MULTIPLE DESCENT: DESIGN YOUR OWN GENERALIZATION CURVE

Anonymous authors
Paper under double-blind review

ABSTRACT

This paper explores the generalization loss of linear regression in variably parameterized families of models, both under-parameterized and over-parameterized. We show that the generalization curve can have an arbitrary number of peaks, and moreover, locations of those peaks can be explicitly controlled. Our results highlight the fact that both classical U-shaped generalization curve and the recently observed double descent curve are not intrinsic properties of the model family. Instead, their emergence is due to the interaction between the properties of the data and the inductive biases of learning algorithms.

1 INTRODUCTION

The main goal of machine learning methods is to provide an accurate out-of-sample prediction, known as generalization. For a fixed family of models, a common way to select a model from this family is through empirical risk minimization, i.e., algorithmically selecting models that minimize the risk on the training dataset. Given a variably parameterized family of models, the statistical learning theory aims to identify the dependence between model complexity and model performance. The empirical risk usually decreases monotonically as the model complexity increases, and achieves its minimum when the model is rich enough to interpolate the training data, resulting in zero (or near-zero) training error. In contrast, the behaviour of the test error as a function of model complexity is far more complicated. Indeed, in this paper we show how to construct a model family for which the generalization curve can be fully controlled (away from the interpolation threshold) in both under-parameterized and over-parameterized regimes. Classical statistical learning theory supports a U-shaped curve of generalization versus model complexity (Geman et al., 1992; Hastie et al., 2009). Under such a framework, the best model is found at the bottom of the U-shaped curve, which corresponds to appropriately balancing under-fitting and over-fitting the training data. From the view of the bias-variance trade-off, a higher model complexity increases the variance while decreasing the bias. A good choice of model complexity achieves a relatively low bias while still keeping the variance under control. On the other hand, a model that interpolates the training data is deemed to over-fit and tends to worsen the generalization performance due to the soaring variance.

Although classical statistical theory suggests a pattern of behavior for the generalization curve up to the interpolation threshold, it does not describe what happens beyond the interpolation threshold, commonly referred to as the over-parameterized regime. This is the exact regime where many modern machine learning models, especially deep neural networks, achieved remarkable success. Indeed, neural networks generalize well even when the models are so complex that they have the potential to interpolate all the training data points (Zhang et al., 2017; Belkin et al., 2018b; Ghorbani et al., 2019; Hastie et al., 2019).

Modern practitioners commonly deploy deep neural networks with hundreds of millions or even billions of parameters. It has become widely accepted that large models achieve performance superior to small models that may be suggested by the classical U-shaped generalization curve (Bengio et al., 2003; Krizhevsky et al., 2012; Szegedy et al., 2015; He et al., 2016; Huang et al., 2019). This indicates that the test error decreases again once model complexity grows beyond the interpolation threshold, resulting in the so called double-descent phenomenon described in (Belkin et al., 2018a), which has been broadly supported by empirical evidence (Neyshabur et al., 2015; Neal et al., 2018; Geiger et al., 2019; 2020). On the theoretical side, this phenomenon has been recently addressed by several works on various model settings. In particular, Belkin et al. (2019a) proved the existence

of double-descent phenomenon for linear regression with random feature selection and analyzed the random Fourier feature model (Rahimi & Recht, 2008). Mei & Montanari (2019) also studied the Fourier model and computed the asymptotic test error which captures the double-descent phenomenon. Bartlett et al. (2020) analyzed and gave explicit conditions for "benign overfitting" in linear regression. In a recent work, Caron & Chretien (2020) provided a finite sample analysis of the nonlinear function estimation and showed that the parameter learned through empirical risk minimization converges to the true parameter with high probability as the model complexity tends to infinity, implying the existence of double descent.

Among all the aforementioned efforts, one particularly interesting question is whether one can observe more than two descents in the generalization curve. In a recent work, d'Ascoli et al. (2020) empirically showed a sample-wise triple-descent phenomenon under the random Fourier feature model. Similar triple-descent was also observed for linear regression (Nakkiran et al., 2020). More rigorously, Liang et al. (2020) presented an upper bound on the risk of the minimum-norm interpolation versus the data dimension in Reproducing Kernel Hilbert Spaces (RKHS), which exhibits multiple descent. However, a multiple-descent upper bound without a properly matching lower bound does not imply the existence of a multiple-descent generalization curve. In this work, we study the multiple descent phenomenon by addressing the following questions:

- Can the existence of a multiple descent generalization curve be rigorously proven?
- Can an arbitrary number of descents occur?
- Can the generalization curve and the locations of descents be designed?

In this paper, we show that the answer to all three of these questions is yes. Further related work is presented in Appendix A.

Our Contribution. We consider the linear regression model and analyze how the risk changes as the dimension of the data grows. In the linear regression setting, the data dimension is equal to the dimension of the parameter space, which reflects the model complexity. Therefore, increasing the data dimension is essentially equivalent to increasing the model complexity. We rigorously show that the multiple descent generalization curve exists under this setting. To our best knowledge, this is the first work proving a multiple descent phenomenon for any learning model.

Our analysis considers both the underparameterized and overparameterized regimes. In both regimes, we show that one can control where a descent or an ascent occurs in the generalization curve. This is realized through our algorithmic construction of a feature-revealing process. To be more specific, we assume that the data is in $\mathbb{R}^D$, where D can be arbitrarily large or even essentially infinite. We view each dimension of the data as a feature. We consider a linear regression problem restricted on the first d features, where $d < D$. New features are revealed by increasing the dimension of the data. We then show that by specifying the distribution of the newly revealed feature to be either a standard Gaussian or a Gaussian mixture, one can determine where an ascent or a descent occurs. In order to create an ascent when a new feature is revealed, it is sufficient that the new feature follows a Gaussian mixture distribution. In order to have a descent *after* an ascent, it is sufficient that the new feature follows a standard Gaussian distribution. Under both the overparamterized and underparameterized regimes, this feature-revealing process allows us to specify the occurrence of ascents of any arbitrary length. For descents, the construction depends on the regime. In the underparameterized regime, our construction allows us to create a descent only after an ascent. However, in the overparameterized regime, we can fully control the occurrence of a descent. Generally speaking, we show that we are able to design the generalization curve.

On the one hand, we show theoretically that the generalization curve is malleable and can be constructed in an arbitrary fashion. On the other hand, we rarely observe complex generalization curves in practice, besides carefully curated constructions. Putting these facts together, we arrive at the conclusion that realistic generalization curves arise from specific interactions between properties of typical data and the inductive biases of algorithms. We should highlight that the nature of these interactions is far from being understood and should be an area of further investigations.

2 Preliminaries and Problem Formulation

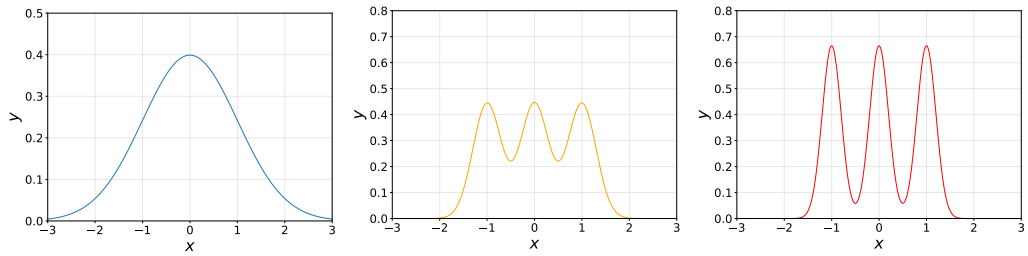
Notation. For $x \in \mathbb{R}^D$ and $d \leq D$, we let $x[1:d] \in \mathbb{R}^d$ denote a d-dimensional vector with $x[1:d]_i = x_i$ for all $1 \leq i \leq d$. For a matrix $A \in \mathbb{R}^{n \times d}$, we denote its Moore-Penrose pseudoinverse by

$A^+ \in \mathbb{R}^{d\times n}$. We use the big O notation $\mathcal{O}$ and write variables in the subscript of $\mathcal{O}$ if the implicit constant depends on them. For example, $\mathcal{O}_{n,d,\sigma}(1)$ is a constant that only depends on n, d, and σ.

Distributions. Let $\mathcal{N}(\mu, \sigma^2)$ ($\mu, \sigma \in \mathbb{R}$) and $\mathcal{N}(\mu, \Sigma)$ ($\mu \in \mathbb{R}^n$, $\Sigma \in \mathbb{R}^{n\times n}$) denote the univariate and multivariate Gaussian distributions, respectively, where $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n\times n}$ is a positive semi-definite matrix. We define a family of *trimodal* Gaussian mixture distributions as follows

$$\mathcal{N}^{\text{mix}}_{\sigma,\mu} \triangleq \frac{1}{3}\mathcal{N}(0,\sigma^2) + \frac{1}{3}\mathcal{N}(-\mu,\sigma^2) + \frac{1}{3}\mathcal{N}(\mu,\sigma^2)\,.$$

For an illustration, please see Fig. 1.



(a) $\mathcal{N}(0,1)$ feature (b) $\mathcal{N}^{\text{mix}}_{\sigma,1}$ feature, ($\sigma = 0.3$) (c) $\mathcal{N}^{\text{mix}}_{\sigma,1}$ feature, ($\sigma = 0.2$)

Figure 1: Density functions of the $\mathcal{N}(0,1)$ and $\mathcal{N}^{\text{mix}}_{\sigma,1}$ feature. A new entry is independently sampled from the 1-dimensional distribution being either a standard Gaussian or trimodal Gaussian mixture. Smaller σ leads to higher concentration around each modes.

Let $\chi^2(k,\lambda)$ denote the noncentral chi-squared distribution with k degrees of freedom and the noncentrality parameter λ. For example, if $X_i \sim \mathcal{N}(\mu_i, 1)$ (for $i = 1, 2, \ldots, k$) are independent Gaussian random variables, we have $\sum_{i=1}^k X_i^2 \sim \chi^2(k,\lambda)$, where $\lambda = \sum_{i=1}^k \mu_i^2$. We also denote by $\chi^2(k)$ the (central) chi-squared distribution with k degrees and the F-distribution by $F(d_1, d_2)$ where d_1 and d_2 are the degrees of freedom.

Problem Setup. Let $x_1, \ldots, x_n \in \mathbb{R}^D$ be column vectors that represent the training data of size n and let $x_{\text{test}} \in \mathbb{R}^D$ be a column vector that represents the test data. We assume that they are all independently drawn from a distribution

$$x_1, \ldots, x_n, x_{\text{test}} \overset{iid}{\sim} \mathcal{D}\,.$$

Let us consider a linear regression problem on the first d features, where $d \le D$ for some arbitrary large D. Here, d can be viewed as the number of features revealed. The *design* matrix A equals $[x_1[1:d], \ldots, x_n[1:d]]^\top \in \mathbb{R}^{n\times d}$. The true linear model is $\beta = 0 \in \mathbb{R}^d$. The noise $\varepsilon \in \mathbb{R}^n$ follows the multivariate standard Gaussian distribution $\mathcal{N}(0, I_{n\times n})$. Let $x = x_{\text{test}}[1:d]$ denote the first d features of the test data.

For the underparameterized regime where $d < n$, the least square solution on the training data is $A^+\varepsilon$. For the overparameterized regime where $d > n$, $A^+\varepsilon$ is the minimum-norm solution. In both regimes we consider the solution $A^+\varepsilon$. The generalization loss on the test data is then given by

$$L'_d \triangleq \mathbb{E}\|x^\top A^+\varepsilon\|^2 = \mathbb{E}\|(A^\top)^+ x\|^2\,.$$

We would like to study the change in the loss caused by the growth in the number of features revealed. Note that the product $(A^+)^\top x$ sums over d dimensions. Once we reveal a new feature, which is equivalent to adding a new row $b^\top$ to $A^\top$ and a new component y to x, the product $\begin{bmatrix} A^\top \\ b^\top \end{bmatrix}^+ \begin{bmatrix} x \\ y \end{bmatrix}$ sums over $d+1$ dimensions. As a result, to compare quantities of different dimensions, we need to normalize the generalization loss by the dimension. We define the dimensionally normalized generalization loss L_d as follows

$$L_d \triangleq \mathbb{E}\left\|\frac{1}{d}(A^\top)^+ x\right\|^2 = \frac{1}{d^2}L'_d\,.$$

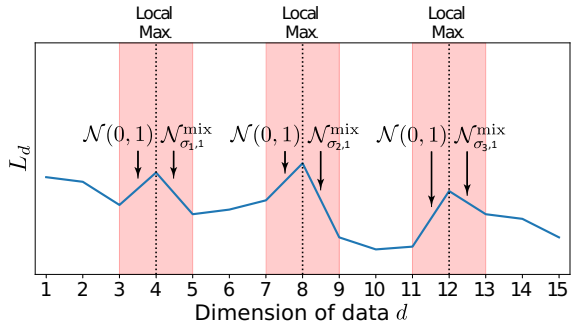


Figure 2: Illustration of multiple descent for the dimension-normalized generalization loss L_d as a function of the dimension d. A local maximum occurs at d if $L_{d-1} < L_d > L_{d+1}$. The triplet $L_{d-1},\ L_d,\ L_{d+1}$ then form an ascent/descent, which is marked by the shaded area. Local maxima are marked by the dotted lines. Adding a new feature with a Gaussian mixture distribution increases the loss, while adding one with a univariate Gaussian distribution decreases the loss. Therefore, a Gaussian mixture feature followed by a Gaussian feature creates one ascent/descent.

Local Maximum and Multiple Descent. We say that a local maximum occurs at a dimension $d \geq 1$ if $L_{d-1} < L_d$ and $L_d > L_{d+1}$. Intuitively, a local maximum occurs if there is an increasing stage of the generalization loss, followed by a decreasing stage, as the dimension d grows. Additionally, we define $L_0 \triangleq -\infty$. If the generalization loss exhibits a single descent, based on our definition, a unique local maximum occurs at $d = 1$. For a double-descent generalization curve, a local maximum occurs at two different dimensions. In general, if we observe a local maximum at K different dimensions we call it a K-descent.

3 Underparameterized Regime

First, we present our main theorem for the underparametrized regime below, whose proof is deferred to the end of Section 3. It states that given any sequence of desired dimensions, we are able to make them a local maximum of the generalization loss.

Theorem 1 (Underparameterized regime)**.** *Let $D + 2 < \sqrt{2n}$. For any $1 < d_1 < d_2 < \cdots < d_K < D$ where $d_{j+1} - d_j \geq 2$, there exists a distribution $\mathcal{D}$ such that a local maximum of the L_d curve occurs at d_j.*

Note that the assumption $d_{j+1} - d_j \geq 2$ is necessary because two local maxima may not be adjacent. We present an example in Fig. 2.

Remark 1 ($\mathcal{D}$ can be a product distribution)**.** As will be clear later in the proof of Theorem 1, the distribution $\mathcal{D}$ can be made as simple as a product distribution $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_D$ such that $x_{i,j} \overset{iid}{\sim} \mathcal{D}_j$ for all $1 \leq i \leq n$, where $\mathcal{D}_j$ is either sampled from $\mathcal{N}(0,1)$ or a Gaussian mixture $\mathcal{N}^{\text{mix}}_{\sigma_j}$ for some $\sigma_j > 0$. As a consequence, by permuting the order of $\mathcal{D}_i$'s in the product distribution, we can change the order of revealing the features.

Remark 2 (Kernel regression on Gaussian data)**.** In light of Remark 1, $\mathcal{D}$ can be chosen to be a product distribution that consists of only $\mathcal{N}(0,1)$ and $\mathcal{N}^{\text{mix}}_{\sigma_j}$. Note that one can simulate $\mathcal{N}^{\text{mix}}_{\sigma,1}$ with $\mathcal{N}(0,1)$ through the inverse transform sampling. To see this, let $F_{\mathcal{N}(0,1)}$ and $F_{\mathcal{N}^{\text{mix}}_{\sigma,1}}$ be the cdf of $\mathcal{N}(0,1)$ and $\mathcal{N}^{\text{mix}}_{\sigma,1}$, respectively. If $X \sim \mathcal{N}(0,1)$, we have $F_{\mathcal{N}(0,1)}(X) \sim \text{Unif}((0,1))$ and therefore $\varphi_\sigma(X) \triangleq F^{-1}_{\mathcal{N}^{\text{mix}}_{\sigma,1}}(F_{\mathcal{N}(0,1)}(X)) \sim \mathcal{N}^{\text{mix}}_{\sigma,1}$. In fact, we can use a multivariate Gaussian $\mathcal{D}' = \mathcal{N}(0, I_{D\times D})$ and a sequence of non-linear kernels $k^{[1:d]}(x,y) \triangleq \langle \phi^{[1:d]}(x), \phi^{[1:d]}(y)\rangle$, where the feature map is $\phi^{[1:d]}(x) \triangleq [\phi_1(x_1), \phi_2(x_2), \ldots, \phi_d(x_d)]^\top \in \mathbb{R}^d$. Here is a simple rule for defining ϕ_j: If $\mathcal{D}_j = \mathcal{N}(0,1)$, we set ϕ_j to the identity function. If $\mathcal{D}_j = \mathcal{N}^{\text{mix}}_{\sigma_j}$, we set ϕ_j to φ_{σ_j}. Thus, the problem becomes a kernel regression problem on the standard Gaussian data.

Getting back Theorem 1, let us discuss how we will construct such a distribution $\mathcal{D}$ inductively. We fix d. Again, denote the first d features of x_{test} by $x \triangleq x_{\text{test}}[1:d]$. Let us consider adding

an additional component to the training data $x_1[1:d], \ldots, x_n[1:d]$ and test data x so that we increment the dimension d by 1. Let $b_i \in \mathbb{R}$ denote the additional component that we add to the vector x_i (so that the new vector is $[x_i[1:d]^\top, b_i]^\top$. Similarly, let $y \in \mathbb{R}$ denote the additional component that we add to the vector x. We form the column vector $b = [b_1, \ldots, b_n]^\top \in \mathbb{R}^n$ that collects all additional components that we add to the training data.

We consider the change in dimension-normalized generalization loss as follows

$$L_{d+1} - L_d = \mathbb{E}\left[\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2 - \left\|\frac{1}{d}(A^+)^\top x\right\|^2\right]. \tag{1}$$

Note that the components $b_1, \ldots, b_n, y$ are i.i.d. Lemma 2 relates the pseudo-inverse of $[A, b]^\top$ to that of $A^\top$.

Lemma 2 (Proof in Appendix B.1)**.** *Let $A \in \mathbb{R}^{n\times d}$ and $0 \neq b \in \mathbb{R}^{n\times 1}$, where $n \geq d+1$. Additionally, let $P = AA^+$ and $Q = bb^+ = \frac{bb^\top}{\|b\|^2}$, and define $z \triangleq \frac{b^\top(I-P)b}{\|b\|^2}$. If $z \neq 0$ and the columnwise partitioned matrix $[A, b]$ has linearly independent columns, we have*

$$\begin{aligned}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+ &= \left[\left(I - \tfrac{bb^\top}{\|b\|^2}\right)\left(I + \tfrac{AA^+bb^\top}{\|b\|^2 - b^\top AA^+ b}\right)(A^+)^\top, \tfrac{(I-AA^+)b}{\|b\|^2 - b^\top AA^+ b}\right]\\ &= \left[(I-Q)(I + \tfrac{PQ}{1-\mathrm{tr}(PQ)})(A^+)^\top, \tfrac{(I-P)b}{b^\top(I-P)b}\right]\\ &= \left[(I-Q)(I + \tfrac{PQ}{z})(A^+)^\top, \tfrac{(I-P)b}{b^\top(I-P)b}\right].\end{aligned}$$

In our construction of $\mathcal{D}$, the components $\mathcal{D}_j$ are all continuous distributions. The matrix $I - P$ is an orthogonal projection matrix and therefore $\mathrm{rank}(I-P) = n - d$. As a result, it holds almost surely that $b \neq 0$, $z \neq 0$, and $[A, b]$ has linearly independent columns. Thus the assumptions of Lemma 2 are satisfied almost surely. In the sequel, we assume that these assumptions are always fulfilled.

Lemma 3 (Proof in Appendix B.2)**.** *Assume d, $n > d+2$ and P are fixed, where $P \in \mathbb{R}^{n\times n}$ is an orthogonal projection matrix whose rank is d. Define $z \triangleq \frac{b^\top(I-P)b}{\|b\|^2}$, where $b = [b_1, \ldots, b_n]^\top \in \mathbb{R}^n$. If $y,\ b_1, \cdots,\ b_n \overset{iid}{\sim} \mathcal{N}^{\text{mix}}_{\sigma,1}$, we have $\mathbb{E}[1/z] = \mathcal{O}_{n,d,\sigma}(1)$ and $\mathbb{E}[y^2/b^\top(I-P)b] = \mathcal{O}_{n,d,\sigma}(1)$.*

Theorem 4 provides an upper bound for the following quantity

$$\mathbb{E}_{b,y}\left[\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2 - \left\|\frac{1}{d}(A^+)^\top x\right\|^2 \middle| A, x\right]$$

if $b_1, \ldots, b_n, y$ are i.i.d. according to $\mathcal{N}(0,1)$ or $\mathcal{N}^{\text{mix}}_{\sigma,1}$. This quantiry is similar to the difference between the dimensionally normalized generalization loss $L_{d+1} - L_d$ but with expectation only over b and y.

Theorem 4 (Proof in Appendix B.3)**.** *Conditioned on A and x, the following statements hold:*

(a) If $d + 2 < \sqrt{2n}$ and $b_1, \ldots, b_n, y \overset{iid}{\sim} \mathcal{N}(0,1)$, we have

$$\mathbb{E}_{b,y}\left[\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2 - \left\|\frac{1}{d}(A^+)^\top x\right\|^2\right] < \frac{d - \|(A^+)^\top x\|^2(2n - (d+2)^2)}{d(d+1)^2(n-d-2)}. \tag{2}$$

(b) If $d + 2 < n$ and $b_1, \ldots, b_n, y \overset{iid}{\sim} \mathcal{N}^{\text{mix}}_{\sigma,1}$, we have

$$\mathbb{E}_{b,y}\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2 \leq \left\|\frac{1}{d}(A^+)^\top x\right\|^2 + \mathcal{O}_{n,d,\sigma}(1),$$

where $\mathcal{O}_{n,d,\sigma}(1)$ is a universal constant that only depends on n, d, and σ.

Corollary 5. *Assume $d+2<\sqrt{2n}$. If either $b_1,\dots,b_n,y \overset{iid}{\sim} \mathcal{N}(0,1)$ or $b_1,\dots,b_n,y \overset{iid}{\sim} \mathcal{N}^{\text{mix}}_{\sigma,1}$, and by taking expectation over all random variables, we have*

$$\mathbb{E}\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2 = \mathcal{O}_{n,d,\sigma}\left(\mathbb{E}\left\|\frac{1}{d}(A^+)^\top x\right\|^2\right).$$

We will use Theorem 4 in two different ways. The first way is presented in Corollary 5. We would like to show inductively (on d) that L_d is finite for every d. Provided that we are able to guarantee finite L_1, Corollary 5 implies that L_d is finite for every d if the components are always sampled from $\mathcal{N}(0,1)$ or $\mathcal{N}^{\text{mix}}_{\sigma,1}$.

Alternatively, we can use Theorem 4 to create a descent, i.e., make $L_{d+1}<L_d$. In light of (2), to make the left-hand side negative, we need

$$d-\|(A^+)^\top x\|^2(2n-(d+2)^2)<0,$$

which is equivalent to

$$\left\|\frac{1}{d}(A^+)^\top x\right\|^2 > \frac{1}{d(2n-(d+2))^2}.$$

One we take expectation over A and x, we need the above equation to hold in expectation in order to create a descent, i.e.,

$$L_d > \frac{1}{d(2n-(d+2))^2}.$$

Provided that L_d can be made sufficiently large, letting L_D satisfy the above inequality and then adding an additional $\mathcal{N}(0,1)$ entry will lead to $L_{d+1}<L_d$. Making a large L_d, in turn, can be achieved by adding an entry sampled from $\mathcal{N}^{\text{mix}}_{\sigma,1}$ when the data dimension increases from $d-1$ to d in the previous step. Indeed, Theorem 6 shows that adding a $\mathcal{N}^{\text{mix}}_{\sigma,1}$ feature can increase the loss by arbitrary amount.

Theorem 6 (Proof in Appendix B.4)**.** *For any $C>0$ and $\mathbb{E}\left\|\frac{1}{d}(A^+)^\top x\right\|^2<+\infty$, there exists a $\sigma>0$ such that if $b_1,\dots,b_n,y \overset{iid}{\sim} \mathcal{N}^{\text{mix}}_{\sigma,1}$, we have*

$$\mathbb{E}\left[\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2-\left\|\frac{1}{d}(A^+)^\top x\right\|^2\right]>C.$$

We are now ready to prove Theorem 1.

Proof of Theorem 1. We construct $\mathcal{D}$ inductively. Let $\mathcal{D}_1=\mathcal{N}(0,1)$. When $d=1$, we have

$$A=[x_1[1:d],\dots,x_n[1:d]]^\top=[x_{1,1},\dots,x_{n,1}]^\top\in\mathbb{R}^n,$$

which is a column vector. Therefore, $A^+=\frac{A^\top}{\|A\|^2}$. As a result, we get

$$L_1=\mathbb{E}\left\|(A^+)^\top x\right\|^2=\mathbb{E}\frac{|x|^2}{\|A\|^2}=\frac{1}{n-2},$$

where $x\sim\mathcal{N}(0,1)$ and $\|A\|^2\sim\chi^2(n)$.

Since we will set $\mathcal{D}_j$ (for $j\geq 2$) to either $\mathcal{N}(0,1)$ or $\mathcal{N}^{\text{mix}}_{\sigma,1}$, by Corollary 5, we have $L_{j+1}=\mathcal{O}_{n,j,\sigma_{j+1}}(L_j)$. By induction, we obtain that L_j is finite for all $1\leq j\leq D$.

We define $d_0\triangleq 0$. Assume that we have determined distributions $\mathcal{D}_1,\dots,\mathcal{D}_{d_j+1}$, where $0\leq j<K$. We set $\mathcal{D}_{d_j+2},\dots,\mathcal{D}_{d_{j+1}-1}$ to $\mathcal{N}(0,1)$. For $\mathcal{D}_{d_{j+1}}$, by Theorem 6, we pick $\sigma_{d_{j+1}}$ such that if $\mathcal{D}_{d_{j+1}}=\mathcal{N}^{\text{mix}}_{\sigma_{d_{j+1}}}$, we have

$$L_{d_{j+1}}>\max\left\{L_{d_{j+1}-1},\frac{1}{d_{j+1}(2n-(d_{j+1}+2)^2)}\right\}. \tag{3}$$

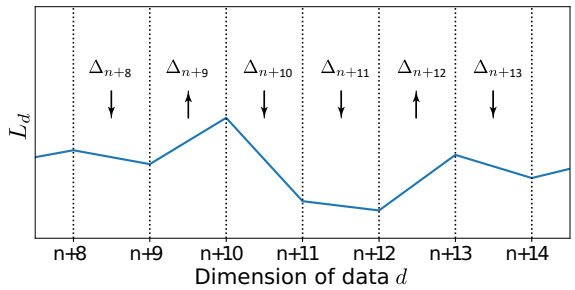


Figure 3: Illustration of the multiple descent phenomenon for the dimension-normalized generalization loss L_d versus the dimension of data d in the overparameterized regime starting from $d = n+8$. One can fully control the generalization curve to increase or decrease as specified by the sequence $\Delta = \{\downarrow, \uparrow, \downarrow, \downarrow, \uparrow, \downarrow, \dots\}$. Adding a new feature with Gaussian mixture distribution increases the loss, while adding one with Gaussian distribution decreases the loss.

Next, we set $\mathcal{D}_{d_{j+1}+1} = \mathcal{N}(0,1)$. Taking the expectation of (2) in Theorem 4 over all random variables, we have

$$L_{d_{j+1}+1} - L_{d_{j+1}} \leq \frac{d_{j+1} - d_{j+1}^2 L_{d_{j+1}}(2n - (d_{j+1}+2)^2)}{d_{j+1}(d_{j+1}+1)^2(n - d_{j+1} - 2)} < 0\,,$$

where the last inequality is due to (3). So far we have constructed a local maximum at d_{j+1}. By induction, we conclude that a local maximum occurs at every d_j. □

Remark 3. From Remark 1 and the proof of Theorem 1 it is clear that $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_D$ is a product distribution. The construction in the proof also shows that the generalization curve is actually determined by the specific choice of the $\mathcal{D}_i$'s. Note that permuting the order of $\mathcal{D}_i$'s is equivalent to changing the order by which the features are being revealed (i.e., permuting the entries of the data x_i's). Therefore, given the same data points $x_1, \cdots, x_n \in \mathbb{R}^D$, we can create many different generalization curves simply by changing the order of the feature-revealing process.

4 Overparameterized Regime

In this section, we study the multiple decent phenomenon in the overparameterized regime. Note that as stated in Section 2, we consider the minimum-norm solution here. As stated in the following theorem, we require $d \geq n+8$, which means d starts at roughly the same order as n. In other words, the result covers almost the entire spectrum of the overparameterized regime.

Theorem 7 (Overparameterized regime)**.** *Let $n < D - 9$. Given any sequence $\Delta_{n+8}, \Delta_{n+9}, \dots, \Delta_{D-1}$ where $\Delta_d \in \{\uparrow, \downarrow\}$, there exists a distribution $\mathcal{D}$ such that for every $n+8 \leq d \leq D-1$, we have*

$$L_{d+1} \begin{cases} > L_d, & \text{if } \Delta_d = \uparrow, \\ < L_d, & \text{if } \Delta_d = \downarrow. \end{cases}$$

In Theorem 7, the sequence Δ_{n+8}, Δ_{n+9}, $\cdots$, Δ_{D-1} is just used to specify the increasing/decreasing behavior of the L_d sequence for $d > n+8$. Compared to Theorem 1 for the underparameterized regime, where one is able to fully control the ascents but only partially control the descents, Theorem 7 indicates that one is able to fully control both ascents and descents in the overparameterized regime by placing an ascent/descent wherever one desires. Fig. 3 illustrates an example.

Lemma 8 gives the pseudo-inverse of A when $d > n$.

Lemma 8 (Proof in Appendix C.1)**.** *Let $A \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^{n \times 1}$, where $n \leq d$. Assume that matrix A and the columnwise partitioned matrix $B \triangleq [A, b]$ have linearly independent rows. Let $G \triangleq (AA^\top)^{-1} \in \mathbb{R}^{n \times n}$ and $u \triangleq \frac{b^\top G}{1 + b^\top G b} \in \mathbb{R}^{1 \times n}$. We have*

$$\begin{bmatrix} A^\top \\ b^\top \end{bmatrix}^+ = \left[(I - bu)^\top (A^+)^\top, u^\top\right]\,.$$

Lemma 9 establishes finite expectation for several random variables. These finite expectation results are necessary for Theorem 10 and Theorem 11 to hold. Technically, they are the dominating random variables needed in Lebesgue's dominated convergence theorem. Lemma 9 indicates that to guarantee these finite expectations, it suffices to set the first $n+8$ distributions to the standard normal distribution and then set $\mathcal{D}_{n+8}, \ldots, \mathcal{D}_D$ to either a Gaussian or a Gaussian mixture distribution. In fact, in Theorem 10 and Theorem 11, we always add a Gaussian distribution or a Gaussian mixture.

Lemma 9 (Proof in Appendix C.2)**.** *Let $\mathcal{D} = \mathcal{D}_1 \times \cdots \times \mathcal{D}_D$ be a product distribution where*

(a) $\mathcal{D}_d = \mathcal{N}(0,1)$ if $d = 1, \ldots, n+8$; and

(b) $\mathcal{D}_d$ is either $\mathcal{N}(0, \sigma_d^2)$ or $\mathcal{N}^{\text{mix}}_{\sigma_d, \mu_d}$ for $d > n+8$.

Let $\mathcal{D}_{[1:d]}$ denote $\mathcal{D}_1 \times \cdots \times \mathcal{D}_d$. Assume that every row of $A \in \mathbb{R}^{n \times d}$ and $x \in \mathbb{R}^{d \times 1}$ are i.i.d. and follow $\mathcal{D}_{[1:d]}$. For any d such that $n+8 \le d \le D$, all of the followings hold:

$$\begin{aligned}
&\mathbb{E}[\|(A^+)^\top x\|^2] < +\infty\,, \\
&\mathbb{E}[\lambda_{\max}^2((AA^\top)^{-1})] < +\infty\,, \\
&\mathbb{E}[\lambda_{\max}((AA^\top)^{-1})\|(A^+)^\top x\|^2] < +\infty\,, \\
&\mathbb{E}[\lambda_{\max}^2((AA^\top)^{-1})\|(A^+)^\top x\|^2] < +\infty\,.
\end{aligned} \tag{4}$$

Theorem 10 shows that in order to have $L_{d+1} < L_d$, it suffices to add a Gaussian feature.

Theorem 10 (Proof in Appendix C.3)**.** *If all equations in* (4) *hold, there exists a $\sigma > 0$ such that if $y, b_1, \ldots, b_n \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$, we have*

$$L_{d+1} - L_d = \mathbb{E}\left\| \frac{1}{d+1} \begin{bmatrix} A^\top \\ b^\top \end{bmatrix}^+ \begin{bmatrix} x \\ y \end{bmatrix} \right\|^2 - \mathbb{E}\left\| \frac{1}{d}(A^+)^\top x \right\|^2 < 0\,.$$

Theorem 11 shows that adding a Gaussian mixture feature can make $L_{d+1} > L_d$.

Theorem 11 (Proof in Appendix C.4)**.** *Assume $\mathbb{E}\|(A^+)^\top x\|^2 < +\infty$. For any $C > 0$, there exist μ, $\sigma > 0$ such that if $y, b_1, \ldots, b_n \overset{iid}{\sim} \mathcal{N}^{\text{mix}}_{\sigma,\mu}$, we have*

$$L_{d+1} - L_d = \mathbb{E}\left\| \frac{1}{d+1} \begin{bmatrix} A^\top \\ b^\top \end{bmatrix}^+ \begin{bmatrix} x \\ y \end{bmatrix} \right\|^2 - \mathbb{E}\left\| \frac{1}{d}(A^+)^\top x \right\|^2 > C\,.$$

The proof of Theorem 7 immediately follows from Theorem 10 and Theorem 11.

Proof of Theorem 7. We construct the product distribution $\mathcal{D} = \prod_{d=1}^{D} \mathcal{D}_d$. We set $\mathcal{D}_d = \mathcal{N}(0,1)$ for $d = 1, \ldots, n+8$. For $n+8 < d \le D$, $\mathcal{D}_d$ is either $\mathcal{N}(0, \sigma_d^2)$ or $\mathcal{N}^{\text{mix}}_{\sigma_d, \mu_d}$ depending on Δ_d being either $\downarrow$ or $\uparrow$. If $\Delta_{d-1} = \downarrow$, by Theorem 10, there exists $\sigma_d > 0$ such that if $\mathcal{D}_d = \mathcal{N}(0, \sigma_d^2)$, then $L_d < L_{d-1}$. Similarly if $\Delta_{d-1} = \uparrow$, by Theorem 11, there exists σ_d and μ_d such that $\mathcal{D}_d = \mathcal{N}^{\text{mix}}_{\sigma_d, \mu_d}$ guarantees $L_d > L_{d-1}$.

□

5 Conclusion

Our work proves that the expected risk of linear regression can manifest multiple descents when the number of features increases and sample size is fixed. This is carried out through an algorithmic construction of a feature-revealing process where the newly revealed feature follows either a Gaussian distribution or a Gaussian mixture distribution. Notably, the construction also enables us to control local maxima in the underparameterized regime and control ascents/descents freely in the overparameterized regime. Overall, this allows us to design the generalization curve away from the interpolation threshold. We conjecture that the same multiple-descent generalization curve can occur in non-linear neural networks and we humbly suggest that entities with infinite computational powers investigate this phenomenon.

REFERENCES

Madhu S Advani and Andrew M Saxe. High-dimensional dynamics of generalization error in neural networks. *arXiv preprint arXiv:1710.03667*, 2017.

Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pp. 242–252, 2019.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. In *Advances in Neural Information Processing Systems*, pp. 8141–8150, 2019a.

Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *ICML*, pp. 477–502, 2019b.

Sanjeev Arora, Simon S Du, Zhiyuan Li, Ruslan Salakhutdinov, Ruosong Wang, and Dingli Yu. Harnessing the power of infinitely wide deep nets on small-data tasks. In *International Conference on Learning Representations*, 2019c.

Jerzy K Baksalary and Oskar Maria Baksalary. Particular formulae for the moore–penrose inverse of a columnwise partitioned matrix. *Linear algebra and its applications*, 421(1):16–23, 2007.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 2020.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine learning and the bias-variance trade-off. *stat*, 1050:28, 2018a.

Mikhail Belkin, Siyuan Ma, and Soumik Mandal. To understand deep learning we need to understand kernel learning. In *International Conference on Machine Learning*, pp. 541–549, 2018b.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571*, 2019a.

Mikhail Belkin, Alexander Rakhlin, and Alexandre B Tsybakov. Does data interpolation contradict statistical optimality? In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1611–1619, 2019b.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Jauvin. A neural probabilistic language model. *Journal of machine learning research*, 3(Feb):1137–1155, 2003.

Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. In *Advances in Neural Information Processing Systems*, pp. 10836–10846, 2019.

Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Emmanuel Caron and Stephane Chretien. A finite sample analysis of the double descent phenomenon for ridge function estimation. *arXiv preprint arXiv:2007.12882*, 2020.

Lin Chen, Yifei Min, Mingrui Zhang, and Amin Karbasi. More data can expand the generalization gap between adversarially robust and standard models. In *ICML*, 2020.

Yehuda Dar, Paul Mayer, Lorenzo Luzi, and Richard G Baraniuk. Subspace fitting meets regression: The effects of supervision and orthonormality constraints on double descent of generalization errors. In *ICML*, 2020.

Stéphane d'Ascoli, Levent Sagun, and Giulio Biroli. Triple descent and the two kinds of overfitting: Where & why do they appear? *arXiv preprint arXiv:2006.03509*, 2020.

Ernesto De Vito, Andrea Caponnetto, and Lorenzo Rosasco. Model selection for regularized least-squares algorithm in learning theory. *Foundations of Computational Mathematics*, 5(1):59–85, 2005.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International Conference on Machine Learning*, pp. 1675–1685, 2019.

Mario Geiger, Stefano Spigler, Stéphane d'Ascoli, Levent Sagun, Marco Baity-Jesi, Giulio Biroli, and Matthieu Wyart. Jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review E*, 100(1):012115, 2019.

Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(2):023401, 2020.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191*, 2019.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Yanping Huang, Youlong Cheng, Ankur Bapna, Orhan Firat, Dehao Chen, Mia Chen, HyoukJoong Lee, Jiquan Ngiam, Quoc V Le, Yonghui Wu, et al. Gpipe: Efficient training of giant neural networks using pipeline parallelism. In *Advances in neural information processing systems*, pp. 103–112, 2019.

Adel Javanmard, Mahdi Soltanolkotabi, and Hamed Hassani. Precise tradeoffs in adversarial training for linear regression. In *Conference on Learning Theory*, 2020.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pp. 1097–1105, 2012.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeles" regression can generalize. *Annals of Statistics*, pp. to appear, 2019.

Tengyuan Liang, Alexander Rakhlin, and Xiyu Zhai. On the multiple descent of minimum-norm interpolants and restricted lower isometry of kernels. In *COLT*, 2020.

Marco Loog, Tom Viering, and Alexander Mey. Minimizers of the empirical risk and risk monotonicity. In *Advances in Neural Information Processing Systems*, pp. 7478–7487, 2019.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

Yifei Min, Lin Chen, and Amin Karbasi. The curious case of adversarially robust models: More data can help, double descend, or hurt generalization, 2020.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *arXiv preprint arXiv:1912.02292*, 2019.

Preetum Nakkiran, Prayaag Venkat, Sham Kakade, and Tengyu Ma. Optimal regularization can mitigate double descent. *arXiv preprint arXiv:2003.01897*, 2020.

Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. In *ICLR (Workshop)*, 2015.

Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pp. 1177–1184, 2008.

Alexander Rakhlin and Xiyu Zhai. Consistency of interpolation with laplace kernels is a high-dimensional phenomenon. In *Conference on Learning Theory*, pp. 2595–2623, 2019.

Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge (less) regression under general source condition. *arXiv preprint arXiv:2006.06386*, 2020.

Alessandro Rudi and Lorenzo Rosasco. Generalization properties of learning with random features. In *Advances in Neural Information Processing Systems*, pp. 3215–3225, 2017.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Dietrich von Rosen. Moments for the inverted wishart distribution. *Scandinavian Journal of Statistics*, pp. 97–109, 1988.

Colin Wei, Jason D Lee, Qiang Liu, and Tengyu Ma. Regularization matters: Generalization and optimization of neural nets vs their induced kernel. In *Advances in Neural Information Processing Systems*, pp. 9712–9724, 2019.

Abraham J Wyner, Matthew Olson, Justin Bleich, and David Mease. Explaining the success of adaboost and random forests as interpolating classifiers. *The Journal of Machine Learning Research*, 18(1):1558–1590, 2017.

Ji Xu and Daniel J Hsu. On the number of variables to use in principal component regression. In *Advances in Neural Information Processing Systems*, pp. 5094–5103, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *ICLR*, 2017.

Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

A FURTHER RELATED WORK

Our work is directly related to the recent line of research in the theoretical understanding of the double descent (Belkin et al., 2019a; Hastie et al., 2019; Xu & Hsu, 2019; Mei & Montanari, 2019) and the multiple descent phenomenon (Liang et al., 2020). Here we briefly discuss some other work that is closely related to this paper.

Least Square Regression. In this paper we focus on the least square linear regression with no regularization. For the regularized least square regression, De Vito et al. (2005) proposed a selection procedure for the regularization parameter. Advani & Saxe (2017) analyzed the generalization of neural networks with mean squared error under the asymptotic regime where both the sample size and model complexity tend to infinity. Richards et al. (2020) proved for least square regression in the asymptotic regime that as the dimension-to-sample-size ratio d/n grows, an additional peak can occur in both the variance and bias due to the covariance structure of the features. As a comparison, in this paper the sample size is fixed and the model complexity increases. Rudi & Rosasco (2017) studied kernel ridge regression and gave an upper bound on the number of the random features to reach certain risk level. Our result shows that there exists a natural setting where by manipulating the random features one can control the risk curve.

Over-Parameterization and Interpolation. The double descent occurs when the model complexity reaches and increases beyond the interpolation threshold. Most previous works focused on proving an upper bound or optimal rate for the risk. Caponnetto & De Vito (2007) gave the optimal rate for least square ridge regression via careful selection of the regularization parameter. Belkin et al. (2019b) showed that the optimal rate for risk can be achieved by a model that interpolates the training data. In a series of work on kernel regression with regularization parameter tending to zero (a.k.a. kernel *ridgeless* regression), Rakhlin & Zhai (2019) showed that the risk is bounded away from zero when the data dimension is fixed with respect to the sample size. Liang & Rakhlin (2019) then considered the case when $d \asymp n$ and proved a risk upper bound that can be small given favorable data and kernel assumptions. Instead of giving a bound, our paper presents an exact computation of risk in the cases of underparameterized and overparameterized linear regression, and proves the existence of the multiple descent phenomenon. Wyner et al. (2017) analyzed AdaBoost and Random Forest from the perspective of interpolation. There has also been a line of work on wide neural networks (Arora et al., 2019a;b;c; Du et al., 2019; Allen-Zhu et al., 2019; Wei et al., 2019; Cao & Gu, 2019; Zou et al., 2020).

Sample-wise Double Descent and Non-monotonicity. There has also been recent development beyond the model-complexity double-descent phenomenon. For example, regarding sample-wise non-monotonicity, Nakkiran et al. (2019) empirically observed the epoch-wise double-descent and sample-wise non-monotonicity for neural networks. Chen et al. (2020) and Min et al. (2020) identified and proved the sample-wise double descent under the adversarial training setting, and Javanmard et al. (2020) discovered double-descent under adversarially robust linear regression. Loog et al. (2019) showed that empirical risk minimization can lead to sample-wise non-monotonicity in the standard linear model setting under various loss functions including the absolute loss and the squared loss, which covers the range from classification to regression. We also refer the reader to their discussion of the earlier work on non-monotonicity of generalization curves. Dar et al. (2020) demonstrated the double descent curve of the generalization errors of subspace fitting problems.

B PROOFS FOR UNDERPARAMETRIZED REGIME

B.1 PROOF OF LEMMA 2

Proof. By (Baksalary & Baksalary, 2007, Theorem 1), we have

$$\begin{bmatrix} A^\top \\ b^\top \end{bmatrix}^+ = \left[(I-Q)A(A^\top(I-Q)A)^{-1}, \frac{(I-P)b}{b^\top(I-P)b} \cdot \right]$$

Define $r \triangleq A^\top b \in \mathbb{R}^d$. Since A has linearly independent columns, the Gram matrix $G = A^\top A$ is non-singular. The Sherman-Morrison formula gives

$$(A^\top(I-Q)A)^{-1} = \left(A^\top A - \frac{rr^\top}{\|b\|^2}\right)^{-1} = G^{-1} + \frac{G^{-1}rr^\top G^{-1}}{\|b\|^2 - r^\top G^{-1} r} = G^{-1} + \frac{G^{-1}rb^\top (A^+)^\top}{\|b\|^2 - r^\top G^{-1} r},$$

where we use the facts $r = A^\top b$ and $AG^{-1} = (A^+)^\top$ in the last equality. Therefore, we deduce

$$\begin{aligned} A(A^\top(I-Q)A)^{-1} &= AG^{-1} + \frac{AG^{-1}rb^\top (A^+)^\top}{\|b\|^2 - r^\top G^{-1} r} \\ &= (A^+)^\top + \frac{AG^{-1}A^\top bb^\top (A^+)^\top}{\|b\|^2 - r^\top G^{-1} r} \\ &= \left(I + \frac{AA^+bb^\top}{\|b\|^2 - r^\top G^{-1} r}\right)(A^+)^\top \\ &= \left(I + \frac{PQ}{1 - \frac{r^\top G^{-1} r}{\|b\|^2}}\right)(A^+)^\top . \end{aligned}$$

Observe that

$$1 - \frac{r^\top G^{-1} r}{\|b\|^2} = 1 - \frac{b^\top A(A^\top A)^{-1}A^\top b}{\|b\|^2} = 1 - \frac{b^\top P b}{\|b\|^2} = z .$$

Therefore, we obtain the desired expression. □

B.2 Proof of Lemma 3

Lemma 12 shows that a noncentral χ^2 distribution first-order stochastically dominates a central χ^2 distribution of the same degree of freedom. It will be needed in the proof of Lemma 3.

Lemma 12. *Assume that random variables $X \sim \chi^2(k, \lambda)$ and $Y \sim \chi^2(k)$, where $\lambda > 0$. For any $c > 0$, we have*

$$\mathbb{P}(X \geq c) > \mathbb{P}(Y \geq c).$$

In other words, the random variable X (first-order) stochastically dominates Y.

Proof. Let $Y_1, X_2, \ldots, X_k \overset{iid}{\sim} \mathcal{N}(0,1)$ and $X_1 \sim \mathcal{N}(\sqrt{\lambda}, 1)$ and all these random variables are jointly independent. Then $X' \triangleq \sum_{i=1}^k X_i^2 \sim \chi^2(k, \lambda)$ and $Y' \triangleq Y_1^2 + \sum_{i=2}^k X_i^2 \sim \chi^2(k)$.

It suffices to show that $\mathbb{P}(X' \geq c) > \mathbb{P}(Y' \geq c)$, or equivalently, $\mathbb{P}(|\mathcal{N}(\mu,1)| \geq c) > \mathbb{P}(|\mathcal{N}(0,1)| \geq c)$ for all $c > 0$ and $\mu \triangleq \sqrt{\lambda} > 0$. Denote $F_c(t) = \mathbb{P}(|\mathcal{N}(\mu,1)| \geq c)$ and we have

$$F_c(\mu) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-c}^{c} \exp\left(-\frac{(x-\mu)^2}{2}\right) dx = 1 - \frac{1}{\sqrt{2\pi}} \int_{-c-\mu}^{c-\mu} \exp\left(-\frac{x^2}{2}\right) dx,$$

and thus

$$\frac{dF_c(\mu)}{d\mu} = \frac{1}{\sqrt{2\pi}} \left[\exp\left(-\frac{(c-\mu)^2}{2}\right) - \exp\left(-\frac{(c+\mu)^2}{2}\right)\right] > 0.$$

This shows $\mathbb{P}(|\mathcal{N}(\mu,1)| \geq c) > \mathbb{P}(|\mathcal{N}(0,1)| \geq c)$ and we are done.

□

Proof of Lemma 3. Since $b_i \overset{iid}{\sim} \mathcal{N}^{\text{mix}}_{\sigma,1}$, we can rewrite $b = u + w$ where $w \sim \mathcal{N}(0, \sigma^2 I_n)$ and the entries of u satisfy $u_i \overset{iid}{\sim} \text{Unif}(\{-1, 0, 1\})$. Furthermore, u and w are independent. Note that for any fixed n and d, the support of u is finite and its cardinality only depends on n. Therefore, we only need to show that conditioning on u, the expectation over w is $\mathcal{O}_{n,d,\sigma}(1)$. In other words, for any fixed u, we want to show $\mathbb{E}_w[1/z \mid u] = \mathcal{O}_{n,d,\sigma}(1)$ and $\mathbb{E}_w\left[\frac{y^2}{b^\top(I-P)b}\middle| u\right] = \mathcal{O}_{n,d,\sigma}(1)$.

Note that since y^2/σ^2 is first-order stochastically dominated by $\chi^2(1,1)$, we have

$$\mathbb{E}[y^2 \mid u] = \mathbb{E}[y^2] \le \sigma^2 \mathbb{E}[\chi^2(1,1)] = 2\sigma^2.$$

Therefore, it remains to show $\mathbb{E}_w[1/z \mid u] = \mathcal{O}_{n,d,\sigma}(1)$ and $\mathbb{E}_w\left[\frac{1}{b^\top(I-P)b}\middle| u\right] = \mathcal{O}_{n,d,\sigma}(1)$.

Note that

$$\frac{1}{z} = \frac{b^\top I b}{b^\top (I-P) b} = 1 + \frac{(u+w)^\top P (u+w)}{(u+w)^\top (I-P)(u+w)}.$$

Since P is an orthogonal projection, there exists an orthogonal transformation O depending only on P such that

$$(u+w)^\top P (u+w) = [O(u+w)]^\top D_d [O(u+w)]$$

where $D_d = \mathrm{diag}([1,\dots,1,0\dots,0])$ with d diagonal entries equal to 1 and the others equal to 0. We denote $\tilde{u} = O(u)$, which is fixed (as u and O are fixed), and $\tilde{w} = O(w) \sim \mathcal{N}(0, \sigma^2 I_n)$. It follows that

$$\frac{1}{z} = 1 + \frac{(\tilde{u}+\tilde{w})^\top D_d (\tilde{u}+\tilde{w})}{(\tilde{u}+\tilde{w})^\top (I-D_d)(\tilde{u}+\tilde{w})} = 1 + \frac{\sum_{i=1}^d (\tilde{u}_i+\tilde{w}_i)^2}{\sum_{i=d+1}^n (\tilde{u}_i+\tilde{w}_i)^2} = 1 + \frac{\sum_{i=1}^d (\tilde{u}_i+\tilde{w}_i)^2/\sigma^2}{\sum_{i=d+1}^n (\tilde{u}_i+\tilde{w}_i)^2/\sigma^2}.$$

Observe that

$$\sum_{i=1}^d (\tilde{u}_i+\tilde{w}_i)^2/\sigma^2 \sim \chi^2\left(d, \sqrt{\sum_{i=1}^d \tilde{u}_i^2}\right)$$
$$\sum_{i=d+1}^n (\tilde{u}_i+\tilde{w}_i)^2/\sigma^2 \sim \chi^2\left(n-d, \sqrt{\sum_{i=d+1}^n \tilde{u}_i^2}\right),$$

and that these two quantities are independent. It follows that

$$\mathbb{E}\left[\sum_{i=1}^d (\tilde{u}_i+\tilde{w}_i)^2/\sigma^2 \middle| u\right] = d + \sqrt{\sum_{i=1}^d \tilde{u}_i^2}.$$

By Lemma 12, the denominator $\sum_{i=d+1}^n (\tilde{u}_i+\tilde{w}_i)^2/\sigma^2$ first-order stochastically dominates $\chi^2(n-d)$. Therefore, we have

$$\mathbb{E}\left[\frac{1}{\sum_{i=d+1}^n (\tilde{u}_i+\tilde{w}_i)^2/\sigma^2}\middle| u\right] \le \mathbb{E}\left[\frac{1}{\chi^2(n-d)}\right] = \frac{1}{n-d-2}.$$

Putting the numerator and denominator together yields

$$\mathbb{E}\left[\frac{1}{z}\middle| u\right] \le 1 + \frac{d + \sqrt{\sum_{i=1}^d \tilde{u}_i^2}}{n-d-2} \le 1 + \frac{d+\sqrt{d}}{n-d-2} = \mathcal{O}_{n,d,\sigma}(1).$$

Similarly, we have

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{b^\top(I-P)b}\middle| u\right] &= \mathbb{E}\left[\frac{1}{[O(u+w)]^\top (I-D_d)[O(u+w)]}\middle| u\right] \\
&= \mathbb{E}\left[\frac{1/\sigma^2}{\sum_{i=d+1}^n (\tilde{u}_i+\tilde{w}_i)^2/\sigma^2}\middle| u\right] \\
&\le \frac{1}{\sigma^2}\mathbb{E}\left[\frac{1}{\chi^2(n-d)}\right] \\
&= \frac{1}{\sigma^2}\cdot\frac{1}{n-d-2} \\
&= \mathcal{O}_{n,d,\sigma}(1).
\end{aligned}$$

□

B.3 PROOF OF THEOREM 4

Proof. First, we rewrite the expression as follows

$$\begin{aligned}&\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2-\left\|\frac{1}{d}(A^+)^\top x\right\|^2\\ =&\frac{1}{(d+1)^2}\left\|(I-Q)(I+PQ/z)(A^+)^\top x+\frac{(I-P)b}{b^\top(I-P)b}y\right\|^2-\frac{1}{d^2}\|(A^+)^\top x\|^2\,,\end{aligned}\tag{5}$$

where P, Q, z are defined in Lemma 2. Since y has mean 0 and is independent of other random variables, so that the cross term vanishes under expectation over b and y:

$$\mathbb{E}_{b,y}\left[\left\langle(I-Q)(I+PQ/z)(A^+)^\top x,\frac{(I-P)b}{b^\top(I-P)b}y\right\rangle\right]=0\,,$$

where $\langle\cdot,\cdot\rangle$ denotes the inner product. Therefore taking the expectation of (5) over b and y yields

$$\mathbb{E}_{b,y}\left[\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2-\left\|\frac{1}{d}(A^+)^\top x\right\|^2\right]\tag{6}$$

$$=\mathbb{E}_{b,y}\left[\frac{1}{(d+1)^2}\|(I-Q)(I+PQ/z)(A^+)^\top x\|^2-\frac{1}{d^2}\|(A^+)^\top x\|^2+\frac{1}{(d+1)^2}\left\|\frac{(I-P)b}{b^\top(I-P)b}y\right\|^2\right]\tag{7}$$

$$=\frac{1}{(d+1)^2}\mathbb{E}_{b,y}\left[\|(I-Q)(I+PQ/z)(A^+)^\top x\|^2-(1+\frac{1}{d})^2\|(A^+)^\top x\|^2+\left\|\frac{(I-P)b}{b^\top(I-P)b}y\right\|^2\right]\,.\tag{8}$$

We simplify the third term. Recall that $I-P=I-AA^+$ is an orthogonal projection matrix and thus idempotent

$$\left\|\frac{(I-P)b}{b^\top(I-P)b}y\right\|^2=\frac{y^2}{(b^\top(I-P)b)^2}\|(I-P)b\|^2=\frac{y^2}{b^\top(I-P)b}\,.\tag{9}$$

Thus we have

$$\mathbb{E}_{b,y}\left[\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2-\left\|\frac{1}{d}(A^+)^\top x\right\|^2\right]\tag{10}$$

$$=\frac{1}{(d+1)^2}\mathbb{E}_{b,y}\left[\|(I-Q)(I+PQ/z)(A^+)^\top x\|^2-(1+\frac{1}{d})^2\|(A^+)^\top x\|^2+\frac{y^2}{b^\top(I-P)b}\right]\,.\tag{11}$$

We consider the first and second terms. We write $v=(A^+)^\top x$ and define $z=\frac{b^\top(I-P)b}{\|b\|^2}$. The sum of the first and second terms equals

$$\|(I-Q)(I+PQ/z)v\|^2-(1+\frac{1}{d})^2\|v\|^2=-v^\top(M+\delta I)v\,,\tag{12}$$

where $\delta=\frac{2}{d}+\frac{1}{d^2}$ and

$$M\triangleq Q-\frac{PQ+QP}{z}+\left(\frac{2}{z}-\frac{1}{z^2}\right)QPQ+\frac{QPQPQ}{z^2}\,.$$

The rank of M is at most 2. To see this, we re-write M in the following way

$$M=\left[Q\left(-\frac{P}{z}+\left(\frac{2}{z}-\frac{1}{z^2}\right)PQ+\frac{PQPQ}{z^2}\right)\right]+\left[-\frac{PQ}{z}\right]\triangleq M_1+M_2\,.$$

Notice that $\mathrm{rank}(M_1) \leq \mathrm{rank}(Q)$, $\mathrm{rank}(M_2) \leq \mathrm{rank}(Q)$, and $\mathrm{rank}(Q) = 1$. It follows that $\mathrm{rank}(M) \leq \mathrm{rank}(M_1) + \mathrm{rank}(M_2) = 2$. The matrix M has at least $n-2$ zero eigenvalues. We claim that M has two non-zero eigenvalues and they are $1 - 1/z < 0$ and 1.

Since

$$\mathrm{rank}(PQ) \leq \mathrm{rank}(Q) = 1$$

and

$$\mathrm{tr}(PQ) = \frac{b^\top P b}{\|b\|^2} = 1 - z,$$

thus PQ has a unique non-zero eigenvalue $1-z$. Let $u \neq 0$ denote the corresponding eigenvector such that $PQu = (1-z)u$. Since $u \in \mathrm{im}\, P$ and P is a projection, we have $Pu = u$. Therefore we can verify that

$$Mu = (1 - \frac{1}{z})u\,.$$

To show that the other non-zero eigenvalue of M is 1, we compute the trace of M

$$\mathrm{tr}(M) = \mathrm{tr}(Q) - \frac{2\,\mathrm{tr}(PQ)}{z} + \left(\frac{2}{z} - \frac{1}{z^2}\right)\mathrm{tr}(PQ) + \frac{\mathrm{tr}((PQ)^2)}{z^2} = 2 - \frac{1}{z}\,,$$

where we use the fact that $\mathrm{tr}(Q) = 1$, $\mathrm{tr}(PQ) = 1 - z$,

$$\mathrm{tr}((PQ)^2) = \mathrm{tr}\left(\frac{Pbb^\top Pbb^\top}{\|b\|^4}\right) = \mathrm{tr}\left(\frac{(b^\top Pb)(b^\top Pb)}{\|b\|^4}\right) = (1-z)^2\,.$$

We have shown that M has eigenvalue $1 - 1/z$ and M has at most two non-zero eigenvalues. Therefore, the other non-zero eigenvalue is $\mathrm{tr}(M) - (1 - 1/z) = 1$.

We are now in a position to upper bound (12) as follows:

$$-v^\top(M + \delta I)v \leq -(1 - 1/z + \delta)\|v\|^2 < -(1 - 1/z + 2/d)\|v\|^2\,.$$

Putting all three terms of the change in the dimension-normalized generalization loss yields

$$\begin{aligned}
&\mathbb{E}_{b,y}\left[\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2 - \left\|\frac{1}{d}(A^+)^\top x\right\|^2\right] \\
&\leq \frac{1}{(d+1)^2}\mathbb{E}_{b,y}\left[-(1 - 1/z + 2/d)\|v\|^2 + \frac{y^2}{b^\top(I-P)b}\right]\,.
\end{aligned}$$

For $b_1, \ldots, b_n, y \overset{iid}{\sim} \mathcal{N}(0,1)$, we have $\mathbb{E}[y^2] = 1$. Moreover, $b^\top(I-P)b$ follows $\chi^2(n-d)$ a distribution. Thus $\frac{1}{b^\top(I-P)b}$ follows an inverse-chi-squared distribution with mean $\frac{1}{n-d-2}$. Therefore the expectation $\mathbb{E}[\frac{y^2}{b^\top(I-P)b}] = \frac{1}{n-d-2}$. Notice that $1/z$ follows a $1 + \frac{d}{n-d}F(d, n-d)$ distribution and thus $\mathbb{E}[1/z] = 1 + \frac{d}{n-d-2}$. As a result, we obtain

$$\begin{aligned}
&\mathbb{E}_{b,y}\left[\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2 - \left\|\frac{1}{d}(A^+)^\top x\right\|^2\right] \\
&\leq \frac{1}{(d+1)^2}\left[\left(\frac{d}{n-d-2} - \frac{2}{d}\right)\|v\|^2 + \frac{1}{n-d-2}\right] \\
&< \frac{d - \|v\|^2(2n - (d+2)^2)}{d(d+1)^2(n-d-2)}\,.
\end{aligned}$$

For $b_1, \ldots, b_n, y \overset{iid}{\sim} \mathcal{N}^{\mathrm{mix}}_{\sigma,1}$, Lemma 3 implies that

$$\mathbb{E}_{b,y}[1/z] < \mathcal{O}_{n,d,\sigma}(1),$$

and

$$\mathbb{E}_{b,y}[\frac{y^2}{b^\top(I-P)b}] < \mathcal{O}_{n,d,\sigma}(1).$$

Therefore, we conclude that

$$\mathbb{E}_{b,y}\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2 \le \left\|\frac{1}{d}(A^+)^\top x\right\|^2 + \mathcal{O}_{n,d,\sigma}(1)\,.$$

□

B.4 Proof of Theorem 6

Proof. We start from (11). Taking expectation over all random variables gives

$$\begin{aligned}
&\mathbb{E}\left[\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2 - \left\|\frac{1}{d}(A^+)^\top x\right\|^2\right]\\
&= \frac{1}{(d+1)^2}\mathbb{E}\left[\|(I-Q)(I+PQ/z)(A^+)^\top x\|^2 - (1+\frac{1}{d})^2\|(A^+)^\top x\|^2 + \frac{y^2}{b^\top(I-P)b}\right]\\
&\ge \frac{1}{(d+1)^2}\left(-(1+\frac{1}{d})^2\mathbb{E}\|(A^+)^\top x\|^2 + \mathbb{E}\left[\frac{y^2}{\sum_{i=1}^n b_i^2}\right]\right)\,.
\end{aligned}$$

Our strategy is to choose σ so that $\mathbb{E}\left[\frac{y^2}{\sum_{i=1}^n b_i^2}\right]$ is sufficiently large. This is indeed possible as we immediately show. Define independent random variables $u \sim \mathrm{Unif}(\{-1,0,1\})$ and $w \sim \mathcal{N}(0,\sigma^2)$. Since y has the same distribution as $u+w$, we have

$$\mathbb{E}[y^2] = \mathbb{E}[(u+w)^2] = \mathbb{E}[u^2] + \mathbb{E}[w^2] \ge \frac{2}{3}\,.$$

On the other hand,

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{\sum_{i=1}^n b_i^2}\right] &\ge \mathbb{P}(\max_i |b_i| \le \sigma)\ \mathbb{E}\left[\frac{1}{\sum_{i=1}^n b_i^2}\middle|\max_i |b_i| \le \sigma\right]\\
&= [\mathbb{P}(|b_1| \le \sigma)]^n\ \mathbb{E}\left[\frac{1}{\sum_{i=1}^n b_i^2}\middle|\max_i |b_i| \le \sigma\right]\\
&\ge \left[\frac{1}{3\sqrt{2\pi\sigma^2}}\int_{-\sigma}^{\sigma}\exp\left(-\frac{t^2}{2\sigma^2}\right)\,dt\right]^n \frac{1}{n\sigma^2}\\
&\ge \frac{1}{5^n n\sigma^2}\,.
\end{aligned}$$

Together we have

$$\mathbb{E}\left[\frac{y^2}{\sum_{i=1}^n b_i^2}\right] \ge \frac{1}{5^{n+1}n\sigma^2}\,.$$

Therefore, for any $C \in \mathbb{R}$ and $\mathbb{E}\left\|\frac{1}{d}(A^+)^\top x\right\|^2$, there exists σ such that

$$\mathbb{E}\left[\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2 - \left\|\frac{1}{d}(A^+)^\top x\right\|^2\right] > C\,.$$

□

C Proofs for Overparametrized Regime

C.1 Proof of Lemma 8

Proof. Since A and B have full row rank, $(AA^\top)^{-1}$ and $(BB^\top)^{-1}$ exist. Therefore we have

$$B^+ = B^\top(BB^\top)^{-1}.$$

The Sherman-Morrison formula gives

$$(BB^\top)^{-1} = (AA^\top + bb^\top)^{-1} = G - \frac{Gbb^\top G}{1 + b^\top Gb} = G - Gbu = G(I - bu)\,.$$

Hence, we deduce

$$B^+ = [A, b]^\top G(I - bu) = \begin{bmatrix} A^\top G(I - bu) \\ b^\top G(I - bu) \end{bmatrix} = \begin{bmatrix} A^+(I - bu) \\ b^\top G(I - bu) \end{bmatrix} = \begin{bmatrix} A^+(I - bu) \\ u \end{bmatrix}\,.$$

Transposing the above equation yields to the promised equation. □

C.2 Proof of Lemma 9

Proof. Let us first denote

$$v \triangleq (A^+)^\top x$$

and

$$G \triangleq (AA^\top)^{-1} \in \mathbb{R}^{n\times n}.$$

First note that by Cauchy-Schwarz inequality, it suffices to show there exists $\mathcal{D}$ such that $\mathbb{E}[\lambda_{\text{max}}^4(G)] < +\infty$ and $\mathbb{E}\|v\|^4 < +\infty$.

We define $A_d \in \mathbb{R}^{n\times d}$ to be the submatrix of A that consists of all n rows and first d columns. Denote

$$G_d \triangleq (A_d A_d^\top)^{-1} \in \mathbb{R}^{n\times n}.$$

We will prove $\mathbb{E}[\lambda_{\text{max}}^4(G)] < +\infty$ by induction.

The base step is $d = n + 8$. Recall $\mathcal{D}_{[1:n+8]} = \mathcal{N}(0, I_{n+8})$. We first show $\mathbb{E}[\lambda_{\text{max}}(G_{n+8})]^4 < +\infty$. Note that since G_{n+8} is almost surely positive definite,

$$\mathbb{E}[\lambda_{\text{max}}^4(G_{n+8})] = \mathbb{E}[\lambda_{\text{max}}(G_{n+8}^4)] \le \mathbb{E}\,\text{tr}(G_{n+8}^4) = \mathbb{E}\,\text{tr}((A_{n+8}A_{n+8}^\top)^{-4}) = \text{tr}(\mathbb{E}[(A_{n+8}A_{n+8}^\top)^{-4}])\,.$$

By our choice of $\mathcal{D}_{[1:n+8]}$, the matrix $(A_{n+8}A_{n+8}^\top)^{-1}$ is an inverse Wishart matrix of size $n \times n$ with $(n + 8)$ degrees of freedom, and thus has finite fourth moment (see, for example, Theorem 4.1 in (von Rosen, 1988)). It then follows that

$$\mathbb{E}[\lambda_{\text{max}}^4(G_{n+8})] \le \text{tr}(\mathbb{E}[(A_{n+8}A_{n+8}^\top)^{-4}]) < +\infty\,.$$

For the inductive step, assume $\mathbb{E}[\lambda_{\text{max}}(G_d)]^4 < +\infty$ for some $d \ge n + 8$. We claim that

$$\lambda_{\text{max}}(G_{d+1}) \le \lambda_{\text{max}}(G_d)\,,$$

or equivalently,

$$\lambda_{\text{min}}(A_d A_d^\top) \le \lambda_{\text{min}}(A_{d+1}A_{d+1}^\top)\,.$$

Indeed, this follows from the fact that

$$A_d A_d^\top \preccurlyeq A_d A_d^\top + bb^\top = A_{d+1}A_{d+1}^\top\,,$$

under the Loewner order, where $b \in \mathbb{R}^{n\times 1}$ is the $(d + 1)$-th column of A. Therefore, we have

$$\mathbb{E}[\lambda_{\text{max}}^4(G_{d+1})] \le \mathbb{E}[\lambda_{\text{max}}^4(G_d)]$$

and by induction, we conclude that $\mathbb{E}[\lambda_{\text{max}}^4(G)] < +\infty$ for all $d \ge n + 8$.

Now we proceed to show $\mathbb{E}\|v\|^4 < +\infty$. We have

$$\|v\|^4 = \|(AA^\top)^{-1}Ax\|^4 \le \|(AA^\top)^{-1}A\|_{op}^4 \cdot \|x\|^4\,,$$

where $\|\cdot\|_{op}$ denotes the $\ell^2 \to \ell^2$ operator norm. Note that

$$\begin{aligned}
\|(AA^\top)^{-1}A\|_{op}^4 &= \lambda_{\text{max}}^2\left(\left((AA^\top)^{-1}A\right)^\top (AA^\top)^{-1}A\right) \\
&= \lambda_{\text{max}}^2\left(A^\top(AA^\top)^{-2}A\right) \\
&= \lambda_{\text{max}}\left(\left(A^\top(AA^\top)^{-2}A\right)^2\right)\,,
\end{aligned}$$

where the last equality uses the fact that $A^\top(AA^\top)^{-2}A$ is positive semidefinite. Moreover, we deduce

$$\begin{aligned}\|(AA^\top)^{-1}A\|_{op}^4 &= \lambda_{\max}\left(A^\top(AA^\top)^{-3}A\right)\\ &\le \operatorname{tr}\left(A^\top(AA^\top)^{-3}A\right)\\ &= \operatorname{tr}\left((AA^\top)^{-3}AA^\top\right)\\ &= \operatorname{tr}\left((AA^\top)^{-2}\right).\end{aligned}$$

Using the fact that $A_dA_d^\top \preccurlyeq A_{d+1}A_{d+1}^\top$ established above, induction gives

$$(AA^\top)^{-2} \preccurlyeq (A_{n+8}A_{n+8}^\top)^{-2}.$$

It follows that

$$\mathbb{E}\left[\|(AA^\top)^{-1}A\|_{op}^4\right] \le \mathbb{E}\left[\operatorname{tr}\left(\left(A_{n+8}A_{n+8}^\top\right)^{-2}\right)\right] = \operatorname{tr}\left(\mathbb{E}\left[\left(A_{n+8}A_{n+8}^\top\right)^{-2}\right]\right) < +\infty\,, \quad (13)$$

where again we use that fact that inverse Wishart matrix $\left(A_{n+8}A_{n+8}^\top\right)^{-1}$ has finite second moment.

Next, we demonstrate $\mathbb{E}\|x\|^4 < +\infty$. Recall that every $\mathcal{D}_i$ is either a Gaussian or a Gaussian mixture distribution. Therefore, every entry of x has a subgaussian tail, and thus $\mathbb{E}\|x\|^4 < +\infty$. Together with (13) and the fact that x and A are independent, we conclude that

$$\mathbb{E}\|v\|^4 \le \mathbb{E}\left[\|(AA^\top)^{-1}A\|_{op}^4\right]\cdot\mathbb{E}\left[\|x\|^4\right] < +\infty\,.$$

□

C.3 Proof of Theorem 10

Proof. The randomness comes from A, x, y and b. We first condition on A and x being fixed.

Let u and G be as defined in Lemma 8. Define

$$v \triangleq (A^+)^\top x$$

and

$$r \triangleq 1 + b^\top Gb.$$

We compute the left-hand side but take the expectation over only y for the moment

$$\begin{aligned}&\mathbb{E}_y\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2 - \left\|\frac{1}{d}(A^+)^\top x\right\|^2\\ &= \frac{1}{(d+1)^2}\left(\mathbb{E}_y\left\|(I-bu)^\top v + u^\top y\right\|^2 - (1+1/d)^2\|v\|^2\right)\\ &= \frac{1}{(d+1)^2}\left(\|(I-bu)^\top v\|^2 + \mathbb{E}_y\|u^\top y\|^2 - (1+1/d)^2\|v\|^2\right) \qquad (\mathbb{E}[y]=0)\\ &= \frac{1}{(d+1)^2}\left(\|(I-bu)^\top v\|^2 + \mathbb{E}_y[y^2]\frac{\|Gb\|^2}{r^2} - (1+1/d)^2\|v\|^2\right). \qquad (14)\end{aligned}$$

Additionally, we define

$$\begin{aligned}\delta &\triangleq 2/d + 1/d^2\,,\\ H &\triangleq bb^\top\,.\end{aligned}$$

Let us first consider the first and third terms of (14),

$$\begin{aligned}\|(I-bu)^\top v\|^2 - (1+1/d)^2\|v\|^2 &= v^\top\left((I-bu)(I-bu)^\top - (1+1/d)^2 I\right)v\\ &= -v^\top\left(\delta I + bu + u^\top b^\top - buu^\top b^\top\right)v\\ &= -v^\top\left(\delta I + \frac{HG+GH}{r} - \frac{HG^2H}{r^2}\right)v\,. \qquad (15)\end{aligned}$$

The next step is to analyze the eigenvalues of

$$N \triangleq \frac{HG + GH}{r} - \frac{HG^2H}{r^2}\,.$$

Note that H has rank 1. The rank of N is at most 2 and thus N has at most two non-zero eigenvalues. To proceed, we need additional definitions

$$M \triangleq \frac{r^2}{\alpha^2}N, \quad c \triangleq \frac{b}{\|b\|}, \quad g \triangleq \frac{Gb}{\|Gb\|}, \quad \alpha \triangleq \|b\|\|Gb\|,$$
$$p \triangleq c^\top g, \quad \beta \triangleq \frac{r}{\alpha}, \quad t \triangleq \mathrm{tr}(HG) = \alpha p = r - 1\,.$$

Using these definitions, we deduce that c and g are both unit vectors and

$$HG = \alpha c g^\top, \quad GH = \alpha g c^\top, \quad M = \frac{r}{\alpha}\left(cg^\top + gc^\top\right) - cc^\top\,.$$

We study the action of M on c and g:

$$Mc = \frac{r}{\alpha}(pc + g) - c = (\beta p - 1)c + \beta g$$
$$Mg = \frac{r}{\alpha}(c + pg) - pc = (\beta - p)c + \beta p g\,.$$

As a result, $\mathrm{span}\{c, g\}$ is an invariant subspace of M and M is zero on the orthogonal complement of $\mathrm{span}\{c, g\}$. Therefore, all non-zero eigenvalues of M on $\mathbb{R}^n$ are the eigenvalues of M on $\mathrm{span}\{c, g\}$.

The vectors c and g are almost surely linearly independent. We can use them as a basis of $\mathrm{span}\{c, g\}$. Hence, M has the following matrix representation under this basis

$$\begin{bmatrix} \beta p - 1\,, & \beta \\ \beta - p\,, & \beta p \end{bmatrix}.$$

Its eigenvalues are

$$\mu_{1,2} = \frac{1}{2}\left(\pm\sqrt{4\beta^2 - 4\beta p + 1} + 2\beta p - 1\right) = \pm\frac{1}{2}\sqrt{\frac{\alpha^2 + 4r^2 - 4rt}{\alpha^2}} + \frac{rt}{\alpha^2} - \frac{1}{2}\,.$$

Therefore, the non-zero eigenvalues of N are

$$\lambda_{1,2} = \frac{\alpha^2}{r^2}\mu_{1,2} = 1 - \frac{1}{r} - \frac{\alpha\left(\alpha \pm \sqrt{\alpha^2 + 4t + 4}\right)}{2r^2} \geq 1 - \frac{1}{r} - \frac{\alpha(\alpha+1)}{r^2}\,.$$

In the last inequality, we use the Cauchy-Schwarz inequality

$$t = b^\top G b \leq \|b\|\|Gb\| = \alpha.$$

It follows that

$$\begin{aligned}
&\mathbb{E}_b\left[\|(I - bu)^\top v\|^2 - (1 + 1/d)^2\|v\|^2\right] \\
\leq\ & -\left(\delta + 1 - \mathbb{E}_b\left[\frac{1}{r}\right] - \mathbb{E}_b\left[\frac{\alpha(\alpha+1)}{r^2}\right]\right)\|v\|^2 \\
\leq\ & -\left(\delta - \mathbb{E}_b\left[\frac{\lambda_{\max}(G)\|b\|^2(\lambda_{\max}(G)\|b\|^2 + 1)}{(1 + \lambda_{\min}(G)\|b\|^2)^2}\right]\right)\|v\|^2 \qquad (16) \\
\leq\ & -\left(\delta - \mathbb{E}_b\left[\lambda_{\max}(G)\|b\|^2(\lambda_{\max}(G)\|b\|^2 + 1)\right]\right)\|v\|^2\,. \qquad (17)
\end{aligned}$$

Recall $\|b\|^2 \sim \sigma^2\chi^2(n)$, and thus $\mathbb{E}[\|b\|^2] < +\infty$ and $\mathbb{E}[\|b\|^4] < +\infty$. For all $\sigma < 1$, we have

$$\lambda_{\max}(G)\|b\|^2(\lambda_{\max}(G)\|b\|^2 + 1)\|v\|^2 \leq \lambda_{\max}^2(G)\frac{\|b\|^4}{\sigma^4}\|v\|^2 + \lambda_{\max}(G)\frac{\|b\|^2}{\sigma^2}\|v\|^2$$

where the expectation of the right hand side equals

$$\mathbb{E}[\lambda_{\max}^2(G)\|v\|^2]\mathbb{E}[(\chi^2)^2] + \mathbb{E}[\lambda_{\max}(G)\|v\|^2]\mathbb{E}[\chi^2] < +\infty$$

by the assumptions. Lebesgue's dominated convergence theorem then implies

$$\lim_{\sigma\to 0^+} \mathbb{E}_{A,x,y}\mathbb{E}_b\left[\lambda_{\max}(G)\|b\|^2(\lambda_{\max}(G)\|b\|^2+1)\|v\|^2\right] = 0\,.$$

Take full expectation of (17) with respect to A, x, y and b, and let $\sigma \to 0^+$, and we have (17) tends to $-\delta\mathbb{E}\|v\|^2$.

Similarly, we get

$$\lim_{\sigma\to 0+} \mathbb{E}_{A,x,y,b}\left[\frac{y^2\|Gb\|^2}{(1+b^\top Gb)^2}\right] = \lim_{\sigma\to 0+} \mathbb{E}_y[y^2]\mathbb{E}_{A,b}\left[\frac{\|Gb\|^2}{(1+b^\top Gb)^2}\right] = \lim_{\sigma\to 0+} \sigma^2\mathbb{E}_{A,b}\left[\frac{\|Gb\|^2}{(1+b^\top Gb)^2}\right] = 0\,,$$

where in the last step we use the dominated convergence theorem and the fact that

$$\frac{\sigma^2\|Gb\|^2}{(1+b^\top Gb)^2} \le \sigma^2\lambda_{\max}^2(G)\|b\|^2\,, \quad \mathbb{E}[\sigma^2\lambda_{\max}^2(G)\|b\|^2] < +\infty\,,$$

by the assumptions. It follows that

$$\begin{aligned}
&\mathbb{E}\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2 - \mathbb{E}\left\|\frac{1}{d}(A^+)^\top x\right\|^2\\
&= \frac{1}{(d+1)^2}\left(\mathbb{E}\left[\|(I-bu)^\top v\|^2 - (1+1/d)^2\|v\|^2\right] + \mathbb{E}[y^2]\mathbb{E}\left[\frac{\|Gb\|^2}{r^2}\right]\right)\\
&\le \frac{1}{(d+1)^2}\left(-(\delta+o(1))\,\mathbb{E}\|v\|^2 + \sigma^2\mathbb{E}_{A,b}\left[\frac{\|Gb\|^2}{(1+b^\top Gb)^2}\right]\right)\\
&\to -\frac{\delta\mathbb{E}\|v\|^2}{(d+1)^2}\,. \qquad \text{(as } \sigma\to 0+)
\end{aligned}$$

Therefore, there exists $\sigma > 0$ such that

$$\mathbb{E}\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2 - \mathbb{E}\left\|\frac{1}{d}(A^+)^\top x\right\|^2 < -\frac{2\mathbb{E}\|(A^+)^\top x\|^2}{d(d+1)^2} < 0\,. \qquad \text{(because } \delta > 2/d)$$

□

C.4 Proof of Theorem 11

Proof. Again we first condition on A and x being fixed. Let u and G be as defined in Lemma 8. We also define the following variables:

$$v \triangleq (A^+)^\top x,$$

and

$$r \triangleq 1 + b^\top Gb.$$

Recall (14)

$$\begin{aligned}
&\mathbb{E}_y\left\|\frac{1}{d+1}\begin{bmatrix}A^\top\\ b^\top\end{bmatrix}^+\begin{bmatrix}x\\ y\end{bmatrix}\right\|^2 - \left\|\frac{1}{d}(A^+)^\top x\right\|^2\\
&= \frac{1}{(d+1)^2}\left(\|(I-bu)^\top v\|^2 + \mathbb{E}_y[y^2]\frac{\|Gb\|^2}{r^2} - (1+1/d)^2\|v\|^2\right)\,. \qquad (18)
\end{aligned}$$

Our strategy is to make $\mathbb{E}[y^2\frac{\|Gb\|^2}{r^2}]$ arbitrarily big. To this end, by the independence of y and b we have

$$\mathbb{E}_{y,b}\left[y^2\frac{\|Gb\|^2}{r^2}\right] = \mathbb{E}_y[y^2]\mathbb{E}_b\left[\frac{\|Gb\|^2}{r^2}\right]\,.$$

By definition of $\mathcal{N}_{\sigma,\mu}^{\text{mix}}$, with probability $2/3$, y is sampled from either $\mathcal{N}(\mu,\sigma^2)$ or $\mathcal{N}(-\mu,\sigma^2)$, which implies $\mathbb{E}[y^2] \ge \frac{1}{3}\mu^2$. For each b_i, we have

$$\mathbb{P}(|b_i| \in [\sigma, 2\sigma]) \ge \frac{1}{3}\times\frac{1}{4}.$$

Also note that G is positive definite. It follows that

$$\mathbb{E}_b\left[\frac{||Gb||^2}{r^2}\right] = \mathbb{E}_b\left[\frac{||Gb||^2}{(1+b^\top Gb)^2}\right] \geq \mathbb{E}_b \frac{(\lambda_{\min}(G)||b||)^2}{(1+\lambda_{\max}(G)||b||^2)^2} \geq \left(\frac{1}{12}\right)^n \frac{\lambda_{\min}^2(G) n\sigma^2}{(1+4\lambda_{\max}(G)n\sigma^2)^2}\,.$$

Altogether we have

$$\mathbb{E}_{y,b}\left[y^2\frac{\|Gb\|^2}{r^2}\right] \geq \frac{1}{3\cdot 12^n}\frac{n\lambda_{\min}^2(G)\mu^2\sigma^2}{(1+4n\lambda_{\max}(G)\sigma^2)^2}\,.$$

Let $\mu = 1/\sigma^2$ and we have

$$\begin{aligned}\lim_{\sigma\to 0^+}\mathbb{E}\left[y^2\frac{\|Gb\|^2}{r^2}\right] &\geq \lim_{\sigma\to 0^+}\mathbb{E}_{A,x}\mathbb{E}_{y,b}\left[\frac{1}{3\cdot 12^n}\frac{n\lambda_{\min}^2(G)}{\sigma^2(1+4n\lambda_{\max}(G)\sigma^2)^2}\right]\\ &= \mathbb{E}_{A,x}\mathbb{E}_{y,b}\lim_{\sigma\to 0^+}\left[\frac{1}{3\cdot 12^n}\frac{n\lambda_{\min}^2(G)}{\sigma^2(1+4n\lambda_{\max}(G)\sigma^2)^2}\right]\\ &= +\infty\,,\end{aligned}$$

where we switch the order of expectation and limit using the monotone convergence theorem. Taking full expectation over A, x, b and y of (18) and using the assumption that $\mathbb{E}\|v\|^2 < +\infty$ we have

$$L_{d+1} - L_d = \frac{1}{(d+1)^2}\left(\mathbb{E}_{A,x,b}\|(I-bu)^\top v\|^2 + \mathbb{E}\left[y^2\frac{\|Gb\|^2}{r^2}\right] - (1+1/d)^2\mathbb{E}_{A,x}\|v\|^2\right) \to +\infty$$

as $\sigma \to 0^+$. □