IN DEFENSE OF UNIFORM CONVERGENCE: GENERALIZATION VIA DERANDOMIZATION WITH AN APPLICATION TO INTERPOLATING PREDICTORS

JEFFREY NEGREA

Dept. of Statistical Sciences, Univ. of Toronto; Vector Institute

GINTARE KAROLINA DZIUGAITE

Element AI

DANIEL M. ROY

Dept. of Statistical Sciences, Univ. of Toronto; Vector Institute

ABSTRACT. We propose to study the generalization error of a learned predictor $\hat{h}$ in terms of that of a surrogate (potentially randomized) predictor that is coupled to $\hat{h}$ and designed to trade empirical risk for control of generalization error. In the case where $\hat{h}$ interpolates the data, it is interesting to consider theoretical surrogate classifiers that are partially derandomized or rerandomized, e.g., fit to the training data but with modified label noise. We also show that replacing $\hat{h}$ by its conditional distribution with respect to an arbitrary σ-field is a convenient way to derandomize. We study two examples, inspired by the work of Nagarajan and Kolter (2019) and Bartlett et al. (2019), where the learned classifier $\hat{h}$ interpolates the training data with high probability, has small risk, and, yet, does not belong to a nonrandom class with a tight uniform bound on two-sided generalization error. At the same time, we bound the risk of $\hat{h}$ in terms of surrogates constructed by conditioning and denoising, respectively, and shown to belong to nonrandom classes with uniformly small generalization error.

1. INTRODUCTION

One of the central problems in learning theory is to explain the statistical performance of deep learning algorithms. There is particular interest in explaining how overparameterized neural networks, trained by simple variants of stochastic gradient descent (SGD), simultaneously achieve low risk and zero empirical risk on benchmark datasets. While certain naive explanations have been ruled out (Zhang et al., 2017), progress has been slow.

The bulk of recent work has approached this problem by arguing that the classifier learned by SGD belongs to a class for which there is a uniform and tight (two-sided) bound on the generalization error (Bartlett, Foster, and Telgarsky, 2017; Golowich, Rakhlin, and Shamir, 2018; Long and Sedghi, 2020; Neyshabur et al., 2017; Wei and Ma, 2019). After raising this observation, Nagarajan and Kolter (2019) argue that this approach may be unable to explain performance observed in overparameterized models. They argue this point by constructing a simple problem where an SGD-like algorithm learns a classifier that achieves low risk and zero empirical risk, yet the learned classifier does not belong (even with high probability) to a class whose generalization error is uniformly small.

In this work, we initiate a response to Nagarajan and Kolter (2019) in defense of the utility of uniform convergence for understanding learning algorithms that obtain zero empirical risk. We will use the term "interpolating" to refer to such learning algorithms and the corresponding learned hypotheses, borrowing the terminology used for functions that achieve zero mean squared error. In problems involving interpolation, the complexity of the task (e.g., model complexity) often must increase with the dataset size in order for interpolation to be possible. This mirrors deep learning practice, where scientists will train larger, more complex models when presented with a larger dataset. Since the complexity of the learning problems in question—and possibly even the sample spaces generating the data—change with the sample size, the traditional notions of uniform convergence (Glivenko–Cantelli classes) are not applicable. Therefore we need to extend the concept of uniform convergence to the setting of sequences of learning problems of increasing complexity, which we do in Section 3 by defining the *structural Glivenko–Cantelli* property. Then in Section 4 we introduce a general approach to relate sequences of learning problems which are not structural Glivenko–Cantelli to ones that may be. The basic idea is to introduce a surrogate learning algorithm that closely mimics the learning algorithm of interest, yet whose output belongs to a class (the surrogate hypothesis class) for which a uniform and vanishing bound on two-sided generalization error holds.

The observations of Nagarajan and Kolter (2019) relate to a number of other empirical learning phenomena that demand explanation. One example is the phenomenon of double descent, brought to light by Advani and Saxe (2017), Belkin et al. (2019), and Geiger et al. (2019). The difficulty of explaining these double descent curves using standard uniform convergence arguments is a central theme of recent talks by Belkin. In a line of work by Hastie et al. (2019) and Mei and Montanari (2019), double descent was observed in unregularized, overparameterized linear regression. Bartlett et al. (2019) show that, for sequences of overparameterized linear regression tasks, the minimum norm interpolating solution to least squares will achieve asymptotically optimal risk with high probability given constraints on the covariate (feature) distribution. In this setting, we show that no class containing the learned hypothesis with high probability can have a vanishing uniform bound on the absolute generalization error. In fact, such a bound cannot be representative of the risk of the learned hypothesis. In Section 5, we show that the analysis of Bartlett et al. (2019) may be viewed as introducing a surrogate classifier. The surrogate in this case is the minimum-norm interpolating solution on the training data *with label noise removed*. We use standard techniques from empirical process theory to demonstrate that the surrogate hypothesis class—the collection of all surrogate hypotheses that could have been learned from the training data—has the structural Glivenko–Cantelli property. We combine our uniform bound based on the structural Glivenko–Cantelli property with other components

of the analysis of Bartlett et al. (2019) to obtain similar bounds on the expected risk of the minimum norm interpolating solution under weaker hypotheses.

In Section 6 we provide a relatively flexible recipe for constructing surrogate classifiers via probabilistic conditioning. The approach produces a probability measure over hypotheses via retraining on data that is equal in distribution to the original training data but has been partially "rerandomized". The approach effectively trades empirical risk for generalization error. Lastly, in Section 7, we apply this recipe to an example, inspired by Nagarajan and Kolter (2019), where an interpolating learning algorithm is constructed for which there is no structural Glivenko–Cantelli class containing the learned hypothesis with high probability. In that example, we construct a surrogate by conditioning with respect to a specific σ-field. We show the corresponding surrogate class is structural Glivenko–Cantelli and can be used to derive risk bounds for the learned classifier, which exhibit a form of double descent.

1.1. **Contibutions.** In this work, we extend our theoretical understanding of generalization, by way of the following contributions:

(1) Defining the structural Glivenko–Cantelli property, a notion of uniform convergence for sequences of learning problems.
(2) Proposing to study generalization error of learning algorithms—including interpolating ones—in terms of surrogate hypotheses that may belong to structural Glivenko–Cantelli classes, even when the original hypotheses do not.
(3) Demonstrating that the hypothesis spaces corresponding to a sequence of unregularized, overparameterized linear regression tasks are not structural Glivenko–Cantelli, but that they can be analyzed by introducing a sequence of surrogates for which the surrogate hypothesis class is structural Glivenko–Cantelli. We further use this fact to provide bounds on the expected risk of the original sequence of tasks under weaker hypotheses than Bartlett et al. (2019).
(4) Introducing a generic technique by which one may introduce surrogate learning algorithms via conditioning, which naturally trades empirical risk for generalization error relative to the original learning algorithm.
(5) Analyzing an example that distills the key features of an example in Nagarajan and Kolter (2019), via a family of surrogates obtained from conditioning. We show that, while the original learning algorithm does not output hypotheses in a sequence of classes with the structural Glivenko–Cantelli property, discarding a few bits of information leads to one that does. We also show that bounds obtained via the surrogate learning algorithm exhibit a form of double descent.

2. Preliminaries

Let $Z, Z_1, \ldots, Z_n$ be i.i.d. random elements in a space $\mathcal{S}$ with common distribution $\mathcal{D}$. Let $S = (Z_1, \ldots, Z_n)$ represent the training set. Fix a loss function $\ell : \mathcal{H} \times \mathcal{S} \to \mathbb{R}_+$ for a space $\mathcal{H}$ of hypotheses. Let $\mathcal{M}_1(\mathcal{H})$ be the space of distributions on $\mathcal{H}$. Note that $\mathcal{H}$ can be embedded into $\mathcal{M}_1(\mathcal{H})$ by the map $h \mapsto \delta_h$ taking a classifier to a Dirac measure degenerating on $\{h\}$. For $Q \in \mathcal{M}_1(\mathcal{H})$, the *(average) loss* and *risk* are defined to be

$$\ell(Q, z) = \int \ell(h, z) Q(\mathrm{d}h), \qquad\qquad L_{\mathcal{D}}(Q) = \int \ell(Q, z) \mathcal{D}(\mathrm{d}z).$$

Let $L_S(Q) = L_{\widehat{\mathcal{D}}_n}(Q)$ denote the *empirical (average) risk*, where $\widehat{\mathcal{D}}_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}$ is the empirical distribution. For $h \in \mathcal{H}$, define $L_{\mathcal{D}}(h) = L_{\mathcal{D}}(\delta_h)$ and $L_S(h) = L_S(\delta_h)$. Let $\hat{h}$ or $\hat{h}(S)$ be a random element in $\mathcal{H}$, representing a learned classifier.

A hypothesis h *interpolates* a dataset S with respect to a non-negative loss ℓ when $L_S(h) = 0$. A learning algorithm $\hat{h}(S)$ is (almost surely) *interpolating* if $L_S(\hat{h}(S)) = 0$ a.s. (or equivalently $\mathbb{E}L_S(\hat{h}_S) = 0$). This extends our geometric intuition that a surface $h : \mathbb{R}^d \to \mathbb{R}$ interpolates points in $\{(x_i, y_i)\}_{i \in [n]} \mathbb{R}^d \times \mathbb{R}$ when $(h(x_i) - y_i)^2 = 0$ for all $i \in [n]$. The surprising properties of interpolating classifiers are explored in Belkin et al. (2019). See also Advani and Saxe (2017) and Geiger et al. (2019).

3. Structural Uniform Convergence

Nagarajan and Kolter (2019) argue that uniform convergence does not explain generalization in several examples that are emblematic of the modern interpolating regime. In those examples, however, the size of the learning problem varies with the cardinality of the training dataset. The standard notion of uniform convergence (i.e., of Glivenko–Cantelli classes, etc.) is not normally defined in this setting. In order to formalize the specific failure of "uniform convergence" in these sequences of learning problems, we introduce a *structural* version of the Glivenko–Cantelli property.

Definition 3.1. Let $\{(\mathcal{S}^{(p)}, \mathcal{F}^{(p)}, \mathcal{D}^{(p)})\}_{p \in \mathbb{N}}$ be a sequence of probability spaces where $\mathcal{S}^{(p)}$ denotes the sample space, $\mathcal{F}^{(p)}$ denotes the σ-field and $\mathcal{D}^{(p)}$ denotes the probability measure. Let $\mathcal{H}^{(p)}$ be a collection of measurable functions on $(\mathcal{S}^{(p)}, \mathcal{F}^{(p)}, \mathcal{D}^{(p)})$ and let $n_p \in \mathbb{N}$ for all $p \in \mathbb{N}$.

Then $\mathcal{H}^{(\cdot)}$ has the *structural $(\mathcal{D}^{(\cdot)}, n_{(\cdot)})$-Glivenko–Cantelli* property, denoted $(\mathcal{D}^{(\cdot)}, n_{(\cdot)})$-SGC, if

$$\lim_{p\to\infty} \mathbb{E}\left[\sup_{h \in \mathcal{H}^{(p)}} \left|\mathcal{D}^{(p)}h - \widehat{\mathcal{D}^{(p)}}_{n_p}h\right|\right] = 0,$$

where $Ph = \int h(x)P(dx)$ and $\widehat{\mathcal{D}^{(p)}}_{n_p}$ is the empirical distribution of an IID sample of size n_p from $\mathcal{D}^{(p)}$.

It is this property which is made to fail in the examples presented by Nagarajan and Kolter (2019). When $\{(\mathcal{S}^{(p)}, \mathcal{F}^{(p)}, \mathcal{D}^{(p)})\}_{p \in \mathbb{N}}$ and $\mathcal{H}^{(p)}$ are constant and $n_p = p$, this reduces to the classical notion of Glivenko–Cantelli.

Remark 3.2 (Relationship between PAC and nonuniform learning). PAC learnability and nonuniform learnability, as defined in (Shalev-Shwartz and Ben-David, 2014), can both be understood in terms of the structural Glivenko–Cantelli property. That PAC learnability implies structural Glivenko–Cantelli follows from the equivalence of PAC learnability with the uniform Glivenko–Cantelli property.

To understand the nonuniform learnability of a class $\mathcal{H}$ in terms of the structural Glivenko–Cantelli property, recall the equivalence that $\mathcal{H}$ is nonuniformly learnable if and only if it is a countable union of VC classes—$\mathcal{H} = \bigcup_{j \in \mathbb{N}} \mathcal{H}_j$ with $\bigcup_{j \in [p]} \mathcal{H}_j$ of finite VC dimension

d_p. Then, for any sequences $\delta_p \searrow 0$ and $\epsilon_p \searrow 0$, take $n_p \geq C_2 \frac{d_p + \log(1/\delta_p)}{\epsilon_p^2}$ where C_2 is the universal constant appearing in (Shalev-Shwartz and Ben-David, 2014, Theorem 6.8, Item 1.). Taking $\{(\mathcal{S}^{(p)}, \mathcal{F}^{(p)}, \mathcal{D}^{(p)})\}_{p \in \mathbb{N}}$ to be constant, and $\mathcal{H}^{(p)} = \bigcup_{j \in [p]} \mathcal{H}_j$, it follows immediately that

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}^{(p)}} \left| \mathcal{D}^{(p)} h - \widehat{\mathcal{D}^{(p)}}_{n_p} h \right|\right] \leq (1 - \delta_p)\epsilon_p + \delta_p \to 0,$$

and hence $\mathcal{H}^{(\cdot)}$ is $(\mathcal{D}^{(\cdot)}, n_{(\cdot)})$-SGC. It would be reasonable in this case to say that $\mathcal{H}^{(\cdot)}$ is $(n_{(\cdot)})$-SGC uniformly over data generating distributions.

The partitioning of the hypothesis space in synchronization with increasing sample size in this derivation is similar to the partitioning of the hypothesis space by sample size occurring in the structural risk minimization algorithm for nonuniform learning. The analysis above tells us that any ERM algorithm restricted to $\mathcal{H}^{(p)}$ when the sample size is n_p will achieve low generalization error. ◁

4. Decompositions of Generalization Error using Surrogate Classifiers

We now describe how one may pass from bounding the generalization error of a learning algorithm to bounding the generalization error of a surrogate and controlling differences in the risk and empirical risk profiles between the original algorithm and the surrogate.

The following result is immediate from the linearity of expectation:

Lemma 4.1 (Surrogate decomposition). *For every random element Q in $\mathcal{M}_1(\mathcal{H})$,*

$$\begin{aligned}
\mathbb{E}[L_\mathcal{D}(\hat{h}) - L_S(\hat{h})] &= \mathbb{E}[L_\mathcal{D}(\hat{h}) - L_\mathcal{D}(Q)] \\
&\quad + \mathbb{E}[L_\mathcal{D}(Q) - L_S(Q)] \\
&\quad + \mathbb{E}[L_S(Q) - L_S(\hat{h})],
\end{aligned}$$

provided the three expectations on the r.h.s. are finite.

This decomposition suggests that one can obtain a bound on the generalization error (and then the risk) of $\hat{h}$ by bounding the three terms individually. We interpret Q here as a (possibly randomized) surrogate hypothesis that is coupled with $\hat{h}$ via some information in the training algorithm and/or the training data. The choice of Q trades off one term for another. In the particular case of a.s. interpolating classifiers (i.e., $\mathbb{E}[L_S(\hat{h})] = 0$), one approach is to trade excess empirical risk, $\mathbb{E}[L_S(Q) - L_S(\hat{h})]$, for less generalization error, $\mathbb{E}[L_\mathcal{D}(Q) - L_S(Q)]$.

One way to control generalization error is to show that Q belongs to a nonrandom class for which there holds a uniform and tight bound on generalization error.

Proposition 4.2 (Bounded loss, two-sided control). *Assume ℓ takes values in an interval of length L. For every random element Q in $\mathcal{M}_1(\mathcal{H})$ and class $G \subseteq \mathcal{M}_1(\mathcal{H})$,*

$$\begin{aligned}
&\mathbb{E}[L_\mathcal{D}(Q) - L_S(Q)] \\
&\quad \leq L\,\mathbb{P}[Q \notin G] + \mathbb{E}\left[\sup_{P \in G} |L_\mathcal{D}(P) - L_S(P)|\right].
\end{aligned}$$

Just as we interpret Q as a surrogate hypothesis that depends on the dataset S, we view G in Proposition 4.2 as a surrogate hypothesis class that contains the surrogate hypothesis with high probability.

The surrogate decomposition may be viewed as similar to a one-step covering argument, where the cover is given by the class of surrogate hypotheses, and the approximation error is given by $\mathbb{E}[L_{\mathcal{D}}(\hat{h}) - L_{\mathcal{D}}(Q)] + \mathbb{E}[L_{\mathcal{D}}(Q) - L_S(Q)]$. In a typical one-step covering argument, the cover is chosen to be sufficiently fine as to have a uniformly small approximation error. The optimal cover density will vary with sample size so that approximation error vanishes as sample size increases. The key difference here is that we may not be able to control the approximation error uniformly or have any hope that it will vanish based on the covering induced by a surrogate. We will only attempt to uniformly control the cover given by the surrogate class. We then can rely on other techniques to handle the approximation error. This allows us to divide the objective of explaining generalization into a portion explained by uniform convergence and portion not explained by uniform convergence.

5. Overparameterized Linear Regression

Our first application of using surrogates and the structural Glivenko–Cantelli property to understand generalization error is inspired by the recent work of Bartlett et al. (2019). They determine necessary and sufficient conditions under which the *minimum norm interpolating linear predictor* generalizes well in mean-squared error for random design linear regression in the overparameterized regime (i.e., more features than observations) for sub-Gaussian random designs with conditionally sub-Gaussian residuals. We chose overparameterized linear regression as a first example to present because of 1) the natural failure of any notion of uniform convergence to explain performance in the problem (in particular we show that the structural Glivenko–Cantelli property fails for any classes containing the learned hypotheses with high probability), and 2) recent work (such as by Hastie et al. (2019) and Mei and Montanari (2019) and others) showing that double descent occurs in variants of this problem such as random feature regression.

We show that the decomposition used by Bartlett et al. (2019) can be naturally related to a decomposition via a surrogate hypothesis, and that said sequence of classes of attainable surrogates is structural Glivenko–Cantelli. We use the uniform convergence of the surrogate to provide bounds on the expected generalization error. We handle only the Gaussian random design / Gaussian response case[1], but note that the results can be extended to the sub-Gaussian with minor modifications. We provide bounds on the expected generalization error only, as the purpose of this example is to illustrate how uniform convergence of a surrogate may be used. However, we would like the reader to be aware that high-probability bounds may be obtained using the high-probability versions of the chaining arguments that appear in the proofs, as is standard in empirical process theory.

---

[1]Bartlett et al. (2019) handled the Gaussian design / Gaussian response case in the first version of their work, and gave extensions to the sub-Gaussian case in a recent update.

Bartlett et al. (2019) define a sequence of covariance matrices $\Sigma_n \in \mathbb{R}^{d_n \times d_n}$ to be *benign* when

$$\lim_{n \to \infty} \left(\|\Sigma_n\| \sqrt{\frac{r_0(\Sigma_n)}{n}} + \frac{k_n^*(\Sigma_n)}{n} + \frac{n}{R_{k_n^*(\Sigma_n)}(\Sigma_n)} \right) = 0,$$

where (for some universal constant $b > 0$ defined in (Bartlett et al., 2019))

$$r_k(\Sigma) = \frac{\sum_{i>k} \lambda_i(\Sigma)}{\lambda_{k+1}(\Sigma)} \qquad\qquad R_k(\Sigma) = \frac{\left(\sum_{i>k} \lambda_i(\Sigma)\right)^2}{\sum_{i>k} \lambda_i^2(\Sigma)},$$

$\{\lambda_i(\Sigma)\}_{i \in [d]}$ are the eigenvalues of Σ in decreasing order (with multiplicity), and

$$k_n^*(\Sigma) = \min \left\{ k \geq 0 : r_k(\Sigma) \geq bn \right\}.$$

Their work shows that it is sufficient that Σ_n be *benign* in order to guarantee that $L_D(\hat{\beta}) \to \sigma^2$. They also show that it is necessary that

$$\lim_{n \to \infty} \left(\frac{k_n^*}{n} + \frac{n}{R_{k_n^*(\Sigma_n)}(\Sigma_n)} \right) = 0$$

in order for $\mathbb{E}L_D(\hat{\beta}) \to \sigma^2$. Note that the risk of the true coefficient vector, β (which is also the global optimizer of the risk) is σ^2, and so we expect the risk of any estimator to be at least that large. We employ instead a weaker hypothesis.

Definition 5.1 (Weakly benign)**.** A sequence of covariance matrix, coefficient vector pairs, $\{(\Sigma_n, \beta_n)\}_{n \in \mathbb{N}}$ is *weakly benign* if

$$\lim_{n \to \infty} \left(\frac{\beta_n' \Sigma_n \beta_n}{\sqrt{n}} + \frac{k_n^*(\Sigma_n)}{n} + \frac{n}{R_{k_n^*(\Sigma_n)}(\Sigma_n)} \right) = 0.$$

Clearly, for a norm-bounded sequence of coefficient vectors $\{\beta_n\}_{n \in \mathbb{N}}$, if $\{\Sigma_n\}_{n \in \mathbb{N}}$ is benign then $\{(\Sigma_n, \beta_n)\}_{n \in \mathbb{N}}$ is weakly benign since $r_0(\Sigma) \geq 1$. The converse may not hold, however, if $\sqrt{r_0(\Sigma_n)}$ is unbounded. Heuristically the growth rate of that quantity determines how "spiked" a covariance matrix is; it is smaller when "a few" eigenvalues are much larger than the rest. That (Σ_n, β_n) form a weakly benign sequence is therefore weaker than the sufficient condition of Bartlett et al. (2019) but stronger than the necessary condition.

5.1. **Construction.** Let $X_i \overset{\text{iid}}{\sim} N_{1 \times d}(0, \Sigma)$ be random row vectors with non-singular $d \times d$ feature covariance matrix Σ. Let $X = (X_1', \ldots X_n')'$ be the corresponding $n \times d$ random design matrix. Let $(Y_i \mid X) \overset{\text{ind}}{\sim} N(X_i \beta, \sigma^2)$ and $Y = (Y_1, \ldots Y_n)'$ be the responses and response vector respectively. Let $Z = Y - X\beta$ be the residual vector. The loss function will be squared error $\ell(\beta, (x, y)) = (x\beta - y)^2$. We want to understand the generalization performance of the *minimum norm interpolating linear predictor* for (X, Y), $\hat{\beta} = (X'X)^+ X'Y$ where A^+ denotes the Moore–Penrose pseudo-inverse for A.

5.2. **Failure of uniform convergence for this problem.**

Lemma 5.2 (Failure of uniform convergence for overparameterized linear regression)**.** *There is no sequence of measurable sets $\{A_n\}_{n\in\mathbb{N}}$ such that $\mathbb{P}((X,Y) \in A_n) > 2/3$ for all $n \in \mathbb{N}$ and for which*

$$\limsup_{n\to\infty} \mathbb{E} \sup_{(\tilde{X},\tilde{Y})\in A_n} \left|L_D(\hat{\beta}(\tilde{X},\tilde{Y})) - L_S(\hat{\beta}(\tilde{X},\tilde{Y}))\right| \leq \frac{3}{2}L_D(\beta).$$

The proof of this result is found in Appendix B.

5.3. **Introducing a surrogate.** We will consider the surrogate given by the minimum norm interpolating predictor for the training data with label noise removed. Mathematically, this can be defined by taking $\hat{\beta}_0 = (X'X)^+ X'X\beta$. Notice that $\hat{\beta}_0 = P(X)\beta$ where $P(X)$ is the projection onto the row-space of X.

The surrogate decomposition of the generalization error for $\hat{\beta}$ is given in the following lemma.

Lemma 5.3 (Surrogate decomposition of $\hat{\beta}$)**.**

$$L_D(\hat{\beta}) - L_S(\hat{\beta})$$
$$= (L_S(\hat{\beta}_0) - L_S(\hat{\beta})) + (L_D(\hat{\beta}) - L_D(\hat{\beta}_0))$$
$$+ (L_D(\hat{\beta}_0) - L_S(\hat{\beta}_0)),$$

with

$$L_S(\hat{\beta}_0) - L_S(\hat{\beta}) = \frac{1}{n} \|Z\|^2$$
$$L_D(\hat{\beta}) - L_D(\hat{\beta}_0) = \text{Tr}(X(X'X)^+\Sigma(X'X)^+X'ZZ')$$
$$L_D(\hat{\beta}_0) - L_S(\hat{\beta}_0) = \sigma^2 - \frac{\|Z\|^2}{n} + \beta'P(X)^\perp\Sigma P(X)^\perp\beta$$

The proof appears in Appendix B. Note that $(L_D(\hat{\beta}) - L_D(\hat{\beta}_0))$ and $(L_S(\hat{\beta}_0) - L_S(\hat{\beta})) + (L_D(\hat{\beta}_0) - L_S(\hat{\beta}_0))$ are exactly the terms which (Bartlett et al., 2019) choose to bound separately. They, however, made this decision using a problem-specific decomposition of the generalization error rather than arriving at the decomposition because it naturally arose from a choice of surrogate.

Lemma 5.4 (The sequence of surrogate hypothesis classes is SGC)**.** *The sequence of implied surrogate hypothesis classes, $\{\hat{\beta}_0(S) : S \in \mathcal{S}^{(n)}\}_{n\in\mathbb{N}}$ is $(\mathcal{D}^{(n)}, n)$-SGC when $(\beta_n'\Sigma_n\beta_n)/\sqrt{n} \to 0$. In particular this holds whenever $\{(\Sigma_n, \beta_n)\}_{n\in\mathbb{N}}$ is weakly benign. Quantitatively, for a universal constant $C > 0$,*

$$\mathbb{E} \sup_{(X_0,Y_0)\in\mathbb{R}^{n\times d}\times\mathbb{R}^n} \left|L_D(\hat{\beta}_0(X_0,Y_0)) - L_S(\hat{\beta}_0(X_0,Y_0))\right|$$
$$\leq C\frac{\sigma^2 + \beta'\Sigma\beta}{\sqrt{n}}$$

The proof appears in Appendix B. Combining Lemma 5.4 with (Bartlett et al., 2019, Lemma 11) (which controls $L_D(\hat{\beta}) - L_D(\hat{\beta}_0)$), we get the following bound on the expected generalization error.

Theorem 5.5 (Expected risk bound for overparameterized linear regression). *For some universal constant $C, c, b > 0$,*

$$\mathbb{E}L_D(\hat{\beta}) \leq \sigma^2 + C\frac{\sigma^2 + \beta'\Sigma\beta}{\sqrt{n}}$$
$$+ c\sigma^2\left(\frac{k_n^*}{n} + \frac{n}{R_{k_n^*}(\Sigma_n)}\right)$$

where b is the same constant appearing in the definition of k^* above. In particular, if $\{(\Sigma_n, \beta_n)\}_{n \in \mathbb{N}}$ is weakly benign then $\mathbb{E}L_D(\hat{\beta}) \to \sigma^2$.

The result is similar to what one would obtain by converting the high-probability bound of Bartlett et al. (2019) into a bound in expectation. However we remove the dependence on $r_0(\Sigma)$ and hence weaken the conditions sufficient to ensure generalization. Furthermore, our approach highlights the role of uniform convergence via a surrogate in this problem. It is note-worthy that, when viewing the surrogate class as a type of one-step covering as discussed in the comments after Proposition 4.2, the approximation error component of the surrogate decomposition does not vanish in this case. Instead, it tends to σ^2 when the covariance matrix, coefficient vector pairs are weakly benign, and may have more erratic behaviour otherwise.

6. Constructing Surrogates by Conditioning

In the overparameterized linear regression example, the surrogate obtained by training on "de-label-noised" data allowed us to construct meaningful generalization bounds. For a generic learning problem, however, there may be no notion of label noise or such an approach may not prove useful. This leads us to seek natural constructions of surrogates in less structured problems.

One generic way to introduce such a surrogate is by conditioning. Let $\mathbb{P}^{\mathcal{F}}$ denote the conditional probability operator given a σ-field $\mathcal{F}$ (or a random variable), taking an event to its conditional probability. For a random variable ψ, let $\mathbb{P}^{\mathcal{F}}[\psi]$ denote the conditional distribution of ψ given $\mathcal{F}$.

Lemma 6.1 (Derandomization via conditioning). *Let $\mathcal{F}$ be a σ-field on (some possible extension of) the underlying probability space upon which S and $\hat{h}$ are defined. Let $Q = \mathbb{P}^{\mathcal{F}}[\hat{h}]$. Then $\mathbb{E}[L_D(\hat{h}) - L_D(Q)] = 0$.*

The following result is then immediate by Lemmas 4.1 and 6.1.

Lemma 6.2 (Surrogate decomposition by conditioning). *Let $\mathcal{F}$ and Q be as in Lemma 6.1. Then*

$$\mathbb{E}[L_D(\hat{h}) - L_S(\hat{h})] = \mathbb{E}[L_S(Q) - L_S(\hat{h})]$$
$$+ \mathbb{E}[L_D(Q) - L_S(Q)].$$

If $\hat{h}$ is a.s. interpolating (i.e., $\mathbb{E}[L_S(\hat{h})] = 0$), then

$$\mathbb{E}[L_D(\hat{h})] = \mathbb{E}[L_S(Q)] + \mathbb{E}[L_D(Q) - L_S(Q)].$$

Every conditional distribution $Q = \mathbb{P}^{\mathcal{F}}[\hat{h}]$ represents a *derandomization* of $\hat{h}$: i.e., by the definition of conditioning, $\hat{h}$ has equal or greater dependence on the data S than Q. There are other ways to achieve derandomization rather than conditioning. However, they may require one to obtain some explicit control on the risk difference, $\mathbb{E}[L_{\mathcal{D}}(\hat{h}) - L_{\mathcal{D}}(Q)]$.

Informally, if $\hat{h}$ interpolates (or more generally overfits), we would expect a derandomized classifier to have excess empirical risk, yet lower generalization error.

Finally, it is important to understand how tautologies can arise from this perspective. If Q is a.s. nonrandom (corresponding, e.g., to conditioning on the trivial σ-algebra), then $Q = \mathbb{P}[\hat{h}]$ a.s., i.e., Q is the distribution of $\hat{h}$. In this case, $\mathbb{E}L_S(Q) = \mathbb{E}L_{\mathcal{D}}(Q) = \mathbb{E}L_{\mathcal{D}}(\hat{h})$, and we obtain the tautology

$$
\begin{aligned}
\mathbb{E}[L_{\mathcal{D}}(\hat{h}) - L_S(\hat{h})] &= \mathbb{E}[L_{\mathcal{D}}(\hat{h}) - L_{\mathcal{D}}(Q)] \\
&+ \mathbb{E}[L_{\mathcal{D}}(Q) - L_S(Q)] \\
&+ \mathbb{E}[L_S(Q) - L_S(\hat{h})] \\
&= 0 + \mathbb{E}[L_{\mathcal{D}}(\hat{h}) - L_S(\hat{h})] + 0.
\end{aligned}
$$

For this extreme example, Q belongs to the singleton class $\{\mathbb{P}[\hat{h}]\}$, which exhibits "uniform convergence" trivially. On the other end of the spectrum, if $Q = \delta_{\hat{h}}$, i.e., we condition on $\mathcal{F} = \sigma(S)$, then we get an equally tautological statement from the decomposition. The idea behind introducing the surrogate classifier Q is that it allows one to conceptually interpolate between these two tautological end points in order to find a (non-tautological) bound on the generalization error of a learning algorithm.

7. Hypercube classifier

The following example is inspired by theoretical and empirical work by Nagarajan and Kolter (2019). Like in their work modelling SGD, we describe an example of a low-risk learned classifier, $\hat{h}$, such that there is no nonrandom class containing $\hat{h}$ almost surely for which one may establish a uniform and nonvacuous bound on generalization error. Using Lemma 6.2 and Proposition 4.2, we show that a derandomization of $\hat{h}$, obtained by conditioning on an explicit σ-field $\mathcal{F}$, yields a tight generalization bound based on uniform convergence of the surrogate.

In this section, we first construct the learning problem we will address. Second, we show that the structural Glivenko–Cantelli property fails on this example even though is has low generalization error. Lastly, we introduce our surrogate learning algorithm, show that it has similar empirical and test performance to the original algorithm, verify that the surrogate has the structural Glivenko–Cantelli property, and finally use this to establish a generalization bound for the original learning algorithms.

7.1. **Construction.** Let $d \gg 1$ index the dimensionality of the feature space and our sequence of learning problems, and let n_d be the sample size for learning problem with index d. Let $\mathcal{X}^{(d)} = \{0,1\}^{2d}$ be the feature space and $\mathcal{Y} = \{0,1\}$ be the label space, and let

$\mathcal{S}^{(d)} = \mathcal{X}^{(d)} \times \mathcal{Y}$. Let $f_d^* : \mathcal{X} \to \mathcal{Y}$ be given by

$$f_d^*(x) = \begin{cases} 1, & \|X\|_1 \leq d, \\ 0, & \text{otherwise.} \end{cases}$$

For $x \in \mathcal{X}^{(d)}$, note that $f_d^*(x) = 1 - f_d^*(1 - x)$. Let $\mathcal{D}^{(d)}$ be the distribution of $(X, f_d^*(X))$, where $X \sim \text{Unif}\left(\mathcal{X}^{(d)}\right)$. Let $S = (Z_1, \dots, Z_{n_d}) \sim (\mathcal{D}^{(d)})^{n_d}$ where $Z_i = (X_i, f_d^*(X_i))$. Let $\hat{f}_d$ be the random element in $\{0,1\}^{2d} \to \{0,1\}$ given by

$$\hat{f}_d^{(S)}(x) = \begin{cases} 1 - f_d^*(x), & x \notin S \text{ and } 1 - x \in S \\ f_d^*(x), & \text{otherwise} \end{cases}.$$

Let $\bar{Z}_i = (1 - X_i, 1 - Y_i)$ and $\bar{S} = (\bar{Z}_1, \dots, \bar{Z}_n)$. We refer to pairs $(Z_i, \bar{Z}_i)$ as antipodes. Our learning algorithm, $\mathcal{A}_d : S \to \hat{f}_d^{(S)}$, only makes a classification error when a test point was not in the training set, but its antipode was.

7.2. **Failure of uniform convergence for this problem.** First, we will note that at every problem size, d, the VC dimension of the collection of accessible decision rules is at least as large as the training dataset. We will not use this fact again, but it does highlight the apparent complexity of the learning problem.

Proposition 7.1 (VC theory not applicable). *For $n_d \leq 2^{2d-1}$, $\mathcal{H}^{(d)} = \{\hat{f}_d^{(S)} : S \in (\mathcal{S}^{(d)})^{n_d}\}$ has VC-dimension at least n_d.*

Proof. For $n_d \leq 2^{2d-1}$, any set of features $(X_1, \dots X_{n_d})$ of size n_d with no antipodal points and no repeated points can be shattered by the subcollection of $\mathcal{H}^{(d)}$ given by $\left\{\hat{f}_d^{(S)} : S \in \prod_{i \in [n_d]} \{Z_i, \bar{Z}_i\}\right\}$. □

Next, notice that this algorithm never makes an error on the training data.

Lemma 7.2 ($\hat{f}_d$ is interpolating.). $L_S(\hat{f}_d) = 0$ *a.s.*

Furthermore, by construction, the learning algorithm cannot return a classifier with high risk, no matter the training data observed, as long as $n_d \in o(2^{2d})$.

Lemma 7.3 ($\hat{f}_d$ has small risk). $L_{\mathcal{D}^{(d)}}(h) \leq n_d 2^{-2d}$ *for all* $h \in \mathcal{H}^{(d)} = \{\hat{f}_d^{(S)} : S \in (\mathcal{S}^{(d)})^{n_d}\}$, *and hence*

$$L_{\mathcal{D}}(\hat{f}_d) - L_S(\hat{f}_d) \leq n_d 2^{-2d} \quad \text{a.s.}$$

The proof appears in Appendix A. The following result demonstrates that uniform convergence (of a class containing $\hat{h}$) *does not* explain the risk. The argument mirrors that of Nagarajan and Kolter (2019).

Theorem 7.4 ($\mathcal{H}^{(\cdot)}$ is not SGC). *If $n_d \in o(2^d)$ then $\left\{\ell \circ \mathcal{H}^{(d)}\right\}_{d \in \mathbb{N}}$ is not $(\mathcal{D}^{(\cdot)}, n_{(\cdot)})$-SGC, in fact*

$$\mathbb{E} \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| = 1 - O(n^2 2^{-2d}).$$

The proof appears in Appendix A. In this example, a generalization error bound was tractable because $L_\mathcal{D}(h)$ was readily bounded for all h, despite the fact that uniform convergence failed for $\mathcal{H}^{(\cdot)}$. One may then ask "how many bits of information do we need to forget about our training data in order for the sequence surrogate hypothesis class obtained by conditioning is structural Glivenko–Cantelli?"

7.3. **Introducing a surrogate classifier.** Let $k_d \leq 2d$. Let $\pi_{k_d} : \{0,1\}^{2d} \to \{0,1\}^{2d}$ satisfy

$$\pi_{k_d}(x_1, \ldots, x_{2d})_j = \begin{cases} 0, & j \leq k_d, \\ x_j, & \text{otherwise.} \end{cases}$$

That is, $\pi_{k_d}(x)$ zeros out the first k_d entries of x. Now, let $\pi_{k_d}(S) = (\pi_{k_d}(X_1), \ldots, \pi_{k_d}(X_{n_d}))$, $\mathcal{G}^{(d)} = \sigma(\pi_{k_d}(S))$, and put $Q(S) = \mathbb{P}^{\mathcal{G}^{(d)}}[\hat{f}_d^{(S)}]$. $Q(S)$ is a Gibbs classifier that is learned from the data, but is less coupled with the data than $\hat{f}_d^{(S)}$. Intuitively, conditioning our learned classifier on $\mathcal{G}^{(d)}$ can be interpreted as redrawing the first k features and the labels associated with the training data independently for each new test point, holding the last $2d - k$ features of each training point fixed. Since $Q(S)$ is $\sigma(\pi_{k_d}(S))$-measurable, then for distinct datasets S and S' with $\pi_{k_d}(S) = \pi_{k_d}(S')$, we have $Q(S) = Q(S')$.

The map $S \mapsto Q(S)$ is the *surrogate learning algorithm* and $Q(S)$ is the *surrogate classifier*. Note that in this example, our chosen surrogate learning algorithm returns a Gibbs classifier, while the original algorithm returns a deterministic classification rule. When the argument S of Q is omitted then it is assumed to be the training dataset, S.

We first evaluate the risk and empirical risk of our surrogate classifier.

Lemma 7.5 (Risk and empirical risk of the surrogate Q)**.** *The following all hold almost surely*

$$L_S(Q) \leq \frac{2^{-k_d}(1 - 2^{-k_d})}{n_d}$$
$$\times \left| \left\{ (i,j) : X_i[k+1 : 2d] = \overline{X_j[k+1 : 2d]} \right\} \right|$$
$$L_{\overline{S}}(Q) \leq \frac{2^{-k}}{n_d} |\{(i,j) : X_i[k+1 : 2d] = X_j[k+1 : 2d]\}| \text{ and}$$
$$L_\mathcal{D}(Q) \leq n_d 2^{-2d}.$$

Furthermore,

$$\mathbb{E}L_S(Q) \leq (n_d - 1)2^{-2d}(1 - 2^{-k_d}),$$
$$\mathbb{E}L_{\overline{S}}(Q) \leq 2^{-k_d} + (n_d - 1)2^{-2d}, \text{ and}$$
$$\mathbb{E}L_\mathcal{D}(Q) \leq n_d 2^{-2d}.$$

The proof appears in Appendix A. As was foreshadowed in Section 4, we have increased the empirical risk by replacing $\hat{f}_d$ with Q. However, at the same time, we have dramatically lowered the empirical risk on the adversarial (antipodal) dataset, and not affected the true risk at all. In fact we are able to trade off empirical risk on the training data with worst case risk on an adversarial dataset explicitly by varying the parameter k_d. Even a small amount of re-randomization in the surrogate (k_d small) can yield very tight control on the adversarial

empirical risk. In this example, the adversarial empirical risk decreases exponentially fast in the number of bits of information lost per example. This allows us to demonstrate that the sequence of surrogate hypothesis classes is structural Glivenko–Cantelli.

Lemma 7.6 (The surrogate is SGC). *Consider the sequence of surrogate hypothesis classes given by $\mathcal{I}^{(d)} = \{Q(S) : S \in \mathcal{S}^{(d)}\}$. If $n_d \in o(2^{k_d}/\sqrt{d})$ then $\{\ell \circ \mathcal{I}^{(d)}\}_{d \in \mathbb{N}}$ is $(\mathcal{D}^{(\cdot)}, n_{(\cdot)})$-SGC. In particular, when $n_d \in o(2^{2d})$ and $k_d \geq \lceil (1 + \epsilon) \log_2(n_d) + \log_2(d)/2 \rceil$ then we have $\{\ell \circ \mathcal{I}^{(d)}\}_{d \in \mathbb{N}}$ is $(\mathcal{D}^{(\cdot)}, n_{(\cdot)})$-SGC. If $\log(n_d) \in \Omega(d)$ this is only possible if $k_d \in \Omega(d)$. If $\log(n_d) \in o(d)$ then this is possible even when $k_d \in o(d)$*

The proof appears in Appendix A. Note that the restrictions upon k_d we provide may be a product of our particular approach to bounding the Rademacher complexity (Massart's Lemma). A more refined approach may yield looser restrictions on k_d. Since the surrogate behaves similarly on training and test data to the original learning algorithm, and since the sequence of classes of achievable surrogates is SGC, we can establish a generalization error bound for the original learning algorithm using the uniform convergence of the surrogate.

Theorem 7.7 (Bounding generalization error via an SGC surrogate). *We have the following bound on the generalization error:*

$$\mathbb{E}[L_{\mathcal{D}}(\hat{f}_d) - L_S(\hat{f}_d)]$$
$$\leq (n_d - 1)2^{-2d} + 2\sqrt{\log(2)}n_d^{1/2}((2d - k_d)n_d + 1)^{1/2}2^{-k_d}.$$

If $n_d \in o\left(2^{2d}\right)$, then for any choice of $\{k_d\}_{d \in \mathbb{N}}$ with $\lim_{d \to \infty}(k_d - \log_2(n_d) - \log_2(d)/2) = \infty$, our surrogate witnesses that the generalization error vanishes

$$\mathbb{E}[L_{\mathcal{D}}(\hat{f}_d) - L_S(\hat{f}_d)] \to 0.$$

The proof appears in Appendix A

7.4. **Interpreting the derandomization bound.** We visualize the dependence of our generalization error bound on k, n and d in Fig. 1. Notice that the bound decays very rapidly in the proportion of randomness removed by conditioning, $k/2d$. When the learning problem is sufficiently complex, even for larger sample size, a small proportion of derandomization leads to a strong control of the generalization error.

7.4.1. *Relationship with Double Descent.* The bound produced by derandomization also exhibits a form of double descent. The rising left branch of each curve in Fig. 2 is a bound on the generalization error based on uniform convergence of a VC-class containing the learned predictor, where the VC dimension is bounded by $\min(n, 2d - 1)$. For $d \ll n$ this gives nonvacuous bounds for the low dimensional setting. Since the classifier is interpolating almost surely, even in the low complexity setting (low dimension d) the first descent begins at 0 and only shows the rising component as complexity increases. The right branch of each curve in Fig. 2, based on the bound of Theorem 7.7, shows the second descent in the high dimensional setting — for sufficiently complex models one may bound the generalization error of learned classifier via uniform convergence of a suitable derandomized classifier. *The bound obtained via derandomization is nonvacuous in the region of the second descent, exactly where standard uniform convergence techniques based on VC theory would give vacuous bounds.*
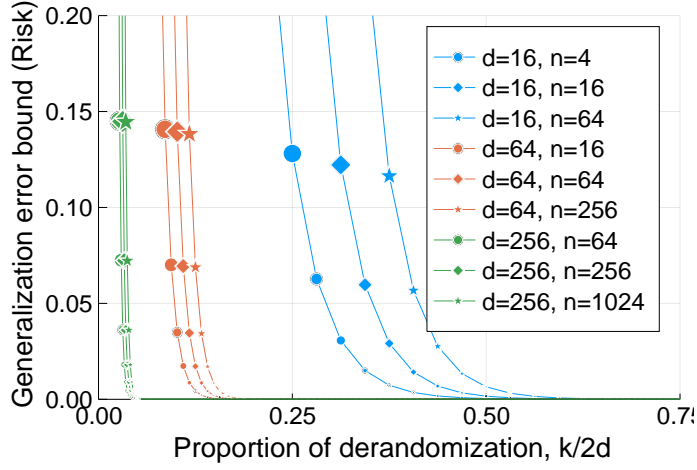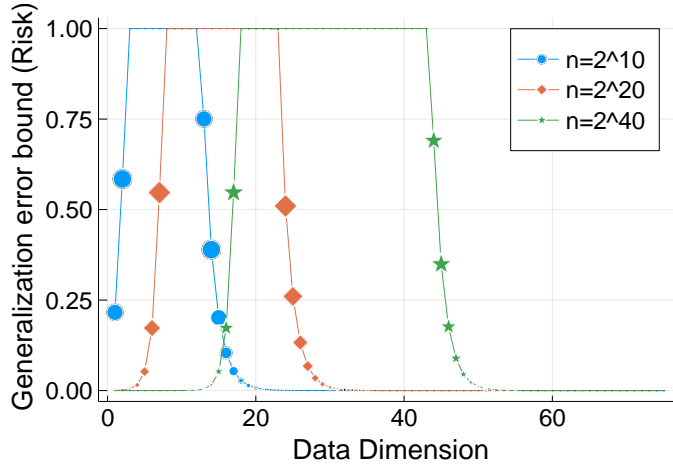
FIGURE 1. Visualizing the bound of Theorem 7.7.

FIGURE 2. Double descent in the bound of Theorem 7.7.

REFERENCES

Advani, M. S. and A. M. Saxe (2017). *High-dimensional dynamics of generalization error in neural networks.* arXiv: `1710.03667`.

Bartlett, P. L., D. J. Foster, and M. J. Telgarsky (2017). "Spectrally-normalized margin bounds for neural networks". In: *Advances in Neural Information Processing Systems*, pp. 6241–6250.

Bartlett, P. L., P. M. Long, G. Lugosi, and A. Tsigler (2019). *Benign overfitting in linear regression.* arXiv: `1906.11300`.

Belkin, M., D. Hsu, S. Ma, and S. Mandal (2019). "Reconciling modern machine-learning practice and the classical bias–variance trade-off". *Proceedings of the National Academy of Sciences* 116.32, pp. 15849–15854.

Geiger, M., S. Spigler, S. d'Ascoli, L. Sagun, M. Baity-Jesi, G. Biroli, and M. Wyart (2019). "Jamming transition as a paradigm to understand the loss landscape of deep neural networks". *Physical Review E* 100.1, p. 012115.

Golowich, N., A. Rakhlin, and O. Shamir (June 2018). "Size-Independent Sample Complexity of Neural Networks". In: *Proceedings of the 31st Conference On Learning Theory*. Ed. by S. Bubeck, V. Perchet, and P. Rigollet. Vol. 75. Proceedings of Machine Learning Research. PMLR, pp. 297–299.

Hastie, T., A. Montanari, S. Rosset, and R. J. Tibshirani (2019). *Surprises in high-dimensional ridgeless least squares interpolation*. arXiv: 1903.08560.

Kuriki, S. and Y. Numata (2010). "On formulas for moments of the Wishart distributions as weighted generating functions of matchings". *Discrete Mathematics & Theoretical Computer Science*.

Long, P. M. and H. Sedghi (2020). "Size-free generalization bounds for convolutional neural networks". In: *International Conference on Learning Representations*.

Mei, S. and A. Montanari (2019). *The generalization error of random features regression: Precise asymptotics and double descent curve*. arXiv: 1908.05355.

Muirhead, R. J. (1982). *Aspects of multivariate statistical theory*. Vol. 197. John Wiley & Sons.

Nagarajan, V. and J. Z. Kolter (2019). "Uniform convergence may be unable to explain generalization in deep learning". In: *Advances in Neural Information Processing Systems (33)*. arXiv: 1902.04742.

Neyshabur, B., S. Bhojanapalli, D. McAllester, and N. Srebro (2017). *A PAC-Bayesian approach to spectrally-normalized margin bounds for neural networks*. arXiv: 1707.09564.

Shalev-Shwartz, S. and S. Ben-David (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.

Wei, C. and T. Ma (2019). "Data-dependent sample complexity of deep neural networks via lipschitz augmentation". In: *Advances in Neural Information Processing Systems*, pp. 9722–9733.

Zhang, C., S. Bengio, M. Hardt, B. Recht, and O. Vinyals (2017). "Understanding deep learning requires rethinking generalization". In: *International Conference on Representation Learning (ICLR)*. arXiv: 1611.03530v2 [cs.LG].

A. Proofs for hypercube classifier

Proof of Lemma 7.3.

$$L_{\mathcal{D}^{(d)}}(\hat{f}_d^{(S)}) \leq \frac{1}{2^{2d}} \sum_{x \in \{0,1\}^{2d}} \mathbf{1}_{x \in \bar{S}_X} \leq n_d 2^{-2d}.$$

$\square$

Proof of Theorem 7.4. We require the following claims and lemmas.

Claim A.1. $S \overset{d}{=} \bar{S}$.

Lemma A.1. *Let E be the event that $\exists i$, $\exists j$, $X_i = 1 - X_j$. Then $\mathbb{P}E \leq \binom{n}{2} 2^{-2d}$, so $\mathbb{P}E \to 0$ as $d \to \infty$ as long as $n_d \in o(2^d)$.*

Proof.

$$\mathbb{P}E = \mathbb{P}(\exists i, \ \exists j, \ X_i = 1 - X_j) \leq \sum_{i \neq j} \mathbb{P}(X_i = 1 - X_j) = \binom{n_d}{2} 2^{-2d}$$

$\square$

Lemma A.2. $L_{\bar{S}}(\hat{f}_d) = 1$ *on E^c.*

Proof.

$$L_{\bar{S}}(\hat{f}_d) = \frac{1}{n_d} \sum_{i=1}^{n_d} \mathbf{1}_{\bar{X}_i \in \bar{S}_X} \mathbf{1}_{\bar{X}_i \notin S_X} = \frac{1}{n_d} \sum_{i=1}^{n_d} \mathbf{1}_{\bar{X}_i \notin S_X} \geq \mathbf{1}_{E^c} \frac{1}{n_d} \sum_{i=1}^{n_d} 1 = \mathbf{1}_{E^c}$$

$\square$

Now, returning to the proof of Theorem 7.4

$$\mathbb{E} \sup_{h \in \mathcal{H}} |L_{\mathcal{D}}(h) - L_S(h)| \geq \mathbb{E} \left| L_{\mathcal{D}}(\hat{f}_d^{(\bar{S})}) - L_S(\hat{f}_d^{(\bar{S})}) \right| \geq \mathbb{E} L_S(\hat{f}_d^{(\bar{S})}) - \mathbb{E} L_{\mathcal{D}}(\hat{f}_d^{(\bar{S})})$$

$$\geq 1 - \left(\binom{n_d}{2} + n_d \right) 2^{-2d} = 1 - \binom{n_d + 1}{2} 2^{-2d}$$

$\square$

Proof of Lemma 7.5. For each $v \in \{0,1\}^{k_d}$ let $g_{k_d}(v, x) = (v_1, \ldots v_{k_d}, x_{k_d+1}, \ldots x_{2d})$, and let $G_{k_d}(V, S_X) = (g_{k_d}(v_1, X_1), \ldots, g_{k_d}(v_{k_d}, X_{n_d}))$

Starting with $L_S(Q)$.

$$L_S(Q) = \frac{1}{n} \sum_{i=1}^{n} (2^{-k})^n \sum_{V \in (\{0,1\}^k)^n} \mathbf{1}_{X_i \in \overline{G_k(V, S_X)}} \mathbf{1}_{X_i \notin G_k(V, S_X)}$$

$$\leq \frac{2^{-nk}}{n} \sum_{i=1}^{n} \sum_{V \in \{0,1\}^{nk}} \mathbf{1}_{X_i \in \overline{G_k(V, S_X)}} \mathbf{1}_{X_i[1:k] \neq v_i}.$$

Now, X_i is in $\overline{G_k(V, S_X)}$ only if $X_i[1:k]$ is antipodal to at least one v, so

$$\frac{2^{-nk}}{n} \sum_{i=1}^{n} \sum_{V \in \{0,1\}^{nk}} \mathbf{1}_{X_i \in \overline{G_k(V,S_X)}} \mathbf{1}_{X_i[1:k] \neq v_i} \leq \frac{2^{-nk}}{n} \sum_{i=1}^{n} \sum_{V \in \{0,1\}^{nk}} \sum_{j=1}^{n} \mathbf{1}_{X_i[1:k]=\overline{v_j} \neq v_i} \mathbf{1}_{X_i[k+1:2d]=\overline{X_j[k+1:2d]}}$$

$$= \frac{2^{-nk}}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{X_i[k+1:2d]=\overline{X_j[k+1:2d]}} \sum_{V \in \{0,1\}^{nk}} \mathbf{1}_{X_i[1:k]=\overline{v_j} \neq v_i}.$$

The inner most summand is constant in all but the ith and jth v, so

$$\frac{2^{-nk}}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{X_i[k+1:2d]=\overline{X_j[k+1:2d]}} \sum_{V \in \{0,1\}^{nk}} \mathbf{1}_{X_i[1:k]=\overline{v_j} \neq v_i}$$

$$\leq \frac{2^{-nk}}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{X_i[k+1:2d]=\overline{X_j[k+1:2d]}} \sum_{(v_i,v_j) \in \{0,1\}^{2k}} 2^{k(n-2)} \mathbf{1}_{X_i[1:k]=\overline{v_j} \neq v_i}$$

$$= \frac{2^{-2k}}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{X_i[k+1:2d]=\overline{X_j[k+1:2d]}} \sum_{(v_i,v_j) \in \{0,1\}^{2k}} \mathbf{1}_{X_i[1:k]=\overline{v_j} \neq v_i}$$

$$= \frac{2^{-2k}}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{X_i[k+1:2d]=\overline{X_j[k+1:2d]}} \sum_{v_j \in \{0,1\}^{k}} (2^k - 1) \mathbf{1}_{X_i[1:k]=\overline{v_j}}$$

$$= \frac{2^{-2k}}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} (2^k - 1) \mathbf{1}_{X_i[k+1:2d]=\overline{X_j[k+1:2d]}}$$

$$= \frac{2^{-k}(1 - 2^{-k})}{n} \left| \left\{ (i,j) \in [n]^2 : X_i[k+1:2d] = \overline{X_j[k+1:2d]} \right\} \right|.$$

Thus

$$L_S(Q) \leq \frac{2^{-k}(1 - 2^{-k})}{n} \left| \left\{ (i,j) \in [n]^2 : X_i[k+1:2d] = \overline{X_j[k+1:2d]} \right\} \right|.$$

Taking expectations, we get

$$\mathbb{E}L_S(Q) \leq \frac{2^{-k}(1 - 2^{-k})}{n} n(n-1) 2^{-(2d-k)} = (n-1)2^{-2d}(1 - 2^{-k}).$$

Next, Looking at $L_{\overline{S}}(Q)$,

$$L_{\overline{S}}(Q) = \frac{1}{n} \sum_{i=1}^{n} (2^{-k})^n \sum_{V \in (\{0,1\}^k)^n} \mathbf{1}_{\overline{X}_i \in \overline{G_k(V,S_X)}} \mathbf{1}_{\overline{X}_i \notin G_k(V,S_X)}$$

$$\leq \frac{2^{-nk}}{n} \sum_{i=1}^{n} \sum_{V \in \{0,1\}^{nk}} \mathbf{1}_{\overline{X}_i \in \overline{G_k(V,S_X)}}.$$

Now, $\overline{X}_i$ is in $\overline{G_k(V, S_X)}$ only if $\overline{X_i[1:k]}$ is antipodal to at least one v, so

$$\frac{2^{-nk}}{n} \sum_{i=1}^{n} \sum_{V \in \{0,1\}^{nk}} \mathbf{1}_{\overline{X}_i \in \overline{G_k(V,S_X)}} \leq \frac{2^{-nk}}{n} \sum_{i=1}^{n} \sum_{V \in \{0,1\}^{nk}} \sum_{j=1}^{n} \mathbf{1}_{\overline{X_i[1:k]}=\overline{v_j}} \mathbf{1}_{X_i[k+1:2d]=X_j[k+1:2d]}$$

$$= \frac{2^{-nk}}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{\overline{X_i[k+1:2d]=X_j[k+1:2d]}} \sum_{V \in \{0,1\}^{nk}} \mathbf{1}_{\overline{X_i[1:k]=\overline{v_j}}}$$

$$= \frac{2^{-nk}}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{X_i[k+1:2d]=X_j[k+1:2d]} \sum_{V \in \{0,1\}^{nk}} \mathbf{1}_{X_i[1:k]=v_j}$$

.

The inner most summand is constant in all but the jth v, so

$$\frac{2^{-nk}}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{X_i[k+1:2d]=X_j[k+1:2d]} \sum_{V \in \{0,1\}^{nk}} \mathbf{1}_{X_i[1:k]=v_j}$$

$$\leq \frac{2^{-nk}}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{X_i[k+1:2d]=X_j[k+1:2d]} \sum_{v_j \in \{0,1\}^{k}} 2^{k(n-1)} \mathbf{1}_{X_i[1:k]=v_j}$$

$$= \frac{2^{-k}}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{X_i[k+1:2d]=X_j[k+1:2d]} \sum_{v_j \in \{0,1\}^{k}} \mathbf{1}_{X_i[1:k]=v_j}$$

$$= \frac{2^{-k}}{n} \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbf{1}_{X_i[k+1:2d]=X_j[k+1:2d]}$$

$$= \frac{2^{-k}}{n} \left| \left\{ (i,j) \in [n]^2 : X_i[k+1:2d] = X_j[k+1:2d] \right\} \right|.$$

Thus

$$L_{\overline{S}}(Q) \leq \frac{2^{-k}}{n} \left| \left\{ (i,j) \in [n]^2 : X_i[k+1:2d] = X_j[k+1:2d] \right\} \right|.$$

Taking expectations, we get

$$\mathbb{E} L_{\overline{S}(Q)} \leq \frac{2^{-k}}{n} (n + n(n-1)2^{-(2d-k)}) = 2^{-k} + (n-1)2^{-2d}.$$

Lastly, for $L_{\mathcal{D}}(Q)$,

$$L_{\mathcal{D}}(Q) = \mathbb{E} L_{\mathcal{D}}(\hat{f}_d) \leq n 2^{-2d}$$

$\square$

Proof of Lemma 7.6. First, using a standard symmetrization argument, we bound the supremum over the empirical process of Curryed losses by twice the Rademacher complexity,

$$\mathbb{E} \left[\sup_{S'} |L_{\mathcal{D}}(Q(S')) - L_S(Q(S'))| \right] \leq 2 \, R(A)$$

Where

$$A = \left\{ (\ell'(Z_i, Q(\hat{f}_d^{(S')})))_{i \in [n_d]} : S' \in (\mathcal{S}^{(d)})^{n_d}, \ell' \in \{\ell, -\ell\} \right\}.$$

and $R(A)$ denotes the Rademacher complexity of A. Then, by Massart's Lemma,

$$R(A) \leq \max_{a \in A} \|a\| \frac{\sqrt{\log(2\,|A|)}}{n_d}$$

$$\leq \max_{a \in A} \|a\| \frac{\sqrt{((2d - k_d)n_d + 1)\log(2)}}{n_d}.$$

Next, by the same arguments as in the proof of Lemma 7.5, $\ell'(Z_i, Q(\hat{f}_d^{(S')})) \leq n_d 2^{-k_d}$ for all i and for all datasets S'. Therefore, for all $a \in A$, $\|a\| \leq n_d^{1/2} n_d 2^{-k_d} = n_d^{3/2} 2^{-k_d}$.

Hence $R(A) \leq \sqrt{\log(2)} n_d^{1/2}((2d - k_d)n_d + 1)^{1/2} 2^{-k_d}$. □

Proof of Theorem 7.7. From Lemmas 6.2 and 7.5 and Proposition 4.2,

$$\mathbb{E}[L_{\mathcal{D}}(\hat{f}_d) - L_S(\hat{f}_d)]$$

$$\leq \mathbb{E}\big[\big|L_S(Q) - L_S(\hat{f}_d)\big|\big]$$

$$+ \mathbb{P}[Q \notin \mathcal{I}^{(d)}] + \mathbb{E}\bigg[\sup_{P \in \mathcal{I}^{(d)}} L_{\mathcal{D}}(P) - L_S(P)\bigg]$$

$$\leq (n_d - 1)2^{-2d}(1 - 2^{-k_d}) + 0 + 0$$

$$+ \mathbb{E}\bigg[\sup_{P \in \mathcal{I}^{(d)}} L_{\mathcal{D}}(P) - L_S(P)\bigg].$$

The last term is controlled using Lemma 7.6 to get

$$\mathbb{E}[L_{\mathcal{D}}(\hat{f}_d) - L_S(\hat{f}_d)]$$

$$\leq (n_d - 1)2^{-2d}(1 - 2^{-k_d})$$

(1)
$$+ 2\sqrt{\log(2)} n_d^{1/2}((2d - k_d)n_d + 1)^{1/2} 2^{-k_d}$$

$$\leq (n_d - 1)2^{-2d}$$

$$+ 2\sqrt{\log(2)} n_d^{1/2}((2d - k_d)n_d + 1)^{1/2} 2^{-k_d}.$$

□

B. Proofs for Overparameterized Linear Regression

Proof of Lemma 5.2. Let $\phi((x, y)) = (x, 2x\beta - y)$. Then (x, y) and $(\phi(x, y))$ are equally probable, and ϕ is its own inverse function. Hence ϕ is measure preserving. Moreover $L_S(\hat{\beta}(\phi(X, Y))) = \frac{4}{n}\|Z\|^2$. Hence for any sequence of sets $A_n \subset \mathcal{S}^n$

$$\mathbb{E} \sup_{(\tilde{X}, \tilde{Y}) \in A_n} \big|L_D(\hat{\beta}(\tilde{X}, \tilde{Y})) - L_S(\hat{\beta}(\tilde{X}, \tilde{Y}))\big|$$

$$\geq \mathbb{E}\mathbf{1}_{\phi(X,Y) \in A_n} \big|L_D(\hat{\beta}(\phi(X, Y))) - L_S(\hat{\beta}(\phi(X, Y)))\big|$$

$$\geq \mathbb{E}\mathbf{1}_{\phi(X,Y) \in A} \max\bigg(0, 4\frac{\|Z\|^2}{n} - L_D(\hat{\beta}(\phi(X, Y)))\bigg)$$

We can couple the spaces for different values of n in any way we choose since no terms in the statement involve multiple probability spaces at once. Hence we can do so in a way that $\mathbb{P}(\{\forall n \in \mathbb{N} : \phi(X, Y) \in A_n\}) \geq 2/3$.

Using Bartlett et al. (2019, Theorem 4) and the weak law of large numbers, we have that $L_D(\hat\beta(\phi(X,Y))) \xrightarrow{\text{P}} \sigma^2$ and $4\frac{\|Z\|^2}{n} \xrightarrow{\text{P}} 4\sigma^2$ when Σ_n is benign. Then there is a subsequence along which this convergence is almost sure. Along this subsequence,

$$(2) \qquad \max\left(0, 4\frac{\|Z\|^2}{n_k} - L_D(\hat\beta(\phi(X,Y)))\right) \xrightarrow{\text{a.s.}} 3\sigma^2.$$

Then by Fatou's lemma along the subsequence

$$\liminf_{k\to\infty} \mathbb{E} \sup_{(\tilde X,\tilde Y)\in A_{n_k}} \left|L_D(\hat\beta(\tilde X,\tilde Y)) - L_S(\hat\beta(\tilde X,\tilde Y))\right|$$
$$\geq \mathbb{E}\liminf_{k\to\infty} \mathbf{1}_{\phi(X,Y)\in A}3\sigma^2$$
$$\geq 2\sigma^2$$

Thus there is a sub-subsequence above $2\sigma^2 - \epsilon$ infinitely often for each $\epsilon > 0$, and hence

$$\limsup_{n\to\infty} \mathbb{E} \sup_{(\tilde X,\tilde Y)\in A_n} \left|L_D(\hat\beta(\tilde X,\tilde Y)) - L_S(\hat\beta(\tilde X,\tilde Y))\right| \geq 2\sigma^2.$$

$\square$

Proof of Lemma 5.3. Let $\hat\beta_0(X_0) = (X_0'X_0)^+X_0'X_0\beta = P(X_0)\beta$. This corresponds to the classifier that solves the learning problem without label noise if the training design matrix was X_0. Let $\hat\beta_0 = \hat\beta_0(X)$ (if no argument is specified then it is the "learned" version).

Let the projection onto the row-span of a matrix A be given by $P(A) = (A'A)^+A'A = A'A(A'A)^+ = A'(AA')^+A$

If $d > n$ then (a.s.) X is of rank n so $P(X)$ is a rank n projection on $\mathbb{R}^d$ and $P(X')$ is a rank n projection on $\mathbb{R}^n$ — so $P(X') = I$ a.s. .

The surrogate decomposition (not in expectation) gives use

$$(3) \qquad L_D(\hat\beta) - L_S(\hat\beta) = L_S(\hat\beta_0) - L_S(\hat\beta) + (L_D(\hat\beta) - L_D(\hat\beta_0)) + (L_D(\hat\beta_0) - L_S(\hat\beta_0)).$$

Then,

$$L_S(\hat\beta_0) = \frac{1}{n}\left\|X(X'X)^+X'X\beta - (X\beta + Z)\right\|^2$$
$$(4) \qquad\qquad = \frac{1}{n}\left\|-Z + (P(X') - I)X\beta\right\|^2$$
$$= \frac{1}{n}\|Z\|^2.$$

Next, for any X_0,

(5)
$$L_D(\hat{\beta}_0(X_0)) - L_S(\hat{\beta}_0(X_0)) = \mathbb{E}_{x,z} \|xP(X_0)\beta - (x\beta + z)\|^2 - \frac{1}{n} \|XP(X_0)\beta - (X\beta + Z)\|^2$$
$$= \mathbb{E}_{x,z} \|xP(X)\beta - (x\beta + z)\|^2 - \frac{1}{n} \|XP(X_0)\beta - (X\beta + Z)\|^2$$
$$= \mathbb{E}_{x,z} \|xP(X_0)\beta - x\beta\|^2 + \sigma^2$$
$$\quad - \frac{1}{n} \|XP(X_0)\beta - X\beta\|^2 + \frac{2}{n} Z'(X(I - P(X_0))\beta) - \frac{1}{n} \|Z\|^2$$
$$= \sigma^2 - \frac{1}{n} \|Z\|^2 + \frac{2}{n} Z'(X(I - P(X_0))\beta)$$
$$\quad + \mathbb{E}_{x,z} \|x[P(X_0) - I]\beta\|^2 - \frac{1}{n} \|X[P(X_0) - I]\beta\|^2$$
$$= \sigma^2 - \frac{1}{n} \|Z\|^2 + \frac{2}{n} Z'(X(I - P(X_0))\beta)$$
$$\quad + \beta' \left[[P(X_0) - I] \left[\Sigma - \frac{1}{n} X'X \right] [P(X_0) - I] \right] \beta.$$

Since $XP(X) = X$, then when $X_0 = X$, this simplifies to

(6)
$$L_D(\hat{\beta}_0(X)) - L_S(\hat{\beta}_0(X)) = \sigma^2 - \frac{1}{n} \|Z\|^2 +$$
$$\quad + \beta' [[P(X) - I]\Sigma[P(X) - I]] \beta.$$

Lastly,

(7)
$$L_D(\hat{\beta}) - L_D(\hat{\beta}_0) = \mathbb{E}_{x,z} \left(x(X'X)^+ X'(X\beta + Z) - (x\beta + z) \right)^2 - \mathbb{E}_{x,z} \left(x(X'X)^+ X'X\beta - (x\beta + z) \right)^2$$
$$= \mathbb{E}_x \left(x(X'X)^+ X'(X\beta + Z) - x\beta \right)^2 - \mathbb{E}_x \left(x(X'X)^+ X'X\beta - x\beta \right)^2$$
$$= \mathbb{E}_x \left(x(X'X)^+ X'Z \right)^2$$
$$= \mathbb{E}_x (Z'X(X'X)^+ x'x(X'X)^+ X'Z)$$
$$= Z'X(X'X)^+ \Sigma (X'X)^+ X'Z$$
$$= \mathrm{Tr}(X(X'X)^+ \Sigma (X'X)^+ X'ZZ')$$

$\square$

Proof of Lemma 5.4. We want to bound

(8)
$$\mathbb{E} \sup_{X_0} \left| L_D(\hat{\beta}_0(X_0)) - L_S(\hat{\beta}_0(X_0)) \right|$$
$$= \mathbb{E} \sup_{X_0} \left| \sigma^2 - \frac{1}{n} \|Z\|^2 + \frac{2}{n} Z'(X(I - P(X_0))\beta) + \beta' \left[[P(X_0) - I] \left[\Sigma - \frac{1}{n} X'X \right] [P(X_0) - I] \right] \beta \right|$$
$$\leq \mathbb{E} \left| \sigma^2 - \frac{1}{n} \|Z\|^2 \right| + \mathbb{E} \sup_{X_0} \left| \frac{2}{n} Z'(X(I - P(X_0))\beta) \right| + \mathbb{E} \sup_{X_0} \left| \beta' \left[[P(X_0) - I] \left[\Sigma - \frac{1}{n} X'X \right] [P(X_0) - I] \right] \beta \right|$$

We will handle each of the three terms separately.

First

$$\mathbb{E}\left|\sigma^2 - \frac{1}{n}\|Z\|^2\right| \leq \sqrt{\mathbb{E}\left[\sigma^2 - \frac{1}{n}\|Z\|^2\right]^2}$$

(9)
$$= \sqrt{\mathbb{E}\left[\sigma^4 - \frac{2\sigma^2}{n}\|Z\|^2 + \frac{1}{n^2}\sum_{i\in[n]}\sum_{j\in[n]}Z_i^2 Z_j^2\right]}$$

$$= \sqrt{\sigma^4 - 2\sigma^4 + \frac{n(n-1)}{n^2}\sigma^4 + 3\frac{n}{n^2}\sigma^4}$$

$$= \sigma^2\sqrt{\frac{2}{n}}.$$

Next, for some universal constants $C_2, C_3 > 0$, the second term is bounded by $C_2\frac{\sigma\sqrt{\beta'\Sigma\beta}}{\sqrt{n}}$ in Lemma B.3 and the third term is bounded by $C_3\frac{\beta'\Sigma\beta}{\sqrt{n}}$ in Lemma B.1.

Putting these all together yields the desired result. $\square$

Lemma B.1 (Third Term). *For some universal constant $C_3 > 0$*
(10)
$$\mathbb{E}\sup_{X_0\in\mathbb{R}^{n\times d}}\left|\beta'[P(X_0) - I]\left[\Sigma - \frac{1}{n}X'X\right][P(X_0) - I]\beta\right| = \mathbb{E}\sup_{P\in\mathbb{OP}(d,d-n)}\left|\beta'P\left[\Sigma - \frac{1}{n}X'X\right]P\beta\right| \leq C_3\frac{\beta'\Sigma\beta}{\sqrt{n}}$$

where $\mathbb{OP}(d, k)$ is the collection of orthogonal projections on $\mathbb{R}^d$ of rank k.

The proofs of Lemmas B.1 and B.3 in this section control the supremum of the respective empirical processes in a bespoke way using the fact that we can identify a particular subexponential random variable on the same space which upper bounds the supremum almost surely. The ability to do this is somewhat specific to the present problem. In Appendix C we provide alternative proofs of the same results using chaining, a standard tool in empirical process theory. This illustrates that underlying principal letting us bound the empirical process is not problem specific. Of course, the chaining approach of Appendix C yields slightly worse constants than the more direct approach in this section.

We will use the following trivial lemma a number of times.

Lemma B.2 (Rank-1 Trick). *If $A \in \mathbb{R}^{m\times m}$ with $\text{rank}(A) = 1$ then*

(11)
$$\|A\| = |\lambda_1(A)| = |\text{Tr}(A)|$$

where $\lambda_1(A)$ is the dominant eigenvalue of A.

Proof of Lemma B.1. Note that $\mathrm{rank}(\beta\beta') = 1$ so then, for any conforming matrix A, $\mathrm{rank}(A\beta\beta') = 1 = \mathrm{rank}(\beta\beta'A')$

$$
\begin{aligned}
(12) \qquad \sup_{P\in\mathbb{OP}(d,d-n)} \left| \beta'P\left[\Sigma - \frac{1}{n}X'X\right]P\beta \right| &= \sup_{P\in\mathbb{OP}(d,d-n)} \left| \mathrm{Tr}\left(\beta'P\left[\Sigma - \frac{1}{n}X'X\right]P\beta\right) \right| \\
&= \sup_{P\in\mathbb{OP}(d,d-n)} \left| \mathrm{Tr}\left(P\left[\Sigma - \frac{1}{n}X'X\right]P\beta\beta'\right) \right| \\
&= \sup_{P\in\mathbb{OP}(d,d-n)} \left\| P\left[\Sigma - \frac{1}{n}X'X\right]P\beta\beta' \right\| \\
&\leq \sup_{P\in\mathbb{OP}(d,d-n)} \|P\| \left\| \left[\Sigma - \frac{1}{n}X'X\right]P\beta\beta' \right\| \\
&= \sup_{P\in\mathbb{OP}(d,d-n)} \left\| \left[\Sigma - \frac{1}{n}X'X\right]P\beta\beta' \right\| \\
&= \sup_{P\in\mathbb{OP}(d,d-n)} \left| \mathrm{Tr}\left(\left[\Sigma - \frac{1}{n}X'X\right]P\beta\beta'\right) \right| \\
&= \sup_{P\in\mathbb{OP}(d,d-n)} \left| \mathrm{Tr}\left(\beta\beta'\left[\Sigma - \frac{1}{n}X'X\right]P\right) \right| \\
&= \sup_{P\in\mathbb{OP}(d,d-n)} \left\| \beta\beta'\left[\Sigma - \frac{1}{n}X'X\right]P \right\| \\
&\leq \sup_{P\in\mathbb{OP}(d,d-n)} \left\| \beta\beta'\left[\Sigma - \frac{1}{n}X'X\right] \right\| \|P\| \\
&= \sup_{P\in\mathbb{OP}(d,d-n)} \left| \mathrm{Tr}\left(\beta\beta'\left[\Sigma - \frac{1}{n}X'X\right]\right) \right| \\
&= \left| \beta'\left[\Sigma - \frac{1}{n}X'X\right]\beta \right| \\
&= \left| \frac{1}{n}\sum_{i=1}^{n}\left[X_{i,\bullet}\beta\beta'X'_{i,\bullet} - \beta'\Sigma\beta\right] \right|
\end{aligned}
$$

where $X_{i,\bullet}$ is the ith row of X, so $X'_{i,\bullet} \overset{\mathrm{iid}}{\sim} \mathrm{N}(0,\Sigma)$. This pointwise inequality can be used derive bounds on the expected supremum as well as high probability bounds on the supremum. It also holds in the non-Gaussian case and can be used to extend the result to the case of sub-Gaussian design and sub-Gaussian conditional responses.

Let $W_i = \beta' X'_{i,\bullet} / \sqrt{\beta'\Sigma\beta}$. Then $W_i \overset{\text{iid}}{\sim} \text{N}(0,1)$. Now,

$$
\mathbb{E} \sup_{P \in \mathbb{OP}(d,d-n)} \left| \beta' P \left[\Sigma - \frac{1}{n} X'X \right] P\beta \right| \leq \beta'\Sigma\beta \mathbb{E} \left| \frac{1}{n} \sum_{i=1}^{n} \left[W_i^2 - 1 \right] \right|
$$

$$
(13) \qquad\qquad\qquad\qquad \leq \beta'\Sigma\beta \sqrt{\mathbb{E} \left(\frac{1}{n} \sum_{i=1}^{n} \left[W_i^2 - 1 \right] \right)^2}
$$

$$
\leq \beta'\Sigma\beta \sqrt{\frac{2}{n}}
$$

And, since $\frac{1}{n}\sum_{i=1}^{n} \left[W_i^2 - 1 \right]$ is sub-Exponential with mean 0, (proxy-variance, moment-generating function range) given by $(2/\sqrt{n}, 1/4)$, then for $\frac{\epsilon}{\beta'\Sigma\beta} \in (0,1)$

$$
\mathbb{P} \left(\sup_{P \in \mathbb{OP}(d,d-n)} \left| \beta' P \left[\Sigma - \frac{1}{n} X'X \right] P\beta \right| > \epsilon \right)
$$

$$
(14) \qquad = \mathbb{P} \left(\left| \frac{1}{n} \sum_{i=1}^{n} \left[W_i^2 - 1 \right] \right| > \frac{\epsilon}{\beta'\Sigma\beta} \right)
$$

$$
\leq 2 \exp \left(-n \frac{\epsilon^2}{8(\beta'\Sigma\beta)^2} \right)
$$

For $\delta \geq 2e^{-n/8}$, taking $\epsilon = \beta'\Sigma\beta \sqrt{\frac{8\log(2/\delta)}{n}}$ and using a standard sub-exponential tail bound we get

$$
(15) \qquad \mathbb{P} \left(\sup_{P \in \mathbb{OP}(d,d-n)} \left| \beta' P \left[\Sigma - \frac{1}{n} X'X \right] P\beta \right| > \beta'\Sigma\beta \sqrt{\frac{8\log(2/\delta)}{n}} \right) \leq \delta
$$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\square$

Lemma B.3 (Second Term). *For some universal constant $C_2 > 0$*

$$
(16) \qquad \mathbb{E} \sup_{X_0 \in \mathbb{R}^{n \times d}} \left| \frac{2}{n} Z'XP(X_0)\beta \right| = \mathbb{E} \sup_{P \in \mathbb{OP}(d,d-n)} \left| \frac{2}{n} Z'XP\beta \right| \leq C_2 \frac{\sigma\sqrt{\beta'\Sigma\beta}}{\sqrt{n}}
$$

where $\mathbb{OP}(d,k)$ is the collection of orthogonal projections on $\mathbb{R}^d$ of rank k.

Proof of Lemma B.3. Note that $\text{rank}(\beta Z') = 1$ so then, for any conforming matrix A, $\text{rank}(A\beta Z') = 1 = \text{rank}(\beta Z'A')$

$$
\begin{aligned}
(17) \qquad \sup_{P \in \mathbb{OP}(d,d-n)} \left| \frac{2}{n} Z'XP\beta \right| &= \frac{2}{n} \sup_{P \in \mathbb{OP}(d,d-n)} \left| \text{Tr}(Z'XP\beta) \right| \\
&= \frac{2}{n} \sup_{P \in \mathbb{OP}(d,d-n)} \left| \text{Tr}(\beta Z'XP) \right| \\
&= \frac{2}{n} \sup_{P \in \mathbb{OP}(d,d-n)} \left\| \beta Z'XP \right\| \\
&\leq \frac{2}{n} \sup_{P \in \mathbb{OP}(d,d-n)} \left\| \beta Z'X \right\| \left\| P \right\| \\
&= \frac{2}{n} \sup_{P \in \mathbb{OP}(d,d-n)} \left\| \beta Z'X \right\| \\
&= \frac{2}{n} \left\| Z'X\beta \right\| \\
&= \frac{2}{n} \left| Z'X\beta \right| \\
&= \frac{2}{n} \left| \sum_{i=1}^{n} Z_i X_{i,\bullet}\beta \right|
\end{aligned}
$$

where $X_{i,\bullet}$ is the ith row of X, so $X'_{i,\bullet} \overset{\text{iid}}{\sim} N(0,\Sigma)$. This pointwise inequality can be used derive bounds on the expected supremum as well as high probability bounds on the supremum. It also holds in the non-Gaussian case and can be used to extend the result to the case of sub-Gaussian design and sub-Gaussian conditional responses.

Let $W_i = \beta'X'_{i,\bullet}/\sqrt{\beta'\Sigma\beta}$. Then $W_i \overset{\text{iid}}{\sim} N(0,1)$. Now,

$$
\begin{aligned}
(18) \qquad \mathbb{E} \sup_{P \in \mathbb{OP}(d,d-n)} \left| \frac{2}{n} Z'XP\beta \right| &\leq \frac{2\sqrt{\beta'\Sigma\beta}}{n} \left| \sum_{i=1}^{n} Z_i W_i \right| \\
&\leq \frac{2\sqrt{\beta'\Sigma\beta}}{n} \sqrt{\mathbb{E}\left(\sum_{i=1}^{n} Z_i W_i \right)^2} \\
&\leq \frac{2\sqrt{\beta'\Sigma\beta}}{n} \sqrt{n\sigma^2} \\
&\leq \frac{2\sigma\sqrt{\beta'\Sigma\beta}}{\sqrt{n}}
\end{aligned}
$$

Moreover, the moment generating function of $Z_1 W_1$ is given by

$$
\begin{aligned}
(19) \qquad \mathbb{E}\exp(\lambda Z_1 W_1) &= \frac{1}{\sqrt{1 - \lambda^2\sigma^2}} \\
&= \exp\left(-\frac{1}{2}\log(1 - \lambda^2\sigma^2) \right)
\end{aligned}
$$

In particular, for $|\lambda| < \frac{1}{\sigma\sqrt{2}}$, using $\log(1-x) \geq \frac{-x}{1-x}$ for $x \in (0,1]$

$$
\begin{aligned}
\mathbb{E}\exp(\lambda Z_1 W_1) &= \exp\left(\frac{\lambda^2\sigma^2}{2(1-\lambda^2\sigma^2)}\right) \\
&\leq \exp\left(\lambda^2\sigma^2\right)
\end{aligned}
\tag{20}
$$

Thus, $Z_1 W_1$ is sub-Exponential with mean 0, (proxy-variance, moment-generating function range) given by $(2\sigma^2, \frac{1}{\sigma\sqrt{2}})$. Then we get that $\frac{1}{n}\sum_{i=1}^{n} Z_i W_i$ is sub-Exponential with mean 0, (proxy-variance, moment-generating function range) given by $(2\sigma^2/\sqrt{n}, \frac{1}{\sigma\sqrt{2}})$. Hence, for $\frac{\epsilon}{\sqrt{\beta'\Sigma\beta}} \in (0, \sigma\sqrt{2})$

$$
\begin{aligned}
\mathbb{P}&\left(\sup_{P\in\mathbb{OP}(d,d-n)}\left|\frac{2}{n}Z'XP\beta\right| > \epsilon\right) \\
&= \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}Z_i W_i\right| > \frac{\epsilon}{2\sqrt{\beta'\Sigma\beta}}\right) \\
&\leq 2\exp\left(-n\frac{\epsilon^2}{4\sigma^2(\beta'\Sigma\beta)^2}\right)
\end{aligned}
\tag{21}
$$

For $\delta \geq 2e^{-n/8}$, taking $\epsilon = \sigma\sqrt{\beta'\Sigma\beta}\sqrt{\frac{4\log(2/\delta)}{n}}$ and using a standard sub-exponential tail bound we get

$$
\mathbb{P}\left(\sup_{P\in\mathbb{OP}(d,d-n)}\left|\beta'P\left[\Sigma - \frac{1}{n}X'X\right]P\beta\right| > \sigma\sqrt{\beta'\Sigma\beta}\sqrt{\frac{4\log(2/\delta)}{n}}\right). \leq \delta
\tag{22}
$$

$\square$

Proof of Theorem 5.5. We only need bounds on $\mathbb{E}(L_S(\hat{\beta}_0) - L_S(\hat{\beta}))$ and $\mathbb{E}(L_D(\hat{\beta}) - L_D(\hat{\beta}_0))$ to combine with Lemma 5.3 and Lemma 5.4.

First,

$$
\mathbb{E}(L_S(\hat{\beta}_0) - L_S(\hat{\beta})) = \mathbb{E}\frac{\|Z\|^2}{n} = \sigma^2
\tag{23}
$$

Second, Bartlett et al. (2019) shows that there is are universal constant, $c, b > 0$, such that for all $\delta < 1$, with probability at least $(1-\delta)$,

$$
\begin{aligned}
L_D(\hat{\beta}) - L_D(\hat{\beta}_0) &= Z'X(X'X)^+\Sigma(X'X)^+X'Z \\
&\leq c\sigma^2\log(1/\delta)\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right),
\end{aligned}
\tag{24}
$$

where $k_* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$, $r_k(\Sigma) = \frac{\sum_{i>k}\lambda_i(\Sigma)}{\lambda_{k+1}(\Sigma)}$ and $R_k(\Sigma) = \frac{\left(\sum_{i>k}\lambda_i(\Sigma)\right)^2}{\sum_{i>k}\lambda_i^2(\Sigma)}$.

We can turn this into a bound in expectation by integrating the tail.

$$
\begin{aligned}
\frac{\mathbb{E}L_D(\hat{\beta}) - L_D(\hat{\beta}_0)}{c\sigma^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)} &= \int_0^\infty \mathbb{P}\left(\frac{L_D(\hat{\beta}) - L_D(\hat{\beta}_0)}{c\sigma^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right)} > t\right) dt \\
&= \int_0^\infty e^{-t} dt \\
&= 1.
\end{aligned}
$$
(25)

Thus, for the same universal constants, c, b,

$$
\mathbb{E}L_D(\hat{\beta}) - L_D(\hat{\beta}_0) \le c\sigma^2\left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)}\right).
$$
(26)

$\square$

C. Alternative proofs of Lemmas B.1 and B.3 using empirical process theory

Alternative proof of Lemma B.1. **Bounding the MGF of the increment**

Notice that $X'X \sim \text{Wishart}_d(\Sigma, n)$ follows a degenerate Wishart distribution.

Consider the empirical process $W_P = \beta'P\left[\frac{1}{n}X'X - \Sigma\right]P\beta$ with $P \in \mathbb{OP}(d, d-n)$

We start by computing the moment generating function of the increments of the process:
(27)
$$
\begin{aligned}
\mathbb{E}\exp\left(\lambda(W_P - W_R)\right) &= \mathbb{E}\exp\left(\lambda\left(\beta'P\left[\frac{1}{n}X'X - \Sigma\right]P\beta - \beta'R\left[\frac{1}{n}X'X - \Sigma\right]R\beta\right)\right) \\
&= \mathbb{E}\exp\left(\lambda\left(\beta'(P-R)\left[\frac{1}{n}X'X - \Sigma\right](P+R)\beta\right)\right) \\
&= \exp(-\lambda\beta'(P-R)\Sigma(P+R)\beta))\mathbb{E}\exp\left(\lambda\left(\beta'(P-R)\left[\frac{1}{n}X'X\right](P+R)\beta\right)\right) \\
&= \exp\left(-\text{Tr}\left(\lambda(P+R)\beta\beta'(P-R)\,\Sigma\right)\right)\mathbb{E}\exp\left(\text{Tr}\left(\frac{\lambda}{n}(P+R)\beta\beta'(P-R)\,X'X\right)\right)
\end{aligned}
$$

Using the MGF formula for Wishart distributions from Kuriki and Numata (2010) (see also Muirhead (1982, Theorem 3.2.3)),
(28)
$$
\begin{aligned}
\mathbb{E}\exp\left(\lambda(W_P - W_R)\right) &= \exp\left(-\text{Tr}\left(\lambda(P+R)\beta\beta'(P-R)\,\Sigma\right)\right)\det\left(I - 2\frac{\lambda}{n}(P+R)\beta\beta'(P-R)\,\Sigma\right)^{-\frac{n}{2}} \\
&= g_0(\lambda(P+R)\beta\beta'(P-R)\,\Sigma)
\end{aligned}
$$

where,

$$
g_0(A) = e^{-\text{Tr}(A)}\det\left(I - \frac{2}{n}A\right)^{-\frac{n}{2}}
$$
(29)

If $\text{rank}(A) = 1$ and the non-zero eigenvalue of A is a then (since $\left(1 - \frac{2}{n}a\right)^{n/2} \leq e^{-a}$)

$$
\begin{aligned}
g_0(A) &= e^{-a}\left(1 - \frac{2}{n}a\right)^{-n/2} \\
&\leq e^{-\|A\|}\left(1 - \frac{2}{n}\|A\|\right)^{-n/2}
\end{aligned}
$$
(30)

Now, using $\log(1 - x) \geq \frac{-x}{1-x}$ for $x \in [0, 1)$, if $\|A\| < n/2$ then

$$
\begin{aligned}
g_0(A) &\leq e^{-\|A\|}\left(1 - \frac{2}{n}\|A\|\right)^{-n/2} \\
&= \exp\left(\left[-\|A\| - \frac{n}{2}\log\left(1 - \frac{2}{n}\|A\|\right)\right]\right) \\
&\leq \exp\left(\left[-\|A\| + \frac{\|A\|}{1 - \frac{2}{n}\|A\|}\right]\right) \\
&= \exp\left(\left[\frac{\frac{2}{n}\|A\|^2}{1 - \frac{2}{n}\|A\|}\right]\right)
\end{aligned}
$$
(31)

In particular, if $\text{rank}(A) = 1$ *and* $\|A\| < \frac{n}{4}$ then

(32) $$g_0(A) \leq \exp\left(\frac{4\|A\|^2}{n}\right)$$

Taking $A = \lambda(P + R)\beta\beta'(P - R)\Sigma$ again, then $\text{rank}(A) = 1$ since $\text{rank}(\beta\beta') = 1$ and $\|A\| \leq 2\lambda\|\beta\beta'(P - R)\Sigma\|$.

Hence, for $\lambda < \frac{n}{8\|\beta\beta'(P-R)\Sigma\|}$

(33) $$\mathbb{E}\exp\left(\lambda(W_P - W_R)\right) \leq \exp\left(\frac{16\lambda^2\|\beta\beta'(P - R)\Sigma\|^2}{n}\right).$$

Let $\rho(P, R) = \|\beta\beta'(P - R)\Sigma\|$. Then ρ is a pseudo-metric on $\mathbb{OP}(d, d - n)$ and when $\lambda < \frac{n}{8\rho(P,R)}$

(34) $$\mathbb{E}\exp\left(\lambda(W_P - W_R)\right) \leq \exp\left(\lambda^2\frac{16\rho(P, R)^2}{n}\right)$$

Covering the space in our pseudometric

Recall again that for a square matrix A with $\text{rank}(A) = 1$ and non-zero eigenvalue a that

(35) $$\|A\| = |a| = |\text{Tr}(A)| = \sqrt{\text{Tr}(A'A)} = \|A\|_F$$

Thus $\rho(P, R) = |\text{Tr}(\beta\beta'(P - R)\Sigma)| = |\beta'(P - R)\Sigma\beta|$.

Let $g_1 : P \mapsto \beta'P\Sigma\beta$. Then $G = g_1(\mathbb{OP}(d, d - n)) \subset I = [-\beta'\Sigma\beta, \beta'\Sigma\beta]$.

For each $\epsilon > 0$ let $R(\epsilon/2)$ be an $\epsilon/2$ cover of I in the usual metric on $\mathbb{R}$. Notice that $R(\epsilon/2)$ may be chosen so that $|R(\epsilon/2)| < \left\lceil\frac{4\beta'\Sigma\beta}{\epsilon}\right\rceil$.

For each $r \in R(\epsilon)$ such that $B_{\epsilon/2} \cap G \neq \emptyset$ select a single $P_r(\epsilon)$ in $\mathbb{OP}(d, d-n)$ such that $g_1(P_r(\epsilon)) \in B_{\epsilon/2}$. Let $\mathcal{O}(\epsilon) = \{P_r(\epsilon) : r \in R(\epsilon)\}$. Then $g_1(\mathcal{O}(\epsilon))$ is an ϵ-net for G in the usual metric on $\mathbb{R}$.

Then $\mathcal{O}(\epsilon)$ is an an ϵ-net for $\mathbb{OP}(d, d-n)$ in ρ with $|\mathcal{O}(\epsilon)| \leq \left\lceil \frac{4\beta'\Sigma\beta}{\epsilon} \right\rceil$.

Hence $N(\mathbb{OP}(d, d-n), \rho, \epsilon) \leq \left\lceil \frac{4\beta'\Sigma\beta}{\epsilon} \right\rceil$

Obtaining a bound via chaining

Let P_0 be any element of $\mathbb{OP}(d, d-n)$ with $P_0\beta = 0$. Then $W_{P_0} = 0$ a.s..

The diameter of $\mathbb{OP}(d, d-n)$ in ρ is at most that $2\beta'\Sigma\beta$. Without loss of generality take the coarsest cover (at scale $\epsilon_0 = 2\beta'\Sigma\beta$) to be $\mathcal{O}(2\beta'\Sigma\beta) = \{P_0\}$.

Let $\pi_A : P \mapsto \arg\min_{R \in A} \rho(P, R)$ with ties broken arbitrarily. Note that $\rho(P, \pi_{\mathcal{O}(2\epsilon)}(P)) \leq 2\epsilon$ for all P, ϵ, and therefore (for $\lambda < \frac{n}{16\epsilon}$)

$$(36) \qquad \mathbb{E} \max_{P \in \mathcal{O}(\epsilon)} \left| W_P - W_{\pi_{\mathcal{O}(2\epsilon)}(P)} \right| \leq \frac{1}{\lambda} \log 2 |\mathcal{O}(\epsilon)| + \lambda \frac{64\epsilon^2}{n}$$

Taking $\lambda \to \min\left(\sqrt{\frac{n \log 2|\mathcal{O}(\epsilon)|}{64\epsilon^2}}, \frac{n}{16\epsilon} \right)$, we get

$$(37) \qquad \mathbb{E} \max_{P \in \mathcal{O}(\epsilon)} \left| W_P - W_{\pi_{\mathcal{O}(2\epsilon)}(P)} \right| \leq \begin{cases} 16\epsilon\sqrt{\frac{\log 2|\mathcal{O}(\epsilon)|}{n}} & : 4 \log 2 |\mathcal{O}(\epsilon)| \leq n \\ 16\epsilon\frac{\log 2|\mathcal{O}(\epsilon)|}{n} + 4\epsilon & : 4 \log 2 |\mathcal{O}(\epsilon)| > n \end{cases}$$

Then for $n > 4\log 2$

(38)

$$\mathbb{E}\sup_{P\in\mathbb{OP}(d,d-n)}|W_P| \leq \mathbb{E}|W_{P_0}| + \sum_{i=1}^{\infty}\mathbb{E}\max_{P\in\mathcal{O}(2^{-i}\epsilon_0)}\left|W_P - W_{\pi_{\mathcal{O}(2^{-(i-1)}\epsilon_0)}}(P)\right|$$

$$\leq \sum_{i=1}^{\infty}\begin{cases}16\epsilon_0 2^{-i}\sqrt{\dfrac{\log 2|\mathcal{O}(\epsilon_0 2^{-i})|}{n}} & : 4\log 2\left|\mathcal{O}(\epsilon_0 2^{-i})\right| \leq n \\[2ex] 16\epsilon_0 2^{-i}\dfrac{\log 2|\mathcal{O}(\epsilon_0 2^{-i})|}{n} + 4\epsilon_0 2^{-i} & : 4\log 2\left|\mathcal{O}(\epsilon_0 2^{-i})\right| > n\end{cases}$$

$$\leq \sum_{i=1}^{\infty}\begin{cases}16\epsilon_0 2^{-i}\sqrt{\dfrac{\log\left(1+\frac{4\beta'\Sigma\beta}{\epsilon_0 2^{-i}}\right)+\log 2}{n}} & : 4\log\left(1+\frac{4\beta'\Sigma\beta}{\epsilon_0 2^{-i}}\right)+4\log 2 \leq n \\[3ex] 16\epsilon_0 2^{-i}\dfrac{\log\left(1+\frac{4\beta'\Sigma\beta}{\epsilon_0 2^{-i}}\right)+\log 2}{n} + 4\epsilon_0 2^{-i} & : 4\log\left(1+\frac{4\beta'\Sigma\beta}{\epsilon_0 2^{-i}}\right)+4\log 2 > n\end{cases}$$

$$\leq 2\int_0^{\epsilon_0/2}\mathbf{1}_{\left[\epsilon\geq\frac{4\beta'\Sigma\beta}{e^{n/4-\log 2}-1}\right]}\left(\frac{16}{\sqrt{n}}\sqrt{\log\left(1+\frac{4\beta'\Sigma\beta}{\epsilon}\right)+\log 2}\right)$$

$$+ \mathbf{1}_{\left[\epsilon<\frac{4\beta'\Sigma\beta}{e^{n/4-\log 2}-1}\right]}\left(\frac{16}{n}\log\left(1+\frac{4\beta'\Sigma\beta}{\epsilon}\right)+\frac{16\log 2}{n}+4\right)d\epsilon$$

$$\leq 2\int_0^{\epsilon_0/2}\mathbf{1}_{\left[\epsilon\geq\frac{2\epsilon_0}{e^{n/4-\log 2}-1}\right]}\left(\frac{16}{\sqrt{n}}\sqrt{\log\left(1+\frac{2\epsilon_0}{\epsilon}\right)+\log 2}\right)$$

$$+ \mathbf{1}_{\left[\epsilon<\frac{2\epsilon_0}{e^{n/4-\log 2}-1}\right]}\left(\frac{16}{n}\log\left(1+\frac{2\epsilon_0}{\epsilon}\right)+\frac{16\log 2}{n}+4\right)d\epsilon$$

If $n \leq 9 < 4\log(10)$ then $\frac{2\epsilon_0}{\frac{1}{2}e^{n/4}-1} \geq \epsilon_0/2$ so only one term appears. We omit this case for brevity, though a quantitative bound may be derived for this case with the same steps as below.

If $n \geq 10 \geq 4\log(10)$ then $\frac{2\epsilon_0}{\frac{1}{2}e^{n/4}-1} \leq \epsilon_0/2$ so both terms appear...

(39)
$$\mathbb{E}\sup_{P\in\mathbb{OP}(d,d-n)}W_P \leq 2\int_{\frac{2\epsilon_0}{e^{n/4-\log 2}-1}}^{\epsilon_0/2}\left(\frac{16}{\sqrt{n}}\sqrt{\log\left(1+\frac{2\epsilon_0}{\epsilon}\right)+\log 2}\right)d\epsilon$$

$$+ 2\int_0^{\frac{2\epsilon_0}{e^{n/4-\log 2}-1}}\left(\frac{16}{n}\log\left(1+\frac{2\epsilon_0}{\epsilon}\right)+\frac{16\log 2}{n}+4\right)d\epsilon$$

The second integral is a bit easier so we will solve it first...

(40)
$$2\int_0^{\frac{2\epsilon_0}{e^{n/4-\log 2}-1}} \left(\frac{16}{n}\log\left(1+\frac{2\epsilon_0}{\epsilon}\right) + \frac{16\log 2}{n} + 4\right) d\epsilon$$

$$= \frac{16\epsilon_0}{\frac{1}{2}e^{n/4}-1}\left(\frac{4\log 2}{n}+1\right) + \frac{32}{n}\int_0^{\frac{2\epsilon_0}{\frac{1}{2}e^{n/4}-1}}\log\left(1+\frac{2\epsilon_0}{\epsilon}\right)d\epsilon$$

$$= \frac{16\epsilon_0}{\frac{1}{2}e^{n/4}-1}\left(\frac{4\log 2}{n}+1\right)$$

$$+ \frac{32}{n}\left[\left(\frac{2\epsilon_0}{\frac{1}{2}e^{n/4}-1}+2\epsilon_0\right)\log\left(\frac{2\epsilon_0}{\frac{1}{2}e^{n/4}-1}+2\epsilon_0\right) - \left(\frac{2\epsilon_0}{\frac{1}{2}e^{n/4}-1}\right)\log\left(\frac{2\epsilon_0}{\frac{1}{2}e^{n/4}-1}\right) - (2\epsilon_0)\log(2\epsilon_0)\right]$$

$$= \frac{16\epsilon_0}{\frac{1}{2}e^{n/4}-1}\left(\frac{4\log 2}{n}+1\right) + \frac{32}{n}\left[\left(\frac{n}{4}-\log 2\right)\frac{2\epsilon_0}{\frac{1}{2}e^{n/4}-1} + 2\epsilon_0\log\left(1+\frac{1}{\frac{1}{2}e^{n/4}-1}\right)\right]$$

$$\le \frac{16\epsilon_0}{\frac{1}{2}e^{n/4}-1}\left(\frac{4\log 2}{n}+1\right) + \frac{32}{n}\left[\left(\frac{n}{4}-\log 2\right)\frac{2\epsilon_0}{\frac{1}{2}e^{n/4}-1} + 2\epsilon_0\frac{1}{\frac{1}{2}e^{n/4}-1}\right]$$

$$\le \frac{32\epsilon_0}{\frac{1}{2}e^{n/16}-1}\left(1+\frac{2}{n}\right)$$

Then, for the first integral, using Cauchy-Schwarz...

(41)
$$2\int_{\frac{2\epsilon_0}{\frac{1}{2}e^{n/4}-1}}^{\epsilon_0/2}\left(\frac{16}{\sqrt{n}}\sqrt{\log\left(1+\frac{2\epsilon_0}{\epsilon}\right)+\log 2}\right)d\epsilon$$

$$\le \frac{32}{\sqrt{n}}\int_0^{\epsilon_0/2}\sqrt{\log\left(1+\frac{2\epsilon_0}{\epsilon}\right)+\log 2}\,d\epsilon$$

$$\le \frac{32}{\sqrt{n}}\sqrt{\frac{\epsilon_0}{2}}\sqrt{\int_0^{\epsilon_0/2}\log\left(1+\frac{2\epsilon_0}{\epsilon}\right)+\log 2\,d\epsilon}$$

$$= 32\sqrt{\frac{\epsilon_0}{2n}}\sqrt{\frac{5\epsilon_0}{2}\log\frac{5\epsilon_0}{2} - \frac{\epsilon_0}{2}\log\frac{\epsilon_0}{2} - 2\epsilon_0\log(2\epsilon_0) + \frac{\epsilon_0}{2}\log 2}$$

$$= 32\sqrt{\frac{\epsilon_0}{2n}}\sqrt{\frac{\epsilon_0}{2}\log 5 + 2\epsilon_0\log\frac{5}{4} + \frac{\epsilon_0}{2}\log 2}$$

$$= \frac{32}{\sqrt{n}}\sqrt{\frac{\epsilon_0}{2}}\sqrt{\frac{5\epsilon_0}{2}\log 5 - \frac{7}{2}\epsilon_0\log 2}$$

$$= \frac{16\epsilon_0}{\sqrt{n}}\sqrt{5\log 5 - 7\log 2}$$

Putting these two parts together with $\epsilon_0 = 2\beta'\Sigma\beta$, for $n \geq 37...$

$$
\begin{aligned}
\mathbb{E} \sup_{P \in \mathbb{OP}(d,d-n)} W_P &= 32\beta'\Sigma\beta \left(\frac{\sqrt{5\log 5 - 7\log 2}}{\sqrt{n}} + \frac{1}{\frac{1}{2}e^{n/4} - 1}\left(1 + \frac{2}{n}\right) \right) \\
&\leq \beta'\Sigma\beta \left(\frac{58}{\sqrt{n}} + \frac{32}{\frac{1}{2}e^{n/4} - 1}\left(1 + \frac{2}{n}\right) \right)
\end{aligned}
\tag{42}
$$

$\square$

Alternative proof of Lemma B.3. **Bounding MGF of increments** Consider the empirical process $V_P = \frac{2}{n}Z'XP\beta$ with $P \in \mathbb{OP}(d, d-n)$

$$
\begin{aligned}
\mathbb{E}\exp\left(\lambda(V_P - V_R)\right) &= \mathbb{E}\exp\left(\frac{2\lambda}{n}Z'X(P-R)\beta\right) \\
&= \mathbb{E}\exp\left(\frac{4\lambda^2\sigma^2}{n^2}\beta'(P-R)X'X(P-R)\beta\right) \\
&= \mathbb{E}\exp\left(\mathrm{Tr}\left(\frac{4\lambda^2\sigma^2}{n^2}(P-R)\beta\beta'(P-R)\ X'X\right)\right) \\
&= \det\left(I - \frac{8\lambda^2\sigma^2}{n^2}(P-R)\beta\beta'(P-R)\Sigma\right)^{-n/2}
\end{aligned}
\tag{43}
$$

Since $\beta\beta'$ is rank-1, then $(P-R)\beta\beta'(P-R)\Sigma$ has rank at most 1, so

$$
\left\|(P-R)\beta\beta'(P-R)\Sigma\right\| = \beta'(P-R)\Sigma(P-R)\beta = \left\|(P-R)\beta\beta'(P-R)\Sigma\right\|_F
\tag{44}
$$

and hence

$$
\begin{aligned}
\mathbb{E}\exp\left(\lambda(V_P - V_R)\right) &\leq \left(1 - \frac{8\lambda^2\sigma^2}{n^2}\beta'(P-R)\Sigma(P-R)\beta\right)^{-n/2} \\
&\leq \exp\left(-\frac{n}{2}\log\left(1 - \frac{8\lambda^2\sigma^2}{n^2}\beta'(P-R)\Sigma(P-R)\beta\right)\right) \\
&\leq \exp\left(\frac{n}{2}\frac{\frac{8\lambda^2\sigma^2}{n^2}\beta'(P-R)\Sigma(P-R)\beta}{1 - \frac{8\lambda^2\sigma^2}{n^2}\beta'(P-R)\Sigma(P-R)\beta}\right)
\end{aligned}
\tag{45}
$$

For $\frac{8\lambda^2\sigma^2}{n^2}\beta'(P-R)\Sigma(P-R)\beta < \frac{1}{2}$, we have

$$
\begin{aligned}
\mathbb{E}\exp\left(\lambda(V_P - V_R)\right) &\leq \exp\left(2\frac{n}{2}\frac{8\lambda^2\sigma^2}{n^2}\beta'(P-R)\Sigma(P-R)\beta\right) \\
&\leq \exp\left(\frac{8\lambda^2\sigma^2}{n}\beta'(P-R)\Sigma(P-R)\beta\right)
\end{aligned}
\tag{46}
$$

Note that using the rank-1 trick again, and defining $\rho(P, R) = \|\Sigma(P - R)\beta\beta'\|$

$$
\begin{aligned}
\beta'(P - R)\Sigma(P - R)\beta &= \left\|(P - R)\Sigma(P - R)\beta\beta'\right\| \\
&= \frac{1}{\beta'\Sigma\beta}\left\|(P - R)\Sigma(P - R)\beta\beta'\Sigma\beta\beta'\right\| \\
&= \frac{1}{\beta'\Sigma\beta}\left|\mathrm{Tr}\left((P - R)\Sigma(P - R)\beta\beta'\Sigma\beta\beta'\right)\right| \\
&= \frac{1}{\beta'\Sigma\beta}\left|\mathrm{Tr}\left(\beta\beta'(P - R)\Sigma(P - R)\beta\beta'\Sigma\right)\right| \\
&= \frac{1}{\beta'\Sigma\beta}\left\|\beta\beta'(P - R)\Sigma(P - R)\beta\beta'\Sigma\right\| \\
&\leq \frac{1}{\beta'\Sigma\beta}\left\|\beta\beta'(P - R)\Sigma\right\|\left\|(P - R)\beta\beta'\Sigma\right\| \\
&= \frac{1}{\beta'\Sigma\beta}\rho(P, R)\left|\mathrm{Tr}\left((P - R)\beta\beta'\Sigma\right)\right| \\
&= \frac{1}{\beta'\Sigma\beta}\rho(P, R)\left|\mathrm{Tr}\left(\Sigma(P - R)\beta\beta'\right)\right| \\
&= \frac{1}{\beta'\Sigma\beta}\rho(P, R)^2
\end{aligned}
$$

(47)

Therefore, for $\lambda^2 \leq \frac{n^2\beta'\Sigma\beta}{8\sigma^2\rho(P,R)^2}$,

(48)
$$
\mathbb{E}\exp\left(\lambda(V_P - V_R)\right) \leq \exp\left(\lambda^2\frac{8\sigma^2\rho(P, R)^2}{n\beta'\Sigma\beta}\right)
$$

Notice that we can use the same covering of $\mathbb{OP}(d, d - n)$ in ρ as in the proof of Lemma B.1, to cover $\mathbb{OP}(d, d - n)$ in ρ at scale ϵ with $|\mathcal{O}(\epsilon)| \leq \left\lceil\frac{4\beta'\Sigma\beta}{\epsilon}\right\rceil$.

Obtaining a bound via chaining Let P_0 be any element of $\mathbb{OP}(d, d - n)$ with $P_0\beta = 0$. Then $V_{P_0} = 0$ a.s..

The diameter of $\mathbb{OP}(d, d - n)$ in ρ is at most that $2\beta'\Sigma\beta$. Without loss of generality take the coarsest cover (at scale $\epsilon_0 = 2\beta'\Sigma\beta$) to be $\mathcal{O}(2\beta'\Sigma\beta) = \{P_0\}$

Let $\pi_A : P \mapsto \arg\min_{R \in A} \rho(P, R)$ with ties broken arbitrarily. Note that $\rho(P, \pi_{\mathcal{O}(2\epsilon)}(P)) \leq 2\epsilon$ for all P, ϵ, and therefore (for $\lambda^2 < \frac{n^2\beta'\Sigma\beta}{32\sigma^2\epsilon^2}$)

(49)
$$
\max_{P \in \mathcal{O}(\epsilon)}\left|V_P - V_{\pi_{\mathcal{O}(2\epsilon)}(P)}\right| \leq \frac{1}{\lambda}\log 2\,|\mathcal{O}(\epsilon)| + \lambda\frac{32\sigma^2\epsilon^2}{n\beta'\Sigma\beta}.
$$

Taking $\lambda \to \min\left(\sqrt{\frac{n^2\beta'\Sigma\beta}{32\,\sigma^2\epsilon^2}}, \sqrt{\frac{n\beta'\Sigma\beta\log(2|\mathcal{O}(\epsilon)|)}{32\sigma^2\epsilon^2}}\right)$, we get

(50)
$$
\max_{P \in \mathcal{O}(\epsilon)}\left|V_P - V_{\pi_{\mathcal{O}(2\epsilon)}(P)}\right| \leq
\begin{cases}
8\sqrt{2}\frac{\sigma\epsilon}{\sqrt{\beta'\Sigma\beta}}\sqrt{\frac{\log(2|\mathcal{O}(\epsilon)|)}{n}} & : n \geq \log(2\,|\mathcal{O}(\epsilon)|) \\
4\sqrt{2}\frac{\sigma\epsilon}{\sqrt{\beta'\Sigma\beta}}\left[\frac{\log(2|\mathcal{O}(\epsilon)|)}{n} + 1\right] & : n < \log(2\,|\mathcal{O}(\epsilon)|)
\end{cases}
$$

Then

(51)

$$\mathbb{E}\sup_{P\in\mathbb{OP}(d,d-n)}V_P \le \mathbb{E}\,|V_{P_0}| + \sum_{i=1}^{\infty}\mathbb{E}\max_{P\in\mathcal{O}(2^{-i}\epsilon_0)}\left|V_P - V_{\pi_{\mathcal{O}(2^{-(i-1)}\epsilon_0)}(P)}\right|$$

$$\le \sum_{i=1}^{\infty}\begin{cases}8\sqrt{2}\frac{\sigma\epsilon_0 2^{-i}}{\sqrt{\beta'\Sigma\beta}}\sqrt{\frac{\log(2|\mathcal{O}(\epsilon_0 2^{-i})|)}{n}} & : n \ge \log(2\,|\mathcal{O}(\epsilon)|)\\[3mm] 4\sqrt{2}\frac{\sigma\epsilon_0 2^{-i}}{\sqrt{\beta'\Sigma\beta}}\left[\frac{\log(2|\mathcal{O}(\epsilon_0 2^{-i})|)}{n}+1\right] & : n < \log(2\,|\mathcal{O}(\epsilon)|)\end{cases}$$

$$\le \sum_{i=1}^{\infty}\begin{cases}8\sqrt{2}\frac{\sigma\epsilon_0 2^{-i}}{\sqrt{\beta'\Sigma\beta}}\sqrt{\frac{\log\left(1+\frac{4\beta'\Sigma\beta}{\epsilon_0 2^{-i}}\right)+\log 2}{n}} & : n \ge \log\left(1+\frac{4\beta'\Sigma\beta}{\epsilon_0 2^{-i}}\right)+\log 2\\[3mm] 4\sqrt{2}\frac{\sigma\epsilon_0 2^{-i}}{\sqrt{\beta'\Sigma\beta}}\left[\frac{\log\left(1+\frac{4\beta'\Sigma\beta}{\epsilon_0 2^{-i}}\right)}{n}+\frac{\log 2}{n}+1\right] & : n < \log\left(1+\frac{4\beta'\Sigma\beta}{\epsilon_0 2^{-i}}\right)+\log 2\end{cases}$$

$$\le 2\int_0^{\epsilon_0/2}\mathbf{1}_{\left[\epsilon\ge\frac{4\beta'\Sigma\beta}{\frac{1}{2}e^n-1}\right]}\left(8\sqrt{2}\frac{\sigma}{\sqrt{\beta'\Sigma\beta}}\sqrt{\frac{\log\left(1+\frac{4\beta'\Sigma\beta}{\epsilon}\right)+\log 2}{n}}\right)$$

$$+\mathbf{1}_{\left[\epsilon<\frac{4\beta'\Sigma\beta}{\frac{1}{2}e^n-1}\right]}\left(4\sqrt{2}\frac{\sigma}{\sqrt{\beta'\Sigma\beta}}\left[\frac{\log\left(1+\frac{4\beta'\Sigma\beta}{\epsilon}\right)}{n}+\frac{\log 2}{n}+1\right]\right)d\epsilon$$

$$\le \frac{16\sigma}{\sqrt{\epsilon_0}}\int_0^{\epsilon_0/2}\mathbf{1}_{\left[\epsilon\ge\frac{2\epsilon_0}{\frac{1}{2}e^n-1}\right]}2\left(\sqrt{\frac{\log\left(1+\frac{2\epsilon_0}{\epsilon}\right)+\log 2}{n}}\right)$$

$$+\mathbf{1}_{\left[\epsilon<\frac{2\epsilon_0}{\frac{1}{2}e^n-1}\right]}\left(\frac{\log\left(1+\frac{2\epsilon_0}{\epsilon}\right)}{n}+\frac{\log 2}{n}+1\right)d\epsilon$$

If $n \le 2 < \log(10)$ then $\frac{2\epsilon_0}{\frac{1}{2}e^n-1}\ge\epsilon_0/2$ so only one term appears. We omit this case for brevity, though a quantitative bound may be derived for this case with the same steps as below.

If $n \ge 3 > \log(10)$ then $\frac{2\epsilon_0}{\frac{1}{2}e^n-1}\le\epsilon_0/2$ so both terms appear...

(52)

$$\mathbb{E}\sup_{P\in\mathbb{OP}(d,d-n)}V_P \le \frac{32\sigma}{\sqrt{\epsilon_0}}\int_{\frac{2\epsilon_0}{\frac{1}{2}e^n-1}}^{\epsilon_0/2}\sqrt{\frac{1}{n}\log\left(1+\frac{2\epsilon_0}{\epsilon}\right)+\log 2}\,d\epsilon$$

$$+\frac{16\sigma}{\sqrt{\epsilon_0}}\int_0^{\frac{2\epsilon_0}{\frac{1}{2}e^n-1}}\left(\frac{1}{n}\log\left(1+\frac{2\epsilon_0}{\epsilon}\right)+\frac{\log 2}{n}+1\right)d\epsilon$$

The second integral is a bit easier so we will solve it first...

(53)
$$\frac{16\sigma}{\sqrt{\epsilon_0}} \int_0^{\frac{2\epsilon_0}{\frac{1}{2}e^n-1}} \left(\frac{1}{n}\log\left(1+\frac{2\epsilon_0}{\epsilon}\right) + \frac{\log 2}{n} + 1\right) d\epsilon$$

$$= \frac{16\sigma}{\sqrt{\epsilon_0}} \frac{2\epsilon_0}{\frac{1}{2}e^n-1}\left(1+\frac{\log 2}{n}\right) + \frac{16\sigma}{n\sqrt{\epsilon_0}} \int_0^{\frac{2\epsilon_0}{\frac{1}{2}e^n-1}} \log\left(1+\frac{2\epsilon_0}{\epsilon}\right) d\epsilon$$

$$= \frac{32\sigma\sqrt{\epsilon_0}}{\frac{1}{2}e^n-1}\left(1+\frac{\log 2}{n}\right) + \frac{16\sigma}{n\sqrt{\epsilon_0}}\left[\left(\frac{2\epsilon_0}{\frac{1}{2}e^n-1}+2\epsilon_0\right)\log\left(\frac{2\epsilon_0}{\frac{1}{2}e^n-1}+2\epsilon_0\right) - 2\epsilon_0\log(2\epsilon_0) - \frac{2\epsilon_0}{\frac{1}{2}e^n-1}\log\frac{2\epsilon_0}{\frac{1}{2}e^n-1}\right]$$

$$= \frac{32\sigma\sqrt{\epsilon_0}}{\frac{1}{2}e^n-1}\left(1+\frac{\log 2}{n}\right) + \frac{16\sigma}{n\sqrt{\epsilon_0}}\left[2\epsilon_0\log\left(1+\frac{1}{\frac{1}{2}e^n-1}\right) + (n-\log 2)\frac{2\epsilon_0}{\frac{1}{2}e^n-1}\right]$$

$$\leq \frac{32\sigma\sqrt{\epsilon_0}}{\frac{1}{2}e^n-1}\left(1+\frac{\log 2}{n}\right) + \frac{16\sigma}{n\sqrt{\epsilon_0}}\left[2\epsilon_0\frac{1}{\frac{1}{2}e^n-1} + (n-\log 2)\frac{2\epsilon_0}{\frac{1}{2}e^n-1}\right]$$

$$\leq \frac{32\sigma\sqrt{\epsilon_0}}{\frac{1}{2}e^n-1}\left(2+\frac{1}{n}\right)$$

Then, for the first integral, using Cauchy-Schwarz

(54)
$$\frac{32\sigma}{\sqrt{\epsilon_0}} \int_{\frac{2\epsilon_0}{\frac{1}{2}e^n-1}}^{\epsilon_0/2} \sqrt{\frac{1}{n}\log\left(1+\frac{2\epsilon_0}{\epsilon}\right) + \log 2}\, d\epsilon$$

$$\leq \frac{32\sigma}{\sqrt{n\epsilon_0}} \int_0^{\epsilon_0/2} \sqrt{\log\left(1+\frac{2\epsilon_0}{\epsilon}\right) + \log 2}\, d\epsilon$$

$$\leq \frac{32\sigma}{\sqrt{n\epsilon_0}} \sqrt{\frac{\epsilon_0}{2}} \sqrt{\int_0^{\epsilon_0/2} \log\left(1+\frac{2\epsilon_0}{\epsilon}\right) + \log 2\, d\epsilon}$$

$$\leq \frac{32\sigma}{\sqrt{n\epsilon_0}} \sqrt{\frac{\epsilon_0}{2}} \sqrt{\frac{\epsilon_0}{2}\left(5\log 5 - 7\log 2\right)}$$

$$\leq \frac{16\sigma\sqrt{\epsilon_0}}{\sqrt{n}} \sqrt{5\log 5 - 7\log 2}$$

Putting these two parts together with $\epsilon_0 = 2\beta'\Sigma\beta$, for $n \geq 3$

(55)
$$\mathbb{E} \sup_{P\in\mathbb{OP}(d,d-n)} V_P = \frac{16\sigma\sqrt{2\beta'\Sigma\beta}}{\sqrt{n}}\sqrt{5\log 5 - 7\log 2} + \frac{32\sigma\sqrt{2\beta'\Sigma\beta}}{\frac{1}{2}e^n-1}\left(2+\frac{1}{n}\right)$$

$$\leq \sigma\sqrt{\beta'\Sigma\beta}\left(\frac{41}{\sqrt{n}} + \frac{46}{\frac{1}{2}e^n-1}\left(2+\frac{1}{n}\right)\right)$$

$\square$