# On Uniform Convergence and Low-Norm Interpolation Learning

**Lijia Zhou**
University of Chicago
zlj@uchicago.edu

**Danica J. Sutherland**
TTI-Chicago
danica@ttic.edu

**Nathan Srebro**
TTI-Chicago
nati@ttic.edu

## Abstract

We consider an underdetermined noisy linear regression model where the minimum-norm interpolating predictor is known to be consistent, and ask: can uniform convergence in a norm ball, or at least (following Nagarajan and Kolter) the subset of a norm ball that the algorithm selects on a typical input set, explain this success? We show that uniformly bounding the difference between empirical and population errors cannot show any learning in the norm ball, and cannot show consistency for any set, even one depending on the exact algorithm and distribution. But we argue we can explain the consistency of the minimal-norm interpolator with a slightly weaker, yet standard, notion: uniform convergence *of zero-error predictors* in a norm ball. We use this to bound the generalization error of low- (but not minimal-) norm interpolating predictors.

## 1 Introduction

In the past several years, it has become empirically clear that – contrary to traditional intuition – it is possible for models which exactly interpolate noisy training data to reliably generalize well on practical problems, especially in deep learning [7, 25, 34]. We refer to this phenomenon as "interpolation learning." It is closely related to the (re-)discovery of the "double descent" phenomenon [1, 5, 23, 29], where many models first improve as their size is increased, then get much worse around the point where they can first interpolate the data, and then improve again as they become more and more overparametrized. Understanding interpolation learning, therefore, seems to be a key step on the path towards better theoretical understanding of the successes of deep learning.

We now know of a few settings where interpolating models can be shown to generalize well [4, 8]. In particular, significant recent attention has been paid to the minimum-norm linear interpolator ("ridgeless" regression) in certain high-dimensional linear regression regimes [3, 6, 14, 21]. This setting is of particular interest not only because it is reasonably accessible to study while exhibiting many of the surprising properties of more complex models, but also because this predictor is the same one found by (stochastic) gradient descent initialized at the origin, and so it seems plausible that its properties may generalize to more complex settings. Much is now understood about the properties of the minimum-norm interpolator for (sub-)Gaussian data, including necessary and sufficient conditions for its consistency. This line of inquiry has proved quite fertile for extensions to related settings and further results [2, 13, 15, 18, 20].

One striking feature of this body of work is that none of it is based on the core workhorse of learning theory, *uniform convergence*; most instead uses various tools, mostly from random matrix theory, to directly analyze the generalization error of a particular predictor. Indeed, some have argued that uniform convergence is unlikely to be able to explain interpolation learning; for instance, Mikhail Belkin has said[1] that "there are no [uniform generalization] bounds" with constants tight enough to

[1] Talk at the Simons Institute for the Theory of Computing, July 2019: simons.berkeley.edu/talks/tbd-65

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

explain interpolation learning, "and no reason they should exist." Meanwhile, Nagarajan and Kolter [22] have also raised significant questions about the ability of uniform convergence arguments to explain learning in certain high-dimensional regimes. Perhaps, then, it is time to wholly abandon uniform convergence in favor of other tools.

We connect these two avenues of work by studying uniform convergence in a particular overparametrized linear regression problem (Section 2) where the minimal-norm interpolator is consistent. We prove that, indeed, uniform convergence bounds based on predictor norm cannot show *any* learning in this setting (Theorem 3.2). We also prove, following Nagarajan and Kolter, that *no* uniform convergence bound can show consistency (Theorem 3.3), not only for the minimal-norm interpolator but even for a wide variety of natural interpolation algorithms.

Yet, even in this setting where the situation looks bleak, we need not abandon uniform convergence entirely. One option would be sidestep the negative results by considering uniform convergence not of our predictor, but of a surrogate separately shown to be not too different [24]. We instead demonstrate that it is possible to show uniform convergence of our predictor directly if we allow ourselves a slightly weaker notion of uniform convergence, one long in common use in realizable PAC analyses: uniform convergence *for predictors with zero error*. Such a bound would be implied by, for example, "optimistic rates" [30], although existing results are not tight enough to show consistency in our setting. Instead we prove (Theorem 4.1) that a tight version of this notion of uniform convergence *does* hold in our setting for low-norm predictors. Our result exactly characterizes the asymptotic worst-case generalization gap for predictors of a given norm via a novel analysis based on strong duality of a particular non-convex problem, and show that while neither having a low norm nor interpolation is sufficient for generalization in our setting, the combination is. By doing so, not only do we prove consistency of the minimal-norm interpolator with a uniform convergence-type argument, we also provide new insight about the behavior of interpolation learning for solutions with low but not minimal norm.

## 2 Problem setting

We begin with a standard linear regression setup, with Gaussian data and errors. Take i.i.d. observations $(x_1, y_1), ..., (x_n, y_n) \sim \mathcal{D}^n$, where the joint distribution $\mathcal{D}$ is given by

**A** $x \in \mathbb{R}^p$ is drawn from $\mathcal{N}(0, \Sigma)$, with $\Sigma \succ 0$, and $\epsilon \in \mathbb{R}$ is independently $\mathcal{N}(0, \sigma^2)$. There is some fixed $w^* \in \mathbb{R}^p$ such that $y = \langle w^*, x \rangle + \epsilon$.

We consider a "junk features" setting, where $x$ decomposes into "signal" and "junk" components, and analysis of interpolation learning is particularly appealing:

**B** In Setting **A**, let $\Sigma = \begin{bmatrix} I_{d_S} & 0_{d_S \times d_J} \\ 0_{d_J \times d_S} & \frac{\lambda_n}{d_J} I_{d_J} \end{bmatrix}$ where $d_S, d_J$ satisfy $d_S + d_J = p$, and $\lambda_n > 0$. In other words, we can write $x = (x_S, x_J)$, where $x_S \sim \mathcal{N}(0, I_{d_S})$ and $x_J \sim \mathcal{N}(0, \frac{\lambda_n}{d_J} I_{d_J})$. Further, the label depends only on $x_S$: $w^* = (w_S^*, 0_{d_J})$ with $w_S^* \in \mathbb{R}^{d_S}$.

Let $Y \in \mathbb{R}^n$ be the vector of responses, $X \in \mathbb{R}^{n \times p}$ the design matrix and $E \in \mathbb{R}^n$ the residual vector, so $Y = Xw^* + E$. The sample covariance is $\hat{\Sigma} = \frac{1}{n} X^\mathsf{T} X$. The population and empirical risks are, respectively,

$$\begin{aligned} L_{\mathcal{D}}(w) &= \mathbb{E}_{(x,y)\sim\mathcal{D}}[(y - \langle w, x \rangle)^2] = L_{\mathcal{D}}(w^*) + \|w - w^*\|_\Sigma^2 \\ L_{\mathbf{S}}(w) &= \frac{1}{n}\|Y - Xw\|^2 = L_{\mathbf{S}}(w^*) + \|w - w^*\|_{\hat{\Sigma}}^2 - \frac{2}{n}\langle X^\mathsf{T} E, w - w^* \rangle, \end{aligned} \tag{1}$$

where $\|x\|_A = \sqrt{x^\mathsf{T} A x}$ denotes the Mahalonobis norm, and $L_{\mathcal{D}}(w^*) = \mathbb{E}\,\epsilon^2 = \sigma^2$.

We will focus on the regime where $d_S$ is fixed, and $d_J \to \infty$ for *finite* values of $n$, e.g. $\lim_{n\to\infty} \lim_{d_J \to \infty} L_{\mathcal{D}}(\hat{w})$. This setting enables relatively easy calculation of many quantities of interest, and can recover many interesting behaviors of overparametrized interpolation, including consistency and the double descent phenomenon.

We will be primarily concerned with the behavior of the minimal-norm interpolator,

$$\hat{w}_{MN} = \operatorname*{arg\,min}_{w \in \mathbb{R}^p \text{ s.t. } Xw = Y} \|w\|_2^2 = X^\mathsf{T}(XX^\mathsf{T})^{-1} Y. \tag{2}$$

This predictor is in fact consistent in Setting **B** when $\lambda_n = o(n)$ and we consider $d_J \to \infty$ for each $n$. We here use a slightly broader notion of *consistency* than is traditional [e.g. in 28]: we mean that

$$\mathbb{E}\left[L_{\mathcal{D}}(\hat{w}_{MN}) - L_{\mathcal{D}}(w^*)\right] \to 0$$

for our *sequence* of learning problems in the given asymptotic regime. Specifically:

**Proposition 2.1.** *In Setting **B** with $\lambda_n = o(n)$,*

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \mathbb{E}\left[L_{\mathcal{D}}(\hat{w}_{MN}) - L_{\mathcal{D}}(w^*)\right] = 0.$$

The proof follows from Lemmas 2.2 and 2.3, which establish first – because the setting was designed exactly to make this true[2] – that $\hat{w}_{MN}$ becomes equivalent to ridge regression on the signal part of $X$ with regularization weight $\lambda_n$, and then that ridge regression is consistent in this setting.

Writing $X = (X_S, X_J)$ with $X_S \in \mathbb{R}^{n\times d_S}$ and $X_J \in \mathbb{R}^{n\times d_J}$, the ridge regression estimate on the signal components with tuning parameter $\lambda$ is given by

$$\begin{aligned}\hat{w}_\lambda &= \operatorname*{arg\,min}_{w\in\mathbb{R}^p} \|Y - X_S w\|^2 + \lambda\|w\|^2 \\ &= (X_S^\mathsf{T} X_S + \lambda I_{d_S})^{-1} X_S^\mathsf{T} Y = X_S^\mathsf{T}(X_S X_S^\mathsf{T} + \lambda I_n)^{-1} Y.\end{aligned}$$

**Lemma 2.2.** *In Setting **B**, $\lim_{d_J\to\infty} \mathbb{E}[L_{\mathcal{D}}(\hat{w}_{MN})] = \mathbb{E}[L_{\mathcal{D}}(\hat{w}_{\lambda_n})]$ for any $n$.*

*Proof.* By the strong law of large numbers, we have that $X_J X_J^\mathsf{T} = \lambda_n \frac{Z_J Z_J^\mathsf{T}}{d_J}$ converges almost surely to $\lambda_n I_n$. Writing $\hat{w}_{MN} = (\hat{w}_{MN,S}, \hat{w}_{MN,J})$, we can easily verify that

- $\hat{w}_{MN,S} = X_S^\mathsf{T}(X_S X_S^\mathsf{T} + X_J X_J^\mathsf{T})^{-1} Y \overset{a.s.}{\to} \hat{w}_{\lambda_n}$ by the continuous mapping theorem.
- $\hat{w}_{MN,J} = X_J^\mathsf{T}(X_S X_S^\mathsf{T} + X_J X_J^\mathsf{T})^{-1} Y$. Drawing a new $x_J \sim \mathcal{N}(0, \frac{\lambda_n}{d_J} I_{d_J})$, $X_J x_J \overset{a.s.}{\to} 0_n$ and so $\langle \hat{w}_{MN,J}, x_J\rangle \overset{a.s.}{\to} 0$.

This implies that for any fixed $x$, $\langle \hat{w}_{MN}, x\rangle \overset{a.s.}{\to} \langle \hat{w}_{\lambda_n}, x_S\rangle$, and hence via continuity we have that $(\langle \hat{w}_{MN}, x\rangle - y)^2 \overset{a.s.}{\to} (\langle \hat{w}_{\lambda_n}, x\rangle - y)^2$. Taking expectations over $(x, y)$ to get $L_{\mathcal{D}}$ and then over the training set, then exchanging the limit with each expectation,[3] we obtain the desired result. ☐

**Lemma 2.3.** *In Setting **B**, if $\lambda_n = o(n)$, then $\lim_{n\to\infty} \mathbb{E}\left[L_{\mathcal{D}}(\hat{w}_{\lambda_n}) - L_{\mathcal{D}}(w^*)\right] = 0$.*

The proof, as for all the following results, is in the appendix. Taking $\lambda_n = o(n)$ ensures the bias due to regularization is negligible; the minimax-optimal scaling would be $\lambda_n \propto \sqrt{n}$ [11].

**Relationship to previous settings** The results of Bartlett et al. [3] apply to our setting, also showing consistency of $\hat{w}_{MN}$. Although they do not require $p \to \infty$ for finite $n$ as we study, their results show that consistency of $\hat{w}_{MN}$ is only possible when the effective $p$ grows much faster than $n$. Muthukumar et al. [21] showed that *no* interpolation method can be consistent in Setting **A** for $p = \mathcal{O}(n)$; we re-derive this (simple) result in Proposition 4.3, since it will also be important for our purposes.

Hastie et al. [14] and various follow-ups, on the other hand, employ the standard asymptotic regime of random matrix theory, where $n/p \to \gamma \in (0, \infty)$, mostly focusing on $\Sigma = I$. Although no interpolator can achieve consistency here, they exactly evaluate $\lim_{(n,d)\to\infty} L_{\mathcal{D}}(\hat{w}_{MN})$. The setting of Belkin et al. [6] is related, with general $(n, p)$ but again with $\Sigma = I$, where $\hat{w}_{MN}$ is not consistent.

[2] If the noise scaling were $\omega(1/d_J)$, then as $d_J \to \infty$, the minimal-norm solution would exploit the exploding magnitude of the noise components, and all of the signal would "bleed" into the noise dimensions [14, 21], giving $\|\hat{w}_{MN}\| \to 0$ and $L_{\mathcal{D}}(\hat{w}_{MN}) \to L_{\mathcal{D}}(0_p)$ – in the ridge regression equivalence, we let the regularization weight go to infinity. On the other hand, if the noise scaling were $o(1/d_J)$, then we would have $\|\hat{w}_{MN}\| \to \infty$, significantly complicating matters. $\Theta(1/d_J)$ is the only scaling in which $\|\hat{w}_{MN}\|$ is bounded but nonzero.

[3] Both exchanges can be justified using dominated convergence thoerem and the techniques from the proof of Proposition 4.6, which independently shows a stronger statement.

## 3 Uniform convergence

We now know, via Proposition 2.1, that $\hat{w}_{MN}$ is consistent in this setting. Could we have discovered this fact directly via uniform convergence? Typically, we would find some class $\mathcal{W}_{n,\delta}$ such that $\Pr(\hat{w}_{MN} \in \mathcal{W}_{n,\delta}) \geq 1-\delta$, and bound the generalization gap

$$\Pr\left(\sup_{w \in \mathcal{W}_{n,\delta}} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) \leq \epsilon_{\mathcal{W}}(n,\delta)\right) \geq 1-\delta. \tag{3}$$

As $L_{\mathbf{S}}(\hat{w}_{MN}) = 0$, this would directly provide an upper bound on $L_{\mathcal{D}}(\hat{w}_{MN})$ with probability $1-2\delta$.

### 3.1 Uniform convergence over norm balls

Our first thought would likely be to find some high-probability upper bound $B_{n,\delta}$ on $\|\hat{w}_{MN}\|$, and take $\mathcal{W}_{n,\delta} = \{w \in \mathbb{R}^p : \|w\| \leq B_{n,\delta}\}$. We can get a rough asymptotic estimate for $B_{n,\delta}$ based on the following, since $\|\hat{w}_{MN}\| = \mathcal{O}_P\left(\sqrt{\mathbb{E}\|\hat{w}_{MN}\|^2}\right)$ by Markov's inequality.

**Proposition 3.1.** *As $n \to \infty$ in Setting **B**, if $\lambda_n$ is both $o(n)$ and $\omega(1)$, then*

$$\lim_{d_J \to \infty} \mathbb{E}\|\hat{w}_{MN}\|^2 = \frac{\sigma^2 n}{\lambda_n} + \mathcal{O}(1) \quad \textit{and} \quad \lim_{d_J \to \infty} \frac{(\mathbb{E}\|\hat{w}_{MN}\|^2)(\mathbb{E}\|x\|^2)}{n} = \sigma^2 + o(1).$$

We could then find $\epsilon_{\mathcal{W}}(n,\delta)$ by studying the Rademacher complexity, given by

$$\mathfrak{R}_n(\mathcal{W}_B) = \mathbb{E}_{\mathbf{S}}\, \mathbb{E}_{\sigma \sim \mathrm{Unif}(\pm 1)^n} \sup_{w : \|w\| \leq B} \frac{1}{n} \sum_{i=1}^n \sigma_i \langle w, x^{(i)} \rangle \leq \sqrt{\frac{1}{n} B^2\, \mathbb{E}\|x\|^2};$$

thus Proposition 3.1 gives us that $\mathfrak{R}_n(\mathcal{W}_{\sqrt{\mathbb{E}\|\hat{w}_{MN}\|^2}}) \leq \sigma + o(1)$.

Standard Rademacher bounds are for Lipschitz losses, which the squared loss is not. If we let $T_n$ be a uniform upper bound on all the labels and $Q_n$ on all the predictions, however, the absolute value of the derivative of the squared loss is at most $2|\hat{y} - y| \leq 2(Q_n + T_n)$, and so we can treat it as $2(Q_n + T_n)$-Lipschitz with high probability. We then obtain in the setting of Proposition 3.1 that

$$\sup_{\|w\|^2 \leq \mathbb{E}\|\hat{w}_{MN}\|^2} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) \leq 4(Q_n + T_n)\left(\sigma + \mathcal{O}_P\left(\frac{1}{\sqrt{n}}\right)\right). \tag{4}$$

To show consistency, we need a bound exactly approaching $\sigma^2$ as $n \to \infty$, i.e. $Q_n + T_n \to \frac{1}{4}\sigma$. But in fact, each of $Q_n$ and $T_n$ diverge to $\infty$ as $n \to \infty$, because we have more and more chances to see a large value. Thus for $n \to \infty$, (4) says nothing at all.

Now, the path to (4) was potentially quite loose, particularly in the Lipschitz step; perhaps, then, we could simply put more effort in to obtain the bound we want. This is not the case: balls which are big enough to contain $\hat{w}_{MN}$ also contain predictors with unbounded generalization gaps as $n \to \infty$.

**Theorem 3.2.** *In Setting **B**, if $\lambda_n = o(n)$ then*

$$\lim_{n \to \infty} \lim_{d_J \to \infty} \mathbb{E}\left[\sup_{\|w\| \leq \|\hat{w}_{MN}\|} |L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)|\right] = \infty.$$

*Proof sketch.* Proposition B.2 shows that the gap is at least $\|\Sigma - \hat{\Sigma}\|(\|\hat{w}_{MN}\| - \|w^*\|)^2 + o(1)$ using (1) and then aligning $w - w^*$ with $\Sigma - \hat{\Sigma}$. By Proposition 3.1, $(\|\hat{w}_{MN}\| - \|w^*\|)^2$ grows like $n/\lambda_n$. Now $\|\Sigma - \hat{\Sigma}\|$ goes to 0, but only at the rate of $\sqrt{\lambda_n/n}$ [17], so the product grows as $\sqrt{n/\lambda_n}$. □

Proposition B.2 also gives a lower bound for $\mathbb{E}\left[\sup_{\|w\| \leq \|\hat{w}_{MN}\|} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)\right]$, the one-sided generalization gap, based on the algebraically largest eigenvalue of $\Sigma - \hat{\Sigma}$ rather than the operator norm. We expect that this eigenvalue should asymptotically behave similarly to the operator norm, and hence the one-sided generalization gap should also diverge.

Norm balls around $w^*$, rather than the origin, fare no better; they would merely remove the asymptotically irrelevant $\|w^*\|$ term from the result of Proposition B.2.

### 3.2 Uniform convergence over algorithm- and distribution-dependent hypothesis classes

Choosing $\mathcal{W}_{n,\delta}$ as a Euclidean norm ball, then, cannot yield the result we want (or, indeed, any meaningful result at all for large $n$). But a norm ball doesn't fully capture everything we know about $\hat{w}_{MN}$: for instance, we know that its norm is not likely to be very small. Perhaps taking a shell rather than a ball would help? Following Nagarajan and Kolter [22], we show that in fact, *no* choice of $\mathcal{W}_{n,\delta}$ can demonstrate consistency using the most common two-sided uniform convergence bounds.

Specifically, let $\mathcal{S}_{n,\delta}$ be a set of typical training examples $\mathbf{S} = (X, Y)$ such that $\Pr(\mathbf{S} \in \mathcal{S}_{n,\delta}) \geq 1 - \delta$, let $\mathcal{A}(X, Y)$ be any learning algorithm, and then take the class of typical outputs of $\mathcal{A}$, $\mathcal{W}_{n,\delta}^{\mathcal{A}} = \{\mathcal{A}(X, Y) : (X, Y) \in \mathcal{S}_{n,\delta}\}$. (Clearly, no bound based on $\mathcal{S}_{n,\delta}$ could choose a smaller $\mathcal{W}_{n,\delta}$.) The *tightest algorithm-dependent uniform convergence bound* [22] is then

$$\sup_{\mathbf{S}\in\mathcal{S}_{n,\delta}} \sup_{w\in\mathcal{W}_{n,\delta}^{\mathcal{A}}} |L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)| \leq \epsilon_{\mathcal{A}}^{\mathcal{D}}(n, \delta), \tag{5}$$

$$\text{implying } \Pr\Big( |L_{\mathcal{D}}(\mathcal{A}(X, y)) - L_{\mathbf{S}}(\mathcal{A}(X, y))| \leq \epsilon_{\mathcal{A}}^{\mathcal{D}}(n, \delta)\Big) \geq 1 - \delta.$$

In interpolation learning, where $L_{\mathbf{S}}$ is zero, we need $\lim_{n\to\infty} \epsilon_{\mathcal{A}}^{\mathcal{D}}(n, \delta) = \sigma^2$ to obtain consistency. Nagarajan and Kolter show that in a particular high-dimensional linear classification setting, stochastic gradient descent has 0 asymptotic loss, but $\epsilon_{\mathcal{A}}^{\mathcal{D}}(n, \delta)$ must be nearly 1 for any $\mathcal{S}_{n,\delta}$. We show a similar result in our setting, not only for $\mathcal{A} = \hat{w}_{MN}$ but indeed for many interpolation methods.[4]

**Theorem 3.3.** *In Setting **B**, let $\mathcal{A}$ be an algorithm outputting interpolators, $X\mathcal{A}(X, Y) = Y$, with*

$$\mathcal{A}\left((X_S, X_J), y\right)_S = \mathcal{A}\left((X_S, -X_J), y\right)_S \quad \textit{and} \quad \lim_{n\to\infty} \lim_{d_J\to\infty} L_{\mathcal{D}}(\mathcal{A}(X, y)) \overset{a.s.}{=} \sigma^2. \tag{6}$$

*For any $\delta \in (0, \frac{1}{2})$ and set of typical training examples $\mathcal{S}_{n,\delta}$ satisfying $\Pr(\mathbf{S} \in \mathcal{S}_{n,\delta}) \geq 1 - \delta$, let $\mathcal{W}_{n,\delta}^{\mathcal{A}} = \{\mathcal{A}(X, Y) : (X, Y) \in \mathcal{S}_{n,\delta}\}$ denote the set of typical outputs. Then*

$$\lim_{n\to\infty} \lim_{d_J\to\infty} \sup_{\mathbf{S}\in\mathcal{S}_{n,\delta}} \sup_{w\in\mathcal{W}_{n,\delta}} |L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)| \overset{a.s.}{\geq} 3\sigma^2. \tag{7}$$

*Proof sketch.* For each $\mathbf{S} = (X, Y) \in \mathcal{S}_{n,\delta}$, let $\tilde{\mathbf{S}} = \left((X_S, -X_J), Y\right)$, which has equal density under $\mathcal{D}$, so that $\mathcal{S}_{n,\delta}$ must contain some $(\mathbf{S}, \tilde{\mathbf{S}})$ pairs. Consider $\tilde{w} = \mathcal{A}(\tilde{\mathbf{S}})$: we know that $L_{\mathcal{D}}(\tilde{w}) \overset{a.s.}{\to} \sigma^2$ by assumption, but we will show $\lim_{n\to\infty} \lim_{d_J\to\infty} L_{\mathbf{S}}(\tilde{w}) \overset{a.s.}{\geq} 4\sigma^2$.

This is easiest to see in the case when $d_S = 0$, so that $y \sim \mathcal{N}(0, \sigma^2)$ is independent of $x$. Then $-X\tilde{w} = Y$, so that $X\tilde{w} = -Y$, and thus $L_{\mathbf{S}}(\tilde{w}) = \frac{1}{n}\|(-Y) - Y\|^2 = \frac{4}{n}\|Y\|^2 \overset{a.s.}{=} 4\sigma^2$.

The general case, in Appendix B.3, shows that since $X_S$ is rank $d_S \ll n$, $\tilde{w}_J$ must be large enough to contribute $4\sigma^2 \frac{n-d_S}{n} \to 4\sigma^2$ to the loss. ☐

From (2), we can see that $\hat{w}_{MN}$ satisfies the symmetry condition in (6). In fact, Proposition B.3 (in Appendix B.3) shows this is also true of many more algorithms, including interpolators which minimize $\|w\|_1$ (basis pursuit) or even $\|w - w^*\|$: any algorithm that picks the interpolator minimizing $f_S(w_S) + f_J(w_J)$, where each function is convex and $f_J(-w) = f_J(w)$.

The attentive reader may have noticed that Theorem 3.3, like Theorem 3.2, applies only to bounds on $|L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)|$, whereas the general argument as in (3) only needs to bound $L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)$. Indeed, the proof of Theorem 3.3 exhibits a hypothesis with low generalization error but high *training* error – not a particularly concerning failure mode. Whenever $\mathcal{A}$ is consistent, it is trivially guaranteed that there is a $\mathcal{W}_{n,\delta}$ where (3) holds with $\epsilon_{\mathcal{W}}(n, \delta) \to L_{\mathcal{D}}(w^*)$, and so Nagarajan and Kolter's approach is not meaningful for one-sided bounds.[5] Thus it is not possible to mathematically rule out that one could prove a one-sided bound on $\sup_{w\in\mathcal{W}} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)$ using a uniform

[4] Lemma 5.2 of Negrea et al. [24] is closely related; it covers Setting **A** in general, but applies only to $\hat{w}_{MN}$ and shows a smaller gap.

[5] Take $\mathcal{S}_{n,\delta} = \{(X, Y) : L_{\mathcal{D}}(X, Y) \leq L_{\mathcal{D}}(w^*) + \epsilon_{n,\delta}\}$; consistency implies that there is a choice of $\epsilon_{n,\delta} \to 0$ such that $\Pr(\mathbf{S} \in \mathcal{S}_{n,\delta}) \geq 1 - \delta$ and

$$\epsilon_{n,\delta} \geq \sup_{\mathbf{S}\in\mathcal{S}_{n,\delta}} \sup_{w\in\mathcal{W}_{n,\delta}^{\mathcal{A}}} L_{\mathcal{D}}(w) \geq \sup_{\mathbf{S}\in\mathcal{S}_{n,\delta}} \sup_{w\in\mathcal{W}_{n,\delta}^{\mathcal{A}}} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w).$$

convergence-type technique. (Again, since one-sided uniform convergence is always a consequence of consistency, this question is essentially one of viewpoint: do you first show uniform convergence and then bound consistency through uniform convergence, or do you establish uniform convergence as a consequence of consistency?) In any case, as argued by Nagarajan and Kolter, existing uniform convergence proofs essentially bound $|L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)|$, not $L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)$.

## 4 Uniform convergence for interpolating predictors

In Setting **B**, we now know it is impossible to prove consistency of $\hat{w}_{MN}$ with a bound on $\sup_{w \in \mathcal{W}} |L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)|$ for any fixed choice of $\mathcal{W}$, and it seems quite unlikely that we can do so with bounds on $\sup_{w \in \mathcal{W}} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)$ either. However, since we are concerned only with zero-training-error predictors, perhaps we should instead look at bounds on

$$\sup_{\|w\| \leq B,\, L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w). \tag{8}$$

Although $L_{\mathbf{S}}(w)$ is identically 0 in (8), we write it to emphasize that this is still fundamentally a bound on the generalization gap as in (3). When $L_{\mathbf{S}}(w) = 0$, of course, one-sided and two-sided convergence become the same. Moreover, when $B = \|\hat{w}_{MN}\|$, (8) becomes identically $L_{\mathcal{D}}(\hat{w}_{MN})$, which we know from Proposition 2.1 is small. Our questions are (a) whether we could have shown this via uniform convergence, and (b) precisely how small $B$ has to be compared to $\|\hat{w}_{MN}\|$ in order to maintain consistency.

The uniform convergence of (8) is a weaker notion than that of Section 3, as the hypothesis set is sample-dependent. But it is still a standard and common form of "uniform convegnce" at the basis of classical learning theory, and is well understood to be necessary for obtaining tight learning guarantees when we expect the training error to be zero. For example, this is the notion used by Valiant [32] to first establish standard (realizable) PAC-learning guarantees, and is the starting point for standard textbooks, as in Section 2.3.1 of Shalev-Shwartz and Ben-David [28], or Theorem 2.1 of Mohri et al. [19] where that book first introduces the term "uniform convergence bound."

A bound on (8) would be implied by bounds with "optimistic rates" [26, 30], which interpolate between a "fast" rate for $L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)$ and a "slow" one depending on $L_{\mathbf{S}}(w)$. For instance, the result of [30] implies that if $\xi_n$ is a high-probability upper bound on $\max_{1 \leq i \leq n} \|x_i\|^2$, we have uniformly over all $w$ with $\|w\| \leq B$ that

$$L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) \leq \tilde{\mathcal{O}}_P \left( \frac{1}{n} B^2 \xi_n + \sqrt{L_{\mathbf{S}}(w) \frac{B^2 \xi_n}{n}} \right). \tag{9}$$

But the hidden constants and logarithmic factors in (9) do not meet our needs: to show consistency (as we discuss shortly) we need an asymptotic coefficient of 1 on $B^2 \xi_n / n$, while [30] showed only an upper bound of $200\,000 \log^3(n)$. It seems likely given their extremely indirect proof technique, though, that a much tighter version holds – especially in the special case of bounded-norm linear predictors for square loss. Given Proposition 3.1, it is reasonable to suspect that something like the following may hold fairly generally:

$$\sup_{\|w\| \leq B,\, L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) \leq \frac{1}{n} B^2 \xi_n + o_P(1), \tag{$\star$}$$

where here $\xi_n$ might refer either to the high-probability upper bound on $\|x\|^2$ or, for sub-Gaussian data, perhaps simply $\mathbb{E}\|x\|^2$. For either choice of $\xi_n$,[6] by taking $B = \|\hat{w}_{MN}\|$ in Setting **B**, applying Proposition 3.1 then gives us (subject to integrability conditions) that for $\lambda_n = \omega(1)$, $\lambda_n = o(n)$,

$$\lim_{d_J \to \infty} \mathbb{E}\, L_{\mathcal{D}}(\hat{w}_{MN}) = \lim_{d_J \to \infty} \mathbb{E} \left[ \sup_{\|w\| \leq \|\hat{w}_{MN}\|,\, L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) \right] \leq \sigma^2 + o(1). \tag{10}$$

But ($\star$) would also do more than this: it makes predictions about the generalization error of interpolators with larger-than-minimal norm, not yet known in the literature. In the setting of Proposition 3.1, ($\star$) would imply that

$$\lim_{d_J \to \infty} \mathbb{E} \left[ \sup_{\|w\| \leq \alpha \|\hat{w}_{MN}\|, L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) \right] \leq \alpha^2 \left[ \sigma^2 + o(1) \right]. \tag{11}$$

[6] If $\xi_n$ is a high-probability upper bound, we further require $\lambda_n = \omega(\log n)$.

These predictions are important in their own right: outside of linear models, we rarely expect to obtain the interpolator with *exactly* minimal norm.

### 4.1 Uniform convergence of low-norm interpolators in Setting B

The predictions made in (11) in fact hold, with equality.

**Theorem 4.1.** *In Setting **B** with $\lambda_n = o(n)$, fix a sequence $(\alpha_n) \to \alpha$, with each $\alpha_n \geq 1$. Then*

$$\lim_{n\to\infty} \lim_{d_J \to \infty} \mathbb{E}\left[ \sup_{\|w\| \leq \alpha_n \|\hat{w}_{MN}\|,\, L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) \right] = \alpha^2 L_{\mathcal{D}}(w^*).$$

The proof of Theorem 4.1 is based on bounding (8) directly, although it will take us several steps to get there which we now outline. Along the way, we provide results, especially Proposition 4.3, which are applicable well beyond Setting **B**.

The first tool we will require in our analysis is the best-conceivable interpolator for a given $X$ and $\mathcal{D}$:

**Definition 4.2.** *The* minimal-risk interpolator *[21, Section 3.3] is*

$$\hat{w}_{MR} = \operatorname*{arg\,min}_{w \text{ s.t. } Xw=Y} L_{\mathcal{D}}(w) = w^* + \Sigma^{-1} X^{\mathsf{T}} (X \Sigma^{-1} X^{\mathsf{T}})^{-1} E. \tag{12}$$

**Proposition 4.3.** *In Setting **A**, the expected risk of the minimal-risk interpolator is*

$$\mathbb{E}\, L_{\mathcal{D}}(\hat{w}_{MR}) = \frac{p-1}{p-1-n} L_{\mathcal{D}}(w^*).$$

Because $\hat{w}_{MR}$ has perfect knowledge of $\Sigma$, its expected risk turns out to be independent of $\Sigma$. As $p$ increases for fixed $n$ (the second of the double descents), $\mathbb{E}\, L_{\mathcal{D}}(\hat{w}_{MR})$ thus improves monotonically: $\hat{w}_{MR}$ can pick among more interpolators.

We use $\hat{w}_{MR}$ as a constructive tool in our proofs: Theorem 4.5 expands the generalization gap around a fixed predictor in terms of that predictor's risk, and so the minimal-risk predictor is an obvious choice for understanding the gap. Proposition 4.3 also provides lower bounds on interpolation methods: if $p = \mathcal{O}(n)$, then $\hat{w}_{MR}$ is not consistent, and hence no interpolator is. For instance, LASSO is minimax-optimal and consistent for sparse linear regression when $n = \Theta(p)$ [10, 12, 27, 31, 33, 35], but no interpolation method can be. Muthukumar et al. [21, Section 3] discuss this type of result in detail, including for non-Gaussian data; see also [15].

Our next tool measures how much energy in $\Sigma$ is missed by the sample $X$.

**Definition 4.4.** *The* restricted eigenvalue under interpolation *for covariance $\Sigma$ and design $X$ is*

$$\kappa_X(\Sigma) = \sup_{\|w\|=1,\, Xw=0} w^{\mathsf{T}} \Sigma w.$$

We now have the tools to show the following result, which holds even more generally than Setting **A**.

**Theorem 4.5.** *The following results hold deterministically, viewing $L_{\mathcal{D}}(w)$ simply as a quadratic function $L_{\mathcal{D}}(w^*) + \|w - w^*\|_\Sigma$, with no distributional assumptions on* **S**.

*(i) It holds that*

$$\sup_{\substack{\|w\| \leq \|\hat{w}_{MR}\| \\ L_{\mathbf{S}}(w)=0}} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) = L_{\mathcal{D}}(\hat{w}_{MR}) + \gamma_n\, \kappa_X(\Sigma) \left[ \|\hat{w}_{MR}\|^2 - \|\hat{w}_{MN}\|^2 \right]$$

*where $1 \leq \gamma_n \leq 4$.*

*If the minimal risk interpolator is consistent, $\mathbb{E}\, L_{\mathcal{D}}(\hat{w}_{MR}) - L_{\mathcal{D}}(w^*) \to 0$, then the class of interpolators with norm less than $\|\hat{w}_{MR}\|$ is uniformly consistent if and only if*

$$\mathbb{E}\, \kappa_X(\Sigma) \cdot \left[ \|\hat{w}_{MR}\|^2 - \|\hat{w}_{MN}\|^2 \right] \to 0.$$

*(ii) Fix a sequence $(B_n)$ such that $B_n \geq \|\hat{w}_{MN}\|$ for all $n$. Then*

$$\sup_{\|w\| \leq B_n,\, L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) = L_{\mathcal{D}}(\hat{w}_{MN}) + \kappa_X(\Sigma)\left[B_n^2 - \|\hat{w}_{MN}\|^2\right] + R_n$$

*where* $0 \leq R_n \leq 2\sqrt{[L_{\mathcal{D}}(\hat{w}_{MN}) - L_{\mathcal{D}}(w^*)]\,\kappa_X(\Sigma)\,[B_n^2 - \|\hat{w}_{MN}\|^2]}$.

*If* $\mathbb{E}\, L_{\mathcal{D}}(\hat{w}_{MN}) - L_{\mathcal{D}}(w^*) \to 0$*, the class of interpolators with norm less than $B_n$ is thus uniformly consistent if and only if*

$$\mathbb{E}\,\kappa_X(\Sigma) \cdot \left[B_n^2 - \|\hat{w}_{MN}\|^2\right] \to 0.$$

The term $\kappa_X(\Sigma)[B^2 - \|\hat{w}_{MN}\|^2]$ appearing in each bound multiplies $\kappa$, essentially "how much" of $\Sigma$ is orthogonal to the data sample, by the amount of excess norm available inside the norm ball. This result makes us expect that ($\star$) should in fact hold fairly generally with $\xi_n = n\,\kappa_X(\Sigma)$.

Notice also that, of course, $\|\hat{w}_{MN}\| \leq \|\hat{w}_{MR}\|$; thus when $\hat{w}_{MR}$ is consistent (e.g. via Proposition 4.3) and $\mathbb{E}\,\kappa_X(\Sigma)[\|\hat{w}_{MR}\|^2 - \|\hat{w}_{MN}\|^2] \to 0$, then (i) implies $\hat{w}_{MN}$ is consistent as well.

*Proof sketch.* Let $\hat{w}$ be any particular predictor that interpolates the data, and $F \in \mathbb{R}^{p \times (p-n)}$ be the matrix whose columns form an orthonormal basis of the kernel of $X$. Then (8) can be rewritten as

$$\sup_{u \in \mathbb{R}^{p-n} : \|\hat{w}+Fu\|^2 \leq B^2} \|\hat{w} + Fu - w^*\|_\Sigma^2. \tag{13}$$

This is a quadratic program with a single quadratic constraint, which enjoys strong duality even though it is a convex *maximization* [9, Appendix B]. We thus need analyze only the (much simpler) one-dimensional dual problem. For (ii), we take $\hat{w} = \hat{w}_{MN}$ in (13) and obtain the dual as

$$\inf_{\lambda > \|F^\mathsf{T}\Sigma F\|} L_{\mathcal{D}}(\hat{w}_{MN}) + \|F^\mathsf{T}\Sigma(\hat{w}_{MN} - w^*)\|^2_{(\lambda I_{p-n} - F^\mathsf{T}\Sigma F)^{-1}} + \lambda\Big[B_n^2 - \|\hat{w}_{MN}\|^2\Big].$$

Given consistency, we can show that the second term's contribution is negligible, as

$$\|F^\mathsf{T}\Sigma(\hat{w}_{MN} - w^*)\|^2 \leq \|F^\mathsf{T}\Sigma F\| \cdot [L_{\mathcal{D}}(\hat{w}_{MN}) - L_{\mathcal{D}}(w^*)],$$

and $(\lambda I_{p-n} - F^\mathsf{T}\Sigma F)^{-1}$ has controlled eigenvalues so that the Mahalanobis norm is similar to the Euclidean norm. Observing that $\kappa_X(\Sigma) = \|F^\mathsf{T}\Sigma F\|$, the conclusion follows by routine calculations.

Case (i) uses a similar strategy, taking $\hat{w} = \hat{w}_{MR}$. The full proof is given in Appendix C.2. ☐

Now, all that remains is to evaluate the relevant quantities in Setting **B**.

**Proposition 4.6.** *In Setting **B** with $\lambda_n = o(n)$,*

$$\lim_{n\to\infty} \lim_{d_J \to \infty} \mathbb{E}\left[\sup_{\|w\| \leq \|\hat{w}_{MR}\|,\, L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)\right] = L_{\mathcal{D}}(w^*).$$

*Proof sketch for Theorem 4.1 and Proposition 4.6.* We apply Theorem 4.5. With probability one,

$$\lim_{d_J \to \infty} \kappa_X(\Sigma) = \frac{\lambda_n}{n}\left\|\left[\frac{X_S^\mathsf{T} X_S}{n} + \frac{\lambda_n}{n} I_{d_S}\right]^{-1}\right\|.$$

As the first term inside the inverse converges to $I_{d_S}$ and the second term vanishes, we can expect $\kappa_X(\Sigma) \approx \lambda_n/n$. We bound the other terms by observing that there exists a sequence $\beta_n \to 1$ with

$$\lim_{d_J\to\infty} \mathbb{E}\|\hat{w}_{MR}\|^2 = \|w_S^*\|^2 + \frac{\sigma^2 n}{\lambda_n}$$

$$\lim_{d_J\to\infty} \mathbb{E}\|\hat{w}_{MN}\|^2 = \|w^*\|^2 + \sigma^2\frac{n - d_S}{\lambda_n} + \beta_n\left(\frac{\sigma^2 d_S - \lambda_n\|w_S^*\|^2}{n}\right),$$

so $\lim_{d_J\to\infty} \mathbb{E}\left[\|\hat{w}_{MR}\|^2 - \mathbb{E}\|\hat{w}_{MN}\|^2\right] = \sigma^2 d_S/\lambda_n + \mathcal{O}\left(\lambda_n\|w^*\|^2/n\right)$.

Because $\hat{w}_{MR}$ is consistent via Proposition 4.3, this proves Proposition 4.6. As $\|\hat{w}_{MN}\| \leq \|\hat{w}_{MR}\|$, this further implies $\hat{w}_{MN}$ is consistent, so that the $R_n$ term of Theorem 4.5 (ii) vanishes. ☐

We can see that $\kappa_X(\Sigma)$ tends to 0 while $\|\hat{w}_{MN}\|$ explodes, and in Setting **B** their product turns out to converge to *exactly* the Bayes risk. Because the other terms of Theorem 4.5 (ii) cancel, this gives us precisely the tight result we need for Theorem 4.1, and further suggests that the speculative upper bound $\kappa_X(\Sigma)B^2$ probably holds in more general settings.

We have at last shown in Theorem 4.1 a uniform convergence bound not only showing consistency of $\hat{w}_{MN}$, but furthermore verifying the predictions of (11). Thus if we obtain an interpolator with norm $1.1\|\hat{w}_{MN}\|$, we will suffer at most $1.21\sigma^2$ asymptotic risk. If we obtain an interpolator with norm no more than a constant amount larger than the minimal norm, we achieve asymptotic consistency.

## 5 Discussion

In this work, we shed new light on uniform convergence and its relationship to interpolation learning. We show that uniform control of the generalization gap cannot explain interpolation learning, for almost *any* interpolator, even in a simple setting. But we argue that when discussing "uniform convergence" in the context of interpolation learning, we should slightly broaden our horizons to include interpolation-specific uniform convergence bounds such as ($\star$), or more generally "optimistic" (training-error-dependent) bounds [26, 30]. We show that despite recent sentiments to the contrary, such bounds *could* in principal explain interpolation learning, by demonstrating this in the "junk features" setting. Doing so requires obtaining very tight bounds, include tight constants – perhaps a difficult task, but not impossible. (For example, for linear predictors with a Lipschitz loss in a non-realizable setting, we do know the exact worst-case bound, with a tight numeric constant [16].)

Our results are also of independent interest in ensuring success with interpolation learning: in settings other than linear regression, where a closed-form solution is available, it is generally unlikely in practice that we find the *exact* minimum-norm solution. (Even gradient descent for linear regression would find this only when initialized exactly in the span of the data; other forms of implicit bias are likewise suboptimal.) Our results give some reassurance that, at least in this simple setting, approximately minimizing the norm is sufficient. The natural next step in this vein would be to study predictors with small but nonzero loss. This could either be done directly in the style of our Theorem 4.1, or by providing an optimistic rate as in (9) with tight constants. Our specific techniques, as well as the general takeaway of considering interpolation-specific bounds, could also be potentially applicable to settings beyond linear regression, especially the idea of studying the generalization gap via the dual problem: although strong duality may not be available in more general settings, upper bounds are always possible with weak duality.

## Broader Impact

Interpolation learning is currently thought to be one of the core mysteries standing between us and a theoretical understanding of modern deep learning. Although there has recently been some key progress, many challenges remain. Our paper, in advancing the study of interpolation learning, makes another step on the path towards understanding the deep learning models that are quickly becoming ubiquitous throughout society, whether we understand them or not. In our view, increased understanding of these models can lead to safer, more reliable, and more controlled deployment, especially in sensitive domains.

In particular, we discuss a key component of statistical learning theory, namely uniform convergence, whose relevance to deep learning in general – and interpolation learning specifically – has recently been questioned. We make an explicit connection between the work on interpolation learning and the recent notion of "algorithmic dependent uniform convergence" [22]. Instead of outright dismissal, we show that a more nuanced view is appropriate. By doing so, we hope to help guide the re-pivoting that statistical learning theory is currently undergoing.

We emphasize that, despite providing some positive theoretical results, we are certainly not advocating for preferring interpolation methods over other approaches. In particular, the increased sensitivity of interpolation methods may have problematic ramifications for robustness or privacy.

## Acknowledgments and Disclosure of Funding

Research supported in part by NSF IIS award 1764032 and NSF HDR TRIPODS award 1934843.

## References

[1] Madhu S Advani and Andrew M Saxe. "High-dimensional dynamics of generalization error in neural networks" (2017). arXiv: 1710.03667.

[2] Jimmy Ba, Murat Erdogdu, Taiji Suzuki, Denny Wu, and Tianzong Zhang. "Generalization of Two-layer Neural Networks: An Asymptotic Viewpoint." *International Conference on Learning Representations*. 2020.

[3] Peter L. Bartlett, Philip M. Long, Gábor Lugosi, and Alexander Tsigler. "Benign overfitting in linear regression." *Proceedings of the National Academy of Sciences* 117.48 (2020), pp. 30063–30070. arXiv: 1906.11300.

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

[5] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. "Reconciling modern machine learning practice and the bias-variance trade-off." *Proceedings of the National Academy of Sciences* 116.32 (2019), pp. 15849–15854. arXiv: 1812.11118.

[6] Mikhail Belkin, Daniel Hsu, and Ji Xu. "Two models of double descent for weak features." *SIAM Journal on Mathematics of Data Science* 2.4 (2020), pp. 1167–1180. arXiv: 1903.07571.

[7] Mikhail Belkin, Siyuan Ma, and Soumik Mandal. "To understand deep learning we need to understand kernel learning." *International Conference on Machine Learning*. 2018. arXiv: 1802.01396.

[8] Mikhail Belkin, Alexander Rakhlin, and Alexandre B. Tsybakov. "Does data interpolation contradict statistical optimality?" *International Conference on Artificial Intelligence and Statistics*. 2019. arXiv: 1806.09471.

[9] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[10] Emmanuel Candes and Terence Tao. "The Dantzig selector: Statistical estimation when $p$ is much larger than $n$." *The Annals of Statistics. Vol. 35, No. 6, 2313–2351* (2005). arXiv: 0506081.

[11] Andrea Caponnetto and Ernesto De Vito. "Optimal rates for regularized least-squares algorithm." *Foundations of Computational Mathematics* 7.3 (2007), pp. 331–368.

[12] Sourav Chatterjee. "Assumptionless consistency of the Lasso" (2013). arXiv: 1303.5817.

[13] Michał Dereziński, Feynman Liang, and Michael W. Mahoney. "Exact expressions for double descent and implicit regularization via surrogate random design." *Advances in Neural Information Processing Systems*. 2020. arXiv: 1912.04533.

[14] Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. "Surprises in High-Dimensional Ridgeless Least Squares Interpolation" (2019). arXiv: 1903.08560.

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

[16] Sham M Kakade, Karthik Sridharan, and Ambuj Tewari. "On the complexity of linear prediction: Risk bounds, margin bounds, and regularization." *Advances in Neural Information Processing Systems*. 2009, pp. 793–800.

[17] Vladimir Koltchinskii and Karim Lounici. "Concentration Inequalities and Moment Bounds for Sample Covariance Operators." *Bernoulli* 23.1 (2017), pp. 110–133. arXiv: 1405.2468.

[18] Song Mei and Andrea Montanari. "The generalization error of random features regression: Precise asymptotics and double descent curve" (2019). arXiv: 1908.05355.

[19] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning*. MIT Press, 2012.

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

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

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

[23] Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. "Deep Double Descent: Where Bigger Models and More Data Hurt." *International Conference on Learning Representations*. 2020. arXiv: 1912.02292.

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

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

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

[27] Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. "Minimax rates of estimation for high-dimensional linear regression over $l_q$-balls." *IEEE Transactions on Information Theory* 57.10 (2011), pp. 6976–6994. arXiv: 0910.2042.

[28] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

[29] Stefano Spigler, Mario Geiger, Stéphane d'Ascoli, Levent Sagun, Giulio Biroli, and Matthieu Wyart. "A jamming transition from under- to over-parametrization affects generalization in deep learning." *Journal of Physics A: Mathematical and Theoretical* 52.47 (2018), p. 474001. arXiv: 1810.09665.

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

[31] Robert Tibshirani. "Regression Shrinkage and Selection via the Lasso." *Journal of the Royal Statistical Society. Series B, Vol. 58, No. 1, pp. 267-288* (1996).

[32] Leslie G Valiant. "A theory of the learnable." *Communications of the ACM* 27.11 (1984), pp. 1134–1142.

[33] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019.

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

[35] Peng Zhao and Bin Yu. "On Model Selection Consistency of Lasso." *Journal of Machine Learning Research 7, 2541-2563* (2006).

## A Proofs for Section 2

**Lemma 2.3.** *In Setting **B**, if $\lambda_n = o(n)$, then* $\lim_{n\to\infty} \mathbb{E}\left[L_{\mathcal{D}}(\hat{w}_{\lambda_n}) - L_{\mathcal{D}}(w^*)\right] = 0$.

*Proof.* We can write

$$\begin{aligned}
\hat{w}_{\lambda_n} - w_S^* &= (X_S^\mathsf{T} X_S + \lambda_n I_{d_S})^{-1} X_S^\mathsf{T}(X_S w_S^* + E) - w_S^* \\
&= ((X_S^\mathsf{T} X_S + \lambda_n I_{d_S})^{-1} X_S^\mathsf{T} X_S - I_{d_S}) w_S^* + (X_S^\mathsf{T} X_S + \lambda_n I_{d_S})^{-1} X_S^\mathsf{T} E \\
&= \left[\left(\frac{X_S^\mathsf{T} X_S}{n} + \frac{\lambda_n}{n} I_{d_S}\right)^{-1} \frac{X_S^\mathsf{T} X_S}{n} - I_{d_S}\right] w_S^* + \left(\frac{X_S^\mathsf{T} X_S}{n} + \frac{\lambda_n}{n} I_{d_S}\right)^{-1} \frac{X_S^\mathsf{T} E}{n}.
\end{aligned}$$

Therefore, by independence of $X_S$ and $E$,

$$\begin{aligned}
&\mathbb{E}[L_{\mathcal{D}}(\hat{w}_{\lambda_n}) - L_{\mathcal{D}}(w^*)] = \mathbb{E}\|\hat{w}_{\lambda_n} - w_S^*\|^2 \\
&= \mathbb{E}\left\|\left[\left(\frac{X_S^\mathsf{T} X_S}{n} + \frac{\lambda_n}{n} I_{d_S}\right)^{-1} \frac{X_S^\mathsf{T} X_S}{n} - I_{d_S}\right] w_S^*\right\|^2 + \mathbb{E}\left\|\left(\frac{X_S^\mathsf{T} X_S}{n} + \frac{\lambda_n}{n} I_{d_S}\right)^{-1} \frac{X_S^\mathsf{T} E}{n}\right\|^2 \\
&= \mathbb{E}\left\|\left[\left(\frac{X_S^\mathsf{T} X_S}{n} + \frac{\lambda_n}{n} I_{d_S}\right)^{-1} \frac{X_S^\mathsf{T} X_S}{n} - I_{d_S}\right] w_S^*\right\|^2 + \sigma^2\, \mathbb{E}\, \frac{1}{n} \operatorname{Tr}\left[\left(\frac{X_S^\mathsf{T} X_S}{n} + \frac{\lambda_n}{n} I_{d_S}\right)^{-2} \frac{X_S^\mathsf{T} X_S}{n}\right].
\end{aligned}$$

Write the SVD for $X_S = UDV^\mathsf{T}$. Since $X_S$ has rank at most $d_S$, we denote its singular values as $\sqrt{\rho_1}, ..., \sqrt{\rho_{d_S}}$, and

$$\|(X_S^\mathsf{T} X_S + \lambda I_{d_S})^{-1} X_S^\mathsf{T} X_S\| = \|(D^\mathsf{T} D + \lambda I_{d_S})^{-1} D^\mathsf{T} D\| = \max_{i\in[p]} \frac{\rho_i}{\lambda_n + \rho_i} \le 1.$$

Thus, we have

$$\left\|\left[\left(\frac{X_S^\mathsf{T} X_S}{n} + \frac{\lambda}{n} I_{d_S}\right)^{-1} \frac{X_S^\mathsf{T} X_S}{n} - I_{d_S}\right] w_S^*\right\|^2 \le (1+1)^2 \|w_S^*\|^2 = 4\|w_S^*\|^2$$

which is clearly integrable.

As $d_S$ stays fixed as $n \to \infty$, by the strong law of large numbers we have $\frac{X_S^\mathsf{T} X_S}{n} \to I_{d_S}$. Assuming that $\frac{\lambda_n}{n} \to \gamma$, then by the continuous mapping and dominated convergence theorems, the first term converges to

$$\mathbb{E} \lim_{n\to\infty} \left\|\left[1 - (1+\gamma)^{-1}\right] w_S^*\right\|^2 = \left(\frac{\gamma}{1+\gamma} \cdot \|w_S^*\|\right)^2,$$

Moreover, it holds that

$$\begin{aligned}
\frac{1}{n} \operatorname{Tr}\left[\left(\frac{X_S^\mathsf{T} X_S}{n} + \frac{\lambda_n}{n} I_{d_S}\right)^{-2} \frac{X_S^\mathsf{T} X_S}{n}\right] &= \sum_{i=1}^{d_S} \left(\frac{\sqrt{\rho_i}}{\rho_i + \lambda_n}\right)^2 \\
&\le \sum_{i=1}^{d_S} \frac{1}{\rho_i} = \operatorname{Tr}\left[\left(X_S^\mathsf{T} X_S\right)^{-1}\right]
\end{aligned}$$

Using the first moment of inverse Wishart distribution, the second term can be controlled by

$$\sigma^2\, \mathbb{E} \operatorname{Tr}\left[\left(X_S^\mathsf{T} X_S\right)^{-1}\right] = \sigma^2 \frac{d_S}{n - d_S - 1} \to 0$$

Note that the first term converges to 0 as long as $\gamma = 0$, and the desired conclusion follows. ∎

## B Proofs for Section 3

### B.1 Size of the minimal-norm interpolator (Proposition 3.1)

**Proposition B.1.** *In Setting **B**, it holds that*

$$\lim_{d_J\to\infty} \mathbb{E}\|\hat{w}_{MR}\|^2 = \|w^*\|^2 + \frac{\sigma^2 n}{\lambda_n}.$$

*Moreover, there exists a sequence $(\beta_n)$ such that $\beta_n \to 1$ and*

$$\lim_{d_J\to\infty} \mathbb{E}\|\hat{w}_{MN}\|^2 = \|w^*\|^2 + \sigma^2\frac{n-d_S}{\lambda_n} + \beta_n\left(\frac{\sigma^2 d_S - \lambda_n\|w_S^*\|^2}{n}\right).$$

*Consequently, we have*

$$\lim_{d_J\to\infty} \mathbb{E}\left[\|\hat{w}_{MR}\|^2 - \|\hat{w}_{MN}\|^2\right] = \frac{\sigma^2 d_S}{\lambda_n} + \beta_n\left(\frac{\lambda_n\|w_S^*\|^2 - \sigma^2 d_S}{n}\right).$$

*Proof.* Let $\{e_i\}$ be the standard basis in $\mathbb{R}^p$ and write $\Sigma = \sum_{i=1}^p \mu_i e_i e_i^T$, with $\mu_i = 1$ for $1 \le i \le d_S$ and $\mu_i = \lambda_n/d_J$ for $i > d_S$. By independence of $X$ and $E$, we have

$$\begin{aligned}
\mathbb{E}\|\hat{w}_{MR}\|^2 &= \|w^*\|^2 + \mathbb{E}\|\Sigma^{-1}X^\mathsf{T}(X\Sigma^{-1}X^\mathsf{T})^{-1}E\|^2 \\
&= \|w^*\|^2 + \sigma^2\,\mathbb{E}\left[\mathrm{Tr}\left((ZZ^\mathsf{T})^{-1}(Z\Sigma^{-1}Z^\mathsf{T})(ZZ^\mathsf{T})^{-1}\right)\right] \\
&= \|w^*\|^2 + \sum_{i=1}^p \frac{\sigma^2}{\mu_i}\,\mathbb{E}\left[\|(ZZ^T)^{-1}Ze_i\|^2\right].
\end{aligned}$$

By rotational invariance of the standard normal distribution for $Z$, we have

$$\mathbb{E}\left[\|(ZZ^T)^{-1}Ze_i\|^2\right] = \frac{\mathbb{E}\,\mathrm{Tr}(Z^T(ZZ^T)^{-2}Z)}{p} = \frac{\mathbb{E}\,\mathrm{Tr}((ZZ^T)^{-1})}{p} = \frac{n}{p(p-n-1)}.$$

Plugging in, we get

$$\begin{aligned}
\mathbb{E}\|\hat{w}_{MR}\|^2 &= \|w^*\|^2 + \left(\sum_{i=1}^p \frac{\sigma^2}{\mu_i}\right)\frac{n}{p(p-n-1)} \\
&= \|w^*\|^2 + \sigma^2\left(d_S + \frac{d_J^2}{\lambda_n}\right)\frac{n}{p(p-n-1)}.
\end{aligned}$$

Sending $d_J \to \infty$ and recalling $p = d_S + d_J$, we obtain

$$\lim_{d_J\to\infty} \mathbb{E}\|\hat{w}_{MR}\|^2 = \|w^*\|^2 + \frac{\sigma^2 n}{\lambda_n}.$$

Moreover, it holds that

$$\begin{aligned}
\|\hat{w}_{MR}\|^2 &= \|w^*\|^2 + \mathrm{Tr}\left((ZZ^\mathsf{T})^{-1}(Z\Sigma^{-1}Z^\mathsf{T})(ZZ^\mathsf{T})^{-1}EE^\mathsf{T}\right) + 2\langle w^*, \Sigma^{-1/2}Z^\mathsf{T}(ZZ^\mathsf{T})^{-1}E\rangle \\
&= \|w^*\|^2 + \mathrm{Tr}\left(\left(\frac{ZZ^\mathsf{T}}{p}\right)^{-1}\left(\frac{Z\Sigma^{-1}Z^\mathsf{T}}{p^2}\right)\left(\frac{ZZ^\mathsf{T}}{p}\right)^{-1}EE^\mathsf{T}\right) + 2\left\langle\frac{Z\Sigma^{-1/2}w^*E^\mathsf{T}}{p}, \left(\frac{ZZ^\mathsf{T}}{p}\right)^{-1}\right\rangle.
\end{aligned}$$

Notice that

$$\lim_{d_J\to\infty}\left(\frac{ZZ^\mathsf{T}}{p}\right)^{-1} \overset{a.s.}{=} I_n$$

$$\lim_{d_J\to\infty}\frac{Z\Sigma^{-1}Z^\mathsf{T}}{p^2} = \lim_{d_J\to\infty}\frac{1}{p^2}\left(Z_S Z_S^\mathsf{T} + \frac{d_J^2}{\lambda_n}\frac{Z_J Z_J^\mathsf{T}}{d_J}\right) \overset{a.s.}{=} \frac{1}{\lambda_n}I_n$$

$$Z\Sigma^{-1/2}w^*E^\mathsf{T} = [Z_S \quad Z_J]\begin{bmatrix} I_{d_S} & 0_{d_S\times d_J} \\ 0_{d_J\times d_S} & \sqrt{\frac{d_J}{\lambda_n}}I_{d_J}\end{bmatrix}\begin{bmatrix} w_S^* \\ 0_{d_J}\end{bmatrix}E^\mathsf{T} = Z_S w_S^* E^\mathsf{T} \implies \frac{Z\Sigma^{-1/2}w^*E^\mathsf{T}}{p} \overset{a.s.}{=} 0.$$

Plugging in, we obtain

$$\lim_{d_J\to\infty}\|\hat{w}_{MR}\|^2 \overset{a.s.}{=} \|w^*\|^2+\frac{\|E\|^2}{\lambda_n},\quad \text{and so}\quad \mathbb{E}\left[\lim_{d_J\to\infty}\|\hat{w}_{MR}\|^2\right]=\lim_{d_J\to\infty}\mathbb{E}\|\hat{w}_{MR}\|^2.$$

Clearly, the sequence of random variables $(\|\hat{w}_{MR}\|^2)$ as we let $d_J\to\infty$ dominates $(\|\hat{w}_{MN}\|^2)$. By the dominated convergence theorem [7]

$$\begin{aligned}
\lim_{d_J\to\infty}\mathbb{E}\|\hat{w}_{MN}\|^2 &= \mathbb{E}\left[\lim_{d_J\to\infty}\|\hat{w}_{MN}\|^2\right]\\
&= \mathbb{E}\left[\lim_{d_J\to\infty}(X_Sw_S^*+E)^\mathsf{T}(XX^\mathsf{T})^{-1}XX^\mathsf{T}(XX^\mathsf{T})^{-1}(X_Sw_S^*+E)\right]\\
&= \mathbb{E}\left[\lim_{d_J\to\infty}(X_Sw_S^*+E)^\mathsf{T}(X_SX_S^\mathsf{T}+X_JX_J^\mathsf{T})^{-1}(X_Sw_S^*+E)\right]\\
&= \mathbb{E}\left[(X_Sw_S^*+E)^\mathsf{T}(X_SX_S^\mathsf{T}+\lambda_nI_n)^{-1}(X_Sw_S^*+E)\right]\\
&= (w_S^*)^\mathsf{T}\,\mathbb{E}[X_S^\mathsf{T}(X_SX_S^\mathsf{T}+\lambda_nI_n)^{-1}X_S]w_S^*+\sigma^2\,\mathbb{E}\,\mathrm{Tr}\left((X_SX_S^\mathsf{T}+\lambda_nI_n)^{-1}\right).
\end{aligned}$$

With probability one, $X_SX_S^\mathsf{T}$ is a $n\times n$ matrix with rank $d_S$, so the eigenvalues of $(X_SX_S^\mathsf{T}+\lambda_nI_n)^{-1}$ consist of the $d_S$ eigenvalues of $(X_S^\mathsf{T}X_S+\lambda_nI_{d_S})^{-1}$ and $(n-d_S)$ copies of $\frac{1}{0+\lambda_n}$. This implies

$$\sigma^2\,\mathbb{E}\,\mathrm{Tr}\left((X_SX_S^\mathsf{T}+\lambda I_n)^{-1}\right)=\sigma^2\,\mathbb{E}\,\mathrm{Tr}\left((X_S^\mathsf{T}X_S+\lambda I_{d_S})^{-1}\right)+\sigma^2\frac{n-d_S}{\lambda_n}.$$

Moreover, by the rotational invariance of $X_S\sim\mathcal{N}(0,I_{d_S})$,

$$\begin{aligned}
(w_S^*)^\mathsf{T}\,\mathbb{E}[X_S^\mathsf{T}(X_SX_S^\mathsf{T}+\lambda_nI_n)^{-1}X_S]w_S^* &= \frac{\|w_S^*\|^2}{d_S}\,\mathbb{E}\,\mathrm{Tr}\left(X_S^\mathsf{T}(X_SX_S^\mathsf{T}+\lambda_nI_n)^{-1}X_S\right)\\
&= \frac{\|w_S^*\|^2}{d_S}\,\mathbb{E}\,\mathrm{Tr}\left(X_S^\mathsf{T}X_S(X_S^\mathsf{T}X_S+\lambda_nI_{d_S})^{-1}\right)\\
&= \frac{\|w_S^*\|^2}{d_S}\,\mathbb{E}\,\mathrm{Tr}\left(I_{d_S}-\lambda_n(X_S^\mathsf{T}X_S+\lambda_nI_{d_S})^{-1}\right)\\
&= \|w_S^*\|^2-\frac{\lambda_n\|w_S^*\|^2}{d_S}\,\mathbb{E}\,\mathrm{Tr}\left((X_S^\mathsf{T}X_S+\lambda_nI_{d_S})^{-1}\right).
\end{aligned}$$

Plugging in, we get

$$\begin{aligned}
\lim_{d_J\to\infty}\mathbb{E}\|\hat{w}_{MN}\|^2 &= \|w^*\|^2+\sigma^2\frac{n-d_S}{\lambda_n}+\left(\sigma^2-\frac{\lambda_n\|w_S^*\|^2}{d_S}\right)\mathbb{E}\,\mathrm{Tr}\left((X_S^\mathsf{T}X_S+\lambda_nI_{d_S})^{-1}\right)\\
&= \|w^*\|^2+\sigma^2\frac{n-d_S}{\lambda_n}+\left(\frac{\sigma^2d_S-\lambda_n\|w_S^*\|^2}{n}\right)\cdot\left[\frac{\mathbb{E}\,\mathrm{Tr}\left(\left(\frac{X_S^\mathsf{T}X_S}{n}+\frac{\lambda_n}{n}I_{d_S}\right)^{-1}\right)}{d_S}\right].
\end{aligned}$$

[7] We use the following version of the theorem, which is slightly more general than the usual one. Suppose there exists a sequence of $l_1$ random variables $Y_n$ such that $Y_n\geq X_n$ and

$$\lim_{n\to\infty}\mathbb{E}\,Y_n=\mathbb{E}\lim_{n\to\infty}Y_n;$$

then we have

$$\lim_{n\to\infty}\mathbb{E}\,X_n=\mathbb{E}\lim_{n\to\infty}X_n.$$

The proof is essentially the same and applies Fatou's lemma to $X_n$ and $Y_n-X_n$.

As $\operatorname{Tr}\left(\left(\frac{X_S^\mathsf{T} X_S}{n}\right)^{-1}\right)$, which has limit $d_S$ in expectation,[8] dominates $\operatorname{Tr}\left(\left(\frac{X_S^\mathsf{T} X_S}{n} + \frac{\lambda_n}{n} I_{d_S}\right)^{-1}\right)$, by the dominated convergence theorem

$$\lim_{n\to\infty} \frac{1}{d_S} \mathbb{E} \operatorname{Tr}\left(\left(\frac{X_S^\mathsf{T} X_S}{n} + \frac{\lambda_n}{n} I_{d_S}\right)^{-1}\right) = 1.$$

Letting the term in brackets be $\beta_n$, we have the result. □

**Proposition 3.1.** *As $n \to \infty$ in Setting **B**, if $\lambda_n$ is both $o(n)$ and $\omega(1)$, then*

$$\lim_{d_J\to\infty} \mathbb{E}\|\hat{w}_{MN}\|^2 = \frac{\sigma^2 n}{\lambda_n} + \mathcal{O}(1) \quad \textit{and} \quad \lim_{d_J\to\infty} \frac{(\mathbb{E}\|\hat{w}_{MN}\|^2)(\mathbb{E}\|x\|^2)}{n} = \sigma^2 + o(1).$$

*Proof.* By Proposition B.1, there exists a sequence $(\beta_n)$ such that $\beta_n \to 1$ and

$$\lim_{d_J\to\infty} \mathbb{E}\|\hat{w}_{MN}\|^2 = \sigma^2 \frac{n}{\lambda_n} + \left[\|w^*\|^2 - \sigma^2 \frac{d_S}{\lambda_n} + \beta_n \left(\frac{\sigma^2 d_S - \lambda_n \|w_S^*\|^2}{n}\right)\right].$$

Moreover, we have

$$\mathbb{E}\|x\|^2 = \operatorname{Tr}(\Sigma) = d_S \cdot 1 + d_J \cdot \frac{\lambda_n}{d_J} = d_S + \lambda_n.$$

Plugging in, we obtain

$$\frac{(\mathbb{E}\|\hat{w}_{MN}\|^2)(\mathbb{E}\|x\|^2)}{n} = \sigma^2 \frac{d_S + \lambda_n}{\lambda_n} + \frac{d_S + \lambda_n}{n}\left[\|w^*\|^2 - \sigma^2 \frac{d_S}{\lambda_n} + \beta_n \left(\frac{\sigma^2 d_S - \lambda_n \|w_S^*\|^2}{n}\right)\right].$$

By assumption, $1/\lambda_n \to 0$ and $\lambda_n/n \to 0$; thus the dominant term inside the brackets is $\|w^*\|^2 = \mathcal{O}(1)$. The conclusion follows by

$$\frac{d_S + \lambda_n}{\lambda_n} \to 1 \quad \text{and} \quad \frac{d_S + \lambda_n}{n} \to 0.$$ □

### B.2 Divergence of the generalization gap of norm balls (Section 3.1)

**Proposition B.2.** *Let $\rho(\Sigma - \hat{\Sigma})$ be the algebraically largest eigenvalue of $\Sigma - \hat{\Sigma}$. It holds that*

$$\sup_{\|w\| \le \|\hat{w}_{MN}\|} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) \ge \rho(\Sigma - \hat{\Sigma}) \cdot (\|\hat{w}_{MN}\| - \|w^*\|)^2 + \left[L_{\mathcal{D}}(w^*) - \frac{1}{n}\|E\|^2\right]$$

*and similarly for two sided uniform convergence, it holds that*

$$\sup_{\|w\| \le \|\hat{w}_{MN}\|} |L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)| \ge \|\Sigma - \hat{\Sigma}\| \cdot (\|\hat{w}_{MN}\| - \|w^*\|)^2 - \left|L_{\mathcal{D}}(w^*) - \frac{\|E\|^2}{n}\right|.$$

*Proof.* Recall from (1) that

$$\begin{aligned} L_{\mathbf{S}}(w) &= \frac{1}{n}\|Xw - Y\|^2 \\ &= \frac{1}{n}\|X(w - w^*) + Xw^* - Y\|^2 \\ &= (w - w^*)^\mathsf{T} \hat{\Sigma} (w - w^*) + \frac{\|E\|^2}{n} - 2\left\langle w - w^*, \frac{X^\mathsf{T} E}{n}\right\rangle. \end{aligned}$$

[8] Using standard properties of the inverse Wishart distribution, we can check that

$$\lim_{n\to\infty} \mathbb{E} \operatorname{Tr}\left(\left(\frac{X_S^\mathsf{T} X_S}{n}\right)^{-1}\right) = d_S = \mathbb{E} \lim_{n\to\infty} \operatorname{Tr}\left(\left(\frac{X_S^\mathsf{T} X_S}{n}\right)^{-1}\right).$$

Therefore, we can decompose the generalization gap as

$$\begin{aligned} L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w) &= L_{\mathcal{D}}(w^*) + (w-w^*)^{\mathsf{T}}\Sigma(w-w^*) - L_{\mathbf{S}}(w) \\ &= \left[L_{\mathcal{D}}(w^*) - \frac{\|E\|^2}{n}\right] + (w-w^*)^{\mathsf{T}}(\Sigma-\hat{\Sigma})(w-w^*) + 2\Big\langle w-w^*, \frac{X^{\mathsf{T}}E}{n}\Big\rangle. \end{aligned}$$

Observe that

$$\begin{aligned} &\sup_{\|w\|\le\|\hat{w}_{MN}\|} (w-w^*)^{\mathsf{T}}(\Sigma-\hat{\Sigma})(w-w^*) + 2\Big\langle w-w^*, \frac{X^{\mathsf{T}}E}{n}\Big\rangle \\ &\qquad\qquad \ge \sup_{\|w\|\le\|\hat{w}_{MN}\|-\|w^*\|} w^{\mathsf{T}}(\Sigma-\hat{\Sigma})w + 2\Big\langle w, \frac{X^{\mathsf{T}}E}{n}\Big\rangle \\ &\qquad\qquad \ge \rho(\Sigma-\hat{\Sigma})\cdot(\|\hat{w}_{MN}\| - \|w^*\|)^2. \end{aligned}$$

The last inequality holds by picking $w$ to be $\pm(\|\hat{w}_{MN}\| - \|w^*\|)$ times the top eigenvector of $\Sigma - \hat{\Sigma}$ for whichever sign makes the linear term nonnegative. By the same reasoning, we have

$$\sup_{\|w\|\le\|\hat{w}_{MN}\|} |L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)| \ge \|\Sigma-\hat{\Sigma}\|\cdot(\|\hat{w}_{MN}\| - \|w^*\|)^2 - \Big|L_{\mathcal{D}}(w^*) - \frac{\|E\|^2}{n}\Big|. \qquad \square$$

**Theorem 3.2.** *In Setting **B**, if $\lambda_n = o(n)$ then*

$$\lim_{n\to\infty}\lim_{d_J\to\infty} \mathbb{E}\left[\sup_{\|w\|\le\|\hat{w}_{MN}\|} |L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)|\right] = \infty.$$

*Proof.* We will show that in Setting **B** as long as $\lambda_n = o(n)$,

$$\lim_{n\to\infty}\lim_{d_J\to\infty} \mathbb{E}\|\Sigma-\hat{\Sigma}\|\cdot\|\hat{w}_{MN}\|^2 = \infty.$$

By Fatou's lemma and the calculation in Proposition B.1,

$$\begin{aligned} \lim_{d_J\to\infty} \mathbb{E}\|\Sigma-\hat{\Sigma}\|\cdot\|\hat{w}_{MN}\|^2 &\ge \mathbb{E}\lim_{d_J\to\infty} \|\Sigma-\hat{\Sigma}\|\cdot\|\hat{w}_{MN}\|^2 \\ &= \mathbb{E}\lim_{d_J\to\infty} \|\Sigma-\hat{\Sigma}\|\cdot\left((X_S w_S^* + E)^{\mathsf{T}}(X_S X_S^{\mathsf{T}} + \lambda_n I_n)^{-1}(X_S w_S^* + E)\right). \end{aligned}$$

By independence of $X$ and $E$, we have

$$\begin{aligned} \lim_{d_J\to\infty} \mathbb{E}\left[\|\Sigma-\hat{\Sigma}\|\cdot\|\hat{w}_{MN}\|^2\right] &\ge \mathbb{E}\lim_{d_J\to\infty} \|\Sigma-\hat{\Sigma}\|\cdot\left(E^{\mathsf{T}}(X_S X_S^{\mathsf{T}} + \lambda_n I_n)^{-1}E\right) \\ &= \sigma^2\,\mathbb{E}\left[\lim_{d_J\to\infty} \|\Sigma-\hat{\Sigma}\|\cdot\operatorname{Tr}\left((X_S X_S^{\mathsf{T}} + \lambda_n I_n)^{-1}\right)\right] \\ &\ge \sigma^2\,\mathbb{E}\left[\lim_{d_J\to\infty} \|\Sigma-\hat{\Sigma}\|\cdot\left(\frac{n-d_S}{\lambda_n}\right)\right] \\ &= \left(\sigma^2\frac{n-d_S}{\lambda_n}\right)\mathbb{E}\left[\lim_{d_J\to\infty} \|\Sigma-\hat{\Sigma}\|\right]. \end{aligned}$$

Next we want to interchange limit and expectation. Note that

$$\begin{aligned} \|\Sigma-\hat{\Sigma}\| &\le \|\Sigma\| + \|\hat{\Sigma}\| \\ &= \|\Sigma\| + \left\|\frac{X_S^{\mathsf{T}}X_S + X_J X_J^{\mathsf{T}}}{n}\right\| \\ &\le \|\Sigma\| + \left\|\frac{X_S^{\mathsf{T}}X_S}{n}\right\| + \operatorname{Tr}\left(\frac{X_J X_J^{\mathsf{T}}}{n}\right) \\ &= \|\Sigma\| + \left\|\frac{X_S^{\mathsf{T}}X_S}{n}\right\| + \frac{\lambda_n}{n}\operatorname{Tr}\left(\frac{Z_J Z_J^{\mathsf{T}}}{d_J}\right). \end{aligned}$$

The first two terms do not depend on $d_J$. It is easy to verify that

$$\lim_{d_J\to\infty}\mathbb{E}\left[\frac{\lambda_n}{n}\operatorname{Tr}\left(\frac{Z_J Z_J^\mathsf{T}}{d_J}\right)\right]=\lambda_n=\mathbb{E}\left[\lim_{d_J\to\infty}\frac{\lambda_n}{n}\operatorname{Tr}\left(\frac{Z_J Z_J^\mathsf{T}}{d_J}\right)\right]$$

as $\frac{Z_J Z_J^\mathsf{T}}{d_J}\overset{a.s.}{\to} I_n$. Therefore, by the dominated convergence theorem

$$\lim_{d_J\to\infty}\mathbb{E}\left[\|\Sigma-\hat{\Sigma}\|\cdot\|\hat{w}_{MN}\|^2\right]\geq\lim_{d_J\to\infty}\left(\sigma^2\frac{n-d_S}{\lambda_n}\right)\mathbb{E}\|\Sigma-\hat{\Sigma}\|.$$

Koltchinskii and Lounici [17] show that, for Gaussian data,

$$\mathbb{E}\|\Sigma-\hat{\Sigma}\|\geq C\max\left(\sqrt{\frac{\operatorname{Tr}(\Sigma)\,\|\Sigma\|}{n}},\frac{\operatorname{Tr}(\Sigma)}{n}\right),$$

where $C$ is a universal constant. Thus, in our case

$$\mathbb{E}\|\Sigma-\hat{\Sigma}\|\geq C\sqrt{\frac{d_S+\lambda_n}{n}}.$$

Since $\lambda_n=o(n)$, this implies

$$\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\left[\|\Sigma-\hat{\Sigma}\|\cdot\|\hat{w}_{MN}\|^2\right]\geq\lim_{n\to\infty}\left(\sigma^2\frac{n-d_S}{\lambda_n}\right)C\sqrt{\frac{d_S+\lambda_n}{n}}=\infty.$$

It is easy to see that the remaining terms in the lower bound of Proposition B.2 are negligible. □

### B.3 Uniform convergence on tighter sets (Section 3.2)

**Theorem 3.3.** *In Setting **B**, let $\mathcal{A}$ be an algorithm outputting interpolators, $X\mathcal{A}(X,Y)=Y$, with*

$$\mathcal{A}\left((X_S,X_J),y\right)_S=\mathcal{A}\left((X_S,-X_J),y\right)_S\quad\text{and}\quad\lim_{n\to\infty}\lim_{d_J\to\infty}L_{\mathcal{D}}(\mathcal{A}(X,y))\overset{a.s.}{=}\sigma^2. \tag{6}$$

*For any $\delta\in(0,\frac{1}{2})$ and set of typical training examples $\mathcal{S}_{n,\delta}$ satisfying $\Pr(\mathbf{S}\in\mathcal{S}_{n,\delta})\geq 1-\delta$, let $\mathcal{W}_{n,\delta}^{\mathcal{A}}=\{\mathcal{A}(X,Y):(X,Y)\in\mathcal{S}_{n,\delta}\}$ denote the set of typical outputs. Then*

$$\lim_{n\to\infty}\lim_{d_J\to\infty}\sup_{\mathbf{S}\in\mathcal{S}_{n,\delta}}\sup_{w\in\mathcal{W}_{n,\delta}}|L_{\mathcal{D}}(w)-L_{\mathbf{S}}(w)|\overset{a.s.}{\geq}3\sigma^2. \tag{7}$$

*Proof.* Fix any $\mathcal{S}_{n,\delta}$ satisfying $\Pr(\mathbf{S}\in\mathcal{S}_{n,\delta})\geq 1-\delta$. For each $\mathbf{S}=((X_S,X_J),Y)$, we define $\tilde{\mathbf{S}}=((X_S,-X_J),Y)$. Note that the marginal distribution of $\tilde{\mathbf{S}}$ is the same as $\mathbf{S}$ because of the isotropic Gaussian distribution. Thus we also have $\Pr(\tilde{\mathbf{S}}\in\mathcal{S}_{n,\delta})\geq 1-\delta$. By a simple union bound

$$\begin{aligned}1-\Pr(\mathbf{S}\in\mathcal{S}_{n,\delta}\cap\tilde{\mathbf{S}}\in\mathcal{S}_{n,\delta})&=\Pr(\mathbf{S}\notin\mathcal{S}_{n,\delta}\cup\tilde{\mathbf{S}}\notin\mathcal{S}_{n,\delta})\\&\leq\Pr(\mathbf{S}\notin\mathcal{S}_{n,\delta})+\Pr(\tilde{\mathbf{S}}\notin\mathcal{S}_{n,\delta})\leq 2\delta.\end{aligned}$$

As $\delta<\frac{1}{2}$, we have $\Pr(\mathbf{S}\in\mathcal{S}_{n,\delta}\cap\tilde{\mathbf{S}}\in\mathcal{S}_{n,\delta})>0$, so the set $\{\mathbf{S}\in\mathcal{S}_{n,\delta}:\tilde{\mathbf{S}}\in\mathcal{S}_{n,\delta}\}$ must be nonempty. Pick any $\mathbf{S}=((X_S,X_J),Y)$ in this set; thus $\hat{w}=\mathcal{A}\left((X_S,X_J),Y\right)\in\mathcal{W}_{n,\delta}$ and $\tilde{w}=\mathcal{A}\left((X_S,-X_J),Y\right)\in\mathcal{W}_{n,\delta}$. As $\mathcal{A}$ outputs interpolators, we have that

$$X_S\hat{w}_S+X_J\hat{w}_J=Y=X_S\tilde{w}_S-X_J\tilde{w}_J,$$

and (6) implies that $\hat{w}_S=\tilde{w}_S$, so then $X_J\hat{w}_J=-X_J\tilde{w}_J$. Thus

$$\begin{aligned}L_{\mathbf{S}}(\tilde{w})=\frac{1}{n}\|X\tilde{w}-Y\|^2&=\frac{1}{n}\|X_S\hat{w}_S-X_J\hat{w}_J-(X_S\hat{w}_S+X_J\hat{w}_J)\|^2=\frac{1}{n}\|-2X_J\hat{w}_J\|^2\\&\geq\frac{4}{n}\|(I_n-\Pi)X_J\hat{w}_J\|^2,\end{aligned}$$

where $\Pi \in \mathbb{R}^{n\times n}$ is the orthogonal projection onto the range of $X_S$. Now,

$$\begin{aligned}(I_n - \Pi)X_J\hat{w}_J &= (I_n - \Pi)(X_S\hat{w}_S + X_J\hat{w}_J)\\ &= (I_n - \Pi)Y\\ &= (I_n - \Pi)(X_S w_S^* + E)\\ &= (I_n - \Pi)E\\ &\sim \mathcal{N}(0, \sigma^2(I_n - \Pi))\end{aligned}$$

using $E \sim \mathcal{N}(0, \sigma^2 I_n)$. As $n \to \infty$, because $X_S$ is almost surely rank $d_S$, $\mathrm{Tr}(I_n - \Pi)$ is almost surely $n - d_S$. Thus we have

$$\frac{1}{n - d_S}\|(I_n - \Pi)X_J\hat{w}_J\|^2 \overset{a.s.}{\to} \sigma^2,$$

and so

$$L_{\mathbf{S}}(\tilde{w}) \overset{a.s.}{\geq} 4\sigma^2\frac{n - d_S}{n} \to 4\sigma^2.$$

The conclusion follows by the observation that

$$\sup_{\mathbf{S}\in\mathcal{S}_{n,\delta}}\sup_{w\in\mathcal{W}_\delta}|L_{\mathcal{D}}(w) - L_{\mathbf{S}}(w)| \geq L_{\mathbf{S}}(\tilde{w}) - L_{\mathcal{D}}(\tilde{w}). \qquad \square$$

**Proposition B.3.** *Let $f_S : \mathbb{R}^{d_S} \to \mathbb{R}$ and $f_J : \mathbb{R}^{d_J} \to \mathbb{R}$ be convex functions, with $f_J$ symmetric, $f_J(-w) = f_J(w)$. Let $\mathcal{A}$ be an interpolation algorithm satisfying*

$$\mathcal{A}(X, y) = \operatorname*{arg\,min}_{w\,\mathrm{s.t.}\,Xw=y} f_S(w_S) + f_J(w_J).$$

*Then negating junk dimensions simply negates the corresponding dimensions of the predictor:*

$$\mathcal{A}\left((X_S, -X_J), Y\right) = \begin{bmatrix} I_{d_S} & 0_{d_S\times d_J} \\ 0_{d_J\times d_S} & -I_{d_J}\end{bmatrix}\mathcal{A}\left((X_S, X_J), Y\right).$$

*(If the minimizer is not unique, the equation holds as an operation on sets.)*

*Proof.* The KKT conditions for $\mathcal{A}(X, y)$, which are both necessary and sufficient in this case, are

$$Xw = X_S w_S + X_J w_J = Y, \qquad 0 \in \partial f_S(w_S) + \nu_S^\mathsf{T} X_S, \qquad 0 \in \partial f_J(w_J) + \nu_J^\mathsf{T} X_J, \tag{14}$$

where $\delta$ denotes the subdifferential, and the dual variables $\nu_S \in \mathbb{R}^{d_S}$ and $\nu_J \in \mathbb{R}^{d_J}$ are otherwise unconstrained. Also note that because $f_J$ is symmetric, if $g \in \partial f_J$ then for any $t$, there is some $g' \in \partial f_J$ such that $g'(-t) = -g(t)$.

Let $(\hat{w}, \nu_S, \nu_J)$ be some solution to (14), and define $\tilde{w} = (\hat{w}_S, -\hat{w}_J)$, $\tilde{X} = (X_S, -X_J)$. Then

$$\begin{gathered}(X_S, -X_J)\,\tilde{w} = X_S\tilde{w}_S - X_J\tilde{w}_J = X_S\hat{w}_S + X_J\hat{w}_J = Y,\\ \partial f_S(\tilde{w}_S) + \nu_S^\mathsf{T}\tilde{X}_S = \partial f_S(\hat{w}_S) + \nu_S^\mathsf{T}X_S \ni 0,\\ \text{and}\quad \partial f_J(\tilde{w}_J) + \nu_J^\mathsf{T}\tilde{X}_J = \partial f_J(-\hat{w}_J) + \nu_J^\mathsf{T}(-X_J) \ni 0 \quad \text{because} \quad 0 \in \partial f_J(\hat{w}_J) + \nu_J^\mathsf{T}X_J.\end{gathered}$$

Thus $(\tilde{w}, \nu_S, \nu_J)$ satisfies the KKT conditions for $\mathcal{A}(\tilde{X}, Y)$. When the minimizer is not unique, the same argument works in reverse, showing that solution sets are related in the same way. $\square$

# C Proofs for Section 4

## C.1 Consistency of the minimal risk interpolator (Proposition 4.3)

**Proposition 4.3.** *In Setting **A**, the expected risk of the minimal-risk interpolator is*

$$\mathbb{E}\, L_{\mathcal{D}}(\hat{w}_{MR}) = \frac{p - 1}{p - 1 - n}L_{\mathcal{D}}(w^*).$$

*Proof.* Recall that

$$\hat{w}_{MR} = w^* + \Sigma^{-1}X^\mathsf{T}(X\Sigma^{-1}X^\mathsf{T})^{-1}E.$$

From this, we can compute

$$\begin{aligned}
L_{\mathcal{D}}(\hat{w}_{MR}) - L_{\mathcal{D}}(w^*) &= (\hat{w}_{MR} - w^*)^\mathsf{T}\Sigma(\hat{w}_{MR} - w^*) \\
&= (\hat{w}_{MR} - w^*)^\mathsf{T}X^\mathsf{T}(X\Sigma^{-1}X^\mathsf{T})^{-1}E \\
&= (X\hat{w}_{MR} - Xw^*)^\mathsf{T}(X\Sigma^{-1}X^\mathsf{T})^{-1}E \\
&= (Y - Xw^*)^\mathsf{T}(X\Sigma^{-1}X^\mathsf{T})^{-1}E \\
&= E^\mathsf{T}(ZZ^\mathsf{T})^{-1}E \\
&= \langle (ZZ^\mathsf{T})^{-1}, EE^\mathsf{T}\rangle.
\end{aligned}$$

By independence of $Z$ and $E$, we get

$$\mathbb{E}[L_{\mathcal{D}}(\hat{w}_{MR}) - L_{\mathcal{D}}(w^*)] = \sigma^2\,\mathbb{E}\,\mathrm{Tr}\left[\left(ZZ^\mathsf{T}\right)^{-1}\right].$$

Note that $\left(ZZ^\mathsf{T}\right)^{-1}$ follows an inverse-Wishart distribution whose expectation is $\frac{I_n}{p-n-1}$. Therefore, we obtain

$$\begin{aligned}
\mathbb{E}[L_{\mathcal{D}}(\hat{w}_{MR})] &= \sigma^2 + \sigma^2\,\mathrm{Tr}\left(\frac{I_n}{p-n-1}\right) \\
&= \sigma^2\left(1 + \frac{n}{p-n-1}\right) = \left(\frac{p-1}{p-n-1}\right)\cdot L_{\mathcal{D}}(w^*).
\end{aligned}$$

□

### C.2 Uniform consistency of low norm interpolators (Section 4.1)

#### C.2.1 General results

Our key lemma is as follows:

**Lemma C.1.** *Let $\hat{w}$ be any predictor that interpolates the data, with $\|\hat{w}\| \le B$, and $F \in \mathbb{R}^{p\times(p-n)}$ be the matrix whose columns form an orthonormal basis of the kernel of $X$. In other words, if $X\hat{w} = Y$, $XF = 0_{n\times(p-n)}$ and $F^\mathsf{T}F = I_{p-n}$, then* (8)*, the worst-case generalization gap for interpolators up to norm $B$, is equal to*

$$L_{\mathcal{D}}(\hat{w}) + \inf_{\lambda > \|F^\mathsf{T}\Sigma F\|} \|F^\mathsf{T}[\lambda\hat{w} - \Sigma(\hat{w} - w^*)]\|_{(\lambda I_{p-n} - F^\mathsf{T}\Sigma F)^{-1}} + \lambda(B^2 - \|\hat{w}\|^2).$$

*Proof.* Observe that $\{w \in \mathbb{R}^p : L_S(w) = 0\} = \{\hat{w} + Fu : u \in \mathbb{R}^{p-n}\}$. Then

$$\begin{aligned}
&\sup_{\substack{\|w\|\le B \\ L_S(w)=0}} L_{\mathcal{D}}(w) - L_S(w) \\
&= L_{\mathcal{D}}(w^*) + \sup_{\substack{\|w\|\le B \\ L_S(w)=0}} L_{\mathcal{D}}(w) - L_{\mathcal{D}}(w^*) \\
&= L_{\mathcal{D}}(w^*) + \sup_{\|\hat{w}+Fu\|^2 \le B^2} (\hat{w} + Fu - w^*)^\mathsf{T}\Sigma(\hat{w} + Fu - w^*) \\
&= L_{\mathcal{D}}(w^*) + \sup_{\|u\|^2 + 2\langle u, F^\mathsf{T}\hat{w}\rangle + \|\hat{w}\|^2 \le B^2} u^\mathsf{T}(F^\mathsf{T}\Sigma F)u + 2\langle u, F^\mathsf{T}\Sigma(\hat{w} - w^*)\rangle + (\hat{w} - w^*)^\mathsf{T}\Sigma(\hat{w} - w^*) \\
&= L_{\mathcal{D}}(\hat{w}) + \sup_{\|u\|^2 + 2\langle u, F^\mathsf{T}\hat{w}\rangle + \|\hat{w}\|^2 \le B^2} u^\mathsf{T}(F^\mathsf{T}\Sigma F)u + 2\langle u, F^\mathsf{T}\Sigma(\hat{w} - w^*)\rangle \\
&= L_{\mathcal{D}}(\hat{w}) - \inf_{\|u\|^2 + 2\langle u, F^\mathsf{T}\hat{w}\rangle + \|\hat{w}\|^2 \le B^2} u^\mathsf{T}(-F^\mathsf{T}\Sigma F)u - 2\langle u, F^\mathsf{T}\Sigma(\hat{w} - w^*)\rangle.
\end{aligned}$$

Although the second term involves a concave minimization problem, it is a quadratic optimization problem with a single quadratic inequality constraint. This is a classical example where strong

duality holds even though the objective is not convex [9, Appendix B]. In order to derive the dual, we write down the Lagrangian:

$$\begin{aligned} L(u,\lambda) &= u^\mathsf{T}(-F^\mathsf{T}\Sigma F)u - 2\langle u, F^\mathsf{T}\Sigma(\hat{w}-w^*)\rangle + \lambda(\|u\|^2 + 2\langle u, F^\mathsf{T}\hat{w}\rangle + \|\hat{w}\|^2 - B^2) \\ &= u^\mathsf{T}(\lambda I_{p-n} - F^\mathsf{T}\Sigma F)u + 2\langle u, F^\mathsf{T}(\lambda\hat{w} - \Sigma(\hat{w}-w^*))\rangle - \lambda(B^2 - \|\hat{w}\|^2); \end{aligned}$$

strong duality tells us that the infimum is equal to $\sup_{\lambda\geq 0}\inf_u L(u,\lambda)$. For $\lambda < \|F^\mathsf{T}\Sigma F\|$, $\lambda I_{p-n} - F^\mathsf{T}\Sigma F$ has strictly negative eigenvalues, and so then $\inf_u L(u,\lambda) = -\infty$. If instead $\lambda > \|F^\mathsf{T}\Sigma F\|$, $\lambda I_{p-n} - F^\mathsf{T}\Sigma F$ is strictly positive definite, and setting the $u$ derivative to zero yields that $\inf_u L(\lambda, u)$ is

$$-\left[F^\mathsf{T}(\lambda\hat{w} - \Sigma(\hat{w}-w^*))\right]^\mathsf{T}(\lambda I_{p-n} - F^\mathsf{T}\Sigma F)^{-1}\left[F^\mathsf{T}(\lambda\hat{w} - \Sigma(\hat{w}-w^*))\right] - \lambda(B^2 - \|\hat{w}\|^2). \quad (15)$$

If instead $\lambda = \|F^\mathsf{T}\Sigma F\|$, we again have $\inf_u L(u,\lambda) = -\infty$ unless $F^\mathsf{T}(\lambda\hat{w} - F^\mathsf{T}\Sigma(\hat{w}-w^*)) = 0$ so that the linear term is identically zero; in this case, the quadratic term is minimized by $u = 0$, and $\inf_u L(u,\lambda) = \lambda(B^2 - \|\hat{w}\|^2)$ agrees with (15), so this case is covered by the strict case as well. Thus the dual problem is to maximize (15) over $\lambda > \|F^\mathsf{T}\Sigma F\|$. The desired result follows by passing the minus sign into the sup of the dual problem. □

We will now prove Theorem 4.5.

**Theorem 4.5.** *The following results hold deterministically, viewing $L_\mathcal{D}(w)$ simply as a quadratic function $L_\mathcal{D}(w^*) + \|w - w^*\|_\Sigma$, with no distributional assumptions on* $\mathbf{S}$.

*(i) It holds that*

$$\sup_{\substack{\|w\|\leq\|\hat{w}_{MR}\| \\ L_\mathbf{S}(w)=0}} L_\mathcal{D}(w) - L_\mathbf{S}(w) = L_\mathcal{D}(\hat{w}_{MR}) + \gamma_n\,\kappa_X(\Sigma)\left[\|\hat{w}_{MR}\|^2 - \|\hat{w}_{MN}\|^2\right]$$

*where $1 \leq \gamma_n \leq 4$.*

*If the minimal risk interpolator is consistent, $\mathbb{E}\,L_\mathcal{D}(\hat{w}_{MR}) - L_\mathcal{D}(w^*) \to 0$, then the class of interpolators with norm less than $\|\hat{w}_{MR}\|$ is uniformly consistent if and only if*

$$\mathbb{E}\,\kappa_X(\Sigma)\cdot\left[\|\hat{w}_{MR}\|^2 - \|\hat{w}_{MN}\|^2\right] \to 0.$$

*(ii) Fix a sequence $(B_n)$ such that $B_n \geq \|\hat{w}_{MN}\|$ for all $n$. Then*

$$\sup_{\|w\|\leq B_n,\, L_\mathbf{S}(w)=0} L_\mathcal{D}(w) - L_\mathbf{S}(w) = L_\mathcal{D}(\hat{w}_{MN}) + \kappa_X(\Sigma)\left[B_n^2 - \|\hat{w}_{MN}\|^2\right] + R_n$$

*where $0 \leq R_n \leq 2\sqrt{[L_\mathcal{D}(\hat{w}_{MN}) - L_\mathcal{D}(w^*)]\,\kappa_X(\Sigma)\,[B_n^2 - \|\hat{w}_{MN}\|^2]}$.*

*If $\mathbb{E}\,L_\mathcal{D}(\hat{w}_{MN}) - L_\mathcal{D}(w^*) \to 0$, the class of interpolators with norm less than $B_n$ is thus uniformly consistent if and only if*

$$\mathbb{E}\,\kappa_X(\Sigma)\cdot\left[B_n^2 - \|\hat{w}_{MN}\|^2\right] \to 0.$$

*Proof.* For case (i), observe that

$$F^\mathsf{T}\Sigma(\hat{w}_{MR} - w^*) = F^\mathsf{T}X^\mathsf{T}(X\Sigma^{-1}X^\mathsf{T})^{-1}E = (XF)^\mathsf{T}(X\Sigma^{-1}X^\mathsf{T})^{-1}E = 0.$$

Thus picking $\hat{w} = \hat{w}_{MR}$ and $B = \|\hat{w}_{MR}\|$ in Lemma C.1 gives that

$$\sup_{\|w\|\leq\|\hat{w}_{MR}\|,\, L_\mathbf{S}(w)=0} L_\mathcal{D}(w) = L_\mathcal{D}(\hat{w}_{MR}) + \inf_{\lambda>\|F^\mathsf{T}\Sigma F\|}\|\lambda F^\mathsf{T}\hat{w}_{MR}\|_{(\lambda I_{p-n} - F^\mathsf{T}\Sigma F)^{-1}}. \quad (16)$$

Since we have

$$\frac{1}{\lambda}I_{p-n} \preceq (\lambda I_{p-n} - F^\mathsf{T}\Sigma F)^{-1},$$

we know that $\sup_{\|w\|\leq\|\hat{w}_{MR}\|,\, L_\mathbf{S}(w)=0} L_\mathcal{D}(w)$ is lower bounded by

$$L_\mathcal{D}(\hat{w}_{MR}) + \inf_{\lambda>\|F^\mathsf{T}\Sigma F\|}\frac{1}{\lambda}\|\lambda F^\mathsf{T}\hat{w}_{MR}\|^2 = L_\mathcal{D}(\hat{w}_{MR}) + \|F^\mathsf{T}\Sigma F\|\cdot\|F^\mathsf{T}\hat{w}_{MR}\|^2.$$

In order to compute $\|F^\mathsf{T}\hat{w}_{MR}\|^2$, we notice that $FF^\mathsf{T}$ is the orthogonal projection onto the kernel of $X$. Using the fact that $\text{im}(X^\mathsf{T}) = \ker(X)^\perp$, we get $I - FF^\mathsf{T}$ is the orthogonal projection onto the image of $X^\mathsf{T}$. Thus,

$$X(I - FF^\mathsf{T})\hat{w}_{MR} = X\hat{w}_{MR} = Y,$$

and left-multiplying both sides by $X^\mathsf{T}(XX^\mathsf{T})^{-1}$ gives that

$$\hat{w}_{MN} = X^\mathsf{T}(XX^\mathsf{T})^{-1}X(I - FF^\mathsf{T})\hat{w}_{MR} = (I - FF^\mathsf{T})\hat{w}_{MR},$$

and so

$$\begin{aligned}
\|F^\mathsf{T}\hat{w}_{MR}\|^2 &= \hat{w}_{MR}^\mathsf{T}FF^\mathsf{T}\hat{w}_{MR} \\
&= \hat{w}_{MR}^\mathsf{T}F(F^\mathsf{T}F)F^\mathsf{T}\hat{w}_{MR} \\
&= \|FF^\mathsf{T}\hat{w}_{MR}\|^2 \\
&= \|\hat{w}_{MR}\|^2 - \|(I - FF^\mathsf{T})\hat{w}_{MR}\|^2 \\
&= \|\hat{w}_{MR}\|^2 - \|\hat{w}_{MN}\|^2
\end{aligned}$$

which establishes the lower bound with a constant of 1.

Similarly, we can use $(\lambda I_{p-n} - F^\mathsf{T}\Sigma F)^{-1} \preceq \frac{1}{\lambda - \|F^\mathsf{T}\Sigma F\|} I_{p-n}$ to upper bound (16) as

$$\begin{aligned}
&L_\mathcal{D}(\hat{w}_{MR}) + \inf_{\lambda > \|F^\mathsf{T}\Sigma F\|} \frac{1}{\lambda - \|F^\mathsf{T}\Sigma F\|}\|\lambda F^\mathsf{T}\hat{w}_{MR}\|^2 \\
=& L_\mathcal{D}(\hat{w}_{MR}) + \inf_{\lambda > 0} \frac{(\lambda + \|F^\mathsf{T}\Sigma F\|)^2}{\lambda}(\|\hat{w}_{MR}\|^2 - \|\hat{w}_{MN}\|^2) \\
=& L_\mathcal{D}(\hat{w}_{MR}) + \inf_{\lambda > 0} \left(\lambda + 2\|F^\mathsf{T}\Sigma F\| + \frac{\|F^\mathsf{T}\Sigma F\|^2}{\lambda}\right)(\|\hat{w}_{MR}\|^2 - \|\hat{w}_{MN}\|^2) \\
=& L_\mathcal{D}(\hat{w}_{MR}) + 4\|F^\mathsf{T}\Sigma F\| \cdot (\|\hat{w}_{MR}\|^2 - \|\hat{w}_{MN}\|^2).
\end{aligned}$$

This gives the desired upper bound with a constant of 4. It follows immediately that (16) converges to $L_\mathcal{D}(w^*)$ if and only if

$$\mathbb{E}\|F^\mathsf{T}\Sigma F\| \cdot (\|\hat{w}_{MR}\|^2 - \|\hat{w}_{MN}\|^2) \to 0.$$

Turning to part (ii), observe that

$$F^\mathsf{T}\hat{w}_{MN} = F^\mathsf{T}X^\mathsf{T}(XX^\mathsf{T})^{-1}Y = (XF)^\mathsf{T}(XX^\mathsf{T})^{-1}Y = 0,$$

so that Lemma C.1 with $\hat{w} = \hat{w}_{MN}$ gives

$$\sup_{\|w\| \le B_n\, L_\mathbf{S}(w) = 0} L_\mathcal{D}(w) = L_\mathcal{D}(\hat{w}_{MN}) + \inf_{\lambda > \|F^\mathsf{T}\Sigma F\|} \|F^\mathsf{T}\Sigma(\hat{w} - w^*)\|_{(\lambda I_{p-n} - F^\mathsf{T}\Sigma F)^{-1}} + \lambda(B_n^2 - \|\hat{w}_{MN}\|^2).$$

Moreover, it is clear that

$$0_{p-n} \prec (\lambda I_{p-n} - F^\mathsf{T}\Sigma F)^{-1} \prec \frac{1}{\lambda - \|F^\mathsf{T}\Sigma F\|} I_{p-n}.$$

Therefore, $\sup_{\|w\| \le B_n,\, L_\mathbf{S}(w) = 0} L_\mathcal{D}(w)$ is lower bounded by, recalling that $\|F^\mathsf{T}\Sigma F\| = \kappa_X(\Sigma)$,

$$L_\mathcal{D}(\hat{w}_{MN}) + \inf_{\lambda > \|F^\mathsf{T}\Sigma F\|} \lambda(B_n^2 - \|\hat{w}_{MN}\|) = L_\mathcal{D}(\hat{w}_{MN}) + \kappa_X(\Sigma) \cdot \left[B_n^2 - \|\hat{w}_{MN}\|^2\right], \quad (17)$$

and we have shown that $R_n \ge 0$ in the result. On the other hand, $\sup_{\|w\| \le B_n,\, L_\mathbf{S}(w) = 0} L_\mathcal{D}(w)$ is upper bounded by

$$\begin{aligned}
&L_\mathcal{D}(\hat{w}_{MN}) + \inf_{\lambda > \|F^\mathsf{T}\Sigma F\|} \frac{1}{\lambda - \|F^\mathsf{T}\Sigma F\|}\|F^\mathsf{T}\Sigma(\hat{w}_{MN} - w^*)\|^2 + \lambda\Big[B_n^2 - \|\hat{w}_{MN}\|^2\Big] \\
=& L_\mathcal{D}(\hat{w}_{MN}) + \inf_{\lambda > 0} \frac{1}{\lambda}\|F^\mathsf{T}\Sigma(\hat{w}_{MN} - w^*)\|^2 + (\lambda + \kappa_X(\Sigma))\Big[B_n^2 - \|\hat{w}_{MN}\|^2\Big] \\
=& L_\mathcal{D}(\hat{w}_{MN}) + \kappa_X(\Sigma) \cdot \Big[B_n^2 - \|\hat{w}_{MN}\|^2\Big] + \inf_{\lambda > 0} \frac{1}{\lambda}\|F^\mathsf{T}\Sigma(\hat{w}_{MN} - w^*)\|^2 + \lambda\Big[B_n^2 - \|\hat{w}_{MN}\|^2\Big] \\
=& L_\mathcal{D}(\hat{w}_{MN}) + \kappa_X(\Sigma) \cdot \Big[B_n^2 - \|\hat{w}_{MN}\|^2\Big] + 2\sqrt{\|F^\mathsf{T}\Sigma(\hat{w}_{MN} - w^*)\|^2 \cdot \Big[B_n^2 - \|\hat{w}_{MN}\|^2\Big]}. \quad (18)
\end{aligned}$$

We can upper bound

$$\begin{aligned}\|F^\mathsf{T}\Sigma(\hat{w}_{MN}-w^*)\|^2 &= (\hat{w}_{MN}-w^*)^\mathsf{T}\Sigma F F^\mathsf{T}\Sigma(\hat{w}_{MN}-w^*)\\ &= [\Sigma^{1/2}(\hat{w}_{MN}-w^*)]^\mathsf{T}(\Sigma^{1/2}FF^\mathsf{T}\Sigma^{1/2})[\Sigma^{1/2}(\hat{w}_{MN}-w^*)]\\ &\le \|\Sigma^{1/2}FF^\mathsf{T}\Sigma^{1/2}\|\cdot\|\Sigma^{1/2}(\hat{w}_{MN}-w^*)\|^2\\ &= \|F^\mathsf{T}\Sigma F\|\cdot[L_\mathcal{D}(\hat{w}_{MN})-L_\mathcal{D}(w^*)]\end{aligned},$$

using the fact that $\|AA^T\| = \|A^TA\|$ with $A = F^T\Sigma^{1/2}$. Plugging into the third term of (18) yields our desired upper bound on $R_n$,

To show the statement about expectations when $\mathbb{E}\, L_\mathcal{D}(\hat{w}_{MN}) - L_\mathcal{D}(w^*) \to 0$, note for one direction that (17) gives

$$\liminf_{n\to\infty}\mathbb{E}\left[\sup_{\substack{\|w\|\le B_n\\ L_\mathbf{S}(w)=0}} L_\mathcal{D}(w)-L_\mathbf{S}(w)\right] \ge L_\mathcal{D}(w^*)+\lim_{n\to\infty}\mathbb{E}\,\kappa_X(\Sigma)\cdot\left[B_n^2-\|\hat{w}_{MN}\|^2\right].$$

For the other direction, we have

$$\begin{aligned}R_n &\le 2\sqrt{\|F^\mathsf{T}\Sigma F\|\cdot[L_\mathcal{D}(\hat{w}_{MN})-L_\mathcal{D}(w^*)]\left[B_n^2-\|\hat{w}_{MN}\|^2\right]}\\ &\le \epsilon\|F^\mathsf{T}\Sigma F\|\cdot\left[B_n^2-\|\hat{w}_{MN}\|^2\right]+\frac{1}{\epsilon}[L_\mathcal{D}(\hat{w}_{MN})-L_\mathcal{D}(w^*)]\end{aligned}$$

for any $\epsilon > 0$. This implies

$$\limsup_{n\to\infty}\mathbb{E}\left[\sup_{\substack{\|w\|\le B_n\\ L_\mathbf{S}(w)=0}} L_\mathcal{D}(w)-L_\mathbf{S}(w)\right] \le L_\mathcal{D}(w^*)+(1+\epsilon)\,\mathbb{E}\left(\lim_{n\to\infty}\kappa_X(\Sigma)\cdot\left[B_n^2-\|\hat{w}_{MN}\|^2\right]\right),$$

showing the desired result. □

### C.2.2 Special case of Setting B

In Setting **B**, we are able to compute $\kappa_X(\Sigma)$.

**Proposition C.2.** *With probability 1, it holds in Setting **B** that*

$$\lim_{d_J\to\infty}\kappa_X(\Sigma) = \frac{\lambda_n}{n}\left\|\left[\frac{X_S^\mathsf{T}X_S}{n}+\frac{\lambda_n}{n}I_{d_S}\right]^{-1}\right\|.$$

*Proof.* Recall that

$$\kappa_X(\Sigma) = \|F^\mathsf{T}\Sigma F\| = \|\Sigma^{1/2}FF^\mathsf{T}\Sigma^{1/2}\| = \|\Sigma^{1/2}(I-X^\mathsf{T}(XX^\mathsf{T})^{-1}X)\Sigma^{1/2}\|.$$

It is a routine calculation to show that

$$\Sigma^{1/2}FF^\mathsf{T}\Sigma^{1/2} = \begin{bmatrix} I_{d_S}-X_S^\mathsf{T}(X_SX_S^\mathsf{T}+X_JX_J^\mathsf{T})^{-1}X_S & -\sqrt{\frac{\lambda_n}{d_J}}X_S^\mathsf{T}(X_SX_S^\mathsf{T}+X_JX_J^\mathsf{T})^{-1}X_J\\ -\sqrt{\frac{\lambda_n}{d_J}}X_J^\mathsf{T}(X_SX_S^\mathsf{T}+X_JX_J^\mathsf{T})^{-1}X_S & \frac{\lambda_n}{d_J}\left[I_{d_J}-X_J^\mathsf{T}(X_SX_S^\mathsf{T}+X_JX_J^\mathsf{T})^{-1}X_J\right]\end{bmatrix}.$$

Intuitively, since only the upper-left block does not vanish as $d_J \to \infty$, we should expect

$$\lim_{d_J\to\infty}\kappa_X(\Sigma) = \|I_{d_S}-X_S^\mathsf{T}(X_SX_S^\mathsf{T}+\lambda_nI_n)^{-1}X_S\|.$$

However, as the dimensions of $\Sigma^{1/2}FF^\mathsf{T}\Sigma^{1/2}$ also increase with $d_J$, the analysis of $\kappa_X(\Sigma)$ requires more care.

It is clear that $\kappa_X(\Sigma) \ge \|I_{d_S}-X_S^\mathsf{T}(X_SX_S^\mathsf{T}+X_JX_J^\mathsf{T})^{-1}X_S\|$, and so

$$\liminf_{d_J\to\infty}\kappa_X(\Sigma) \ge \|I_{d_S}-X_S^\mathsf{T}(X_SX_S^\mathsf{T}+\lambda_nI_n)^{-1}X_S\|.$$

To upper bound the limit, fix any $v = (v_1, v_2)$ such that $v_1 \in \mathbb{R}^{d_S}$, $v_2 \in \mathbb{R}^{d_J}$ and $\|v\| = 1$. We can write

$$\begin{aligned} v^\mathsf{T}\Sigma^{1/2}FF^\mathsf{T}\Sigma^{1/2}v = v_1^\mathsf{T}(I_{d_S} - X_S^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1}X_S)v_1 \\ + \frac{\lambda_n}{d_J}v_2^\mathsf{T}\left[I_{d_J} - X_J^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1}X_J\right]v_2 \\ - 2\sqrt{\frac{\lambda_n}{d_J}}v_1^\mathsf{T}X_S^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1}X_Jv_2. \end{aligned} \tag{19}$$

The first term is upper bounded by

$$\|I_{d_S} - X_S^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1}X_S\| \cdot \|v_1\| \le \|I_{d_S} - X_S^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1}X_S\|,$$

and the second term is upper bounded by $\lambda_n/d_J$, because

$$v_2^\mathsf{T}v_2 \le 1 \qquad \text{and} \qquad v_2^\mathsf{T}X_J^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1}X_Jv_2 \ge 0.$$

For any $\epsilon > 0$, we have

$$\begin{aligned} &- 2\sqrt{\frac{\lambda_n}{d_J}}v_1^\mathsf{T}X_S^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1}X_Jv_2 \\ &\le 2\|v_1^\mathsf{T}X_S^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1/2}\| \cdot \left\|\sqrt{\frac{\lambda_n}{d_J}}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1/2}X_Jv_2\right\| \\ &\le \epsilon\|v_1^\mathsf{T}X_S^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1/2}\|^2 + \frac{1}{\epsilon}\left\|\sqrt{\frac{\lambda_n}{d_J}}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1/2}X_Jv_2\right\|^2 \\ &\le \epsilon\|X_S^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1}X_S\| + \frac{\lambda_n}{\epsilon\, d_J}\|X_J^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1}X_J\| \\ &= \epsilon\|X_S^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1}X_S\| + \frac{\lambda_n}{\epsilon\, d_J}\|(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1/2}X_JX_J^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1/2}\|. \end{aligned}$$

Taking a supremum over $v$ in (19), we get

$$\begin{aligned} \kappa_X(\Sigma) \le \|I_{d_S} - X_S^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1}X_S\| + \epsilon\|X_S^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1}X_S\| \\ + \frac{\lambda_n}{d_J}\left[1 + \frac{1}{\epsilon}\|(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1/2}X_JX_J^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1/2}\|\right]. \end{aligned}$$

Note that

$$\begin{aligned} \lim_{d_J\to\infty}\|(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1/2}X_JX_J^\mathsf{T}(X_SX_S^\mathsf{T} + X_JX_J^\mathsf{T})^{-1/2}\| \\ = \lambda_n\|(X_SX_S^\mathsf{T} + \lambda_nI_n)^{-1}\| < \infty, \end{aligned}$$

so for any $\epsilon > 0$,

$$\limsup_{d_J\to\infty}\kappa_X(\Sigma) \le \|I_{d_S} - X_S^\mathsf{T}(X_SX_S^\mathsf{T} + \lambda_nI_n)^{-1}X_S\| + \epsilon\|X_S^\mathsf{T}(X_SX_S^\mathsf{T} + \lambda_nI_n)^{-1}X_S\|.$$

Sending $\epsilon \to 0$ matches the $\liminf$ and $\limsup$. Finally, because

$$(X_SX_S^\mathsf{T} + \lambda_nI_n)^{-1}X_S = X_S(X_S^\mathsf{T}X_S + \lambda_nI_{d_S})^{-1},$$

we have

$$\begin{aligned} I_{d_S} - X_S^\mathsf{T}(X_SX_S^\mathsf{T} + \lambda_nI_n)^{-1}X_S &= I_{d_S} - X_S^\mathsf{T}X_S(X_S^\mathsf{T}X_S + \lambda_nI_{d_S})^{-1} \\ &= \lambda_n(X_S^\mathsf{T}X_S + \lambda_nI_{d_S})^{-1} \\ &= \frac{\lambda_n}{n}\left[\frac{X_S^\mathsf{T}X_S}{n} + \frac{\lambda_n}{n}I_{d_S}\right]^{-1} \end{aligned}$$

and the proof is concluded. □

**Proposition C.3.** *In Setting **B**, it holds that*

$$\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\,\kappa_X(\Sigma)\cdot\|\hat{w}_{MN}\|^2 = L_{\mathcal{D}}(w^*),$$
$$\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\,\kappa_X(\Sigma)\cdot\left[\|\hat{w}_{MR}\|^2 - \|\hat{w}_{MN}\|^2\right] = 0.$$

*Proof.* Notice that $\kappa_X(\Sigma)\cdot\|\hat{w}_{MN}\|^2$ can be dominated by $\|\Sigma\|\cdot\|\hat{w}_{MR}\|^2$ and Proposition B.1 showed that $\|\hat{w}_{MR}\|^2$ is integrable, so by the dominated convergence theorem,

$$\lim_{d_J\to\infty}\mathbb{E}\;\kappa_X(\Sigma)\cdot\|\hat{w}_{MN}\|^2 = \mathbb{E}\lim_{d_J\to\infty}\kappa_X(\Sigma)\cdot\|\hat{w}_{MN}\|^2.$$

Similarly, $\lim_{d_J\to\infty}\kappa_X(\Sigma)\cdot\|\hat{w}_{MN}\|^2$ can be dominated by

$$\lim_{d_J\to\infty}\kappa_X(\Sigma)\cdot\|\hat{w}_{MR}\|^2 \overset{a.s.}{=} \frac{\lambda_n}{n}\left\|\left[\frac{X_S^\mathsf{T}X_S}{n}+\frac{\lambda_n}{n}I_{d_S}\right]^{-1}\right\|\cdot\left(\|w^*\|^2+\frac{\|E\|^2}{\lambda_n}\right)$$

according to Propositions B.1 and C.2.

As $\left\|\left[\frac{X_S^\mathsf{T}X_S}{n}+\frac{\lambda_n}{n}I_{d_S}\right]^{-1}\right\| \overset{a.s.}{\to} 1$ and $\frac{\|E\|^2}{n}\overset{a.s.}{\to}\sigma^2$, we have

$$\lim_{n\to\infty}\lim_{d_J\to\infty}\kappa_X(\Sigma)\cdot\|\hat{w}_{MR}\|^2 \overset{a.s.}{=} \sigma^2.$$

Moreover, by independence of $X_S$ and $E$

$$\mathbb{E}\lim_{d_J\to\infty}\kappa_X(\Sigma)\cdot\|\hat{w}_{MR}\|^2 = \left(\frac{\lambda_n\|w^*\|^2}{n}+\sigma^2\right)\cdot\mathbb{E}\left\|\left[\frac{X_S^\mathsf{T}X_S}{n}+\frac{\lambda_n}{n}I_{d_S}\right]^{-1}\right\|.$$

Again, $\left\|\left[\frac{X_S^\mathsf{T}X_S}{n}+\frac{\lambda_n}{n}I_{d_S}\right]^{-1}\right\|$ can be dominated by $\mathrm{Tr}\left(\left(\frac{X_S^\mathsf{T}X_S}{n}\right)^{-1}\right)$, so that

$$\lim_{n\to\infty}\mathbb{E}\lim_{d_J\to\infty}\kappa_X(\Sigma)\cdot\|\hat{w}_{MR}\|^2 = \sigma^2 = \mathbb{E}\lim_{n\to\infty}\lim_{d_J\to\infty}\kappa_X(\Sigma)\cdot\|\hat{w}_{MR}\|^2.$$

It is also straightforward to check that

$$\lim_{n\to\infty}\mathbb{E}\,\frac{\lambda_n}{n}\left(\lim_{d_J\to\infty}\|\hat{w}_{MR}\|^2\right) = \sigma^2 = \mathbb{E}\lim_{n\to\infty}\frac{\lambda_n}{n}\cdot\left(\lim_{d_J\to\infty}\|\hat{w}_{MR}\|^2\right).$$

Another application of DCT shows that

$$\begin{aligned}
\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\;\kappa_X(\Sigma)\cdot\|\hat{w}_{MN}\|^2 &= \lim_{n\to\infty}\mathbb{E}\lim_{d_J\to\infty}\kappa_X(\Sigma)\cdot\|\hat{w}_{MN}\|^2\\
&= \mathbb{E}\lim_{n\to\infty}\lim_{d_J\to\infty}\kappa_X(\Sigma)\cdot\|\hat{w}_{MN}\|^2\\
&= \mathbb{E}\lim_{n\to\infty}\frac{\lambda_n}{n}\left\|\left[\frac{X_S^\mathsf{T}X_S}{n}+\frac{\lambda_n}{n}I_{d_S}\right]^{-1}\right\|\cdot\left(\lim_{d_J\to\infty}\|\hat{w}_{MN}\|^2\right)\\
&= \mathbb{E}\lim_{n\to\infty}\frac{\lambda_n}{n}\cdot\left(\lim_{d_J\to\infty}\|\hat{w}_{MN}\|^2\right).
\end{aligned}$$

Using the fact that

$$\frac{\lambda_n}{n}\cdot\left(\lim_{d_J\to\infty}\|\hat{w}_{MN}\|^2\right) \le \frac{\lambda_n}{n}\cdot\left(\lim_{d_J\to\infty}\|\hat{w}_{MR}\|^2\right)$$

and $\|\hat{w}_{MN}\|^2\le\|\hat{w}_{MR}\|^2$, two final applications of DCT give

$$\begin{aligned}
\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\;\kappa_X(\Sigma)\cdot\|\hat{w}_{MN}\|^2 &= \lim_{n\to\infty}\frac{\lambda_n}{n}\left(\mathbb{E}\lim_{d_J\to\infty}\|\hat{w}_{MN}\|^2\right)\\
&= \lim_{n\to\infty}\frac{\lambda_n}{n}\left(\lim_{d_J\to\infty}\mathbb{E}\|\hat{w}_{MN}\|^2\right)\\
&= \lim_{n\to\infty}\frac{\lambda_n}{n}\left[\|w^*\|^2+\sigma^2\frac{n-d_S}{\lambda_n}+\beta_n\left(\frac{\sigma^2 d_S-\lambda_n\|w_S^*\|^2}{n}\right)\right]\\
&= \sigma^2.
\end{aligned}$$

by Proposition B.1. Consequently, we have established

$$\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\left[\kappa_X(\Sigma)\cdot\left(\|\hat{w}_{MR}\|^2-\|\hat{w}_{MN}\|^2\right)\right]=0. \qquad \square$$

We are finally ready to prove Theorem 4.1 and Proposition 4.6.

**Proposition 4.6.** *In Setting **B** with $\lambda_n = o(n)$,*

$$\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\left[\sup_{\|w\|\le\|\hat{w}_{MR}\|,\, L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w)-L_{\mathbf{S}}(w)\right]=L_{\mathcal{D}}(w^*).$$

*Proof.* Recall in the proof of Theorem 4.5, it is shown that

$$\sup_{\|w\|\le\|\hat{w}_{MR}\|,\, L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w)\le L_{\mathcal{D}}(\hat{w}_{MR})+4\,\kappa_X(\Sigma)\cdot\left[\|\hat{w}_{MR}\|^2-\|\hat{w}_{MN}\|^2\right].$$

Proposition 4.3 implies that

$$\lim_{d_J\to\infty}\mathbb{E}\; L_{\mathcal{D}}(\hat{w}_{MR})=L_{\mathcal{D}}(w^*).$$

Combined with Proposition C.3, we have shown

$$\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\left[\sup_{\|w\|\le\|\hat{w}_{MR}\|,\, L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w)-L_{\mathbf{S}}(w)\right]\le L_{\mathcal{D}}(w^*).$$

On the other hand, we have the trivial lower bound

$$\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\left[\sup_{\substack{\|w\|\le\|\hat{w}_{MR}\|\\ L_{\mathbf{S}}(w)=0}} L_{\mathcal{D}}(w)-L_{\mathbf{S}}(w)\right]\ge\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\,L_{\mathcal{D}}(\hat{w}_{MR})=L_{\mathcal{D}}(w^*). \qquad \square$$

**Theorem 4.1.** *In Setting **B** with $\lambda_n = o(n)$, fix a sequence $(\alpha_n)\to\alpha$, with each $\alpha_n\ge 1$. Then*

$$\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\left[\sup_{\|w\|\le\alpha_n\|\hat{w}_{MN}\|,\, L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w)-L_{\mathbf{S}}(w)\right]=\alpha^2 L_{\mathcal{D}}(w^*).$$

*Proof.* In the proof of Theorem 4.5, it is shown for every $\epsilon\ge 0$ that

$$\sup_{\substack{\|w\|\le B_n\\ L_{\mathbf{S}}(w)=0}} L_{\mathcal{D}}(w)-L_{\mathbf{S}}(w)\le L_{\mathcal{D}}(\hat{w}_{MN})+(1+\epsilon)\kappa_X(\Sigma)\cdot\left[B_n^2-\|\hat{w}_{MN}\|^2\right]+\frac{1}{\epsilon}[L_{\mathcal{D}}(\hat{w}_{MN})-L_{\mathcal{D}}(w^*)].$$

Proposition 4.6 implies that $\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\,L_{\mathcal{D}}(\hat{w}_{MN})=L_{\mathcal{D}}(w^*)$. Thus, plugging in $B_n=\alpha_n\|\hat{w}_{MN}\|$ and taking expectations and limits on both sides gives

$$\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\left[\sup_{\substack{\|w\|\le\alpha_n\|\hat{w}_{MN}\|\\ L_{\mathbf{S}}(w)=0}} L_{\mathcal{D}}(w)\right]\le L_{\mathcal{D}}(w^*)+(1+\epsilon)\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}(\alpha_n^2-1)\kappa_X(\Sigma)\|\hat{w}_{MN}\|^2;$$

further applying Proposition C.3 yields

$$\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\left[\sup_{\substack{\|w\|\le\alpha_n\|\hat{w}_{MN}\|\\ L_{\mathbf{S}}(w)=0}} L_{\mathcal{D}}(w)\right]\le L_{\mathcal{D}}(w^*)+(1+\epsilon)(\alpha^2-1)L_{\mathcal{D}}(w^*).$$

Sending $\epsilon\to 0$ yields the upper bound $\alpha^2 L_{\mathcal{D}}(w^*)$.

To get the lower bound, in the proof of Theorem 4.5 it is also shown

$$\sup_{\substack{\|w\|\le B_n\\ L_{\mathbf{S}}(w)=0}} L_{\mathcal{D}}(w)-L_{\mathbf{S}}(w)\ge L_{\mathcal{D}}(\hat{w}_{MN})+\kappa_X(\Sigma)\cdot\left[B_n^2-\|\hat{w}_{MN}\|^2\right].$$

By Proposition C.3, letting $B_n=\alpha_n\|\hat{w}_{MN}\|$ we obtain

$$\lim_{n\to\infty}\lim_{d_J\to\infty}\mathbb{E}\left[\sup_{\|w\|\le\alpha_n\|\hat{w}_{MN}\|,\, L_{\mathbf{S}}(w)=0} L_{\mathcal{D}}(w)\right]\ge L_{\mathcal{D}}(w^*)+(\alpha^2-1)L_{\mathcal{D}}(w^*)=\alpha^2 L_{\mathcal{D}}(w^*)$$

and the proof is concluded. $\square$