\|_2$ is uniform on the sphere). Then*

$$\frac{\|P_S V\|_2}{\|V\|_2} \le \sqrt{d/n} + 2\sqrt{\log(2/\delta)/n}. \tag{54}$$

with probability at least $1 - \delta$. Conditional on this inequality holding, we therefore have uniformly for all $s \in S$ that

$$|\langle s, V\rangle| = |\langle s, P_S V\rangle| \le \|s\|_2 \|P_S V\|_2 \le \|s\|_2 \|V\|_2 \left(\sqrt{d/n} + 2\sqrt{\log(2/\delta)/n}\right). \tag{55}$$

Theorem 16 ((Convex) Gaussian Minmax Theorem; [17, 43]). *Let $Z : n \times d$ be a matrix with i.i.d. $N(0, 1)$ entries and suppose $G \sim N(0, I_n)$ and $H \sim N(0, I_d)$ are independent of Z and each other. Let S_w, S_u be compact sets and $\psi : S_w \times S_u \to \mathbb{R}$ be an arbitrary continuous function. Define the* Primary Optimization (PO) *problem*

$$\Phi(Z) := \min_{w \in S_w} \max_{u \in S_u} \langle u, Zw\rangle + \psi(w, u) \tag{56}$$

and the Auxiliary Optimization (AO) *problem*

$$\phi(G, H) := \min_{w \in S_w} \max_{u \in S_u} \|w\|_2 \langle G, u\rangle + \|u\|_2 \langle H, w\rangle + \psi(w, u). \tag{57}$$

Under these assumptions, $\Pr(\Phi(Z) < c) \le 2\Pr(\phi(G, H) \le c)$ *for any $c \in \mathbb{R}$.*

Furthermore, if we suppose that S_w, S_u are convex sets and $\psi(w, u)$ is convex in w and concave in u, then $\Pr(\Phi(Z) > c) \le 2\Pr(\phi(G, H) \ge c)$.

B PROOFS FOR SECTION 3

B.1 Proof of Theorem 1

To apply the Gaussian Minimax Theorem, we first formulate the quantity of interest as an optimization problem in terms of a random matrix with $N(0, 1)$ entries.

Lemma 4. *Under the model assumptions in* (1), *let F be an arbitrary function and β be any positive real number. Define the primary optimization problem (PO) as*

$$\Phi = \sup_{w} \inf_{\|\lambda\|_2 = 1} \langle Zw, \lambda\rangle + \sqrt{\frac{1}{1+\beta}\left(n\sigma^2 + n\|w\|_2^2\right)} - \langle \xi, \lambda\rangle - F(\Sigma^{-1/2} w + w^*) \tag{58}$$

where Z is an $n \times d$ random matrix with i.i.d. standard normal entries independent of ξ and each other. Then it holds that

$$\sup_w \sqrt{\frac{1}{1+\beta} \cdot L(w)} - \left(\sqrt{\hat{L}(w)} + \frac{F(w)}{\sqrt{n}} \right) \overset{\mathcal{D}}{=} \frac{1}{\sqrt{n}} \Phi \tag{59}$$

Proof. By our definition of population and empirical loss, we have

$$\begin{aligned}
&\sup_w \sqrt{\frac{1}{1+\beta} \cdot L(w)} - \left(\sqrt{\hat{L}(w)} + \frac{F(w)}{\sqrt{n}} \right) \\
&= \sup_w \sqrt{\frac{1}{1+\beta} \left(\sigma^2 + \|w - w^*\|_\Sigma^2 \right)} - \left(\frac{1}{\sqrt{n}} \|Y - Xw\|_2 + \frac{F(w)}{\sqrt{n}} \right) \\
&= \sup_w \inf_{\|\lambda\|_2 = 1} \sqrt{\frac{1}{1+\beta} \left(\sigma^2 + \|w - w^*\|_\Sigma^2 \right)} - \left(\frac{1}{\sqrt{n}} \langle Y - Xw, \lambda \rangle + \frac{F(w)}{\sqrt{n}} \right)
\end{aligned}$$

By equality in distribution, we can write $X = Z\Sigma^{1/2}$. Using a change of variables, the above becomes

$$\begin{aligned}
&\sup_w \inf_{\|\lambda\|_2 = 1} \sqrt{\frac{1}{1+\beta} \left(\sigma^2 + \|w\|_2^2 \right)} - \left(\frac{1}{\sqrt{n}} \langle \xi - Zw, \lambda \rangle + \frac{F(\Sigma^{-1/2} w + w^*)}{\sqrt{n}} \right) \\
&= \frac{1}{\sqrt{n}} \sup_w \inf_{\|\lambda\|_2 = 1} \langle Zw, \lambda \rangle + \sqrt{\frac{1}{1+\beta} \left(n\sigma^2 + n \|w\|_2^2 \right)} - \langle \xi, \lambda \rangle - F(\Sigma^{-1/2} w + w^*)
\end{aligned}$$

□

To apply Theorem 16, we will use a truncation argument. The following result is an exercise in real analysis, which we include for completeness.

Lemma 5. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be an arbitrary function, then it holds that*

$$\lim_{r \to \infty} \sup_{\|w\|_2 \le r} f(w) = \sup_w f(w) \tag{60}$$

Proof. We consider two cases:

(1) Suppose that $\sup_w f(w) = \infty$, then for any $M > 0$, there exists x_M such that $f(x_M) > M$. Hence for any $r > \|x_M\|_2$, it holds that

$$\sup_{\|w\|_2 \le r} f(w) > M \implies \liminf_{r \to \infty} \sup_{\|w\|_2 \le r} f(w) \ge M$$

As the choice of M is arbitrary, we have $\lim_{r \to \infty} \sup_{\|w\|_2 \le r} f(w) = \infty$ as desired.

(2) Suppose that $\sup_w f(w) = M < \infty$, then for any $\epsilon > 0$, there exists x_ϵ such that $f(x_\epsilon) > M - \epsilon$. Hence for any $r > \|x_\epsilon\|_2$, it holds that

$$\sup_{\|w\|_2 \le r} f(w) > M - \epsilon \implies \liminf_{r \to \infty} \sup_{\|w\|_2 \le r} f(w) \ge M - \epsilon$$

As the choice of ϵ is arbitrary, we have $\liminf_{r \to \infty} \sup_{\|w\|_2 \le r} f(w) \ge M$. On the other hand, it must be the case (by definition of supremum) that

$$\sup_{\|w\|_2 \le r} f(w) \le M \implies \limsup_{r \to \infty} \sup_{\|w\|_2 \le r} f(w) \le M$$

Consequently, the limit of $\sup_{\|w\|_2 \le r} f(w)$ exists and equals M. □

Lemma 6. *Let $G \sim N(0, I_n), H \sim N(0, I_d)$ be Gaussian vectors independent of Z, ξ and each other. Define the auxiliary problem (AO) as*

$$\phi = \sup_{w} \sqrt{\frac{1}{1+\beta}\left(n\sigma^2 + n\|w\|_2^2\right)} - \|G\|w\|_2 - \xi\|_2 + \langle H, w\rangle - F(\Sigma^{-1/2}w + w^*). \tag{61}$$

Suppose that F is continuous, then it holds that for any $t \in \mathbb{R}$

$$\Pr(\Phi > t \mid \xi) \leq 2\Pr(\phi \geq t \mid \xi), \tag{62}$$

and taking expectations we have

$$\Pr(\Phi > t) \leq 2\Pr(\phi \geq t). \tag{63}$$

Proof. First, by (58) define the truncated PO as

$$\Phi_r = \sup_{\|w\| \leq r} \inf_{\|\lambda\|_2 = 1} \langle Zw, \lambda\rangle + \sqrt{\frac{1}{1+\beta}\left(n\sigma^2 + n\|w\|_2^2\right)} - \langle \xi, \lambda\rangle - F(\Sigma^{-1/2}w + w^*), \tag{64}$$

and the corresponding AO is

$$\begin{aligned}\phi_r = \sup_{\|w\| \leq r} \inf_{\|\lambda\|_2 = 1} &\|w\|_2 \langle G, \lambda\rangle + \|\lambda\|_2 \langle H, w\rangle + \sqrt{\frac{1}{1+\beta}\left(n\sigma^2 + n\|w\|_2^2\right)} \\ &- \langle \xi, \lambda\rangle - F(\Sigma^{-1/2}w + w^*) \qquad (65)\\ = \sup_{\|w\| \leq r} &\langle H, w\rangle - \|G\|w\|_2 - \xi\|_2 + \sqrt{\frac{1}{1+\beta}\left(n\sigma^2 + n\|w\|_2^2\right)} - F(\Sigma^{-1/2}w + w^*).\end{aligned}$$

By Lemma 5, with probability one, we have Φ_r and ϕ_r monotonically increase to Φ and ϕ as $r \longrightarrow \infty$, respectively. By continuity of measure (from below), it holds that

$$\begin{aligned}\Pr(\Phi > t \mid \xi) &= \Pr\left(\lim_{r\to\infty} \Phi_r > t \mid \xi\right) \\ &\leq \Pr\left(\cup_{r\in\mathbb{N}} \cap_{R \geq r} \Phi_R > t \mid \xi\right) \\ &= \lim_{r\to\infty} \Pr\left(\cap_{R\geq r}\Phi_R > t \mid \xi\right) = \lim_{r\to\infty} \Pr\left(\Phi_r > t \mid \xi\right)\end{aligned}$$

By Theorem 16, it follows that

$$\Pr\left(\Phi_r > t \mid \xi\right) = \Pr\left(-\Phi_r < -t \mid \xi\right) \leq 2\Pr\left(-\phi_r < -t \mid \xi\right) \leq 2\Pr(\phi > t \mid \xi)$$

Plugging in the bound above yields the desired conclusion. □

Lemma 7. *Let F satisfy the condition in Theorem 1 and $n \geq 196\log(12/\delta)$, then there exists $\beta \leq 14\sqrt{\frac{\log(12/\delta)}{n}}$ such that*

$$\Pr(\phi \geq 0) \leq \delta' + \delta \tag{66}$$

Proof. For notational simplicity, define

$$\alpha := 2\sqrt{\frac{\log(12/\delta)}{n}}$$
$$\rho := \sqrt{\frac{1}{n}} + 2\sqrt{\frac{\log(6/\delta)}{n}}.$$

By a union bound, the following collection of events occur with probability at least $1 - \delta - \delta'$

(1) By Lemma 2, it holds that

$$1 - \alpha \leq \frac{1}{\sqrt{n}} \|G\|_2 \tag{67}$$

and

$$1 - \alpha \leq \frac{1}{\sqrt{n}\sigma} \|\xi\|_2 \tag{68}$$

(2) By Lemma 3, it holds that

$$\langle \xi, G \rangle \leq \rho \|\xi\|_2 \|G\|_2 \tag{69}$$

(3) By our assumption on F, it holds that uniformly over all $w \in \mathbb{R}^d$

$$\langle H, w \rangle \leq F(\Sigma^{-1/2} w + w^*) \tag{70}$$

Equations (67), (68) and (69) implies that

$$\begin{aligned} \|G \|w\|_2 - \xi\|_2^2 &\geq (1 - \rho) \left(\|G\|_2^2 \|w\|_2^2 + \|\xi\|_2^2 \right) \\ &\geq (1 - \rho)(1 - \alpha)^2 n \left(\|w\|_2^2 + \sigma^2 \right) \end{aligned}$$

Therefore, if we take $1 + \beta = (1 - \rho)^{-1}(1 - \alpha)^{-2}$, combining with (70) shows that $\phi \leq 0$. To simplify the expression of β, observe that

$$(1 - \rho)(1 - \alpha)^2 \geq 1 - 2\alpha - \rho.$$

Finally, it is routine to check that $\beta \leq 14\sqrt{\frac{\log(12/\delta)}{n}}$. □

Theorem 1. *Under the model assumption in* (1), *let* $F : \mathbb{R}^d \to [0, \infty]$ *be a function such that for* $x \sim N(0, \Sigma)$, *with probability at least* $1 - \delta'$, *it holds uniformly over all* $w \in \mathbb{R}^d$ *that*

$$\langle w - w^*, x \rangle \leq F(w). \tag{5}$$

For any $\delta > 0$, *assume* $n \geq 196 \log(12/\delta)$. *Then there exists* $\beta_1 \leq 14\sqrt{\frac{\log(12/\delta)}{n}}$ *such that with probability at least* $1 - 2(\delta' + \delta)$, *it holds uniformly over all* $w \in \mathbb{R}^d$ *that*

$$L(w) \leq (1 + \beta_1) \left(\sqrt{\hat{L}(w)} + \frac{F(w)}{\sqrt{n}} \right)^2. \tag{6}$$

Proof. First, we prove the result under the temporary additional assumption that F is continuous. By Lemma 4 and Lemma 6, we have

$$\begin{aligned} &\Pr\left(\exists w \in \mathbb{R}^d,\, L(w) > (1 + \beta) \left(\sqrt{\hat{L}(w)} + \frac{F(w)}{\sqrt{n}} \right)^2 \right) \\ = &\Pr\left(\sup_w \sqrt{\frac{1}{1 + \beta} \cdot L(w)} - \left(\sqrt{\hat{L}(w)} + \frac{F(w)}{\sqrt{n}} \right) > 0 \right) \\ = &\Pr(\Phi > 0) \leq 2 \Pr(\phi \geq 0) \end{aligned}$$

Then Lemma 7 shows that $\Pr(\phi \geq 0) \leq \delta' + \delta$ and so the desired event occurs with probability at least $1 - 2(\delta' + \delta)$.

Now we describe how to remove the extraneous assumption that F is continuous. Let $\delta'' > 0$ be arbitrary. With probability at least $1 - \delta''$ and using that x is equal in law to $\Sigma^{1/2} z$ for $z \sim N(0, I_d)$, we have that $\langle w - w^*, x \rangle \leq$

$\|w - w^*\|_\Sigma(\sqrt{d} + 2\sqrt{\log(4/\delta')})$ by Lemma 2. So by the union bound, with probability at least $1 - \delta' - \delta''$ we have that

$$\langle w - w^*, x\rangle \le \min\{F(w), \|w - w^*\|_\Sigma(\sqrt{d} + 2\sqrt{\log(4/\delta')})\}. \tag{71}$$

Let $F_{\delta''}$ be the greatest convex minorant of the right hand side of Equation (71). (As discussed below in Remark 7, the greatest convex minorant of a function f is the largest convex function which is everywhere at most f. It always exists since the supremum of convex functions is convex.) If (71) holds for all x then $\langle w - w^*, x\rangle \le F_{\delta''}(x)$ as well, by the defining property of the greatest convex minorant. By Corollary 10.1.1 of [37], $F_{\delta''}$ is continuous since it is convex and finite. Therefore, we can derive the desired bound with $F_{\delta''}$, which is no larger than F, and this holds with probability at least $1 - 2(\delta + \delta' + \delta'')$. Taking $\delta'' \to 0$ proves the desired result. □

Remark 7. In Theorem 1, if the assumption (5) is satisfied for a function F then it is also satisfied for its greatest convex minorant $\operatorname{conv}(F)$, which is the largest convex function such that $\operatorname{conv}(F)(w) \le F(w)$ for all w, and replacing F by $\operatorname{conv}(F)$ only makes the conclusion stronger. Also, we note the conclusion can be written in terms of the population measure μ and empirical measure μ_n from n samples as

$$\|Y - \langle w, X\rangle\|_{L_2(\mu)} \le (1+\beta)^{1/2}\left(\|Y - \langle w, X\rangle\|_{L_2(\mu_n)} + F(w)/\sqrt{n}\right)$$

so it can be interpreted as a lower isometry estimate for the empirical L_2 metric about the point Y.

B.2 Proof of Theorem 2

For convenience, we restate the theorem below:

Theorem 2. *Under the model assumptions in* (1), *let $\mathcal{K}$ be an arbitrary compact set, and take any covariance splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$. Fixing $\delta \le 1/4$, let $\beta_2 = 32\left(\sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\operatorname{rank}(\Sigma_1)}{n}}\right)$. If n is large enough that $\beta_2 \le 1$, then the following holds with probability at least $1-\delta$ for all $w \in \mathcal{K}$:*

$$L(w) \le (1+\beta_2)\left(\sqrt{\hat{L}(w)} + \frac{W_{\Sigma_2}(\mathcal{K})}{\sqrt{n}} + \left[\|w^*\|_{\Sigma_2} + \operatorname{rad}(\Sigma_2^{1/2}\mathcal{K})\right]\sqrt{\frac{2\log(32/\delta)}{n}}\right)^2. \tag{8}$$

Moreover, a stronger version of the above is also true: it holds that uniformly over all dilation factors $\alpha \ge 0$ and $w \in \alpha\mathcal{K}$, we have

$$L(w) \le (1+\beta_2)\left(\sqrt{\hat{L}(w)} + \frac{\alpha W_{\Sigma_2}(\mathcal{K})}{\sqrt{n}} + \left[\|w^*\|_{\Sigma_2} + \alpha\operatorname{rad}(\Sigma_2^{1/2}\mathcal{K})\right]\sqrt{\frac{2\log(32/\delta)}{n}}\right)^2. \tag{9}$$

Proof. First, we show how to choose the complexity function F in Theorem 1 and show the result without dilations. We can write $x = \Sigma^{1/2}H$ where $H \sim N(0, I_d)$. For any splitting $\Sigma = \Sigma_1 \oplus \Sigma_2$, let H_1 be the orthogonal projection of H onto the span of Σ_1. Similarly, we let H_2 be the orthogonal projection of H onto the span of Σ_2. Then observe that

$$\begin{aligned}\langle w^* - w,\, x\rangle &= \langle w^* - w,\, \Sigma_1^{1/2}H\rangle + \langle w^* - w,\, \Sigma_2^{1/2}H\rangle\\ &= \langle w^* - w,\, \Sigma_1^{1/2}H_1\rangle + \langle w^* - w,\, \Sigma_2^{1/2}H_2\rangle\\ &\le \|\Sigma_1^{1/2}(w - w^*)\|_2 \cdot \|H_1\|_2 + |\langle \Sigma_2^{1/2}w^*, H_2\rangle| + \sup_{w \in \Sigma_2^{1/2}\mathcal{K}} |\langle w, H_2\rangle|\end{aligned}$$

where the equality is by orthogonality of the split and the inequality is by Cauchy-Schwarz and the definition of supremum. Next, observe by Lemma 2 that with probability at least $1 - \delta/8$,

$$\|H_1\| \le \sqrt{\operatorname{rank}\Sigma_1} + 2\sqrt{\log(32/\delta)},$$

and by Theorem 15 with probability at least $1 - \delta/8$

$$\sup_{w \in \Sigma_2^{1/2}\mathcal{K}} |\langle w, H_2\rangle| \le W_{\Sigma_2}(\mathcal{K}) + \operatorname{rad}(\Sigma_2^{1/2}\mathcal{K})\sqrt{2\log(32/\delta)}$$

and by the standard Gaussian tail bound $\Pr_{Z\sim N(0,1)}(|Z| \ge t) \le 2e^{-t^2/2}$, it holds that

$$|\langle \Sigma_2^{1/2} w^*, H\rangle| \le \|w^*\|_{\Sigma_2}\sqrt{2\log(32/\delta)} \tag{72}$$

because the marginal law of $\langle \Sigma_2^{1/2} w^*, H\rangle$ is $N(0, \|w^*\|_{\Sigma_2}^2)$. Hence, by the union bound we have that with probability at least $1 - 3\delta/8$,

$$\langle w^* - w,\, x\rangle \le F(w) := \|w^* - w\|_{\Sigma_1}\left(\sqrt{\operatorname{rank}\Sigma_1} + 2\sqrt{\log(32/\delta)}\right) + W_{\Sigma_2}(\mathcal{K}) + [\|w^*\|_{\Sigma_2} + \operatorname{rad}(\Sigma_2^{1/2}\mathcal{K})]\sqrt{2\log(32/\delta)}.$$

Now applying Theorem 1 with $F(w) = \infty$ outside of $\mathcal{K}$ gives, where β_1 is as defined in the statement of that result,

$$\sqrt{\frac{L(w)}{1+\beta_1}} \le \sqrt{\hat{L}(w)} + \|w^* - w\|_{\Sigma_1}\left(\sqrt{\frac{\operatorname{rank}\Sigma_1}{n}} + 2\sqrt{\frac{\log(32/\delta)}{n}}\right) + \frac{W_{\Sigma_2}(\mathcal{K})}{\sqrt{n}} + [\|w^*\|_{\Sigma_2} + \operatorname{rad}(\Sigma_2^{1/2}\mathcal{K})]\sqrt{\frac{2\log(32/\delta)}{n}}.$$

Observe that $\|w^* - w\|_{\Sigma_1} \le \|w^* - w\|_{\Sigma} \le \sqrt{L(w)}$ so we have

$$\left((1+\beta_1)^{-1/2} - \sqrt{\frac{\operatorname{rank}\Sigma_1}{n}} - 2\sqrt{\frac{\log(32/\delta)}{n}}\right)\sqrt{L(w)} \le \sqrt{\hat{L}(w)} + \frac{W_{\Sigma_2}(\mathcal{K})}{\sqrt{n}} + [\|w^*\|_{\Sigma_2} + \operatorname{rad}(\Sigma_2^{1/2}\mathcal{K})]\sqrt{\frac{2\log(32/\delta)}{n}}$$

and by solving for $\sqrt{L(w)}$, we just need to consider β_2 such that

$$\left((1+\beta_1)^{-1/2} - \sqrt{\frac{\operatorname{rank}\Sigma_1}{n}} - 2\sqrt{\frac{\log(32/\delta)}{n}}\right)^{-2} \le 1 + \beta_2.$$

The above establishes the result when there is no dilation ($\alpha = 1$). Clearly, the same argument also shows the bound uniformly over all $\alpha \ge 0$ if we take

$$F(w) := \|w^* - w\|_{\Sigma_1}\left(\sqrt{\operatorname{rank}\Sigma_1} + 2\sqrt{\log(32/\delta)}\right) + \alpha(w) W_{\Sigma_2}(\mathcal{K}) + [\|w^*\|_{\Sigma_2} + \alpha(w)\operatorname{rad}(\Sigma_2^{1/2}\mathcal{K})]\sqrt{2\log(32/\delta)}$$

where $\alpha(w)$ is the infimum over all α such that $w \in \alpha\mathcal{K}$. □

B.3 Proof of Theorem 3

The following Lemma abstracts the key deterministic argument from the setting of Theorem 3 to essentially any application of Theorem 1; the key insight is that a generalization bound of the form (73) is exactly of the right form to explain flatness along the regularization path. Note that in the below Lemma, the function F is assumed to be convex which is always without loss of generality when applying Theorem 1, see Remark 7.

Lemma 8. *Suppose there exist a convex function F and $\epsilon \in (0,1)$ such that:*

(1) for all $w \in \mathbb{R}^d$, it holds that

$$\sqrt{L(w)} \le (1+\epsilon)\left(\sqrt{\hat{L}(w)} + \frac{F(w)}{\sqrt{n}}\right). \tag{73}$$

(2) ϵ is sufficiently large that

$$\sqrt{\hat{L}(w^*)} = \frac{\|\xi\|_2}{\sqrt{n}} \le (1+\epsilon)\sigma \quad \textit{and} \quad \frac{F(w^*)}{\sqrt{n}} \le \epsilon. \tag{74}$$

(3) for some $w' \in \mathbb{R}^d$, it holds that

$$\hat{L}(w') = 0 \quad \textit{and} \quad \frac{F(w')}{\sqrt{n}} \leq (1+\epsilon)\sigma + \epsilon. \tag{75}$$

Then for all R between $F(w^*)$ and $F(w')$ and any constrained empirical risk minimizer of the form

$$\hat{w}_R := \arg\min_{F(w) \leq R} \hat{L}(w),$$

we have $L(\hat{w}_R) \leq (\sigma + 5\epsilon(\sigma \vee 1))^2$.

Proof. For any R between $F(w^*)$ and $F(w')$, we can write

$$R = (1-\alpha)F(w^*) + \alpha F(w')$$

for some $\alpha \in [0,1]$. If we define $w_\alpha := (1-\alpha)w^* + \alpha w'$ accordingly, then by convexity, we have

$$F(w_\alpha) \leq (1-\alpha)F(w^*) + \alpha F(w') = R$$

and

$$\hat{L}(w_\alpha) = \frac{1}{n}\|Y - Xw_\alpha\|_2^2 = \frac{1}{n}(1-\alpha)^2\|Y - Xw^*\|_2^2 = (1-\alpha)^2\hat{L}(w^*).$$

By the definition of w_R, it must be the case that $\hat{L}(w_R) \leq \hat{L}(w_\alpha)$ and so by (73), (74) and (75)

$$\begin{aligned}
\sqrt{L(w_R)} &\leq (1+\epsilon)\left(\sqrt{\hat{L}(w_R)} + \frac{F(w_R)}{\sqrt{n}}\right) \\
&\leq (1+\epsilon)\left(\sqrt{\hat{L}(w_\alpha)} + \frac{R}{\sqrt{n}}\right) \\
&= (1+\epsilon)\left((1-\alpha)\sqrt{\hat{L}(w^*)} + \frac{(1-\alpha)F(w^*) + \alpha F(w')}{\sqrt{n}}\right) \\
&\leq (1+\epsilon)\left((1-\alpha)(1+\epsilon)\sigma + (1-\alpha)\epsilon + \alpha\left((1+\epsilon)\sigma + \epsilon\right)\right) \\
&= (1+\epsilon)^2\sigma + \epsilon(1+\epsilon) \\
&\leq \sigma + 5\epsilon(\sigma \vee 1).
\end{aligned}$$

□

Theorem 3. *Under the model assumptions in* (1)*, let $\|\cdot\|$ be an arbitrary norm on $\mathbb{R}^d$ and consider the complexity functional C_Σ and the constrained ERM $\hat{w}_R$ given by* (13) *and* (14)*. Suppose there is a split $\Sigma = \Sigma_1 \oplus \Sigma_2$ and $\epsilon > 0$ such that with probability at least $1-\delta$, it holds that*

$$\sqrt{\hat{L}(w^*)} \leq (1+\epsilon)\sigma \quad \textit{and} \quad C_{\Sigma_2}(\|w^*\|) \leq \epsilon \tag{15}$$

and there exists $w' \in \mathbb{R}^d$ such that

$$\hat{L}(w') = 0 \quad \textit{and} \quad C_{\Sigma_2}(\|w'\|) \leq (1+\epsilon)\sigma + \epsilon. \tag{16}$$

Then, with probability at least $1-2\delta$, it holds uniformly over any $R \geq \|w^*\|$ that

$$L(\hat{w}_R) \leq (\sigma + 5(\epsilon + \beta_2)(\sigma \vee 1))^2. \tag{17}$$

for the same choice of β_2 as in Theorem 2.

Proof. Notice that C_{Σ_2} is a monotone increasing function in $\|w\|$, so without loss of generality we can assume that w' is the minimal norm interpolator. By (9) of Theorem 2 and condition (16), we have

$$L(w') \le (1+\beta_2)\left((1+\epsilon)\sigma + \epsilon\right)^2$$

and it is easy to see this upper bound is no larger than the desired upper bound. By our convention, if $R > \|w'\|$ then $\hat{w}_R = w'$ and we are done. So we only need to consider the case when $\|w^*\| \le R \le \|w'\|$.

To apply Lemma 8, consider $F(w) = \sqrt{n}C_{\Sigma_2}(\|w\|)$ and let $\epsilon + \beta_2$ plays the role of ϵ. Clearly, (73), (74) and (75) are satisfied by Theorem 2 and our assumptions (15) and (16). Since C_{Σ_2} is monotone increasing, the condition that $\|w\| \le R$ is the same as $F(w) \le \sqrt{n}C_{\Sigma_2}(R)$ and $F(w^*) \le \sqrt{n}C_{\Sigma_2}(R) \le F(w')$. We can conclude the proof by a union bound. □

Corollary 2. *Let $\sigma > 0$ be fixed. Under the assumptions of Theorem 3 with $\|\cdot\|$ as the Euclidean norm, suppose that $\Sigma = \Sigma(n)$ is a sequence of covariance matrices with splits $\Sigma = \Sigma_1 \oplus \Sigma_2$ satisfying the benign overfitting conditions* (12). *Then it holds that*

$$\sup_{R \ge \|w^*\|_2} L(\hat{w}_R) \longrightarrow \sigma^2 \quad \textit{in probability.} \tag{18}$$

Proof. By Lemma 2, with probability at least $1-\delta/2$, we have $\sqrt{\hat{L}(w^*)} \le \left(1 + 2\sqrt{\frac{\log(8/\delta)}{n}}\right)\sigma$. Theorem 2 and 3 of Koehler et al. [21] shows that we can pick w' to be the minimal ℓ_2 norm interpolator, and there exists

$$\gamma \lesssim \sqrt{\frac{\log(1/\delta)}{n}} + \sqrt{\frac{\log(1/\delta)}{r(\Sigma_2)}} + \frac{n\log(1/\delta)}{R(\Sigma_2)}$$

such that with probability at least $1-\delta/2$, we have

$$C_{\Sigma_2}(\|w'\|_2) \le (1+\gamma)\left(\sigma + \|w^*\|_2\sqrt{\frac{\text{Tr}(\Sigma_2)}{n}}\right).$$

So we can take ϵ to be the maximum of $2\sqrt{\frac{\log(8/\delta)}{n}}$, $C_{\Sigma_2}(\|w^*\|_2)$, γ and $(1+\gamma)\|w^*\|_2\sqrt{\frac{\text{Tr}(\Sigma_2)}{n}}$. We can apply Theorem 3 and observe that $\epsilon + \beta_2 \longrightarrow 0$ under the benign overfitting conditions (12). □

C PROOFS FOR SECTION 4

C.1 Optimally-tuned regularized regression

Corollary 3. *Under the assumptions of Theorem 2, consider the regularized regression estimators $\hat{w}_\lambda$ as in* (20) *with an arbitrary norm $\|\cdot\|$. With probability at least $1-\delta$, there exists a $\lambda^* \ge 0$ such that*

$$L(\hat{w}_{\lambda^*}) \le (1+3\beta_2)\left(\sigma + \frac{\|w^*\|}{\sqrt{n}}\left(\mathop{\mathbb{E}}_{x\sim\mathcal{N}(0,\Sigma_2)}\|x\|_* + \sup_{\|u\|\le 1}\|u\|_{\Sigma_2}\cdot\sqrt{8\log(36/\delta)}\right)\right)^2. \tag{21}$$

Hence, we have $L(\hat{w}_{\lambda^*}) \longrightarrow \sigma^2$ in probability if

$$\frac{\text{rank}(\Sigma_1)}{n} \longrightarrow 0, \quad \frac{\|w^*\|\cdot\mathbb{E}_{x\sim\mathcal{N}(0,\Sigma_2)}\|x\|_*}{\sqrt{n}} \longrightarrow 0, \quad \textit{and} \quad \frac{\|w^*\|\cdot\sup_{\|u\|\le 1}\|u\|_{\Sigma_2}}{\sqrt{n}} \longrightarrow 0. \tag{22}$$

Proof. By comparing the KKT conditions, it is easy to see that there is some choice of λ^* such that

$$\hat{w}_{\lambda^*} = \mathop{\arg\min}_{\hat{L}(w)\le\|\xi\|^2/n} \|w\|.$$

Since $\hat{L}(w^*) = \|\xi\|^2/n$, it follows that $\|\hat{w}_{\lambda^*}\| \le \|w^*\|$. To apply Theorem 2, we consider $\mathcal{K} = \{w : \|w\| \le 1\}$ and observe that

$$\operatorname{rad}(\Sigma_2^{1/2}\mathcal{K}) = \sup_{\|w\|\le 1} \|w\|_{\Sigma_2} \quad \text{and} \quad \|w^*\|_{\Sigma_2} \le \|w^*\| \cdot \sup_{\|w\|\le 1} \|w\|_{\Sigma_2}.$$

Plugging in (9), by Lemma 2 and a union bound, we obtain

$$\begin{aligned} L(\hat{w}_{\lambda^*}) &\le (1+\beta_2)\left(\sqrt{\hat{L}(\hat{w}_{\lambda^*})} + \frac{\|\hat{w}_{\lambda^*}\| \cdot \mathbb{E}\,\|x\|_*}{\sqrt{n}} + \left[\|w^*\|_{\Sigma_2} + \|\hat{w}_{\lambda^*}\| \cdot \sup_{\|w\|\le 1}\|w\|_{\Sigma_2}\right]\sqrt{\frac{2\log(36/\delta)}{n}}\right)^2 \\ &\le (1+\beta_2)\left(\frac{\|\xi\|_2}{\sqrt{n}} + \frac{\|w^*\| \cdot \mathbb{E}\,\|x\|_*}{\sqrt{n}} + \|w^*\| \cdot \sup_{\|w\|\le 1}\|w\|_{\Sigma_2} \cdot \sqrt{\frac{8\log(36/\delta)}{n}}\right)^2 \\ &\le (1+\beta_2)\left(\left(1 + 2\sqrt{\frac{\log(36/\delta)}{n}}\right)\sigma + \frac{\|w^*\| \cdot \mathbb{E}\,\|x\|_*}{\sqrt{n}} + \|w^*\| \cdot \sup_{\|u\|\le 1}\|u\|_{\Sigma_2} \cdot \sqrt{\frac{8\log(36/\delta)}{n}}\right)^2 \end{aligned}$$

It is routine to check that $(1+\beta_2)\left(1 + 2\sqrt{\frac{\log(36/\delta)}{n}}\right)^2 \le 1 + 3\beta_2$ and the proof is complete. □

C.2 LASSO

Lemma 1. *Suppose w^* is k-sparse, i.e. supported on coordinate set $S \subset [d]$ with $|S| \le k$. Every w with $\|w\|_1 \le \|w^*\|_1$ satisfies*

$$\|(w - w^*)_{S^C}\|_1 \le \|(w^* - w_S)\|_1. \tag{25}$$

Proof. Note that over this set, we have

$$\|(w - w^*)_{S^C}\|_1 = \|w_{S^C}\|_1 = \|w\|_1 - \|w_S\|_1 \le \|w^*\|_1 - \|w_S\|_1 \le \|(w^* - w_S)\|_1$$

where the first inequality uses $\|w\|_1 \le \|w^*\|_1$ and the second inequality is the triangle inequality. □

Theorem 4. *Under the model assumptions in* (1), *additionally assume that:*

(1) w^* is a k-sparse vector.
(2) For $S \subset [d]$ the support of w^*, the covariance matrix Σ satisfies the S-compatibility condition.
(3) The number of samples n satisfies

$$n > \frac{32 \max_i \Sigma_{ii}}{\phi^2(\Sigma, S)} \cdot k \log\left(\frac{32d}{\delta}\right).$$

Then, for all w satisfying $\|w\|_1 \le \|w^*\|_1$ and $\hat{L}(w) \le (1+\epsilon)\sigma^2$ for an arbitrary ϵ, we have

$$L(w) - \sigma^2 \lesssim (\beta_1 + \epsilon)\sigma^2 + (1+\epsilon)\,\frac{\max_i \Sigma_{ii}}{\phi(\Sigma, S)^2} \cdot \frac{\sigma^2 k \log(32d/\delta)}{n}, \tag{27}$$

where $\beta_1 = O(\sqrt{\log(1/\delta)/n})$ is as defined in Theorem 1. In particular, when $\sigma = 0$ we have that $\|w - w^*\|_\Sigma = 0$, and so if Σ is positive definite then we have $w = w^*$ (exact recovery).

Proof. We start with the application of Theorem 1 as in Example 2. Observe that for $x \sim N(0, \Sigma)$ we have by Lemma 1, the compatibility condition, the standard Gaussian tail bound and the union bound that with probability at least $1 - \delta/8$,

$$\begin{aligned} \langle w - w^*, x \rangle &\le \|w - w^*\|_1 \|x\|_\infty \le 2\|(w - w^*)_S\|_1 \|x\|_\infty \\ &\le 2\frac{k^{1/2}\|w - w^*\|_\Sigma}{\phi(\Sigma, S)} \max_i \sqrt{2\Sigma_{ii}\log(16d/\delta)} \end{aligned} \tag{76}$$

so applying Theorem 1 with $F(w)$ equal to the right hand side of (76) gives

$$\begin{aligned} \sigma^2 + \|w - w^*\|_\Sigma^2 = L(w) &\le (1 + \beta_1)\left(\sqrt{\hat{L}(w)} + \frac{2k^{1/2}}{\phi(\Sigma, S)}\|w - w^*\|_\Sigma \max_i \sqrt{2\Sigma_{ii}\log(32d/\delta)/n}\right)^2 \\ &\le (1 + \beta_1)\left(\sigma\sqrt{1 + \epsilon} + \frac{2k^{1/2}}{\phi(\Sigma, S)}\|w - w^*\|_\Sigma \max_i \sqrt{2\Sigma_{ii}\log(32d/\delta)/n}\right)^2 \end{aligned}$$

For a sufficiently large n, we have $\beta_1 \le 1$. Expanding the square and rearranging gives

$$\begin{aligned} \|w - w^*\|_\Sigma^2 \le [\beta_1 + \epsilon + \epsilon\beta_1]\sigma^2 &+ 8\sigma\frac{\sqrt{k(1+\epsilon)}}{\phi(\Sigma, S)}\|w - w^*\|_\Sigma \max_i \sqrt{2\Sigma_{ii}\log(32d/\delta)/n} \\ &+ \frac{16k \max_i \Sigma_{ii}\log(32d/\delta)}{\phi(\Sigma, S)^2} \cdot \frac{\|w - w^*\|_\Sigma^2}{n} \end{aligned}$$

and using the assumption on n to rearrange the last term gives

$$\begin{aligned} \|w - w^*\|_\Sigma^2 &\le 2[\beta_1 + \epsilon + \epsilon\beta_1]\sigma^2 + 16\sigma\frac{\sqrt{k(1+\epsilon)}}{\phi(\Sigma, S)}\|w - w^*\|_\Sigma \max_i \sqrt{2\Sigma_{ii}\log(32d/\delta)/n} \\ &\le 4[\beta_1 + \epsilon]\sigma^2 + \sqrt{\frac{512\sigma^2 k(1+\epsilon)\max_i \Sigma_{ii}\log(32d/\delta)}{\phi(\Sigma, S)^2 n}} \cdot \|w - w^*\|_\Sigma. \end{aligned}$$

Solving this quadratic equation, it is not to difficult to check that

$$\|w - w^*\|_\Sigma^2 \le 8[\beta_1 + \epsilon]\sigma^2 + \frac{512(1+\epsilon)\max_i \Sigma_{ii}}{\phi(\Sigma, S)^2}\frac{\sigma^2 k \log(32d/\delta)}{n}$$

which is the desired result. □

Remark 8 (Generalization Bound for Larger Cones)**.** For simplicity, in the above analysis we gave a generalization bound for predictors w satisfying $\|w\|_1 \le \|w^*\|_1$, or more generally $\|(w - w^*)_{S^C}\|_1 \le \|(w - w^*)_S\|_1$, which covers the case of the LASSO with oracle regularization commonly considered in the literature [see, e.g., 48]. In situations where adaptivity to the unknown value of $\|w^*\|_1$ is important, the relevant predictor w may only be guaranteed to satisfy the weaker bound $\|(w - w^*)_{S^C}\|_1 \le C\|(w - w^*)_S\|_1$ for some $C > 1$ and the analogous version of the compatibility condition/restricted eigenvalue condition over this cone is assumed [see, e.g., 10, 36, 45, 50]; adopting the analysis to predictors in this larger cone is straightforward and we omit the details.

C.3 OLS

The following training error bounds are standard, which we include for completeness.

Lemma 9. *Under the model assumptions in* (1) *with* $d \le n$, *consider the ordinary least square estimator* $\hat{w}_{\text{OLS}} = (X^T X)^{-1} X^T Y$. *With probability at least* $1 - \delta$, *it holds that*

$$\sqrt{\hat{L}(\hat{w}_{\text{OLS}})} \le \sigma\left(\sqrt{1 - \frac{d}{n}} + 2\sqrt{\frac{\log(4/\delta)}{n}}\right) \tag{77}$$

Similarly, with probability at least $1-\delta$*, it holds that*

$$\left\|\hat{w}_{\mathrm{OLS}} - w^*\right\|_{\hat{\Sigma}} \le \sigma\left(\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(4/\delta)}{n}}\right) \tag{78}$$

Proof. By our model assumptions, we can write $\hat{w}_{\mathrm{OLS}} = w^* + (X^TX)^{-1}X^T\xi$, and so $Y - X\hat{w}_{\mathrm{OLS}} = (I - X(X^TX)^{-1}X^T)\xi$. Since $(I - X(X^TX)^{-1}X^T)$ is almost surely an idempotent matrix with rank $n-d$, it follows that the distribution of

$$\frac{n\hat{L}(\hat{w}_{\mathrm{OLS}})}{\sigma^2} = \frac{1}{\sigma^2}\,\xi^T(I - X(X^TX)^{-1}X^T)\xi,$$

is a Chi-square distribution with $n-d$ degrees of freedom. By the same reasoning, the distribution of

$$\frac{n\,\|\hat{w}_{\mathrm{OLS}} - w^*\|_{\hat{\Sigma}}^2}{\sigma^2} = \frac{1}{\sigma^2}\,\xi^T X(X^TX)^{-1}X^T\xi$$

is a Chi-square distribution with d degrees of freedom. By Lemma 2, with probability at least $1-\delta$, it holds that

$$\frac{\sqrt{n}}{\sigma}\sqrt{\hat{L}(\hat{w}_{\mathrm{OLS}})} \le \sqrt{n-d} + 2\sqrt{\log(4/\delta)}.$$

Similarly, we have

$$\frac{\sqrt{n}}{\sigma}\left\|\hat{w}_{\mathrm{OLS}} - w^*\right\|_{\hat{\Sigma}} \le \sqrt{d} + 2\sqrt{\log(4/\delta)}.$$

Rearranging the terms conclude the proof. □

Theorem 5. *Under the model assumptions in* (1)*, let* $\gamma = d/n < 1$*. There exists some* $\epsilon \lesssim \left(\frac{\log(36/\delta)}{n}\right)^{1/2}$ *such that for all sufficiently large n, with probability* $1-\delta$ *it holds uniformly for all* $w \in \mathbb{R}^d$ *that*

$$\left|\sqrt{L(w) - \sigma^2} - \sqrt{\frac{\gamma\hat{L}(w)}{(1-\gamma)^2}}\right| \le \epsilon\sqrt{\hat{L}(w)} + \sqrt{\frac{1}{1-\gamma}\left(\frac{\hat{L}(w)}{1-\gamma} - \sigma^2\right) + \epsilon\hat{L}(w)}. \tag{29}$$

For the empirical risk minimizer $\hat{w}_{\mathrm{OLS}} = (X^TX)^{-1}X^TY$*, the right hand side of* (29) *is approximately zero because we also have*

$$\hat{L}(\hat{w}_{\mathrm{OLS}}) \le \sigma^2(1-\gamma) + \sigma^2\epsilon\sqrt{1-\gamma}. \tag{30}$$

Therefore, we obtain the following generalization bound:

$$L(\hat{w}_{\mathrm{OLS}}) - \frac{\sigma^2}{1-\gamma} \lesssim \sigma^2\left(\frac{\log(36/\delta)}{n}\right)^{1/4}. \tag{31}$$

Proof. By Lemma 2, we can pick

$$\begin{aligned} F(w) &= \left(\sqrt{d} + 2\sqrt{\log(4/\delta')}\right)\|\Sigma^{1/2}(w^* - w)\|_2 \\ &= \left(\sqrt{d} + 2\sqrt{\log(4/\delta')}\right)\sqrt{L(w) - \sigma^2}. \end{aligned}$$

Let $\delta' = \delta/9$ and replace δ by $\delta/3$ in Theorem 1, plug in the estimates from Lemma 9 using confidence level $\delta/9$, then by a union bound with $\gamma = \frac{d}{n}$ and $\epsilon = \sqrt{\frac{\log(36/\delta)}{n}}$, we have

$$\sqrt{\hat{L}(\hat{w}_{\mathrm{OLS}})} \le \sigma\sqrt{1-\gamma} + 2\sigma\epsilon \tag{79}$$

and the bound (6) becomes

$$L(w) \leq (1 + 14\epsilon)\left(\sqrt{\hat{L}(w)} + (\sqrt{\gamma} + 2\epsilon)\sqrt{L(w) - \sigma^2}\right)^2.$$

We can simplify this by expanding the square

$$(1 + 14\epsilon)^{-1}L(w) \leq \hat{L}(w) + (\sqrt{\gamma} + 2\epsilon)^2(L(w) - \sigma^2) + 2(\sqrt{\gamma} + 2\epsilon)\sqrt{\hat{L}(w)}\sqrt{L(w) - \sigma^2}.$$

Rearranging, we arrive at

$$\left[(1 + 14\epsilon)^{-1} - (\sqrt{\gamma} + 2\epsilon)^2\right](L(w) - \sigma^2) \leq \hat{L}(w) - (1 + 14\epsilon)^{-1}\sigma^2 + 2(\sqrt{\gamma} + 2\epsilon)\sqrt{\hat{L}(w)}\sqrt{L(w) - \sigma^2}.$$

Note that this is a quadratic equation in terms of $\sqrt{L(w) - \sigma^2}$

$$(L(w) - \sigma^2) - 2\frac{(\sqrt{\gamma} + 2\epsilon)\sqrt{\hat{L}(w)}}{(1 + 14\epsilon)^{-1} - (\sqrt{\gamma} + 2\epsilon)^2}\sqrt{L(w) - \sigma^2} \leq \frac{\hat{L}(w) - (1 + 14\epsilon)^{-1}\sigma^2}{(1 + 14\epsilon)^{-1} - (\sqrt{\gamma} + 2\epsilon)^2}.$$

We can complete the square, which leads to the following

$$\left[\sqrt{L(w) - \sigma^2} - \frac{(\sqrt{\gamma} + 2\epsilon)\sqrt{\hat{L}(w)}}{(1 + 14\epsilon)^{-1} - (\sqrt{\gamma} + 2\epsilon)^2}\right]^2 \leq \frac{(1 + 14\epsilon)^{-1}}{(1 + 14\epsilon)^{-1} - (\sqrt{\gamma} + 2\epsilon)^2}\left(\frac{\hat{L}(w)}{(1 + 14\epsilon)^{-1} - (\sqrt{\gamma} + 2\epsilon)^2} - \sigma^2\right)$$

Observe that $(1 + 14\epsilon)^{-1} - (\sqrt{\gamma} + 2\epsilon)^2 = 1 - \gamma - O(\epsilon)$ and so

$$\frac{\sqrt{\gamma}}{1 - \gamma} \leq \frac{(\sqrt{\gamma} + 2\epsilon)}{(1 + 14\epsilon)^{-1} - (\sqrt{\gamma} + 2\epsilon)^2} \leq \frac{\sqrt{\gamma}}{1 - \gamma} + O(\epsilon).$$

We can handle the other terms similarly. Plugging in (79) concludes the proof. □

C.4 Minimum-Norm Interpolation with Isotropic Covariance

Lemma 10. *Let w^*, w be arbitrary vectors with $w^* \neq 0$, let V be the (one-dimensional) span of w^*, and let P_V be the orthogonal projection onto V. Then for any vector x,*

$$\langle w - w^*, x\rangle \leq \|w - w^*\|_2 \cdot \|P_V x\|_2 + \|x\|_2\sqrt{\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2}}.$$

Proof. Observe that by expanding the square, we have

$$\|w - w^*\|_2^2 = \|w\|_2^2 + \|w^*\|_2^2 - 2\langle P_V w, w^*\rangle$$

and so rearranging gives the Parallelogram identity

$$\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2 = 2\langle P_V w, w^*\rangle.$$

Taking absolute value of both sides and using that $P_V w$ and w^* are colinear gives

$$\left|\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right| = 2\|P_V w\|_2\|w^*\|_2.$$

Combining this with the Pythagorean Theorem, we find

$$\|P_{V^\perp} w\|_2^2 = \|w\|_2^2 - \|P_V w\|_2^2 = \|w\|_2^2 - \left(\frac{\left|\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right|}{2\|w^*\|}\right)^2.$$

Thus, applying the Cauchy-Schwarz inequality and the above gives

$$\begin{aligned}
\langle w - w^*, x\rangle &= \langle P_V(w - w^*), x\rangle + \langle P_{V^\perp} w, x\rangle \\
&\le \langle P_V(w - w^*), x\rangle + \|P_{V^\perp} w\|_2 \|x\|_2 \\
&= \langle w - w^*, P_V x\rangle + \|x\|_2 \sqrt{\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2}} \\
&\le \|w - w^*\|_2 \cdot \|P_V x\|_2 + \|x\|_2 \sqrt{\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2}}.
\end{aligned}$$

which is the desired inequality. □

Lemma 11. *Under the assumptions of Theorem 1 with $\gamma = d/n > 1$ and the further assumption that the data has isotropic covariance $\Sigma = I_d$, there exists $\epsilon \lesssim \sqrt{\frac{\log(18/\delta)}{n}}$ such that with probability at least $1 - \delta$, we have*

$$\|w - w^*\|_2^2 + \sigma^2 \le (1 + \epsilon) \left(\sqrt{\hat{L}(w)} + \sqrt{\gamma} \cdot \sqrt{\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2}} \right)^2.$$

Proof. Observe that $\frac{\langle w^*, x\rangle}{\|w^*\|_2} \sim \mathcal{N}(0, 1)$ and so by a standard Gaussian tail bound, Lemma 2 and a union bound, with probability at least $1 - \delta$, it holds that

$$\|P_V x\|_2 = \left\| \frac{w^*(w^*)^T}{\|w^*\|_2^2} x \right\|_2 = \frac{|\langle w^*, x\rangle|}{\|w^*\|_2} \le \sqrt{2\log(6/\delta)}$$

and

$$\|x\|_2 \le \sqrt{d} + 2\sqrt{\log(6/\delta)}.$$

Combining Lemma 10 with Theorem 1 and another union bound gives

$$\begin{aligned}
&\frac{1}{\sqrt{1 + \beta_1}} \sqrt{\|w - w^*\|_2^2 + \sigma^2} \\
&\le \sqrt{\hat{L}(w)} + \|w - w^*\|_2 \sqrt{\frac{2\log(18/\delta)}{n}} + \left(\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(18/\delta)}{n}} \right) \sqrt{\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2}}.
\end{aligned}$$

Using the fact that $\|w - w^*\|_2 \le \sqrt{\|w - w^*\|_2^2 + \sigma^2}$ and $d > n$, we have

$$\begin{aligned}
&\left(1 + 2\sqrt{\frac{\log(18/\delta)}{n}}\right)^{-1} \left(\frac{1}{\sqrt{1 + \beta_1}} - \sqrt{\frac{2\log(18/\delta)}{n}} \right) \sqrt{\|w - w^*\|_2^2 + \sigma^2} \\
&\le \sqrt{\hat{L}(w)} + \sqrt{\gamma} \cdot \sqrt{\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2}}.
\end{aligned}$$

To simplify, there exists $\epsilon \lesssim \sqrt{\frac{\log(18/\delta)}{n}}$ such that

$$\frac{1}{\sqrt{1+\epsilon}}\sqrt{\|w - w^*\|_2^2 + \sigma^2} \le \sqrt{\hat{L}(w)} + \sqrt{\gamma} \cdot \sqrt{\|w\|_2^2 - \frac{\left(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2\right)^2}{4\|w^*\|_2^2}}.$$

and rearranging concludes the proof. □

The generalization bound from Lemma 11 holds for all w; we now show what happens when we specialize it to interpolators.

Theorem 6. *Under the model assumptions in* (1) *with* $\gamma = d/n > 1$ *and* $\Sigma = I_d$*, there exists* $\epsilon \lesssim \left(\frac{\log(18/\delta)}{n}\right)^{1/2}$ *such that with probability at least* $1 - \delta$*, the following holds uniformly over all* w *such that* $\hat{L}(w) = 0$*:*

$$\left| L(w) - \left[\sigma^2 + \|w\|_2^2 + \left(1 - \frac{2}{(1+\epsilon)\gamma}\right)\|w^*\|_2^2\right]\right| \le 2\|w^*\|_2\sqrt{\left(1 - \frac{1}{\gamma}\right)\left(\|w\|_2^2 - \frac{\|w^*\|_2^2}{\gamma}\right) - \frac{\sigma^2}{\gamma} + 3\epsilon\|w\|_2^2}. \tag{32}$$

Proof. By Lemma 11, there exists some $\epsilon \lesssim \sqrt{\frac{\log(18/\delta)}{n}}$ such that with probability at least $1 - \delta$, for all w such that $\hat{L}(w) = 0$ it holds that

$$\begin{aligned}\|w - w^*\|_2^2 + \sigma^2 &\le (1+\epsilon)\gamma\left(\|w\|_2^2 - \frac{(\|w\|_2^2 + \|w^*\|_2^2 - \|w - w^*\|_2^2)^2}{4\|w^*\|_2^2}\right)\\ &= (1+\epsilon)\gamma\left(\|w\|_2^2 - \frac{(\|w\|_2^2 + \|w^*\|_2^2)^2 - 2(\|w\|_2^2 + \|w^*\|_2^2)\|w - w^*\|_2^2 + \|w - w^*\|_2^4}{4\|w^*\|_2^2}\right)\end{aligned}$$

Rearranging, we have

$$4\|w^*\|_2^2 \cdot \frac{\|w - w^*\|_2^2 + \sigma^2}{(1+\epsilon)\gamma} \le 4\|w^*\|_2^2 \cdot \|w\|^2 - (\|w\|_2^2 + \|w^*\|_2^2)^2 + 2(\|w\|_2^2 + \|w^*\|_2^2)\|w - w^*\|_2^2 - \|w - w^*\|_2^4$$

Grouping the terms with $\|w - w^*\|_2^2$, we see that

$$\|w - w^*\|_2^4 + \left(\frac{4\|w^*\|_2^2}{(1+\epsilon)\gamma} - 2(\|w\|_2^2 + \|w^*\|_2^2)\right) \cdot \|w - w^*\|_2^2 + 4\|w^*\|_2^2 \cdot \frac{\sigma^2}{(1+\epsilon)\gamma} \le 4\|w^*\|_2^2 \cdot \|w\|^2 - (\|w\|_2^2 + \|w^*\|_2^2)^2$$

which is equivalent to

$$\|w - w^*\|_2^4 - 2\left(\|w\|_2^2 + \left(1 - \frac{2}{(1+\epsilon)\gamma}\right)\|w^*\|_2^2\right) \cdot \|w - w^*\|_2^2 + (\|w\|_2^2 - \|w^*\|_2^2)^2 + 4\|w^*\|_2^2 \cdot \frac{\sigma^2}{(1+\epsilon)\gamma} \le 0.$$

To complete the square, we compute

$$\begin{aligned}&\left(\|w\|_2^2 + \left(1 - \frac{2}{(1+\epsilon)\gamma}\right)\|w^*\|_2^2\right)^2 - (\|w\|_2^2 - \|w^*\|_2^2)^2 - 4\|w^*\|_2^2 \cdot \frac{\sigma^2}{(1+\epsilon)\gamma}\\ =&\left(1 - \frac{2}{(1+\epsilon)\gamma}\right)^2\|w^*\|_2^4 + 2\left(1 - \frac{2}{(1+\epsilon)\gamma}\right)\|w\|_2^2\|w^*\|_2^2 - \|w^*\|_2^4 + 2\|w\|_2^2\|w^*\|_2^2 - 4\|w^*\|_2^2 \cdot \frac{\sigma^2}{(1+\epsilon)\gamma}\\ =&\left(\frac{4}{(1+\epsilon)^2\gamma^2} - \frac{4}{(1+\epsilon)\gamma}\right)\|w^*\|_2^4 + 4\left(1 - \frac{1}{(1+\epsilon)\gamma}\right)\|w\|_2^2\|w^*\|_2^2 - 4\|w^*\|_2^2 \cdot \frac{\sigma^2}{(1+\epsilon)\gamma}\\ =&4\|w^*\|_2^2\left[\left(1 - \frac{1}{(1+\epsilon)\gamma}\right)\left(\|w\|_2^2 - \frac{\|w^*\|_2^2}{(1+\epsilon)\gamma}\right) - \frac{\sigma^2}{(1+\epsilon)\gamma}\right]\\ \le&4\|w^*\|_2^2\left[\left(1 + \epsilon - \frac{1}{\gamma}\right)\left(\|w\|_2^2 + \epsilon\|w\|_2^2 - \frac{\|w^*\|_2^2}{\gamma}\right) - \frac{\sigma^2}{\gamma}\right]\end{aligned}$$

where in the last step we use $(1+\epsilon)^2 \ge 1$ and $\frac{\sigma^2(1+\epsilon)}{\gamma} \ge \frac{\sigma^2}{\gamma}$. To simplify, it is routine to check that

$$\left(1 + \epsilon - \frac{1}{\gamma}\right)\left(\|w\|_2^2 + \epsilon\|w\|_2^2 - \frac{\|w^*\|_2^2}{\gamma}\right) - \left(1 - \frac{1}{\gamma}\right)\left(\|w\|_2^2 - \frac{\|w^*\|_2^2}{\gamma}\right) \le 3\epsilon\|w\|_2^2$$

and so we can conclude that

$$\left| \|w - w^*\|_2^2 - \left[\|w\|_2^2 + \left(1 - \frac{2}{(1+\epsilon)\gamma}\right) \|w^*\|_2^2 \right] \right| \leq 2\|w^*\|_2 \sqrt{\left(1 - \frac{1}{\gamma}\right)\left(\|w\|_2^2 - \frac{\|w^*\|_2^2}{\gamma}\right) - \frac{\sigma^2}{\gamma} + 3\epsilon\|w\|_2^2}.$$

as desired. □

Theorem 7. *Under the model assumptions in* (1) *with* $\gamma = d/n > 1$ *and* $\Sigma = I_d$, *there exists* $\epsilon \lesssim \left(\frac{\log(40/\delta)}{n}\right)^{1/2}$ *such that with probability at least* $1 - \delta$, *it holds that*

$$\min_{w:Xw=Y} \|w\|_2^2 \leq (1+\epsilon)\left(\frac{\|w^*\|_2^2}{\gamma} + \frac{\sigma^2}{\gamma - 1}\right). \tag{33}$$

Thus, by Theorem 6, we have

$$L(\hat{w}) - \left[\left(1 - \frac{1}{\gamma}\right)\|w^*\|_2^2 + \sigma^2 \frac{\gamma}{\gamma - 1}\right] \leq \epsilon\left(\frac{\|w^*\|_2^2}{\gamma} + \frac{\sigma^2}{\gamma - 1}\right) + \|w^*\|_2 \sqrt{\epsilon\left(\frac{\|w^*\|_2^2}{\gamma} + \frac{\sigma^2}{\gamma - 1}\right)} \tag{34}$$

where $\hat{w}$ *is the minimal-*ℓ_2 *norm interpolator. If we fix* σ^2, γ *and* $\|w^*\|_2$, *then as* $n \to \infty$

$$L(\hat{w}) \to \left(1 - \frac{1}{\gamma}\right)\|w^*\|_2^2 + \sigma^2 \frac{\gamma}{\gamma - 1} \quad \textit{in probability.} \tag{35}$$

Proof. The proof strategy here follows the same lines as in Theorem 2 of Koehler et al. [21], but handles the w^* term more carefully. First, we introduce the Lagrangian and apply a change of variable

$$\begin{aligned} \min_{Xw=Y} \|w\|^2 &= \min_{w} \max_{\lambda} \langle \lambda, Xw - Y\rangle + \|w\|^2 \\ &= \min_{w} \max_{\lambda} \langle \lambda, Xw - \xi\rangle + \|w + w^*\|^2 \end{aligned}$$

To apply CGMT (Theorem 16), we need a double truncation argument. For any $r, t > 0$, introduce the following problem:

$$\Phi_r(t) = \min_{\|w+w^*\|^2 \leq 2t} \max_{\|\lambda\| \leq r} \langle \lambda, Xw - \xi\rangle + \|w + w^*\|^2. \tag{80}$$

We also introduce

$$\begin{aligned} \Phi(t) &= \min_{\|w+w^*\|^2 \leq 2t} \max_{\lambda} \langle \lambda, Xw - \xi\rangle + \|w + w^*\|^2 \\ &= \min_{\substack{Xw=\xi \\ \|w+w^*\|^2 \leq 2t}} \|w + w^*\|^2 \end{aligned} \tag{81}$$

and claim that $\Phi_r(t) \to \Phi(t)$ as $r \to \infty$. By definition, $\Phi_r(t) \leq \Phi_s(t)$ for $r \leq s$. We consider two cases:

(1) $\Phi(t) = \infty$, i.e. the minimization problem defining $\Phi(t)$ is infeasible. In this case, we know that for all $\|w+w^*\|^2 \leq 2t$

$$\|Xw - \xi\|_2 > 0.$$

By compactness, there exists $\mu = \mu(X, \xi) > 0$ (in particular, independent of r) such that

$$\|Xw - \xi\|_2 \geq \mu.$$

Therefore, considering λ along the direction of $Xw - \xi$ shows that

$$\Phi_r(t) = \min_{\|w+w^*\|^2 \leq 2t} \max_{\|\lambda\|_2 \leq r} \langle \lambda, Xw - \xi\rangle + \|w + w^*\|^2 \geq r\mu$$

so $\Phi_r(t) \to \infty$ as $r \to \infty$.

(2) Otherwise $\Phi(t) < \infty$, i.e. the minimization problem defining $\Phi(t)$ is feasible. In this case, we can let $w(r)$ be an arbitrary minimizer achieving the objective $\Phi_r(t)$ for each $r \geq 0$ by compactness. By compactness again, the sequence $\{w(r)\}_{r=1}^{\infty}$ at positive integer values of r has a subsequential limit $w(\infty)$ such that $\|w(\infty) + w^*\| \leq 2t$. Equivalently, there exists an increasing sequence r_n such that $\lim_{n\to\infty} w(r_n) = w(\infty)$.

Suppose for the sake of contradiction that $Xw(\infty) \neq \xi$, then by continuity, there exists $\mu > 0$ and a sufficiently small $\epsilon > 0$ such that for all $\|w - w(\infty)\|_2 \leq \epsilon$

$$\|Xw - \xi\|_2 \geq \mu.$$

This implies that for sufficiently large n, we have

$$\|Xw(r_n) - \xi\|_2 \geq \mu$$

and by the same argument as in the previous case

$$\Phi_{r_n}(t) = \max_{\|\lambda\|_2 \leq r} \langle \lambda, Xw(r_n) - \xi\rangle + \|w(r_n) + w^*\|^2 \geq r\mu$$

so $\Phi_{r_n} \to \infty$, but this is impossible since $\Phi_r(t) \leq \Phi(t) < \infty$. By contradiction, it must be the case that $Xw(\infty) = \xi$. By taking $\lambda = 0$ in the definition of $\Phi_r(t)$, we have

$$\Phi_{r_n}(t) \geq \|w(r_n) + w^*\|^2.$$

By continuity, we show that

$$\liminf_{n\to\infty} \Phi_{r_n}(t) \geq \lim_{n\to\infty} \|w(r_n) + w^*\|^2 = \|w(\infty) + w^*\|^2 \geq \Phi(t).$$

Since $\Phi_{r_n}(t) \leq \Phi(t)$, the limit of $\Phi_{r_n}(t)$ exists and equals $\Phi(t)$. We can conclude that $\lim_{r\to\infty} \Phi_r(t) = \Phi(t)$ because $\Phi_r(t)$ is an increasing function of r.

In both cases, we have $\Phi_r(t) \to \Phi(t)$ as $r \to \infty$. The auxiliary problem corresponding to $\Phi_r(t)$ is

$$\phi_r(t) = \min_{\|w+w^*\|^2 \leq 2t} \max_{\|\lambda\|_2 \leq r} \|\lambda\|\langle H, w\rangle + \|w\|\langle G, \lambda\rangle - \langle \lambda, \xi\rangle + \|w + w^*\|^2 \tag{82}$$

which is upper bounded by

$$\begin{aligned}\phi(t) &= \min_{\|w+w^*\|^2 \leq 2t} \max_{\lambda} \|\lambda\|\langle H, w\rangle + \|w\|\langle G, \lambda\rangle - \langle \lambda, \xi\rangle + \|w + w^*\|^2 \\ &= \min_{\substack{\langle H, w\rangle + \|G\|w\| - \xi\| \leq 0 \\ \|w+w^*\|^2 \leq 2t}} \|w + w^*\|^2.\end{aligned} \tag{83}$$

Applying CGMT and the fact that $\Phi_r(t)$ monotonically increases to $\Phi(t)$ almost surely, we can conclude

$$\begin{aligned}\Pr\left(\min_{Xw=Y} \|w\|^2 > t \mid \xi\right) &= \Pr\left(\Phi(t) > t \mid \xi\right) = \Pr\left(\lim_{r\to\infty} \Phi_r(t) > t \mid \xi\right) \\ &\leq \lim_{r\to\infty} \Pr\left(\Phi_r(t) > t \mid \xi\right) \\ &\leq 2 \cdot \lim_{r\to\infty} \Pr\left(\phi_r(t) > t \mid \xi\right) \\ &\leq 2 \cdot \Pr\left(\phi(t) > t \mid \xi\right) = 2 \cdot \Pr\left(\min_{\langle H, w\rangle + \|G\|w\| - \xi\| \leq 0} \|w + w^*\|^2 > t \mid \xi\right)\end{aligned}$$

By tower law, we have shown that

$$\Pr\left(\min_{Xw=Y}\|w\|^2 > t\right) \leq 2 \cdot \Pr\left(\min_{\|G\|w\|-\xi\| \leq \langle H,w\rangle} \|w+w^*\|^2 > t\right).$$

To upper bound the minimum, we consider w of the form $\alpha w^* + \beta PH$ where $P = I - \frac{w^*(w^*)^T}{\|w^*\|^2}$. For the simplicity of notation, define

$$\epsilon = 2\sqrt{\frac{\log(40/\delta)}{n}} \quad \text{and} \quad \rho = \sqrt{\frac{1}{n}} + 2\sqrt{\frac{\log(20/\delta)}{n}}.$$

By a union bound, the following collection of events occurs with probability at least $1-\delta/2$:

(1) By Lemma 3, it holds that

$$|\langle \xi, G\rangle| \leq \rho\|\xi\| \cdot \|G\|$$

(2) By Lemma 2, it holds that

$$(1-\epsilon)\sigma\sqrt{n} \leq \|\xi\| \leq (1+\epsilon)\sigma\sqrt{n}$$
$$(1-\epsilon)\sqrt{n} \leq \|G\| \leq (1+\epsilon)\sqrt{n}$$
$$\left(\sqrt{\frac{d-1}{n}} - \epsilon\right)\sqrt{n} \leq \|PH\| \leq \left(\sqrt{\frac{d-1}{n}} + \epsilon\right)\sqrt{n}$$

(3) By standard Gaussian tail bound, it holds that

$$|\langle H, w^*\rangle| \leq \|w^*\|\epsilon\sqrt{n}$$

The above bounds imply that

$$\begin{aligned}\|G\|w\| - \xi\|^2 &= \|G\|^2\|w\|^2 + \|\xi\|^2 - 2\|w\|\langle G, \xi\rangle \\ &\leq (1+\rho)(\|G\|^2\|w\|^2 + \|\xi\|^2) \\ &\leq (1+\rho)(1+\epsilon)^2 n(\|w\|^2 + \sigma^2).\end{aligned}$$

By orthogonality, observe that

$$\|w\|^2 = \alpha^2\|w^*\|^2 + \beta^2\|PH\|^2$$
$$\langle H, w\rangle = \alpha\langle H, w^*\rangle + \beta\|PH\|^2,$$

and so to ensure that $\|G\|w\| - \xi\| \leq \langle H, w\rangle$, we can choose β such that

$$(1+\rho)^{1/2}(1+\epsilon)\sqrt{n(\alpha^2\|w^*\|^2 + \beta^2\|PH\|^2 + \sigma^2)} + \alpha\|w^*\|\epsilon\sqrt{n} \leq \beta\|PH\|^2.$$

Note that it suffices to have

$$(1+\rho)^{1/2}(1+2\epsilon)\sqrt{n(\alpha^2\|w^*\|^2 + \beta^2\|PH\|^2 + \sigma^2)} \leq \beta\|PH\|^2$$

$$\iff \alpha^2\frac{\|w^*\|^2}{(1+\rho)^{-1}(1+2\epsilon)^{-2}\frac{\|PH\|^2}{n} - 1} + \frac{\sigma^2}{(1+\rho)^{-1}(1+2\epsilon)^{-2}\frac{\|PH\|^2}{n} - 1} \leq \beta^2\|PH\|^2$$

Again, by orthogonality, we have

$$\|w + w^*\|^2 = (1+\alpha)^2\|w^*\|^2 + \beta^2\|PH\|^2$$

and so

$$\begin{aligned}&\min_{\|G\|w\|-\xi\| \le \langle H, w\rangle} \|w + w^*\|^2 \\ \le & \frac{\sigma^2}{(1+\rho)^{-1}(1+2\epsilon)^{-2}\frac{\|PH\|^2}{n} - 1} + \min_{\alpha} (1+\alpha)^2\|w^*\|^2 + \alpha^2 \frac{\|w^*\|^2}{(1+\rho)^{-1}(1+2\epsilon)^{-2}\frac{\|PH\|^2}{n} - 1} \\ = & \frac{\sigma^2}{(1+\rho)^{-1}(1+2\epsilon)^{-2}\frac{\|PH\|^2}{n} - 1} + \frac{\|w^*\|^2}{(1+\rho)^{-1}(1+2\epsilon)^{-2}\frac{\|PH\|^2}{n}}\end{aligned}$$

Finally, we can plug in the high probability lower bound for $\|PH\|\sqrt{n}$ and the proof is complete after some routine calculations. □

C.5 LASSO with Isotropic Covariance

Theorem 8. *Using the notation of Theorem 5, we have with probability at least* $1-\delta$ *that for all* w *with* $\|w\|_1 \le \|w^*\|_1$,

$$\left|\sqrt{L(w) - \sigma^2} - \sqrt{\frac{\gamma \hat{L}(w)}{(1-\gamma)^2}}\right| \le \epsilon\sqrt{\hat{L}(w)} + \sqrt{\frac{1}{1-\gamma}\left(\frac{\hat{L}(w)}{1-\gamma} - \sigma^2\right) + \epsilon \hat{L}(w)} \tag{36}$$

provided $\gamma + 2\epsilon/\sqrt{n} < 1$, *where*

$$\mathcal{K}' := \{u : \exists \delta > 0, \|w^* + \delta u\|_1 \le \|w^*\|_1\} \quad \textit{and} \quad \gamma := \frac{1}{n} \cdot W(\mathcal{K}' \cap S^{n-1})^2.$$

Proof. We use that for $\mathcal{K}' := \{u : \exists \delta > 0, \|w^* + \delta u\|_1 \le \|w^*\|_1\}$

$$\langle w^* - w, x\rangle \le \|w^* - w\| \sup_{u \in \mathcal{K}' \cap S^{n-1}} \langle u, x\rangle$$

where S^{n-1} is the unit sphere. Recall that $\omega := W(\mathcal{K}' \cap S^{n-1})$ denotes the Gaussian width of the intersection of the tangent cone $\mathcal{K}'$ with the unit sphere. Let $\epsilon = \Theta\left(\frac{\log(36/\delta)}{n}\right)^{1/2}$ as in Theorem 5, then with this notation Theorem 1 gives

$$\sigma^2 + \|w^* - w\|_2^2 \le (1+\beta)\left(\sqrt{\hat{L}(w)} + \|w^* - w\|_2(\omega + 2\epsilon)/\sqrt{n}\right)^2 \le (1+14\epsilon)\left(\sqrt{\hat{L}(w)} + \|w^* - w\|_2(\omega + 2\epsilon)/\sqrt{n}\right)^2.$$

This is a quadratic equation in $\|w^* - w\|_2$ which is of exactly the same form as the quadratic equation that arose in the analysis of Ordinary Least Squares (proof of Theorem 5), if we define $\gamma = \omega^2/n$. So solving the quadratic equation in the exact same way, we find that under the assumption $\gamma + 2\epsilon/\sqrt{n} < 1$ that

$$\left|\sqrt{L(w) - \sigma^2} - \sqrt{\frac{\gamma \hat{L}(w)}{(1-\gamma)^2}}\right| \le \epsilon\sqrt{\hat{L}(w)} + \sqrt{\frac{1}{1-\gamma}\left(\frac{\hat{L}(w)}{1-\gamma} - \sigma^2\right) + \epsilon \hat{L}(w)}. \tag{84}$$

□

D PROOFS FOR SECTION 5

We start with the following result, which lets us upper bound the training error of the ERM in a convex set $\mathcal{K}$ and is proved using a direct application of the Convex Gaussian Minmax Theorem.

Theorem 9. *Suppose that $\mathcal{K}$ is a convex set and consider the upper summary function ψ_δ^+ as defined in* (38). *It holds with probability at least* $1 - \delta$,

$$\min_{w \in \mathcal{K}} \sqrt{\hat{L}(w)} \leq \min_{r \geq 0} \psi_\delta^+(r) \tag{40}$$

Proof. Observe that

$$\min_{w \in \mathcal{K}} \sqrt{\hat{L}(w)} = \frac{1}{\sqrt{n}} \min_{w \in \mathcal{K}} \max_{\|\lambda\|_2 \leq 1} \langle \xi + Z\Sigma^{1/2}(w^* - w), \lambda \rangle \tag{85}$$

which is a minimax optimization problem over a convex-conave function on a convex set. Hence by the Convex Gaussian Minmax Theorem (Theorem 16) and the same kind of truncation argument based on Lemma 5, to get a probability at least $1 - \delta$ upper bound on the Primary Optimization (85), it suffices to prove a probability at least $1 - \delta/2$ upper bound on the following auxillary problem:

$$\begin{aligned}
&\frac{1}{\sqrt{n}} \min_{w \in \mathcal{K}} \max_{\|\lambda\|_2 \leq 1} \langle \xi, \lambda \rangle + \|\lambda\|_2 \langle H, \Sigma^{1/2}(w^* - w) \rangle + \|w^* - w\|_\Sigma \langle G, \lambda \rangle \\
&= \frac{1}{\sqrt{n}} \min_{w \in \mathcal{K}} \max_{\|\lambda\|_2 \leq 1} \langle \xi + \|w^* - w\|_\Sigma G, \lambda \rangle + \|\lambda\|_2 \langle H, \Sigma^{1/2}(w^* - w) \rangle \\
&= \frac{1}{\sqrt{n}} \min_{w \in \mathcal{K}} \max \left\{ \|\xi + \|w^* - w\|_\Sigma G\|_2 + \langle H, \Sigma^{1/2}(w^* - w) \rangle, 0 \right\} \\
&\leq \frac{1}{\sqrt{n}} \min_{w \in \mathcal{K}} \max \left\{ (1 + \beta_1)\sqrt{\sigma^2 n + \|w^* - w\|_\Sigma^2 n} - \langle H, \Sigma^{1/2}(w - w^*) \rangle, 0 \right\}
\end{aligned}$$

where in the last equality, we used that the maximum is attained along the direction $\xi + \|w^* - w\|_\Sigma G$ and is attained either at $\|\lambda\| = 0$ or $\|\lambda\| = 1$. Also, consider two cases: either there exists $w \in \mathcal{K}$ such that the non-zero quantity inside the max is negative, in which case the minimum is just zero, or for all $w \in \mathcal{K}$, this quantity is positive and so we can drop the max inside the minimum. In either case, we see that this is not larger than

$$\max \left\{ 0, \min_{w \in \mathcal{K}} (1 + \beta_1)\sqrt{\sigma^2 + \|w^* - w\|_\Sigma^2} - \frac{1}{\sqrt{n}} \langle H, \Sigma^{1/2}(w - w^*) \rangle \right\}.$$

For any particular $r \geq 0$, we can control it by restricting to $\mathcal{K}_r$

$$\begin{aligned}
&\min_{w \in \mathcal{K}} (1 + \beta_1)\sqrt{\sigma^2 + \|w^* - w\|_\Sigma^2} - \frac{1}{\sqrt{n}} \langle H, \Sigma^{1/2}(w - w^*) \rangle \\
\leq &\min_{w \in \mathcal{K}_r} (1 + \beta_1)\sqrt{\sigma^2 + r^2} - \frac{1}{\sqrt{n}} \langle H, \Sigma^{1/2}(w - w^*) \rangle \\
= &(1 + \beta_1)\sqrt{\sigma^2 + r^2} - \frac{1}{\sqrt{n}} \sup_{\|w^* - w\|_\Sigma \leq r} \langle H, \Sigma^{1/2}(w - w^*) \rangle
\end{aligned}$$

and so by Gaussian concentration (Theorem 15)

$$\min_{w \in \mathcal{K}} \sqrt{\hat{L}(w)} \leq \max \left\{ 0, (1 + \beta_1)\sqrt{\sigma^2 + r^2} - W_\Sigma(\mathcal{K}_r)/\sqrt{n} + O(r\sqrt{\log(2/\delta)/n}) \right\}.$$

In particular, we can choose the r that minimizes the right hand side, which concludes the proof. □

Lemma 12. *For any $\sigma \geq 0$, the function $r \mapsto \sqrt{\sigma^2 + r^2}$ is strictly increasing, convex, and 1-Lipschitz on $\mathbb{R}_{\geq 0}$, and also strictly convex if $\sigma > 0$.*

Proof. Let $f(r) := \sqrt{\sigma^2 + r^2}$, then

$$f'(r) = \frac{r}{\sqrt{\sigma^2 + r^2}} \in (0, 1]$$

and

$$f''(r) = \frac{1}{\sqrt{\sigma^2 + r^2}} - \frac{r^2}{(\sigma^2 + r^2)^{3/2}} = \frac{\sigma^2}{(\sigma^2 + r^2)^{3/2}}$$

which is nonnegative, and positive if $\sigma > 0$. □

Lemma 13. *If $\mathcal{K}$ is a convex set in $\mathbb{R}^d$ and*

$$\mathcal{K}_r := \mathcal{K} \cap \{w : \|w - w^*\|_\Sigma \leq r\}$$

then for any $x \in \mathbb{R}^d$, the function

$$g(r) := \sup_{w \in \mathcal{K}_r} \langle x, w - w^* \rangle$$

is increasing and concave. In particular, the function $\omega(r) := W_\Sigma(\mathcal{K}_r)$ is increasing and concave.

Proof. Without loss of generality we may assume the set $\mathcal{K}$ is closed, since replacing $\mathcal{K}$ by its closure does not change the value of $g(r)$. The fact that it is increasing is obvious from the definition. Let $r = (1-\lambda)s+\lambda t$ and let $w_s \in \mathcal{K}_s, w_t \in \mathcal{K}_t$. Then $w_r := (1-\lambda)w_s + \lambda w_t$ lies in $\mathcal{K}_r$ by convexity of $\mathcal{K}$, and because $\|w_r - w^*\|_\Sigma \leq (1-\lambda)\|w_s - w^*\|_\Sigma + \lambda\|w_t - w^*\|_\Sigma$ by the triangle inequality. Since

$$\langle w_r - w^*, x \rangle = (1-\lambda)\langle w_s - w^*, x \rangle + \lambda \langle w_t - w^*, x \rangle$$

and w_s, w_t were arbitrary vectors in $\mathcal{K}_s, \mathcal{K}_t$, taking the maximum over w_s, w_t shows

$$\max_{w \in \mathcal{K}_r} \langle w - w^*, x \rangle \geq (1-\lambda)g(s) + \lambda g(t).$$ □

We now give the main arguments used in the proof of Theorem 10. The following lemma shows how to derive lower bounds on the generalization error of the constrained Empirical Risk Minimizer, by formalizing the informal argument from Section 5. To avoid having to perform a union bound over all localization radiuses r, we show how to get the conclusion by applying Theorem 1 for a few carefully chosen values of sets $\mathcal{K}_r$; this is equivalent to applying Theorem 1 once with a simplified version of the "optimal complexity functional" described before.

Lemma 14. *Suppose that $\mathcal{K}$ is a convex set and we are under the model assumptions* (1) *and recall summary functionals $\psi_\delta^+, \psi_\delta^-$ as defined in* (38) *and* (39). *Let $\delta > 0$ be arbitrary, let $\mu^* := \min_{r \geq 0} \psi_\delta^+(r)$, and suppose that $r_- \geq 0, \mu > \mu^*$ and $\eta > 0$ are such that we have $\eta K \leq \delta$ for $K := \left\lceil \frac{r_-}{\mu - \mu^*} \right\rceil$ and for all $r \in [0, r_-]$*

$$\min_{r \in [0, r_-]} \psi_\eta^-(r) > \mu.$$

Then with probability at least $1 - 2\delta$, the constrained empirical risk minimizer $\hat{w} = \arg\min_{w \in \mathcal{K}} \hat{L}(w)$ satisfies

$$\|\hat{w} - w^*\|_\Sigma > r_-.$$

Proof. Observe that for any fixed value of $r \leq r_-$, it follows from Theorem 1 that with probability at least $1 - \eta$ for all $w \in \mathcal{K}_r$ where $\eta = \delta + \tau/r$

$$\sqrt{\hat{L}(w)} > (1-\beta_1)\sqrt{\sigma^2 + \|w - w^*\|^2} - W_\Sigma(\mathcal{K}_r)/\sqrt{n} - Cr\sqrt{\log(2/\eta)/n} \tag{86}$$

$$\geq \psi_\eta^-(r) - (1-\beta_1)(r - \|w - w^*\|_\Sigma) \tag{87}$$

$$\geq \mu - (1-\beta_1)(r - \|w - w^*\|_\Sigma) \tag{88}$$

where we used the Lipschitz property from Lemma 12. We apply this argument for a grid on $[0, r_-]$ which includes the right end point r_- with spacing $\mu - \mu^* < (\mu - \mu^*)/(1 - \beta_1)$, i.e. with $\lceil \frac{r_-}{\mu - \mu^*} \rceil \leq K$ many grid points and apply the union bound, it follows that with probability at least $1 - \eta K \geq 1 - \delta$ that for all w with $\|w - w^*\|_\Sigma \leq r_-$ that

$$\sqrt{\hat{L}(w)} > \mu^*.$$

Recall from Theorem 9 that with probability at least $1 - \delta$ the constrained ERM satisfies $\sqrt{\hat{L}(\hat{w})} \leq \mu^*$. Thus, by applying the union bound we show that $\|w' - w^*\|_\Sigma > r_-$ with probability at least $1 - 2\delta$. □

Theorem 10. *Suppose that $\mathcal{K}$ is a convex set and consider the summary functional $\psi_\delta^+, \psi_\delta^-$ as defined in* (38) *and* (39)*. Let $\delta > 0$ and μ be arbitrary such that $\mu > \mu^* := \min_{r \geq 0} \psi_\delta^+(r)$ and define $r^* := \inf\{r : \psi_\delta^+(r) = \mu^*\}$. Then with probability at least $1 - 4\delta$, it holds that uniformly over all $w \in \mathcal{K}$ such that $\sqrt{\hat{L}(w)} \leq \mu$ that:*

$$\|w - w^*\|_\Sigma \leq r_+ := \sup\{r \geq 0 : \psi_\delta^-(r) \leq \mu\} \tag{41}$$

and also

$$\|w - w^*\|_\Sigma \geq r_- := \inf\left\{r \geq 0 : \psi_\tau^-(r) \leq \mu\right\} \tag{42}$$

where $\tau := \delta / \lceil \frac{\mu - \mu^*}{r^*} \rceil$.

Proof. We first show the upper bound $\|w - w^*\|_\Sigma \leq r_+$ for all $w \in \mathcal{K}$ with $\sqrt{\hat{L}(w)} \leq \mu$. If $r_+ = \infty$, then the upper bound is trivial. Otherwise, we have

$$\psi_\delta^-(r_+) = \mu \tag{89}$$

by continuity. Observe that the conclusion of Theorem 1 can be written as

$$(1 + \beta)^{-1/2}\sqrt{L(w)} - \frac{F(w)}{\sqrt{n}} \leq \sqrt{\hat{L}(w)} \tag{90}$$

so taking $F(w) = W(\mathcal{K}_{r_+}) + Cr_+\sqrt{\log(2/\delta)/n}$ for $w \in \mathcal{K}_{r_+}$ and ∞ outside, applying Theorem 1, and recalling the definition of ψ_δ^- from (39) and using (89) gives

$$\min_{w \in \mathcal{K}, \|w - w^*\|_\Sigma = r_+} \sqrt{\hat{L}(w)} \geq \psi_\delta^-(r_+) = \mu \tag{91}$$

Also, by definition if $r \geq r^+$, then $\psi_\delta^+(r) > \psi_\delta^-(r) \geq \mu > \mu^*$ and so r cannot be the minimizer of ψ_δ^+, i.e. we have shown $r^* < r^+$, where r^* is the minimizer of ψ_δ^+ so

$$\mu^* = \psi_\delta^+(r^*) = \min_{r \geq 0} \psi_\delta^+(r).$$

Note that since the minimizer $r^* < r^+$, by applying Theorem 9 we have with probability at least $1 - \delta$ that

$$\min_{w \in \mathcal{K}, \|w - w^*\|_\Sigma < r_+} \sqrt{\hat{L}(w)} \leq \psi_\delta^+(r^*) = \mu^* < \mu. \tag{92}$$

This establishes the claim $\|w - w^*\|_\Sigma \leq r$ by convexity: suppose for contradiction there exists $w \in \mathcal{K}$ such that $\|w - w^*\|_\Sigma > r_+$ and $\sqrt{\hat{L}(w)} \leq \mu$. By (92), there exists $w' \in \mathcal{K}$ with $\|w' - w^*\|_\Sigma < r_+$ and $\sqrt{\hat{L}(w)} < \mu$. Therefore, by convexity we conclude that there exists w'' which is a convex combination of w, w' such that $\sqrt{\hat{L}(w'')} < \mu$ and $\|w - w^*\|_\Sigma = r_+$, but this contradicts (91).

Now we show that $\|w - w^*\|_\Sigma \geq r_-$ for all $w \in \mathcal{K}$ with $\sqrt{\hat{L}(w)} \leq \mu$. If $r_- = -\infty$ then the bound is trivial. Otherwise, by continuity

$$\psi^-_{\tau/r^*}(r_-) = \mu$$

and by definition for all $r < r^*$ we have $\psi^-_{\tau/r^*}(r_-) \geq \mu$. Also, since $\psi^-_{\tau/r^*}(r^*) \leq \psi^+_\delta(r^*) = \mu^* < \mu$ from the definition, we know that $r^* > r_-$, hence $\tau/r_- > \tau/r^*$ and so

$$\min_{r \in [0, r_-]} \psi^-_{\tau/r_-}(r) \geq \min_{r \in [0, r_-]} \psi^-_{\tau/r^*}(r) = \mu.$$

Therefore, we can apply Lemma 14 to conclude that with probability at least $1 - \delta$, the constrained ERM $\hat{w} = \arg\min_{w \in \mathcal{K}} \hat{L}(w)$ satisfies

$$\|\hat{w} - w^*\|_\Sigma > r_-.$$

By applying Theorem 1 analogously to the r_+ case, we know that with probability at least $1 - \tau \geq 1 - \delta$,

$$\min_{\|w - w^*\|_\Sigma = r_-, w \in \mathcal{K}} \sqrt{\hat{L}(\hat{w})} > \mu \tag{93}$$

and since $\mu^* < \mu$, it follows by a convexity argument that for all w with $\|w - w^*\|_\Sigma \leq r_-$,

$$\sqrt{\hat{L}(w)} > \mu \tag{94}$$

which establishes the desired conclusion as the contrapositive. The convexity argument is symmetrical to the r_+ case: if (94) is false for some w, then interpolating between w and $\hat{w}$ and observes that there exists a convex combination w'' such that $\sqrt{\hat{L}(w)} \leq \mu$ and $\|w'' - w^*\|_\Sigma = r_-$, which contradicts (93). □

E PROOFS FOR SECTION 6

E.1 Faster Rates for Low-Complexity Classes

Lemma 15. *Under the assumptions of Theorem 1 and with the definition of β_1 there, with probability at least $1 - 4(\delta + \delta')$*

$$L(\hat{w}) \le \sigma^2 + (1 + 2\beta_1)\left(\sqrt{\sigma F(\hat{w})/\sqrt{n}} + F(\hat{w})/\sqrt{n}\right)^2$$

where $\hat{w}$ is any empirical risk minimizer over a closed convex set $\mathcal{K}$ containing w^*, i.e. $\hat{L}(\hat{w}) = \min_{w \in \mathcal{K}} \hat{L}(w)$.

Proof. Write $X = Z\Sigma^{1/2}$ with Z a matrix of i.i.d. Gaussians, and observe

$$\frac{1}{n}\langle Z^T\xi, \Sigma^{1/2}(w - w^*)\rangle = \frac{1}{n}\langle \xi, Z\Sigma^{1/2}(w - w^*)\rangle = \frac{1}{n}\langle \xi, X(w - w^*)\rangle$$

Note that conditional on ξ, $Z^T\xi$ is just a standard Gaussian $N(0, \|\xi\|_2^2 I_d)$. So with probability at least $1 - \delta'$ (recalling the defining property of the complexity functional F) we have

$$\frac{1}{n}\langle Z^T\xi, \Sigma^{1/2}(w - w^*)\rangle \le \frac{\|\xi\|_2}{n}F(w). \tag{95}$$

Observe that

$$\nabla_w \hat{L}(w) = \frac{1}{n}\nabla_w\|Y - Xw\|_2^2 = -\frac{2}{n}X^T(Y - Xw) = -\frac{2}{n}(X^T\xi + X^T X(w^* - w))$$

so from the KKT condition $\langle w^* - \hat{w}, \nabla_w \hat{L}(\hat{w})\rangle \ge 0$ we have

$$\langle w^* - \hat{w}, X^T\xi\rangle + \langle w^* - w, X^T X(w^* - w)\rangle \le 0$$

so rearranging gives the first inequality, and using (95) gives the second inequality in

$$\|w^* - \hat{w}\|_{\hat{\Sigma}} \le \sqrt{\frac{1}{n}\langle \xi, X(\hat{w} - w^*)\rangle} \le \sqrt{\frac{\|\xi\|_2}{n}F(w)}.$$

By Theorem 1 (defining $F(w) = \infty$ outside of $\mathcal{K}$), for all $w \in \mathcal{K}$

$$\|w^* - w\|_{\Sigma} \le (1 + \beta_1)\left[\|w^* - w\|_{\hat{\Sigma}} + F(w)/\sqrt{n}\right]$$

and so for $\hat{w}$ we have

$$\begin{aligned}\|w^* - \hat{w}\|_{\Sigma} &\le (1 + \beta_1)\left[\|w^* - \hat{w}\|_{\hat{\Sigma}} + F(\hat{w})/\sqrt{n}\right] \\ &\le (1 + \beta_1)\left[\sqrt{\frac{\|\xi\|_2}{n}F(\hat{w})} + F(\hat{w})/\sqrt{n}\right]\end{aligned}$$

and using the fact that the norm $\|\xi\|_2$ concentrates about $\sigma\sqrt{n}$ by Lemma 2 and recalling the definition of β_1, we have

$$\|w^* - \hat{w}\|_{\Sigma}^2 \le (1 + 2\beta_1)\left(\sqrt{\sigma F(\hat{w})/\sqrt{n}} + F(\hat{w})/\sqrt{n}\right)^2.$$

Finally, recalling that $L(\hat{w}) = \sigma^2 + \|w - \hat{w}\|_{\Sigma}^2$ gives the bound as claimed. □

Theorem 11. *Let $\mathcal{K}$ be a closed convex set in $\mathbb{R}^d$ containing w^* and suppose $\delta' \ge 0, p \ge 0$ are such that with probability at least $1 - \delta'$ over the randomness of $x \sim N(0, \Sigma)$, uniformly over all $w \in \mathcal{K}$ we have*

$$\langle w - w^*, x\rangle \le \|w - w^*\|_{\Sigma}\sqrt{p}. \tag{43}$$

Suppose that $\hat{w} = \arg\min_{w \in \mathcal{K}} \hat{L}(w)$ and $p/n \le 0.999$, then for all $n \ge C \log(2/\delta)$ for some absolute constant $C > 0$, it holds with probability at least $1 - (\delta + \delta')$ that

$$L(\hat{w}) - \sigma^2 \le (1 + \tau)\sigma^2 \cdot \frac{p}{n}. \tag{44}$$

where $\tau = \tau(p, n, \delta)$ is upper bounded by an absolute constant and satisfies $\tau(p, n, \delta) \to 1$ in any joint limit $[p + \log(2/\delta)]/n \to 0, n \to \infty$.

Proof. Defining $\rho := \sqrt{p/n}$ and Lemma 15 gives

$$\|w^* - \hat{w}\|_\Sigma \le (1 + 2\beta_1)^{1/2} \left(\sqrt{\sigma F(\hat{w})/\sqrt{n}} + F(\hat{w})/\sqrt{n} \right) = (1 + 2\beta_1)^{1/2} \left(\sqrt{\sigma\rho\|w - w^*\|_\Sigma} + \rho\|w - w^*\|_\Sigma \right)$$

hence

$$(1 - (1 + 2\beta_1)^{1/2}\rho)\|w - w^*\|_\Sigma \le (1 + 2\beta_1)^{1/2}\sqrt{\sigma\rho\|w - w^*\|_\Sigma}$$

which is equivalent to

$$\|w - w^*\|_\Sigma \le \frac{(1 + 2\beta_1)\sigma\rho}{(1 - (1 + 2\beta_1)^{1/2}\rho)^2}$$

and this in turn is equivalent to the final result. □

Corollary 4. *Under the model assumptions* (1) *with $d < n$ and assuming a sufficiently large n, it holds with probability at least $1 - \delta$ that*

$$L(\hat{w}_{\text{OLS}}) - \sigma^2 \lesssim \sigma^2 \left(\sqrt{\frac{d}{n}} + 2\sqrt{\frac{\log(36/\delta)}{n}} \right)^2 \tag{45}$$

Proof. Recall from the proof of Theorem 5 that with probability at least $1 - \delta'$ we have

$$\langle w - w^*, x \rangle \le \left(\sqrt{d} + 2\sqrt{\log(4/\delta')} \right) \|\Sigma^{1/2}(w^* - w)\|_2$$

where $\delta' = \delta/9$ so the result follows from Theorem 11 with $\mathcal{K} = \mathbb{R}^d$. □

Corollary 5. *Applying Theorem 11 with $\mathcal{K} = \{\|w\|_1 \le \|w^*\|_1\}$ the rescaled ℓ_1-ball and under the sparsity and compatability condition assumptions of Theorem 4, we have with probability at least $1 - \delta$ that the LASSO solution*

$$\hat{w}_{LASSO} = \underset{w: \|w\|_1 \le \|w^*\|_1}{\arg\min} \hat{L}(w)$$

satisfies

$$L(\hat{w}_{LASSO}) - \sigma^2 \lesssim \frac{\max_i \Sigma_{ii}}{\phi(\Sigma, S)^2} \cdot \frac{\sigma^2 k \log(16d/\delta)}{n} \tag{46}$$

provided n is sufficiently large that

$$\sqrt{\frac{\max_i \Sigma_{ii}}{\phi(\Sigma, S)^2} \cdot \frac{8k \log(16d/\delta)}{n}} \le 0.999.$$

Proof. Recall from the proof of Theorem 4, more specially (76), that with probability at least $1 - \delta/8$

$$\langle w - w^*, x \rangle \le \|w - w^*\|_1 \|x\|_\infty \le 2\|(w - w^*)_S\|_1 \|x\|_\infty \le \frac{2k^{1/2}}{\phi(\Sigma, S)} \|w - w^*\|_\Sigma \max_i \sqrt{2\Sigma_{ii} \log(16d/\delta)}.$$

so the result follows from Theorem 11. □

E.2 Precise Rates for OLS

Theorem 12. *Under the model assumptions in* (1)*, fix $\gamma = d/n$ to be some value in $(0, 1)$ and pick any $c > 0$. Then there exists another absolute constant $c' > 0$ such that for all sufficiently large n, with probability at least $1 - \delta$, there exists a $w \in \mathbb{R}^d$ such that*

$$\hat{L}(w) - \hat{L}(\hat{w}_{\text{OLS}}) \leq c \cdot \frac{\sigma^2}{n^{1/2}}, \tag{47}$$

but the population error satisfies

$$L(w) - L(\hat{w}_{\text{OLS}}) \geq c' \cdot \frac{\sigma^2}{n^{1/4}}. \tag{48}$$

Proof. Consider the following estimator:

$$\begin{aligned} w_\alpha &= w^* + \alpha\left(\hat{w}_{\text{OLS}} - w^*\right) \\ &= w^* + \alpha(X^TX)^{-1}X^T\xi \end{aligned}$$

Then the training error is

$$\begin{aligned} \hat{L}(w_\alpha) &= \frac{1}{n}\|Y - Xw_\alpha\|^2 \\ &= \frac{1}{n}\|\xi - \alpha X(X^TX)^{-1}X^T\xi\|^2 \\ &= \frac{1}{n}\|\left(I - X(X^TX)^{-1}X^T\right)\xi + (1-\alpha)X(X^TX)^{-1}X^T\xi\|^2 \\ &= \frac{1}{n}\|\left(I - X(X^TX)^{-1}X^T\right)\xi\|^2 + (1-\alpha)^2\frac{1}{n}\|X(X^TX)^{-1}X^T\xi\|^2 \\ &= \hat{L}(\hat{w}_{\text{OLS}}) + (1-\alpha)^2\left\|\hat{w}_{\text{OLS}} - w^*\right\|_{\hat{\Sigma}}^2 \end{aligned}$$

By Lemma 9, with probability at least $1 - \delta$, it holds that

$$\left\|\hat{w}_{\text{OLS}} - w^*\right\|_{\hat{\Sigma}}^2 \leq \sigma^2\left(\sqrt{\gamma} + 2\sqrt{\frac{\log(4/\delta)}{n}}\right)^2$$

which can again be upper bounded by, for example, $4\sigma^2\gamma$ for a sufficiently large n. Therefore, we can let

$$(1-\alpha)^2 4\sigma^2\gamma = c \cdot \frac{\sigma^2}{\sqrt{n}}$$

and it suffices to pick

$$\alpha = 1 + \sqrt{\frac{c}{4\gamma}} \cdot \frac{1}{n^{1/4}}.$$

So if we define $c' = 2\sqrt{\frac{c}{4\gamma}}$, then the excess error of w_α satisfies

$$\begin{aligned} L(w_\alpha) - \sigma^2 &= \|\Sigma^{1/2}(w_\alpha - w^*)\|^2 \\ &= \alpha^2\|\Sigma^{1/2}(\hat{w}_{\text{OLS}} - w^*)\|^2 \\ &\geq \left(1 + \frac{c'}{n^{1/4}}\right) \cdot L(\hat{w}_{\text{OLS}}). \end{aligned}$$

The last inequality follows from the fact that $L(\hat{w}_{\text{OLS}}) \geq \sigma^2$. □

Theorem 13. *Under the model assumptions in* (1) *with $d \leq n$, consider the ordinary least square estimator $\hat{w}_{\text{OLS}} = (X^TX)^{-1}X^TY$. It holds that*

$$\begin{aligned}\mathbb{E}\, L(\hat{w}_{\text{OLS}}) &= \sigma^2 \frac{n-1}{n-d-1} \\ \operatorname{Var}(L(\hat{w}_{\text{OLS}})) &= 2\sigma^4 \frac{d(n-1)}{(n-d-1)^2(n-d-3)}\end{aligned} \tag{49}$$

Hence as $d/n \longrightarrow \gamma$, it holds that

$$\mathbb{E}\, L(\hat{w}_{\text{OLS}}) \longrightarrow \frac{\sigma^2}{1-\gamma} \quad \textit{and} \quad \frac{n}{\sigma^4} \operatorname{Var}(L(\hat{w}_{\text{OLS}})) \longrightarrow \frac{2\gamma}{(1-\gamma)^3}. \tag{50}$$

If d is held constant, as $n \longrightarrow \infty$, we have

$$n\, \mathbb{E}[L(\hat{w}_{\text{OLS}}) - \sigma^2] \longrightarrow \sigma^2 d \quad \textit{and} \quad \frac{n^2}{\sigma^4} \operatorname{Var}(L(\hat{w}_{\text{OLS}})) \longrightarrow 2d. \tag{51}$$

Proof. Write $X = Z\Sigma^{1/2}$ and recall that

$$\begin{aligned}L(\hat{w}_{\text{OLS}}) - \sigma^2 &= \left\|\hat{w}_{\text{OLS}} - w^*\right\|_\Sigma^2 = \|\Sigma^{1/2}(X^TX)^{-1}X^T\xi\|_2^2 \\ &= \xi^T Z(Z^TZ)^{-2}Z^T\xi.\end{aligned}$$

First, we compute the expectation. By the tower law, we have

$$\begin{aligned}\mathbb{E}\, L(\hat{w}_{\text{OLS}}) - \sigma^2 &= \mathbb{E}\left[\mathbb{E}\left[\xi^T Z(Z^TZ)^{-2}Z^T\xi \,|\, Z\right]\right] \\ &= \sigma^2\, \mathbb{E} \operatorname{Tr}((Z^TZ)^{-1}) \\ &= \sigma^2 \operatorname{Tr}(\mathbb{E}\left[(Z^TZ)^{-1}\right])\end{aligned}$$

Proposition 2.1 of von Rosen [49] shows that

$$\mathbb{E}[(Z^TZ)^{-1}] = \frac{1}{n-d-1} I_d,$$

and so

$$\mathbb{E}\, L(\hat{w}_{\text{OLS}}) = \sigma^2 + \sigma^2 \frac{d}{n-d-1} = \sigma^2 \frac{n-1}{n-d-1}.$$

To compute the variance, by the law of total variance, we have

$$\begin{aligned}\operatorname{Var}(L(\hat{w}_{\text{OLS}})) &= \operatorname{Var}(L(\hat{w}_{\text{OLS}}) - \sigma^2) \\ &= \mathbb{E} \operatorname{Var}(\xi^T Z(Z^TZ)^{-2}Z^T\xi \,|\, Z) + \operatorname{Var}(\mathbb{E}(\xi^T Z(Z^TZ)^{-2}Z^T\xi \,|\, Z))\end{aligned}$$

By the variance formula of Gaussian quadratic form, we have

$$\operatorname{Var}(\xi^T Z(Z^TZ)^{-2}Z^T\xi \,|\, Z) = 2\sigma^4 \operatorname{Tr}((Z^TZ)^{-2})$$

Proposition 2.1 of von Rosen [49] shows that

$$\mathbb{E}[(Z^TZ)^{-2}] = \frac{n-1}{(n-d)(n-d-1)(n-d-3)} I_d,$$

and so

$$\mathbb{E} \operatorname{Var}(\xi^T Z(Z^TZ)^{-2}Z^T\xi \,|\, Z) = \frac{2\sigma^4 d(n-1)}{(n-d)(n-d-1)(n-d-3)}.$$

To compute the second term, observe that

$$\begin{aligned}\operatorname{Var}(\mathbb{E}(\xi^T Z(Z^T Z)^{-2} Z^T \xi \mid Z)) &= \sigma^4 \operatorname{Var}(\operatorname{Tr}((Z^T Z)^{-1})) \\ &= \sigma^4 \operatorname{Var}(\operatorname{vec}(I_d)^T \operatorname{vec}((Z^T Z)^{-1})) \\ &= \sigma^4 \operatorname{vec}(I_d)^T \operatorname{Var}(\operatorname{vec}((Z^T Z)^{-1})) \operatorname{vec}(I_d)\end{aligned}$$

Proposition 2.1 of von Rosen [49] shows that

$$\operatorname{Var}(\operatorname{vec}((Z^T Z)^{-1})) = \frac{I_{d^2} + \sum_{i,j}(e_i \otimes e_j)(e_j^T \otimes e_i^T)}{(n-d)(n-d-1)(n-d-3)} + 2\frac{\operatorname{vec}(I_d)\operatorname{vec}(I_d)^T}{(n-d)(n-d-1)^2(n-d-3)}$$

and so

$$\begin{aligned}\frac{1}{\sigma^4}\operatorname{Var}(\mathbb{E}(\xi^T Z(Z^T Z)^{-2} Z^T \xi \mid Z)) &= \frac{2d}{(n-d)(n-d-1)(n-d-3)} + \frac{2d^2}{(n-d)(n-d-1)^2(n-d-3)} \\ &= \frac{2d(n-1)}{(n-d)(n-d-1)^2(n-d-3)}.\end{aligned}$$

Finally, we have shown that

$$\operatorname{Var}(L(\hat{w}_{\mathrm{OLS}})) = 2\sigma^4 \frac{d(n-1)}{(n-d-1)^2(n-d-3)}.$$

□

Theorem 14. *Under the model assumptions in* (1) *with $d \le n$, consider the ordinary least square estimator $\hat{w}_{\mathrm{OLS}} = (X^T X)^{-1} X^T Y$ and denote $\gamma = d/n$. Assume that $\gamma \le 0.999$, then with probability at least $1-\delta$, it holds that*

$$L(\hat{w}_{\mathrm{OLS}}) - \frac{\sigma^2}{1-\gamma} \lesssim \sigma^2 \sqrt{\frac{\gamma \log(36/\delta)}{n}}.$$

Proof. We are interested in the excess risk:

$$L(\hat{w}_{\mathrm{OLS}}) - \sigma^2 = \|\Sigma^{1/2}(\hat{w}_{\mathrm{OLS}} - w^*)\|^2 = \|(Z^T Z)^{-1} Z^T \xi\|^2.$$

Notice that

$$\|(Z^T Z)^{-1} Z^T \xi\|^2 = \left((Z^T Z)^{-1/2} Z^T \xi\right)^T (Z^T Z)^{-1} \left((Z^T Z)^{-1/2} Z^T \xi\right)$$

and we have the following equality:

$$\begin{aligned}b^T (Z^T Z)^{-1} b &= \max_u -\|Zu\|^2 + 2\langle u, b\rangle \\ &= \max_u \min_v \|v\|^2 + 2\langle v, Zu\rangle + 2\langle u, b\rangle.\end{aligned}$$

We can plug in $(Z^T Z)^{-1/2} Z^T \xi$ into b. The b term may seem a bit complicated, but the key observation is that conditioned on Z, the distribution of $(Z^T Z)^{-1/2} Z^T \xi \sim \mathcal{N}(0, \sigma^2 I_d)$ actually does not depend on Z, and so they are independent. Therefore, we can condition on $b = (Z^T Z)^{-1/2} Z^T \xi$ and the law of Z remains unchanged. To apply Theorem 16, we need use a truncation argument. Define the truncated problem as

$$\Phi_r = \max_{\|u\| \le r} \min_v \|v\|^2 + 2\langle v, Zu\rangle + 2\langle u, b\rangle, \tag{96}$$

then by Lemma 5, we have

$$\begin{aligned}&\Pr\left(L(\hat{w}_{\mathrm{OLS}}) - \sigma^2 > t \mid (Z^T Z)^{-1/2} Z^T \xi = b\right) \\ =&\Pr\left(\lim_{r\to\infty} \Phi_r > t\right) \le \lim_{r\to\infty} \Pr\left(\Phi_r > t\right).\end{aligned}$$

Given u, the minimizer $v = -Zu$ satisfies $\|v\| \leq r\|Z\|$ and so for any $M > 0$, we have

$$\begin{aligned}
\Pr\left(\Phi_r > t\right) &\leq \Pr\left(\max_{\|u\|\leq r}\min_{\|v\|\leq rM} \|v\|^2 + 2\langle v, Zu\rangle + 2\langle u, b\rangle > t\right) + \Pr(\|Z\| \geq M) \\
&\leq 2\Pr\left(\max_{\|u\|\leq r}\min_{\|v\|\leq rM} \|v\|^2 + 2\|v\|\langle H, u\rangle + 2\|u\|\langle G, v\rangle + 2\langle u, b\rangle > t\right) + \Pr(\|Z\| \geq M) \\
&= 2\Pr\left(\max_{\|u\|\leq r}\min_{\|v\|\leq rM} \|v\|^2 + 2\|v\|\left(\langle H, u\rangle - \|G\|\|u\|\right) + 2\langle u, b\rangle > t\right) + \Pr(\|Z\| \geq M)
\end{aligned}$$

by Gaussian minimax theorem. On the event that $\|G\| \geq \|H\|$, the minimizer is

$$\|v\| = \|G\|\|u\| - \langle H, u\rangle \geq (\|G\| - \|H\|)\|u\| > 0.$$

At the same time, we have

$$\|v\| \leq r(\|G\| + \|H\|)$$

and so

$$\begin{aligned}
\Pr\left(\Phi_r > t\right) \leq 2\Pr\left(\max_{\|u\|\leq r} 2\langle u, b\rangle - (\langle H, u\rangle - \|G\|\|u\|)^2 > t, \|G\| > \|H\|\right) + 2\Pr(\|G\| \leq \|H\|) \\
+ 2\Pr(\|G\| + \|H\| \geq M) + \Pr(\|Z\| \geq M).
\end{aligned}$$

As the max over $\{u : \|u\| \leq r\}$ is always smaller than the overall max, taking $M \to \infty$, we have

$$\Pr\left(\Phi_r > t\right) \leq 2\Pr\left(\max_{u} 2\langle u, b\rangle - (\|G\|\|u\| - \langle H, u\rangle)^2 > t, \|G\| > \|H\|\right) + 2\Pr(\|G\| \leq \|H\|)$$

Observe that any u can be decomposed into two parts: one part spanned by b and the other part in the orthogonal complement of b. Formally, we write $u = \alpha b + k$ where $\langle k, b\rangle = 0$, and the problem becomes

$$\max_{\alpha\in\mathbb{R}, \langle k,b\rangle=0} 2\alpha\|b\|^2 - \left(\|G\| \cdot \sqrt{\alpha^2\|b\|^2 + \|k\|^2} - \langle H, k\rangle - \alpha\langle H, b\rangle\right)^2.$$

Define $P = I_d - \frac{bb^T}{\|b\|^2}$. On the event that $\|G\| > \|H\|$, the quantity inside the square is always positive and so we want to choose the direction of k that make $\langle H, k\rangle$ as large as possible:

$$\begin{aligned}
&\max_{\alpha\in\mathbb{R}} 2\alpha\|b\|^2 - \min_{\langle k,b\rangle=0}\left(\|G\| \cdot \sqrt{\alpha^2\|b\|^2 + \|k\|^2} - \langle H, k\rangle - \alpha\langle H, b\rangle\right)^2 \\
=&\max_{\alpha\in\mathbb{R}} 2\alpha\|b\|^2 - \left(\min_{\langle k,b\rangle=0} \|G\| \cdot \sqrt{\alpha^2\|b\|^2 + \|k\|^2} - \langle H, k\rangle - \alpha\langle H, b\rangle\right)^2 \\
=&\max_{\alpha\in\mathbb{R}} 2\alpha\|b\|^2 - \left(\min_{\beta\geq 0} \|G\| \cdot \sqrt{\alpha^2\|b\|^2 + \beta^2} - \beta\|PH\| - \alpha\langle H, b\rangle\right)^2 \\
=&\max_{\alpha\in\mathbb{R}} 2\alpha\|b\|^2 - \left(|\alpha| \cdot \|b\|\sqrt{\|G\|^2 - \|PH\|^2} - \alpha\langle H, b\rangle\right)^2 \\
\leq&\max_{\alpha\in\mathbb{R}} 2\alpha\|b\|^2 - \alpha^2\|b\|^2\left(\sqrt{\|G\|^2 - \|PH\|^2} - \frac{|\langle H, b\rangle|}{\|b\|}\right)^2 = \frac{\|b\|^2}{\left(\sqrt{\|G\|^2 - \|PH\|^2} - \frac{|\langle H,b\rangle|}{\|b\|}\right)^2}
\end{aligned}$$

By the tower law, we have shown that

$$\begin{aligned}\Pr\left(L(\hat{w}_{\text{OLS}}) - \sigma^2 > \frac{\|b\|^2}{t}\right) &\leq 2\Pr\left(\|G\| \leq \|H\|\right) + 2\Pr\left(\sqrt{\|G\|^2 - \|PH\|^2} - \frac{|\langle H, b\rangle|}{\|b\|} < \sqrt{t}, \|G\| > \|H\|\right) \\ &= 2\Pr\left(\|G\| \leq \|H\| \quad \text{or} \quad \sqrt{\|G\|^2 - \|PH\|^2} - \frac{|\langle H, b\rangle|}{\|b\|} < \sqrt{t}, \|G\| > \|H\|\right)\end{aligned}$$

For the simplicity of notation, denote

$$\epsilon = 2\sqrt{\frac{\log(32/\delta)}{n}}.$$

By a union bound, with probability at least $1 - \delta/2$, the following occurs:

(1) by Lemma 2 and the fact that $b \sim \mathcal{N}(0, \sigma^2 I_d)$, it holds that

$$\|G\|^2 \geq n(1-\epsilon)^2$$

$$\|PH\|^2 \leq n(\sqrt{\gamma} + \epsilon)^2 \quad \text{and} \quad \|b\|^2 \leq \sigma^2 n(\sqrt{\gamma} + \epsilon)^2$$

(2) As $\frac{\langle H, b\rangle}{\|b\|} \sim \mathcal{N}(0, 1)$, by standard Gaussian concentration, it holds that

$$\frac{|\langle H, b\rangle|}{\|b\|} \leq \epsilon\sqrt{n}$$

Therefore, for sufficiently large n, we have $\|G\| > \|H\|$ and we can pick t by setting

$$\sqrt{t} = \sqrt{n(1-\epsilon)^2 - n(\sqrt{\gamma}+\epsilon)^2} - \epsilon\sqrt{n}$$

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

$$L(\hat{w}_{\text{OLS}}) - \sigma^2 \leq \frac{\sigma^2(\sqrt{\gamma}+\epsilon)^2}{\left(\sqrt{(1-\epsilon)^2 - (\sqrt{\gamma}+\epsilon)^2} - \epsilon\right)^2}.$$

It is then routine to check the desired bound. □