Revisiting Optimism and Model Complexity in the Wake of Overparameterized Machine Learning

Pratik Patil[†]
pratikpatil@berkeley.edu

Jin-Hong Du[‡§]
jinhongd@andrew.cmu.edu

Ryan J. Tibshirani[†]
ryantibs@berkeley.edu

Abstract

Common practice in modern machine learning involves fitting a large number of parameters relative to the number of observations. These overparameterized models can exhibit surprising generalization behavior, e.g., "double descent" in the prediction error curve when plotted against the raw number of model parameters, or another simplistic notion of complexity. In this paper, we revisit model complexity from first principles, by first reinterpreting and then extending the classical statistical concept of (effective) *degrees of freedom*. Whereas the classical definition is connected to fixed-X prediction error (in which prediction error is defined by averaging over the same, nonrandom covariate points as those used during training), our extension of degrees of freedom is connected to random-X prediction error (in which prediction error is averaged over a new, random sample from the covariate distribution). The random-X setting more naturally embodies modern machine learning problems, where highly complex models, even those complex enough to interpolate the training data, can still lead to desirable generalization performance under appropriate conditions. We demonstrate the utility of our proposed complexity measures through a mix of conceptual arguments, theory, and experiments, and illustrate how they can be used to interpret and compare arbitrary prediction models.

1 Introduction

Model complexity is a key concept in statistics and machine learning, and is a core consideration in prediction problems—a higher complexity allows for a better fit to the training data, but may result in overfitting, whereas a lower complexity may lack the ability to capture sufficiently rich behavior, and hence lead to underfitting. There are numerous different ways to quantify the complexity of a prediction model. One such way is called the (effective) *degrees of freedom* (Efron, 1983, 1986; Hastie and Tibshirani, 1987) of a model, which is a classical concept in statistics, and will play a central role in our paper. This is often interpreted as the number of "free parameters" in the fitted model.

Meanwhile, driven by the enormous practical successes of neural networks and deep learning, there has recently been great interest in the community in studying *overparameterized models*, where the number of parameters is large relative to the number of observations. Overparameterized models can exhibit surprising generalization behavior, in that they can generalize well even if they perfectly (or nearly) interpolate noisy training data (Zhang et al., 2017; Belkin et al., 2019). As we will explain later (Section 2.3), classical degrees of freedom fails to adequately explain this phenomenon. For example, it is not able to distinguish between interpolating models: the degrees of freedom of any interpolator is exactly n, the number of training observations.

[†] Department of Statistics, University of California, Berkeley.
[‡] Department of Statistics and Data Science, Carnegie Mellon University.
[§] Machine Learning Department, Carnegie Mellon University.

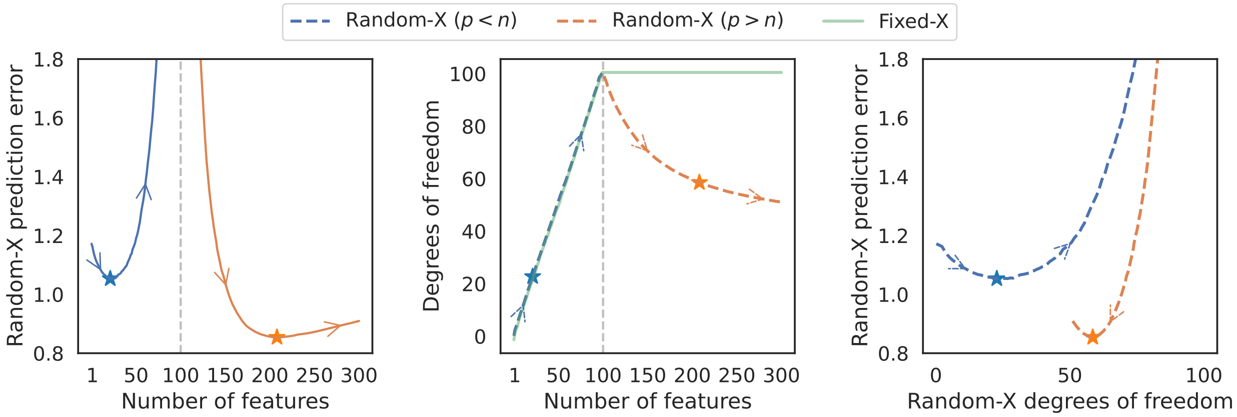

Figure 1: An illustration using ridgeless least squares regression as the prediction model, trained on $n = 100$ samples and p features, where p ranges from 1 to 300. The true conditional mean is a nonlinear function in the features, and hence adding more features to the working linear model helps its approximation capacity (The precise details are given in Appendix C.1). In the left panel, we can see that the random-X prediction error curve exhibits "double descent" in p. In the middle panel, the classical (fixed-X) definition of degrees of freedom increases linearly for $p \leqslant n$, but then it flattens out at the trivial answer of n degrees of freedom for all $p > n$. The "intrinsic" random-X degrees of freedom, one of two basic versions of random-X degrees of freedom to be defined later in Section 3, is *decreasing* when $p > n$, indicating that the ridgeless interpolator is becoming *less* complex as the dimensionality grows. In the right panel, we plot the random-X prediction error as a function of random-X degrees of freedom. The interpretation: our proposed complexity measure maps every overparameterized model onto an equivalent underparameterized model, and the best-predicting model (which lies in the overparameterized regime) actually has relatively low complexity.

The underlying limitation of degrees of freedom, as classically defined, is that it is tied to a measure of prediction error which we refer to (following Rosset and Tibshirani 2020) as *fixed-X* prediction error. In this measure, prediction error is defined by averaging over the same fixed set of covariate points as those used during training. In certain problem settings—that is, low-dimensional, smooth prediction problems—this measure is a good proxy for *random-X* prediction error, which is given by averaging over a new random sample from the covariate distribution. Yet, in high-dimensional and/or nonsmooth prediction problems, fixed-X and random-X errors can behave quite differently. A generalizing interpolator epitomizes this difference (Section 2.1): as $n \to \infty$, it has fixed-X excess error converging to the noise level but random-X excess error converging to zero.

In nearly all modern machine learning prediction problems, random-X error is the perspective of interest. Given its connection to fixed-X error, it should not be surprising that classical degrees of freedom can break down for prediction models such as interpolators, where random-X and fixed-X errors diverge. In this paper, we propose a new measure of degrees of freedom that connects directly to random-X prediction error, and allows us to reason about complexity in a nontrivial way for *any* predictive model, including interpolators. We provide a simple illustration in Figure 1.

1.1 Summary and outline

We provide a summary of our contributions and outline the structure of the paper below.

New random-X measures of degrees of freedom. After we review preliminary materials in Section 2, we present new measures of model complexity in Section 3. In particular, we extend the classical notion of degrees of freedom to the setting of random-X. We do so by first reinterpreting the classical construction of degrees of freedom in a new light, then translating this to random-X prediction error. We propose two basic versions of random-X degrees of freedom: one to capture both bias and variance components of the error, and another based on variance alone.

Basic properties and theory for random-X degrees of freedom. In Section 4, we describe basic properties of the proposed random-X degrees of freedom measures, and draw connections to related ideas in the literature. Section 5 derives theory for a few standard prediction models, such as ridge regression and the lasso, and demonstrates that degrees of freedom typically decreases as the regularization strength increases, and typically increases as the number of features increases.

Numerical experiments for a diverse set of prediction models. In Section 6, we illustrate the versatility of our complexity measures by presenting results from numerical experiments using the lasso, k-nearest neighbors regression, and random forests.

Decomposing degrees of freedom under distribution shift. In Section 7, we discuss how to decompose the random-X degrees of freedom of a prediction model into constituent parts, so as to quantify the contribution of various components—such as bias, variance, and covariate shift—to the final measure of model complexity. This is based on borrowing ideas from Shapley values.

1.2 Related work

There is a lot of literature related to the topic of our paper, which we discuss in two groups.

Model optimism and degrees of freedom. Optimism and (effective) degrees of freedom are classical concepts and well-studied in statistics, with important references being Efron (1983, 1986, 2004). Degrees of freedom for linear regression and linear smoothers have a particular simple form, as the trace of the smoother matrix, and have a long history of study, for example, Mallows (1973); Craven and Wahba (1978); Hastie and Tibshirani (1987, 1990). Broadly related to this is the topic of estimating risk for model selection, which is widely studied and itself carries quite a rich literature, for example, Sclove (1969); Hocking (1976); Akaike (1973); Schwarz (1978); Thompson (1978a,b); Golub et al. (1979); Breiman and Freedman (1983); Breiman and Spector (1992), and many others.

A landmark contribution in the study of degrees of freedom and unbiased risk estimation is known as *Stein's unbiased risk estimator* (SURE), due to Stein (1981). This has enabled the development of numerous closed-form unbiased estimators of degrees of freedom (and fixed-X prediction error) for methods such as wavelet denoising, shape-constrained regression, quantile regression, lasso and various generalizations, and low-rank matrix factorization; see, for example, Donoho and Johnstone (1995); Cai (1999); Meyer and Woodroofe (2000); Zou et al. (2007); Zou and Yuan (2008); Tibshirani and Taylor (2012); Candès et al. (2013); Tibshirani (2015); Mikkelsen and Hansen (2018); Chen et al. (2020), among others. For an alternative perspective based on auxiliary randomization (which reduces to SURE in a limiting case), see Oliveira et al. (2021, 2022).

The above literature is all rooted in the fixed-X setting, which (as we will explain precisely in the next section) measures prediction error at the same fixed covariate points as those used in training. Rosset and Tibshirani (2020) compare and contrast the bias-variance tradeoff, prediction error, and other core concepts in statistical decision theory in the fixed-X and random-X settings. Our work builds on theirs and introduces a notion of random-X degrees of freedom. Though we believe that this should be of general interest, it is of particular interest for interpolators.

Closely related to our proposed complexity measure is the recent work of Luan et al. (2021, 2022); Curth et al. (2023). They propose a measure of random-X degrees of freedom that is suitable for linear smoothers. It is related to our approach in this special case, and Section 4.4 provides details. Broadly speaking, our approach is more general (accommodates arbitrary prediction models), and also, allows for both bias and variance components of the random-X optimism to enter into the complexity measure, whereas the previous proposals focus on variance alone.

Other complexity measures. There are many other criteria for measuring the complexity of a model or an object. Broadly, this includes ideas from information theory and theoretical computer science, such as Kolmogorov complexity (Kolmogorov, 1963), minimum message length (Wallace and Boulton, 1968), and minimum description length (Rissanen, 1978). Closer to our study, coming from machine learning theory, are Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 1971) and Rademacher complexity (Bartlett and Mendelson, 2002). For a discussion of these concepts and their role in generalization theory, see, for example, Shalez-Shwartz and Ben-David (2014) or Mohri et al. (2018). An important point to clarify is that VC dimension and Rademacher complexity differ from degrees of freedom in the following sense: the former measures apply to a *class* of prediction models, whereas the latter applies to a particular *fitted* prediction model. In other words, degrees of freedom as complexity measure is more finely-tuned to the *way* in which a given model is trained, incorporating the action of the fitting algorithm, and the distribution of the underlying data. As an example, a linear model trained via least squares and ridge regression (using strong regularization) will have the same Rademacher complexity, but different degrees of freedom.

2 Preliminaries

We start with a review of fixed-X and random-X prediction error, and classical (fixed-X) optimism and degrees of freedom. Then we discuss the limitations of classical degrees of freedom with respect to understanding overparameterized models.

2.1 Fixed-X and random-X prediction error

Consider a standard regression setup, with independent and identically distributed (i.i.d.) training samples $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, which follow the relationship

$$y_i = f(x_i) + \varepsilon_i, \quad i \in [n], \tag{1}$$

for $f(x) = \mathbb{E}[y_i | x_i = x]$, and i.i.d. mean zero stochastic errors ε_i, $i \in [n]$. We assume that each ε_i is independent of x_i. Here and throughout, we abbreviate $[n] = \{1, \dots, n\}$. Also, let $\sigma^2 = \mathrm{Var}[\varepsilon_i] > 0$ denote the error variance, let $X \in \mathbb{R}^{n \times p}$ denote the feature matrix (with i^{th} row x_i), and let $y \in \mathbb{R}^n$ denote the response vector (with i^{th} entry y_i).

Suppose that we have a model fitting procedure $\widehat{f}$ which produces the predictor $\widehat{f}(\cdot; X, y) : \mathbb{R}^p \to \mathbb{R}$ when trained on the data (X, y). Thus, $\widehat{f}(x; X, y)$ is an estimate of $f(x)$. When the training data is clear from the context, we will simply write this as $\widehat{f}(x)$.

In *fixed-X* prediction error, we measure the error of $\widehat{f}$ at a set of new response values y_i^*, $i \in [n]$, where each y_i^* and y_i are i.i.d. conditional on x_i. Formally, this is

$$\mathsf{err}_{\mathrm{F}}(\widehat{f}) = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^{n} \big(y_i^* - \widehat{f}(x_i)\big)^2 \,\Big|\, X\right], \tag{2}$$

In *random-X* prediction error, we measure the error of $\widehat{f}$ at a new sample $(x_0, y_0) \in \mathbb{R}^p \times \mathbb{R}$, which is i.i.d. to the training samples (x_i, y_i), $i \in [n]$. Formally, this is

$$\mathsf{err}_{\mathrm{R}}(\widehat{f}) = \mathbb{E}\big[\big(y_0 - \widehat{f}(x_0)\big)^2\big]. \tag{3}$$

To be clear, the expectation in (2) is taken with respect to y, y^*, and is conditional on X, whereas that in (3) is taken with respect to X, y, x_0, y_0.

While random-X prediction error is the central object of interest in machine learning theory and in many modern statistics problems, fixed-X prediction error has a long history of study in statistics; we refer to Rosset and Tibshirani (2020) (and references therein) for an in-depth discussion. For our purposes, to motivate our study, it suffices to make only high-level comments to compare them. For smooth functions $f, \widehat{f}$ in low dimensions (i.e., n large compared to p), one can generally expect $\mathsf{err}_{\mathrm{F}}(\widehat{f})$ and $\mathsf{err}_{\mathrm{R}}(\widehat{f})$ to behave similarly. For example, empirical process theory offers uniform control on the deviation between the L^2 norms based on taking a sample average over i.i.d. draws x_i, $i \in [n]$, and taking an expectation with respect to $x_0 \sim P_x$. Such results can be used to derive an asymptotic equivalence (and nonasymptotic bounds) between $\mathsf{err}_{\mathrm{F}}(\widehat{f})$ and $\mathsf{err}_{\mathrm{R}}(\widehat{f})$ in certain settings.

However, for nonsmooth functions and/or high-dimensional problem settings, the two metrics can behave quite differently. Consider, as an example, a generalizing interpolator: here, we would have random-X excess error $\mathsf{err}_{\mathrm{R}}(\widehat{f}) - \mathsf{err}_{\mathrm{R}}(f) \to 0$ as $n \to \infty$, but fixed-X excess error

$$\mathsf{err}_{\mathrm{F}}(\widehat{f}) - \mathsf{err}_{\mathrm{F}}(f) = \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^n (y_i^* - y_i)^2 \,\bigg|\, X\bigg] - \sigma^2 = \sigma^2,$$

where recall $\sigma^2 = \mathrm{Var}[\varepsilon_i]$ in the data model (1). This represents a huge difference between the two metrics: one vanishing, and the other pinned at the noise level.

2.2 Fixed-X optimism and degrees of freedom

The (effective) *degrees of freedom* of $\widehat{f}$ is defined as

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}) = \frac{1}{\sigma^2}\sum_{i=1}^n \mathrm{Cov}[y_i, \widehat{f}(x_i) \,|\, X]. \tag{4}$$

This is often motivated intuitively as follows: the more complex the fitting procedure $\widehat{f}$, the more "self-influence" each response y_i will have on the corresponding fitted value $\widehat{f}(x_i)$ (and hence the higher the degrees of freedom in total). An important property of degrees of freedom is its intimate connection to *fixed-X* optimism, which is defined as

$$\mathsf{opt}_{\mathrm{F}}(\widehat{f}) = \mathsf{err}_{\mathrm{F}}(\widehat{f}) - \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^n \big(y_i - \widehat{f}(x_i)\big)^2 \,\bigg|\, X\bigg]. \tag{5}$$

The second quantity on the right-hand side above is simply the training error (conditional on X). The precise connection between (4) and (5) is given by what is sometimes called *Efron's optimism theorem*, attributed to Efron (1986, 2004):

$$\mathsf{opt}_{\mathrm{F}}(\widehat{f}) = \frac{2\sigma^2}{n}\mathsf{df}_{\mathrm{F}}(\widehat{f}). \tag{6}$$

This holds without any assumptions on $\widehat{f}$, and can be checked via simple algebra (add and subtract y_i^* within the square in each summand in $\mathsf{err}_{\mathrm{F}}(\widehat{f})$ in (2), then expand and simplify).

The rest of this subsection can be skipped without interrupting the flow of main ideas. We use it as an opportunity to provide general context about classical interest in degrees of freedom, as alluded to in the related work subsection. *Stein's lemma* (Stein, 1981) says if $\widehat{f}$ is weakly differentiable as a function of y, and we assume Gaussian errors ε_i, $i \in [n]$ in (1), then

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}) = \mathbb{E}\bigg[\sum_{i=1}^n \frac{\partial \widehat{f}(x_i)}{\partial y_i} \,\bigg|\, X\bigg]. \tag{7}$$

Based on (7), we are able to form an unbiased estimate of $\mathsf{df}_{\mathrm{F}}(\widehat{f})$, namely, $\widehat{\mathsf{df}}_{\mathrm{F}} = \sum_{i=1}^n \partial \widehat{f}(x_i)/\partial y_i$ (if we are able to compute it). From (5) and (6), we see that this in turn provides an unbiased estimate of fixed-X prediction error, namely, $\frac{1}{n}\sum_{i=1}^n (y_i - \widehat{f}(x_i))^2 + 2\sigma^2 \widehat{\mathsf{df}}_{\mathrm{F}}$.

Thus we can see that there is a clear interest in estimating degrees of freedom, and utilizing Stein's formula, in order to estimate fixed-X prediction error. However, this is not really aligned with the general focus of our paper henceforth, and our paper actually proceeds in the opposite direction: we will presume an estimate of prediction error in order to estimate degrees of freedom. As we will see in Section 3, this is a fruitful way to extend degrees of freedom past the fixed-X setting.

2.3 Limitations of classical degrees of freedom

A critical limitation of classical (fixed-X) degrees of freedom, as defined in (4), is straightforward to state. For any interpolator, satisfying $\widehat{f}(x_i) = y_i$, $i \in [n]$, we have the trivial answer:

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}) = \frac{1}{\sigma^2}\sum_{i=1}^n \mathrm{Cov}[y_i, y_i \,|\, X] = n. \tag{8}$$

If characterizing fixed-X optimism is truly the end goal of degrees of freedom, then we should not be bothered by this (seemingly) obvious fact since any interpolator has zero training error and the same fixed-X prediction error. Yet, if we are to think of degrees of freedom as a general measure of model complexity, then (8) leaves a lot to be desired. As we know from the recent wave of work in machine learning and statistics (for example, see the review articles Belkin (2021); Bartlett et al. (2021) and references therein), some interpolators—in particular, implicitly regularized ones—are actually quite well-behaved and can generalize well to unseen data. In classical degrees of freedom, thus, we are lacking a complexity measure that can distinguish between well-behaved interpolators, which are smooth in between the covariate points, and wild ones, which are arbitrarily nonsmooth.

The next section develops an extension of the classical notion of degrees of freedom which connects to random-X (rather than fixed-X) prediction error. As we will see, the extension will overcome the limitation just described—the new notion will assign a meaningful complexity measure to every prediction model, including interpolators.

3 Random-X degrees of freedom

In this section, we first present a fresh reinterpretation of fixed-X degrees of freedom. Then we show how this leads to a generalization of degrees of freedom in the random-X setting.

3.1 Reinterpreting fixed-X degrees of freedom

We first recall a standard fact about fixed-X degrees of freedom: if the feature matrix $X \in \mathbb{R}^{n \times p}$ has linearly independent columns, then least squares regression of y on X, given by $\widehat{f}^{\mathrm{ls}}(x) = x^\top \widehat{\beta}^{\mathrm{ls}}$ where

$\widehat{\beta}^{\mathrm{ls}} = (X^\top X)^{-1} X^\top y$, has degrees of freedom exactly p. This is simply the number of parameters in $\widehat{\beta}^{\mathrm{ls}}$. This fact is easily verified from (4), abbreviating $P_X = X(X^\top X)^{-1} X^\top$:

$$\begin{aligned} \mathsf{df}_{\mathrm{F}}(\widehat{f}^{\mathrm{ls}}) &= \frac{1}{\sigma^2} \operatorname{tr}(\operatorname{Cov}[X\widehat{\beta}^{\mathrm{ls}}, y \,|\, X]) \\ &= \frac{1}{\sigma^2} \operatorname{tr}(\operatorname{Cov}[P_X y, y \,|\, X]) \\ &= \operatorname{tr}(P_X) \qquad (9) \\ &= p, \qquad (10) \end{aligned}$$

where we used $\operatorname{Cov}[P_X y, y \,|X] = P_X \operatorname{Cov}[y|X] = \sigma^2 P_X$ in the second-to-last line, and we used the cyclic property $\operatorname{tr}(P_X) = \operatorname{tr}(X^\top X (X^\top X)^{-1}) = p$ in the last line.

Now we show that the fact about least squares in (9), which is well-known in the literature, can be used to reinterpret fixed-X degrees of freedom in a new light. Recalling Efron's optimism formula (6), the least squares regression predictor $\widehat{f}^{\mathrm{ls}}$ has fixed-X optimism

$$\mathsf{opt}_{\mathrm{F}}(\widehat{f}^{\mathrm{ls}}) = \frac{2\sigma^2}{n} p.$$

Given an arbitrary predictor $\widehat{f}$, we know that it still satisfies (copying (6) here for convenience)

$$\mathsf{opt}_{\mathrm{F}}(\widehat{f}) = \frac{2\sigma^2}{n} \mathsf{df}_{\mathrm{F}}(\widehat{f}).$$

Comparing the last two displays, we see that we may hence interpret the degrees of freedom of $\widehat{f}$ as the value of $d \in [0, \infty]$ for which least squares predictor on d linearly independent features has the same fixed-X optimism as $\mathsf{opt}_{\mathrm{F}}(\widehat{f})$. This is simply a reformulation of the original definition (4), and the next proposition records this idea precisely.

Proposition 1. *For each fixed $d \leqslant n$, let $\widetilde{X}_d \in \mathbb{R}^{n \times d}$ be an arbitrary feature matrix having linearly independent columns, and consider $\widehat{f}^{\mathrm{ls}}(\cdot; \widetilde{X}_d, y)$, the predictor from least squares regression of y on $\widetilde{X}_d$, which we call our "reference" model, and abbreviate as $\widehat{f}^{\mathrm{ref}}_d$. This satisfies*

$$\mathsf{opt}_{\mathrm{F}}(\widehat{f}^{\mathrm{ref}}_d) = \frac{2\sigma^2}{n} d, \quad d = 1, \dots, n. \qquad (11)$$

Let us extend these reference values so that we may write for all nonnegative d,

$$\overline{\mathsf{opt}}_{\mathrm{F}}(\widehat{f}^{\mathrm{ref}}_d) = \frac{2\sigma^2}{n} d, \quad d \in [0, \infty]. \qquad (12)$$

Given an arbitrary predictor $\widehat{f} = \widehat{f}(\cdot; X, y)$, define d to be the unique nonnegative number for which

$$\mathsf{opt}_{\mathrm{F}}(\widehat{f}) = \overline{\mathsf{opt}}_{\mathrm{F}}(\widehat{f}^{\mathrm{ref}}_d). \qquad (13)$$

Then $\mathsf{df}_{\mathrm{F}}(\widehat{f}) = d$.

Proof. The proof is immediate. The left-hand side in (13) equals $(2\sigma^2/n)\mathsf{df}_{\mathrm{F}}(\widehat{f})$ and the right-hand side equals $(2\sigma^2/n)d$. Cancelling the common factor of $2\sigma^2/n$ gives the result. □

Next we show how to lift this idea to the random-X setting.

3.2 Defining random-X degrees of freedom

The idea behind Proposition 1 is both fairly natural and fairly general. To cast the core idea at a high level, in order to define the complexity of a given prediction model $\widehat{f}$, we require two things:

i. a *metric* met, which we assume (without loss of generality) is negatively-oriented: the lower the value of $\mathsf{met}(\widehat{f})$, the less complex we deem $\widehat{f}$;

ii. a *reference class* $\{\widehat{f}^{\mathrm{ref}}_d : d \in D\}$, which is a class of models indexed by a number of parameters d, assumed to be "canonical" in some sense to the prediction task at hand.

We then assign to $\widehat{f}$ a complexity of d where d is smallest value in D for which $\mathsf{met}(\widehat{f}) \leqslant \mathsf{met}(\widehat{f}^{\mathrm{ref}}_d)$. In other words, it is defined to be the number of parameters in the smallest reference model whose metric value is at least that of $\widehat{f}$.

Fixed-X degrees of freedom is a special case of this general recipe, in which the metric is implicitly taken to be fixed-X optimism—but suitably extended so that this metric ranges over the full set of nonnegative reals, and we can always achieve equality: $\mathsf{met}(\widehat{f}) = \mathsf{met}(\widehat{f}^{\mathrm{ref}}_d)$ for some $d \geqslant 0$. The reference class is taken to be least squares regression on an arbitrary full rank feature matrix.

Towards a random-X extension, a natural inclination would be to maintain least squares regression as the reference class, and simply replace fixed-X optimism (5) with random-X optimism, defined as

$$\mathsf{opt}_{\mathrm{R}}(\widehat{f}) = \mathsf{err}_{\mathrm{R}}(\widehat{f}) - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\big(y_i - \widehat{f}(x_i)\big)^2\right]. \tag{14}$$

This is now the random-X prediction error (rather than the fixed-X error) minus the training error. Before we pursue a random-X extension, it is important to note that the classical definition, which uses least squares and fixed-X optimism in the equivalent characterization given in Proposition 1, is special for two reasons. The metric assigned to the reference model here, i.e., the fixed-X optimism (11) of least squares, depends neither on X nor on the law of $y|X$, beyond assuming isotropic errors (as we have done throughout, i.e., $\mathrm{Cov}[y|X] = \sigma^2 I$, with I being the $n \times n$ identity matrix).

In comparison, the random-X optimism (14) of least squares regression of y on X depends on both the distribution of X and of $y|X$. This means that we will have to be more precise in defining the distribution of the data on which we measure the random-X optimism of least squares, so that this quantity becomes well-defined. The next definition provides details.

Definition 1. *Assume that $n \geqslant 2$. For each fixed $d \leqslant n-1$, let $\widetilde{X}_d \in \mathbb{R}^{n\times d}$ have i.i.d. rows from $\mathcal{N}(0,\Sigma)$, with $\Sigma \in \mathbb{R}^{d\times d}$ an arbitrary deterministic positive definite covariance matrix. Let*

$$\widetilde{y}|\widetilde{X}_d \sim \mathcal{N}(\widetilde{X}_d\beta, \sigma^2 I), \tag{15}$$

with $\beta \in \mathbb{R}^d$ an arbitrary deterministic coefficient vector. Consider $\widehat{f}^{\mathrm{ls}}(\cdot;\widetilde{X}_d,\widetilde{y})$, the predictor from least squares regression of $\widetilde{y}$ on $\widetilde{X}_d$, as our reference model, which we abbreviate as $\widehat{f}^{\mathrm{ref}}_d$. We have

$$\mathsf{opt}_{\mathrm{R}}(\widehat{f}^{\mathrm{ref}}_d) = \sigma^2\left(\frac{d}{n} + \frac{d}{n-d-1}\right), \quad d = 1,\dots,n-1. \tag{16}$$

Let us extend these reference values so that we may write

$$\overline{\mathsf{opt}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ref}}_d) = \sigma^2\left(\frac{d}{n} + \frac{d}{n-d-1}\right), \quad d \in [0, n-1]. \tag{17}$$

Then, given an arbitrary predictor $\widehat{f} = \widehat{f}(\cdot; X, y)$, we define $\mathsf{df}_{\mathrm{R}}(\widehat{f}) = d$ as the unique $d \in [0, n-1]$ for which

$$\mathsf{opt}_{\mathrm{R}}(\widehat{f}) = \overline{\mathsf{opt}}_{\mathrm{R}}(\widehat{f}_d^{\mathrm{ref}}). \tag{18}$$

The result in (16) is driven by the random-X prediction error of least squares regression for jointly Gaussian data, which is well-known, and can be found in, e.g., Stein (1960); Tukey (1967); Hocking (1976); Thompson (1978a,b); Dicker (2013); Rosset and Tibshirani (2020), among others. We give a derivation in Appendix A.1 for completeness.

Several remarks are in order, to discuss random-X degrees of freedom as defined in Definition 1 and compare it to the classical notion of fixed-X degrees of freedom.

- Fixed-X degrees of freedom ranges from 0 to ∞.[1] That is, we cannot rule out arbitrarily large values of fixed-X degrees of freedom, a property that has been criticized by some authors (e.g., Janson et al. (2015)). In contrast, random-X degrees of freedom ranges from 0 to $n-1$. The reason for this is that the random-X optimism of least squares diverges at $d = n-1$, whereas the fixed-X optimism does not (and only diverges as $d \to \infty$). In other words, the random-X optimism of least squares sweeps the entire range of possible optimism values as we vary the number of features from 0 to $n-1$, and this places a finite upper limit on random-X degrees of freedom of $n-1$, achieved when the given predictor has infinite random-X optimism.

- The two metrics used in defining fixed-X and random-X degrees of freedom, namely, fixed-X and random-X optimism, scale differently with the number of parameters d in the underlying reference model, least squares regression. As we can see, (12) scales linearly with d, whereas (17) scales nonlinearly. For large d (close to n), the latter demonstrates "diminishing returns": large increases in random-X optimism only contribute small increases in random-X degrees of freedom. Figure 2 gives an illustration.

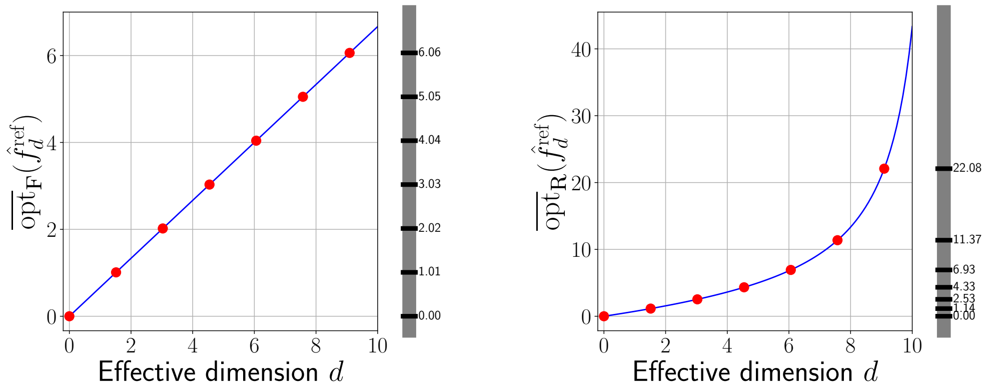

Figure 2: An illustration of the metrics that underlie fixed-X and random-X degrees of freedom: fixed-X and random-X optimism of least squares regression on d features.

- The choice of Gaussian features $\widetilde{X}_d$ in Definition 1 facilitates the calculation of the random-X optimism of least squares regression (16), since we can leverage well-known properties of the (inverse) Wishart distribution. Interestingly, we can see that the result (16) does not depend on the feature covariance Σ. By standard arguments in random matrix theory, as explained in Section 3.4, the formula (16) remains asymptotically valid (as $d/n \to \xi < 1$) for a broad class of feature models.

[1] In fact, negative values are also allowed, but we implicitly rule this out in Proposition 1.

- The linear mean $\mathbb{E}[\widetilde{y}|\widetilde{X}_d] = \widetilde{X}_d\beta$ in Definition 1 is important, but the assumption of Gaussian errors in (15) is not. The calculations in Appendix A.1 actually only assume isotropic errors (i.e., $\widetilde{y} = \widetilde{X}_d\beta + v$, where $v|\widetilde{X}_d$ has mean zero and covariance $\sigma^2 I$). Moreover, the random-X optimism (16) does not depend on the underlying signal vector β (due to the unbiasedness of underparameterized least squares regression), and only depends on the noise level σ^2.

3.3 An intrinsic version of model complexity

The reference model we use in Definition 1 is least squares regression on *well-specified* data, where the mean is linear in the covariates, as can be seen in (15). As previously commented (and verified in Appendix A.1), the least squares predictor is unbiased in this case, and its random-X prediction error and thus random-X optimism is comprised of pure variance.

Therefore, when we match the observed optimism to the reference one in (18), we are comparing $\mathsf{opt}_{\mathrm{R}}(\widehat{f})$—which is generically comprised of both bias and variance, to $\overline{\mathsf{opt}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ref}}_d)$—which is made up of variance alone. This is intentional. The notion of random-X degrees of freedom from Definition 1 determines the complexity of the given predictor $\widehat{f}$ by incorporating the "full effect" of the data at hand, allowing for potential model misspecification to enter into the calculation of optimism. To emphasize, we will sometimes refer to this as the *emergent* random-X degrees of freedom.

Alternatively, we might want to match variance to variance in determining degrees of freedom, i.e., we might want to exclude bias effects in calculating the random-X optimism of the given model $\widehat{f}$. This gives rise to a different notion of model complexity, which we define next.

Definition 2. *Under the exact same setup as in Definition 1, draw $v \sim \mathcal{N}(0, \sigma^2 I)$, independent of everything else. We define $\mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}) = d$ to be the unique $d \in [0, n-1]$ for which*

$$\mathsf{opt}_{\mathrm{R}}(\widehat{f}(\cdot; X, v)) = \overline{\mathsf{opt}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ref}}_d). \tag{19}$$

The difference between (18), (19) is that the latter measures the random-X optimism of $\widehat{f}$ when it is being trained and tested on "pure noise" $v \sim \mathcal{N}(0, \sigma^2 I)$. Because the random-X optimism of least squares does not depend on β in (15), note that we may set $\beta = 0$ and write (19) equivalently as

$$\mathsf{opt}_{\mathrm{R}}(\widehat{f}(\cdot; X, v)) = \overline{\mathsf{opt}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ls}}(\cdot; \widetilde{X}_d, v)).$$

We call the quantity $\mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f})$ in Definition 2 the *intrinsic* random-X degrees of freedom of $\widehat{f}$. It can be interpreted as the model complexity that is intrinsic or inherent to the model $\widehat{f}$, a reflection of its ability to overfit to pure noise (calibrated to that of least squares).

In what follows, we will further examine the relationship between emergent and intrinsic random-X degrees of freedom, and learn through theory and experiments that the emergent notion is generally larger than the intrinsic one. In short, the presence of bias generally "adds complexity".

3.4 Universality of random-X optimism for least squares

As is well-known to those versed in random matrix theory, the random-X prediction error of least squares regression, for well-specified, underparameterized data models, displays a remarkable degree of universality. This is studied in, e.g., Girko (1990, 1995); Verdu and Shamai (1997); Verdu (1998); Tse and Hanly (1999); Tse and Zeitouni (2000); Serdobolskii (2001, 2002), among others. Thus, the random-X optimism also has a universal limit under proportional asymptotics, as noted in Rosset and Tibshirani (2020). For completeness, we relay this precisely below.

Theorem 2. *Assume $\widetilde{X}_d = Z\Sigma^{1/2}$ where $Z \in \mathbb{R}^{n\times d}$ has i.i.d. entries with zero mean, unit variance, and bounded moments up to order $4+\delta$ for some $\delta > 0$, and $\Sigma \in \mathbb{R}^{d\times d}$ is an arbitrary deterministic positive definite covariance matrix. Also assume for an arbitrary deterministic signal vector $\beta \in \mathbb{R}^d$,*

$$\widetilde{y} = \widetilde{X}_d \beta + v, \quad \text{where } \mathbb{E}[v|\widetilde{X}_d] = 0 \text{ and } \mathrm{Cov}[v|\widetilde{X}_d] = \sigma^2 I.$$

Then as $n, d \to \infty$ such that $d/n \to \xi \in (0,1)$, we have, almost surely with respect to $\widetilde{X}_d$,

$$\mathsf{opt}_{\mathrm{R}}(\widehat{f}^{\mathrm{ls}}(\cdot; \widetilde{X}_d, \widetilde{y}) \,|\, \widetilde{X}_d) \to \sigma^2\left(\xi + \frac{\xi}{1-\xi}\right),$$

where $\mathsf{opt}_{\mathrm{R}}(\widehat{f}^{\mathrm{ls}}(\cdot; \widetilde{X}_d, \widetilde{y}) \,|\, \widetilde{X}_d) = \mathbb{E}[(\widetilde{y}_0 - \widetilde{x}_0^\top \widehat{\beta}^{\mathrm{ls}})^2 - \|\widetilde{y} - \widetilde{X}_d \widehat{\beta}^{\mathrm{ls}}\|_2^2/n \,|\, \widetilde{X}_d]$ denotes the random-X optimism conditional on $\widetilde{X}_d$ (and $(\widetilde{x}_0, \widetilde{y}_0)$ is a test point that is i.i.d. to the training data $(\widetilde{X}_d, \widetilde{y})$).

Proof. Following the calculations in Appendix A.1 leads to

$$\begin{aligned} \mathsf{opt}_{\mathrm{R}}(\widehat{f}^{\mathrm{ls}}(\cdot; \widetilde{X}_d, \widetilde{y}) \,|\, \widetilde{X}_d) &= \sigma^2\big(d/n + \mathrm{tr}[\Sigma(\widetilde{X}_d^\top \widetilde{X}_d)^{-1}]\big) \\ &= \sigma^2\big(d/n + \mathrm{tr}[(Z^\top Z)^{-1}]\big). \end{aligned}$$

Under the assumptions in the theorem, the quantity

$$\mathrm{tr}[(Z^\top Z)^{-1}] = \frac{d}{n} \cdot \frac{1}{d}\,\mathrm{tr}\left[\left(\frac{Z^\top Z}{n}\right)^{-1}\right]$$

has a universal limit, almost surely with respect to Z; see, e.g., Theorem 3.10 of Bai and Silverstein (2010). Again from the calculations in Appendix A.1, if the entries of Z are i.i.d. standard Gaussian, then

$$\mathbb{E}\big[\,\mathrm{tr}[(Z^\top Z)^{-1}]\big] = \frac{d}{n-d-1}.$$

This converges to $\xi/(1-\xi)$ as $d/n \to \xi$, which must thus also be the universal almost sure limit in the general case, regardless of the distribution of entries of Z. This yields the almost sure limit of the conditional optimism

$$\mathsf{opt}_{\mathrm{R}}(\widehat{f}^{\mathrm{ls}}(\cdot; \widetilde{X}_d, \widetilde{y}) \,|\, \widetilde{X}_d) \to \sigma^2\left(\xi + \frac{\xi}{1-\xi}\right),$$

as claimed. □

Theorem 2 reveals that the choice of Gaussian features in the reference optimism calculation, for either Definition 1 or Definition 2, is in a certain sense unimportant, because all feature models of the form described in the theorem lead to the same asymptotic answer anyway.

3.5 Practical calculation of random-X degrees of freedom

The concept of random-X degrees of freedom, from Definition 1, is a population-level quantity—it depends on the random-X optimism $\mathsf{opt}_{\mathrm{R}}(\widehat{f})$, which of course itself depends on the (unknown) joint distribution of the features and response. To estimate $\mathsf{df}_{\mathrm{R}}(\widehat{f})$ in practice, we need to first estimate $\mathsf{opt}_{\mathrm{R}}(\widehat{f})$, which we can do by estimating random-X prediction error using (say) cross-validation and then subtracting off the observed training error. We also need to estimate the noise level σ^2, which is an equally (if not more) difficult task, but as a proxy we can use the random-X prediction error of the best-predicting model we have for the task at hand. Given such estimates $\widehat{\mathsf{opt}}_{\mathrm{R}}(\hat{f})$ and $\hat{\sigma}^2$, we set up the sample analog of the matching equation (18),

$$\widehat{\mathsf{opt}}_{\mathrm{R}}(\hat{f}) = \hat{\sigma}^2\bigg(\frac{d}{n} + \frac{d}{n-d-1}\bigg), \tag{20}$$

solve for d, and set $\widehat{\mathsf{df}}_{\mathrm{R}}(\hat{f}) = d$.

To estimate intrinsic random-X degrees of freedom, from Definition 2, we can follow the analogous steps. The only difference is that we train the predictor $\hat{f}$ on pure noise $v \sim \mathcal{N}(0, \hat{\sigma}^2 I)$ (instead of the original response y) which alters our estimates of both random-X prediction error and training error. We set up the sample analog of the matching equation (19),

$$\widehat{\mathsf{opt}}_{\mathrm{R}}(\hat{f}(\cdot; X, v)) = \hat{\sigma}^2\bigg(\frac{d}{n} + \frac{d}{n-d-1}\bigg), \tag{21}$$

solve for d, and set $\widehat{\mathsf{df}}^{\mathsf{i}}_{\mathrm{R}}(\hat{f}) = d$.

Lastly, just to emphasize, we do not require the (estimated) random-X degrees of freedom to be an integer in any of (18), (19), (20), (21). If desired, then one could of course achieve this taking the integer ceiling $\lceil d \rceil$ of the solution d to the given matching equation. We find this unnecessary; note that fixed-X degrees of freedom as originally defined in (4) is also not restricted to be an integer.

4 Properties and connections

We develop some basic properties of the random-X degrees of freedom proposals from the previous section, and make connections to related ideas in the literature.

4.1 Mapping optimism to degrees of freedom

Reflecting on the matching equations (18), (19), (20), (21), each one is an equation of the form

$$x = \frac{d}{n} + \frac{d}{n-d-1}.$$

The above is a quadratic equation in d. It is straightforward to check that it has a unique solution in $[0, n-1]$ which we can write as $d = \omega_n(x)$, where

$$\omega_n(x) = \frac{2n - 1 + nx - \sqrt{(2n-1+nx)^2 - 4(n-1)nx}}{2}. \tag{22}$$

The function $\omega_n(x)$ is a map from normalized optimism x to degrees of freedom d. It is increasing, concave, and ranges from 0 (at $x = 0$) to $n-1$ (as $x \to \infty$). Each of the definitions of (estimated) random-X degrees of freedom from the last section, given by solving (18), (19), (20), or (21), can be written concisely in terms of ω_n, and differ only in the form of normalized optimism that they use:

$$\begin{aligned} \mathsf{df}_{\mathrm{R}}(\hat{f}) &= \omega_n\big(\mathsf{opt}_{\mathrm{R}}(\hat{f})/\sigma^2\big), & \mathsf{df}^{\mathsf{i}}_{\mathrm{R}}(\hat{f}) &= \omega_n\big(\mathsf{opt}^{\mathsf{i}}_{\mathrm{R}}(\hat{f})/\sigma^2\big), \\ \widehat{\mathsf{df}}_{\mathrm{R}}(\hat{f}) &= \omega_n\big(\widehat{\mathsf{opt}}_{\mathrm{R}}(\hat{f})/\hat{\sigma}^2\big), & \widehat{\mathsf{df}}^{\mathsf{i}}_{\mathrm{R}}(\hat{f}) &= \omega_n\big(\widehat{\mathsf{opt}}^{\mathsf{i}}_{\mathrm{R}}(\hat{f})/\hat{\sigma}^2\big). \end{aligned}$$

Here and henceforth we write $\mathsf{opt}^{\mathsf{i}}_{\mathrm{R}}(\hat{f}) = \mathsf{opt}_{\mathrm{R}}(\hat{f}(; \cdot, \tilde{X}_d, v))$ for convenience, and will refer to this as intrinsic random-X optimism (and similarly for the estimated version).

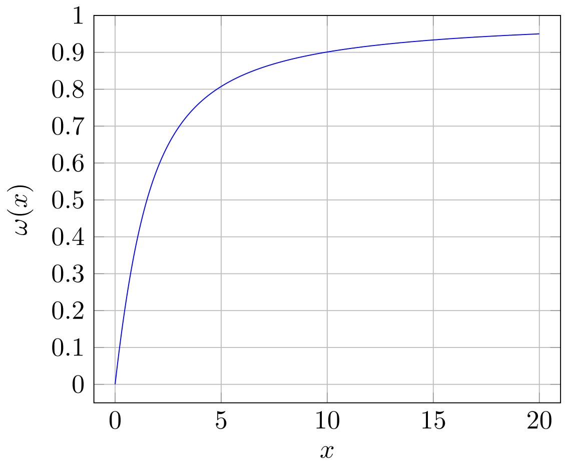

Figure 3: Plot of ω in (23), which maps from normalized optimism (optimism divided by σ^2) to normalized degrees of freedom (degrees of freedom divided by $n-1$).

For large n, the function ω_n in (23) is well-approximated by $\omega_n(x) \approx n \cdot \omega(x)$, where

$$\omega(x) = 1 + \frac{x}{2} - \sqrt{1 + \frac{x^2}{4}}. \tag{23}$$

This function is increasing, concave, and ranges from 0 (at $x = 0$) to 1 (as $x \to \infty$). See Figure 3 for a visualization. The precise relationship between ω_n and ω is that, for any fixed x,

$$|\omega_n(x)/n - \omega(x)| \to 0, \quad \text{as } n \to \infty, \tag{24}$$

which is verified in Appendix A.2.

Finally, a calculation involving L'Hôpital's rule can be used to show $\omega(x)/(x/2) \to 1$ as $x \to 0^+$. In other words, for small values of normalized optimism x and large n we have $d = \omega_n(x) \approx n\omega(x) \approx nx/2$, which mirrors the relationship in the fixed-X setting (6).

4.2 Linear smoothers

Let $\widehat{f}$ be a linear smoother, which means that we can write

$$\widehat{f}(x; X, y) = L_X(x)^\top y, \tag{25}$$

for a weight function $L_X : \mathbb{R}^p \to \mathbb{R}^n$ that is allowed to depend on the training features X, but not the training response y. For convenience, we will write

$$L_X(X) = \begin{bmatrix} L_X(x_1)^\top \\ \vdots \\ L_X(x_n)^\top \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Similarly, for a function $g : \mathbb{R}^p \to \mathbb{R}$, we will write $g(X) = (g(x_1), \dots, g(x_n)) \in \mathbb{R}^n$ for the row-wise application of g to X. In this notation, we can rewrite the data model (1) more compactly as

$$y = f(X) + \varepsilon, \tag{26}$$

where $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Cov}[\varepsilon] = \sigma^2 I$.

The following proposition provides closed-form expressions for random-X optimism and degrees of freedom for linear smoothers.

Proposition 3. *For the linear smoother (25), its intrinsic and emergent random-X optimism are*

$$\mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}) = \sigma^2 \mathbb{E}\bigg[\frac{2}{n}\operatorname{tr}[L_X(X)] + \mathbb{E}[L_X(x_0)^\top L_X(x_0) \,|\, X] - \frac{1}{n}\operatorname{tr}[L_X(X)^\top L_X(X)]\bigg], \tag{27}$$

$$\mathsf{opt}_{\mathrm{R}}(\widehat{f}) = \mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}) + \mathbb{E}\bigg[\mathbb{E}\big[(f(x_0) - L_X(x_0)^\top f(X))^2 \,|\, X\big] - \frac{1}{n}\|(I - L_X(X))f(X)\|_2^2\bigg]. \tag{28}$$

Consequently, the intrinsic and emergent random-X degrees of freedom are given by dividing by σ^2 and applying ω_n in (22).

The calculations to derive (27), (28) are standard; they are based on the bias-variance decomposition of random-X prediction error for linear smoothers, which is found in many places in the literature. In the next subsection, we draw a connection to Rosset and Tibshirani (2020), whose work provides a framework that allows us to easily verify the optimism results (27), (28).

It is worth noting that the intrinsic optimism for a linear smoother (27) is directly proportional to σ^2. As a result, the intrinsic random-X degrees of freedom does not depend on σ^2.

It is also worth noting that for an interpolating linear smoother, we have $L_X(X) = I$. In this case, intrinsic and emergent optimism simplify to

$$\mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}) = \sigma^2\big(1 + \mathbb{E}[L_X(x_0)^\top L_X(x_0)]\big),$$
$$\mathsf{opt}_{\mathrm{R}}(\widehat{f}) = \mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}) + \mathbb{E}\big[(f(x_0) - L_X(x_0)^\top f(X))^2\big].$$

As a result we can see that intrinsic and emergent random-X degrees of freedom (given by dividing by σ^2 and applying ω_n) are each able to distinguish between interpolating linear smoothers, unlike fixed-X degrees of freedom, which always equals n for an interpolator, recalling (8).

4.3 Connection to Rosset and Tibshirani (2020)

Rosset and Tibshirani (2020) proposed the following decomposition of random-X optimism, for an arbitrary predictor $\widehat{f}$:

$$\mathsf{opt}_{\mathrm{R}}(\widehat{f}) = \mathbb{E}[\mathsf{opt}_{\mathrm{F}}(\widehat{f})] + B^+(\widehat{f}) + V^+(\widehat{f}). \tag{29}$$

The first expectation on the right-hand side above is with respect to the training covariates X, and the next two terms $B^+(\widehat{f}), V^+(\widehat{f})$ are called the *excess bias* and *excess variance* of $\widehat{f}$, respectively, defined as:

$$B^+(\widehat{f}) = \mathbb{E}\big[(f(x_0) - \overline{f}(x_0))^2\big] - \mathbb{E}\bigg[\frac{1}{n}\|f(X) - \overline{f}(X)\|_2^2\bigg], \tag{30}$$

$$V^+(\widehat{f}) = \mathbb{E}\big[\operatorname{Var}[\widehat{f}(x_0)|X, x_0]\big] - \mathbb{E}\bigg[\frac{1}{n}\operatorname{tr}(\operatorname{Cov}[\widehat{f}(X)|X])\bigg], \tag{31}$$

where we abbreviate $\overline{f}(X) = \mathbb{E}[\widehat{f}(X)|X]$ and $\overline{f}(x_0) = \mathbb{E}[\widehat{f}(x_0)|X, x_0]$. The relationship (29) follows from expressing the random-X and fixed-X prediction errors of $\widehat{f}$ into bias and variance terms, and then comparing the two decompositions: $B^+(\widehat{f})$ represents the difference in random-X and fixed-X squared bias, and $V^+(\widehat{f})$ the difference in random-X and fixed-X variance.

Though the decomposition (29) is general, we now describe its implications for linear smoothers in particular. For $\widehat{f}$ as in (25), fixed-X degrees of freedom is simple to compute:

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}) = \frac{1}{\sigma^2}\operatorname{tr}(\operatorname{Cov}[L_X(X)y, y \,|\, X] = \operatorname{tr}[L_X(X)].$$

Based on (6), this gives a simple formula for fixed-X optimism: $\mathsf{opt}_{\mathrm{F}}(\widehat{f}) = (2\sigma^2/n)\,\mathrm{tr}[L_X(X)]$. We can plug this into (29) (after integrating over X), along with excess bias and variance calculations, to verify the random-X optimism claims in (27), (28): beginning with the intrinsic case, where we set $f = 0$, it is not hard to see the excess bias is zero and we only need to compute $V^+(\widehat{f})$, which is given by the latter two terms in (27); as for the emergent case, we add in $B^+(\widehat{f})$, which is given by the latter two terms in (28). This completes the proof of Proposition 3.

It is worth emphasizing a result that appears in passing in the arguments from the last paragraph: for a linear smoother,

$$\mathsf{opt}_{\mathrm{R}}(\widehat{f}) = \mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}) + B^+(\widehat{f}). \tag{32}$$

This is not true for a general predictor $\widehat{f}$. For linear smoothers, it holds for *any* distribution of the error vector ε in the original data model (26) (provided we maintain $\mathbb{E}[\varepsilon] = 0$ and $\mathrm{Cov}[\varepsilon] = \sigma^2 I$), even though the pure noise model used for intrinsic optimism in Definition 2 specifies $v \sim \mathcal{N}(0, \sigma^2 I)$. This is because the random-X optimism for a linear smoother depends only on σ^2, the noise level, and not the distribution of the error ε itself.

The fact in (32) is important because, together with monotonicity of the map ω_n in (22), it tells us when we should expect emergent degrees of freedom to be larger than intrinsic degrees of freedom:

$$\begin{aligned} \mathsf{df}_{\mathrm{R}}(\widehat{f}) \geqslant \mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}) \iff& \mathsf{opt}_{\mathrm{R}}(\widehat{f}) \geqslant \mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}) \\ \iff& B^+(\widehat{f}) \geqslant 0. \end{aligned}$$

Rosset and Tibshirani (2020) established nonnegativity of $B^+(\widehat{f})$ for various predictors $\widehat{f}$; the next proposition summarizes these results and their implications for random-X degrees of freedom.

Proposition 4. *For any linear smoother defined by minimizing a penalized least squares criterion, excess bias is always nonnegative, and hence emergent random-X degrees of freedom always larger than intrinsic random-X degrees of freedom. This includes:*

- *least squares regression (underparameterized case);*
- *ridgeless least squares regression (overparameterized case);*
- *ridge regression, for any regularization strength $\lambda \geqslant 0$;*
- *kernel ridge, smoothing splines, and thin-plate splines, for any regularization strength $\lambda \geqslant 0$.*

For nonlinear smoothers, such as the lasso, direct analysis of $\mathsf{df}_{\mathrm{R}}(\widehat{f}) - \mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f})$ (or its sign) does not appear to be as generally tractable. However, as we will see later in Sections 5.3 to 5.5, it is possible to prove the excess bias is nonnegative asymptotically, under certain assumptions on the feature matrix and response model.

4.4 Connection to Luan et al. (2021)

Luan et al. (2021) proposed an extension of classical fixed-X degrees of freedom to the random-X setting, which they called "predictive model" degrees of freedom. Their proposal is limited to linear smoothers. In the notation of the Section 4.2 above, it can be expressed as:

$$\mathsf{df}^{\mathrm{pm}}_X(\widehat{f}) = \mathrm{tr}[L_X(X)] + \frac{n}{2}\bigg(\mathbb{E}[L_X(x_0)^\top L_X(x_0) \,|\, X] - \frac{1}{n}\mathrm{tr}[L_X(X)^\top L_X(X)]\bigg). \tag{33}$$

Comparing this to (27), we note that

$$\mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}) = \frac{2\sigma^2}{n}\mathbb{E}[\mathsf{df}^{\mathrm{pm}}_X(\widehat{f})],$$

where the expectation on the right-hand side is with respect to X. Thus we can see that, for linear smoothers, Luan et al. (2021) define a notion of model complexity in terms of intrinsic random-X optimism by reusing the same functional form that connects fixed-X degrees of freedom to fixed-X optimism (6). (Their follow-up work Luan et al. (2022) considers a weighted version of (33) which allows for heteroscedastic noise.)

There are three differences worth pointing out, to the ideas in the current paper. First, restricting our attention to intrinsic optimism for linear smoothers, Luan et al. (2021) transform normalized intrinsic optimism $x = \mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f})/\sigma^2$ to degrees of freedom via the linear map $x \mapsto nx/2$, whereas we use the nonlinear map $x \mapsto \omega_n(x)$, with ω_n as defined in (22), for what we call intrinsic random-X degrees of freedom. Recalling the discussion in Section 4.1, we have $\omega_n(x) \approx nx/2$ for small values of x, but for large values of ω_n behaves quite differently, and it saturates at $n-1$.

Second, still restricting our attention to linear smoothers, we also consider another (usually larger) notion of model complexity that stems from incorporating bias into random-X optimism, which we call emergent random-X degrees of freedom.

Third, the concepts of emergent and intrinsic random-X degrees of freedom in Definitions 1 and 2 do not require $\widehat{f}$ to be a linear smoother and allow it to be arbitrary. This is possible because the core motivation for these proposals is to match random-X optimism between the given model and a reference model, which we take to be least squares. Being able to carry out this matching does not require special knowledge of any sort about the given predictor $\widehat{f}$ (beyond being able to estimate its random-X optimism, in practice).

5 Case studies: theory

In this section, we pass through various standard prediction models, and develop some theory on random-X degrees of freedom in each case.

5.1 Ridge regression

Recall the ridge regression predictor, given a response vector y and feature matrix X, is defined as $\widehat{f}^{\mathrm{ridge}}_\lambda(x) = x^\top \widehat{\beta}^{\mathrm{ridge}}_\lambda$, where $\widehat{\beta}^{\mathrm{ridge}}_\lambda = (X^\top X/n + \lambda I)^{-1} X^\top y/n$ and $\lambda > 0$ is a tuning parameter. The coefficient vector $\widehat{\beta}^{\mathrm{ridge}}_\lambda$ equivalently solves the following ℓ_2-regularized least squares problem:

$$\widehat{\beta}^{\mathrm{ridge}}_\lambda = \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \frac{1}{n}\|y - Xb\|_2^2 + \lambda\|b\|_2^2. \tag{34}$$

The ridge predictor is a linear smoother, with $L_X(x) = X(X^\top X/n + \lambda I)^{-1}x/n$. Hence, the results in Proposition 3 and Proposition 4 apply. Recall, these results explicitly characterize its random-X degrees of freedom, and assert the nonnegativity of the amount of degrees of freedom "due to bias" $\mathsf{df}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_\lambda) - \mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_\lambda)$, respectively.

In this subsection, we derive two further characterizations, one finite-sample and one asymptotic. The first, finite-sample property concerns the behavior of intrinsic random-X degrees of freedom as a function of the regularization parameter λ.

Proposition 5. *For the ridge predictor $\widehat{f}^{\mathrm{ridge}}_\lambda$ with tuning parameter $\lambda > 0$, its intrinsic random-X degrees of freedom $\mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_\lambda)$ is monotonically decreasing in λ, and $\mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_\lambda) \to 0$ as $\lambda \to \infty$.*

The proof of Proposition 5 is elementary, and deferred to Appendix B.1. Numerical illustrations of the results can be found in Appendix C.3.

The second property gives asymptotic equivalents for emergent and intrinsic random-X degrees of freedom. In preparation for this, we first state our assumptions on the distribution of the features and response variable. These assumptions are similar to those used in Theorem 2, and to those used in the literature on analyzing ridge regression under proportional asymptotics.

Assumption A.

1. *The features satisfy $X = Z\Sigma^{1/2}$, where $Z \in \mathbb{R}^{n\times p}$ is a random matrix with i.i.d. entries having zero mean, unit variance, and bounded moments up to order $4+\delta$ for some $\delta > 0$, and where $\Sigma \in \mathbb{R}^{p\times p}$ is a deterministic positive definite covariance matrix whose eigenvalues are bounded above and below by $r_{\max} < \infty$ and $r_{\min} > 0$, respectively.*

2. *The response vector satisfies $y = f(X) + \varepsilon$, where f is centered (which means $\mathbb{E}[f(x)] = 0$ for a draw x from the feature distribution) with bounded $L^{4+\delta}$ norm (which means $\mathbb{E}[|f(x)|^q]^{1/q}$ is bounded for $q = 4+\delta$) for some $\delta > 0$, and the noise vector $\varepsilon \in \mathbb{R}^n$ has i.i.d. entries with zero mean, variance σ^2, and bounded moments up to order $4+\eta$ for some $\eta > 0$.*

Note that we can always decompose the regression function as

$$f(x) = x^\top \beta + f_{\mathrm{NL}}(x). \tag{35}$$

Here $x^\top\beta$ is the projection of f onto the space of functions linear in x, i.e., it minimizes $\mathbb{E}[(f(x) - x^\top b)^2]$ over $b \in \mathbb{R}^p$, where recall we use x for a draw from the feature distribution. By construction, the components $x^\top\beta$ and $f_{\mathrm{NL}}(x)$ are uncorrelated, though in general they are dependent. We denote the variance of the nonlinear component by $\sigma^2_{\mathrm{NL}} = \mathbb{E}[|f_{\mathrm{NL}}(x)|^2]$.

To introduce some additional notation, let $\gamma_n = p/n$, and for given $\lambda, \gamma_n > 0$, let $\mu_n = \mu(\lambda; \gamma_n)$ be the unique solution to the fixed point equation:

$$\mu_n = \lambda + \gamma_n \mu_n \,\overline{\mathrm{tr}}[\Sigma(\Sigma + \mu_n I)^{-1}], \tag{36}$$

where here and in what follows, we abbreviate $\overline{\mathrm{tr}}(A) = \mathrm{tr}(A)/p$ for $A \in \mathbb{R}^{p\times p}$. We are now ready to state our asymptotic results.

Theorem 6. *Consider the ridge predictor $\widehat{f}^{\mathrm{ridge}}_\lambda$ with tuning parameter $\lambda > 0$, and assume*

$$0 < \liminf_{n\to\infty} \gamma_n \leqslant \limsup_{n\to\infty} \gamma_n < \infty,$$

where recall $\gamma_n = p/n$. Under Assumption A1 for fixed-X degrees of freedom and intrinsic random-X degrees of freedom, and additionally Assumption A2 for emergent random-X degrees of freedom, we have the following asymptotic equivalences, where recall ω is the function in (23):

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}^{\mathrm{ridge}}_\lambda)/n \simeq 1 - \lambda/\mu_n, \tag{37}$$

$$\mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_\lambda)/n \simeq \omega\big((1 - \lambda^2/\mu_n^2)(V_n/D_n + 1)\big), \tag{38}$$

$$\mathsf{df}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_\lambda)/n \simeq \omega\big((1 - \lambda^2/\mu_n^2)(B_n/D_n + (V_n/D_n + 1)(1 + \sigma^2_{\mathrm{NL}}/\sigma^2))\big). \tag{39}$$

Here we use $a_n \simeq b_n$ to mean $|a_n - b_n| \to 0$ as $n \to \infty$ (almost surely, if a_n, b_n are random). Also,

$$V_n = \gamma_n \,\overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu_n I)^{-2}], \tag{40}$$

$$B_n = \mu_n^2 \beta^\top(\Sigma + \mu_n I)^{-1}\Sigma(\Sigma + \mu_n I)^{-1}\beta/\sigma^2, \tag{41}$$

$$D_n = 1 - \gamma_n \,\overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu_n I)^{-2}]. \tag{42}$$

The proof of Theorem 6 is given in Appendix B.2. It is based on the exact asymptotic analysis the of training and prediction errors of ridge regression in various settings (fixed-X, intrinsic random-X, emergent random-X), which can be done following techniques developed and employed previously in Dobriban and Wager (2018); Hastie et al. (2022); Patil and Du (2023); Bach (2024); LeJeune et al. (2024); Patil et al. (2024), among others. Numerical examination of the results in Theorem 6 can be found in Appendix C.3.

We now reflect on the interpretation of the asymptotic equivalences for ridge degrees of freedom in Theorem 6. Inspecting the result for fixed-X degrees of freedom in (37), observe that by (36) we can write its asymptotic (and deterministic) equivalent as

$$1 - \lambda/\mu_n = \gamma_n \,\overline{\mathrm{tr}}[\Sigma(\Sigma + \mu_n I)^{-1}] = \mathrm{tr}[\Sigma(\Sigma + \mu_n I)^{-1}]/n.$$

We can see this as a (normalized) "population-level" degrees of freedom for ridge regression, where we replace $\widehat{\Sigma} = X^\top X/n$ by Σ in the usual "sample-level" formula, $\mathrm{tr}[L_X(X)]/n = \mathrm{tr}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)]/n$. Furthermore, in the population-level formula in the last display, we can see that the regularization level has been changed from λ to μ_n. In other words, each regularization level λ at the sample-level induces a corresponding regularization level μ_n at the population-level, determined by solving a fixed point equation (36). If $\gamma_n = p/n \to 0$, then one can check that $\mu_n \to \lambda$, as would be expected in the low-dimensional regime. In general, we have that $\mu_n \geqslant \lambda$, with strict inequality in the proportional asymptotic regime $\gamma_n \to \gamma > 0$. Further properties of μ_n can be found in Patil et al. (2024).

It is interesting to note that the inflation ratio in the regularization level, $(\mu_n - \lambda)/\mu_n = 1 - \lambda/\mu_n$, is precisely the asymptotic equivalent for fixed-X degrees of freedom in (37). This relationship is not limited to ridge regression and in fact it holds more generally for regularized estimators with convex penalties, as we will see in Section 5.5.

The asymptotic equivalents for random-X degrees of freedom in (38), (39) also have nice interpretations. Note from (37) that λ/μ_n is asymptotically equivalent to $1 - \mathsf{df}_{\mathrm{F}}(\widehat{f}_\lambda^{\mathrm{ridge}})/n$. Thus the factor of $1 - \lambda^2/\mu_n^2$ in both (38), (39) is $1 - (1 - \mathsf{df}_{\mathrm{F}}(\widehat{f}_\lambda^{\mathrm{ridge}})/n)^2$. The other terms in these expressions V_n and B_n in (40), (41) are asymptotic (and deterministic) equivalents for prediction variance and squared bias (scaled by the noise level σ^2) for population ridge regression, at a regularization level μ_n. The final factor that makes this work is D_n, which we interpret next.

The quantity $1 - D_n = \gamma_n \,\overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu_n I)^{-2}] = \mathrm{tr}[\Sigma^2(\Sigma + \mu_n I)^{-2}]/n$, from (42), is a related notion of a (normalized) "population-level" degrees of freedom of a linear smoother, where we square the smoothing matrix before taking the trace. This has appeared in classic literature on additive models (Buja et al., 1989; Hastie and Tibshirani, 1990), and in later analyses of linear and ridge regression generalization (Zhang, 2005; Caponnetto and De Vito, 2007; Hsu et al., 2014). The link between the prediction error of ridge regression at regularization level λ, and a population ridge estimator at an induced level μ_n through the factor D_n, was first derived (using a heuristic argument) by Sollich (2001) in the context of Gaussian processes. It has been recently rederived using the replica method (again heuristic) in Bordelon et al. (2020), and using random matrix theory in Hastie et al. (2022); Cheng and Montanari (2022); Bach (2024), among others.

In our discussion above, we restricted $\lambda > 0$ for simplicity. But, as we can see from (36), if $\mu_n > \lambda$ and we want to keep μ_n small (yet still positive), then we can actually set $\lambda < 0$. The greater the degree of overparameterization (higher γ_n), the more the flexibility we have. This is at the heart of why small (i.e., zero or even negative) values of λ can lead to favorable prediction accuracy in the overparameterized regime. Let us define $\mu_{\min}$, as in LeJeune et al. (2024), to be the unique solution that satisfies $\mu_{\min} > -r_{\min}$ to the fixed point equation:

$$1 = \gamma_n \,\overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu_{\min} I)^{-2}]. \tag{43}$$

Note from (42), (43) that $\mu_{\min}$ is the value at which $D_n = 0$, i.e., both $\mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_{\lambda})$ and $\mathsf{opt}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_{\lambda})$ diverge to ∞ (equivalently, both $\mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_{\lambda})/n$ and $\mathsf{df}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_{\lambda})/n$ converge to 1). We will revisit the relation between D_n and overfitting soon, in the context of ridgeless regression.

5.2 Ridgeless regression

Next we study a special case of ridge regression when $\lambda \to 0^+$, also known as "ridgeless" regression. This is defined by $\widehat{f}^{\mathrm{ridge}}_0(x) = x^\top \widehat{\beta}^{\mathrm{ridge}}_0$, where

$$\widehat{\beta}^{\mathrm{ridge}}_0 = \lim_{\lambda \to 0^+} \widehat{\beta}^{\mathrm{ridge}}_\lambda = (X^\top X)^\dagger X^\top y,$$

and $A^\dagger$ is the usual (Moore-Penrose) pseudoinverse of a matrix A. In the underparameterized case where $p \leqslant n$ (and $\mathrm{rank}(X) = p$), this reduces to the ordinary least squares estimator. However, in the overparameterized case where $p > n$ (and $\mathrm{rank}(X) = n$), there are infinitely many solutions in the least squares problem, each achieving perfect training error, and the ridgeless solution $\widehat{\beta}^{\mathrm{ridge}}_0$ can be interpreted as the interpolator with minimum ℓ_2 norm:

$$\widehat{\beta}^{\mathrm{ridge}}_0 = \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \{\|b\|_2 : y = Xb\}.$$

Ridgeless regression has been thrust into the spotlight, due to recent interest in overparameterized machine learning and the study of double descent. See Bartlett et al. (2020); Belkin et al. (2020); Hastie et al. (2022), among many others.

Continuing in the vein ridge analysis from the last subsection, we will study the degrees of freedom of the ridgeless predictor in an asymptotic regime where we let the sample size n and the feature size p diverge, while keeping their ratio bounded. In preparation for this, for a given $\gamma_n > 1$, define $\mu_n = \mu_n(0; \gamma_n)$ be the unique solution to the fixed point equation:

$$1 = \gamma_n \,\overline{\mathrm{tr}}[\Sigma(\Sigma + \mu_n I)^{-1}]. \tag{44}$$

Observe that (44) is the limiting case of (36) as $\lambda \to 0^+$. We are now ready to state our asymptotic results on ridgeless degrees of freedom.

Theorem 7. *For the ridgeless predictor $\widehat{f}^{\mathrm{ridge}}_0$, under the same assumptions as Theorem 6, we have the following asymptotic equivalences:*

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}^{\mathrm{ridge}}_0)/n \simeq \begin{cases} \gamma_n & \text{for } \gamma_n \leqslant 1 \\ 1 & \text{for } \gamma_n > 1, \end{cases} \tag{45}$$

$$\mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_0)/n \simeq \begin{cases} \gamma_n & \text{for } \gamma_n \leqslant 1 \\ \omega(V_n/D_n + 1) & \text{for } \gamma_n > 1, \end{cases} \tag{46}$$

$$\mathsf{df}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_0)/n \simeq \begin{cases} \omega\big((\gamma_n + \gamma_n/(1-\gamma_n))(1 + \sigma^2_{\mathrm{NL}}/\sigma^2)\big) & \text{for } \gamma_n \leqslant 1 \\ \omega\big(B_n/D_n + (V_n/D_n + 1)(1 + \sigma^2_{\mathrm{NL}}/\sigma^2)\big) & \text{for } \gamma_n > 1, \end{cases} \tag{47}$$

where μ_n is as defined in (36), and all other quantities are as defined in Theorem 6.

Note: if the response model is well-specified, or in other words, $f_{\mathrm{NL}}(x) = 0$ in (35), then $\sigma^2_{\mathrm{NL}} = 0$, so the emergent random-X degrees of freedom in (47) reduces to (as expected):

$$\omega(\gamma_n + \gamma_n/(1-\gamma_n)) = \gamma_n.$$

The check the equality above, recall $\omega(x)$ in (23) is the value of u that solves $x = u + u/(1-u)$.

The proof of Theorem 7 is given in Appendix B.3, and numerical examination of the results can be found in Appendix C.4. It is interesting to note that each of the intrinsic and emergent normalized random-X degrees of freedom curves are continuous at $\gamma_n = 1$, even though the prediction error of ridgeless regression blows up at $\gamma_n = 1$ (it has an essential discontinuity at this point).

Our next result develops monotonicity properties of the asymptotic equivalents for intrinsic and emergent random-X degrees of freedom for ridgeless regression.

Proposition 8. *The following properties hold for the asymptotic equivalents from Theorem 7.*

1. *The asymptotic equivalent for intrinsic random-X degrees of freedom $\mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_0)/n$ in (46) is increasing in γ_n on $(0,1)$, maximized at $\gamma_n = 1$, and decreasing in γ_n on $(1,\infty)$.*
2. *The asymptotic equivalent for emergent random-X degrees of freedom $\mathsf{df}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_0)/n$ in (47) is increasing in γ_n on $(0,1)$, and maximized at $\gamma_n = 1$.*

The proof of Proposition 8 is in Appendix B.4, and numerical illustrations are in Appendix C.4.

Beyond what we discussed in the last subsection, we provide one more connection between D_n and overfitting in ridgeless regression. Mallinar et al. (2022) defined three categories: *benign*, *catastrophic*, and *tempered* overfitting, based on whether the excess random-X prediction error goes to 0, ∞, or is bounded away from 0 and ∞, respectively. Zhou et al. (2023) then showed that these regimes can be characterized in terms of the spectrum of Σ, which recovers the results of Bartlett et al. (2020), by connecting this to the notion of effective rank. In terms of D_n, these regimes correspond to whether $1/D_n$ goes to 1, ∞, or is bounded away from 1 and ∞, respectively.

5.3 Lasso regression

Many of the qualitative properties and relationships we observed for ridge and ridgeless regression degrees of freedom carry over to nonlinear smoothers too. To see this, we first study lasso regression (Tibshirani, 1996), which recall, is defined by $\widehat{f}^{\mathrm{lasso}}_\lambda(x) = x^\top \widehat{\beta}^{\mathrm{lasso}}_\lambda$, where $\widehat{\beta}^{\mathrm{lasso}}_\lambda$ solves the following ℓ_1-regularized least squares optimization problem, for a tuning parameter $\lambda > 0$:

$$\widehat{\beta}^{\mathrm{lasso}}_\lambda \in \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \frac{1}{2}\|y - Xb\|_2^2 + \lambda\|b\|_1. \tag{48}$$

The element notation above is used to emphasize the fact that the minimizer in (48) is not unique in general. However, it is unique under weak conditions, for example, if the columns of the feature matrix X are in general position (Tibshirani, 2013).

As indicated in Section 1.2, the fixed-X degrees of freedom of the lasso and various generalizations have been studied extensively. When the lasso solution is unique, the fixed-X degrees of freedom of the lasso predictor is the expected number of nonzero coefficients in the lasso solution (Zou et al., 2007; Tibshirani and Taylor, 2012). Here, we will derive exact formulae for the limiting random-X degrees of freedom of the lasso predictor under proportional asymptotics, where n, p both diverge, and their ratio converges to a constant, $p/n \to \gamma \in (0,\infty)$.

We begin by stating our assumptions on the training data; these will be more restrictive than those in Assumption A, used for ridge, but are standard when using approximate message passing (AMP) or the convex Gaussian minimax theorem (CGMT) to analyze regularized M-estimators.

Assumption B.

1. *The feature matrix X has i.i.d. entries from $\mathcal{N}(0, 1/n)$.*
2. *The response vector follows $y = X\beta + \varepsilon$, where the signal vector $\beta \in \mathbb{R}^p$ has i.i.d. entries from a distribution F with bounded second moment, and the noise vector $\varepsilon \in \mathbb{R}^n$ has i.i.d. entries with zero mean and variance σ^2.*

Note: under Assumption B1, the columns of X will be in general position almost surely, and hence the lasso solution will be unique almost surely.

The limiting degrees of freedom of the lasso, under the assumptions stated above, is determined by the solution of a nonlinear system. We introduce some relevant notation. First, define

$$\mathsf{soft}(u;t) = \begin{cases} u - t & \text{if } u > t \\ 0 & \text{if } u \in [-t, t]] \\ u + t & \text{if } u < -t. \end{cases}$$

Next, for a fixed $\gamma \in (0, \infty)$, define $(\tau, \mu) \in \mathbb{R}^2$ as the unique solution to the nonlinear system:

$$\tau^2 = \sigma^2 + \gamma \mathbb{E}[(\mathsf{soft}(B + \tau H; \mu) - B)^2], \tag{49}$$

$$\mu = \lambda + \gamma\mu \mathbb{E}[\mathsf{soft}'(B + \tau H; \mu)], \tag{50}$$

where $B \sim F$ and $H \sim \mathcal{N}(0, 1)$ are independent. This system is from Bayati and Montanari (2011), who show that its solution determines the limiting behavior of the lasso estimator. (We modify the form of the system slightly in order to unify our presentation of ridge, lasso, and convex penalties.) Moreover, we use (τ_0, μ_0) to denote the solution in (49), (50) when we replace F by a point mass at 0 (i.e., we set $B = 0$). We are ready to state our asymptotic results.

Theorem 9. *Consider the lasso predictor $\widehat{f}_\lambda^{\text{lasso}}$ with tuning parameter $\lambda > 0$. Under Assumption B, the following asymptotic equivalences hold, as $n, p \to \infty$ such that $p/n \to \gamma \in (0, \infty)$, where recall ω is the function in (23):*

$$\mathsf{df}_{\text{F}}(\widehat{f}_\lambda^{\text{lasso}})/n \simeq 1 - \lambda/\mu, \tag{51}$$

$$\mathsf{df}_{\text{R}}^{\text{i}}(\widehat{f}_\lambda^{\text{lasso}})/n \simeq \omega\big((1 - \lambda^2/\mu^2)\tau_0^2/\sigma^2\big), \tag{52}$$

$$\mathsf{df}_{\text{R}}(\widehat{f}_\lambda^{\text{lasso}})/n \simeq \omega\big((1 - \lambda^2/\mu^2)\tau^2/\sigma^2\big). \tag{53}$$

Here we use $a_n \simeq b_n$ to mean $|a_n - b_n| \to 0$ as $n \to \infty$ (in probability, if a_n, b_n are random).

Theorem 9 is a special case of a more general result on regularized least squares estimators that we derive later in Theorem 13. Numerical verification of Theorem 9 is given in Appendix C.5.

We pause to interpret the lasso results above, and compare them to those on ridge regression from Theorem 6. It is instructive to rewrite (51) using (50), which gives

$$1 - \lambda/\mu = \gamma \mathbb{E}[\mathsf{soft}'(B + \tau H; \mu)] = p\mathbb{E}[\mathsf{soft}'(B + \tau H; \mu)]/n.$$

In this reformulation, the factor μ again plays the role of an induced regularization amount at the "population level", analogous to μ_n in the ridge regression analysis. The right-hand side in the last display can thus be viewed as the normalized (scaled by n) fixed-X degrees of freedom of the lasso, with regularization parameter μ, when it is fit on a population model with orthogonal features and responses drawn according to the original linear model, but with noise variance τ^2.

As with ridge regression, if the aspect ratio diminishes: $p/n \to 0$ (i.e., $\gamma = 0$) then we have $\mu = \lambda$ and thus the induced regularization level μ matches the original one λ in the low-dimensional regime. In general, however, we have $\mu > \lambda$ when $\gamma \in (0, \infty)$, which mirrors the inflation of the effective regularization level in ridge regression in the high-dimensional regime.

Just as with ridge regression, the fixed-X degrees of freedom of the lasso is asymptotically (51) the inflation ratio in effective regularization, $(\mu - \lambda)/\mu = 1 - \lambda/\mu$. However, the following is a notable difference between the ridge and lasso fixed point equations. For ridge regression, we can solve for μ_n in (36) based on knowledge of Σ only. In particular, this means that μ does not depend on the signal, through either its linear β or nonlinear f_{NL} parts. For lasso, we must solve for (τ, μ) jointly in (49), (50), which depends on the signal distribution F (via the draw $B \sim F$). A consequence of this difference is that the fixed-X degrees of freedom for ridge (37) does not actually depend on the signal, whereas for lasso (51) it does.

The expressions for random-X degrees of freedom in (52), (53) also have interesting interpretations, which we leave to Section 5.5, when we cover regularized least squares estimators more generally.

Unlike ridge regression (which is a linear smoother), it is not possible to establish monotonicity of intrinsic random-X degrees of freedom as a function of λ (recall Proposition 5) or nonnegativity of the random-X degrees of freedom "due to bias" (recall Proposition 4) for the lasso, via elementary arguments. However, the results in Theorem 9 allow us to infer such properties asymptotically.

Proposition 10. *The following properties hold for the asymptotic equivalents from Theorem 9.*

1. *The asymptotic equivalent for intrinsic random-X degrees of freedom $\mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{lasso}}_{\lambda})/n$ in (52) is monotonically decreasing in λ, and converges to 0 as $\lambda \to \infty$.*
2. *The asymptotic equivalent for emergent random-X degrees of freedom $\mathsf{df}_{\mathrm{R}}(\widehat{f}^{\mathrm{lasso}}_{\lambda})/n$ in (47) is always larger or equal to that for intrinsic random-X degrees of freedom in (52).*

The proof of Proposition 10 is in Appendix B.5, and numerical illustrations are in Appendix C.5.

5.4 Lassoless regression

We now study a special case of lasso regression when $\lambda \to 0^+$, which we term "lassoless" regression. The simplest way to define this estimator is to assume that the lasso solution is unique for each λ (which recall, is implied almost surely under Assumption B1), and define the lassoless solution as

$$\widehat{\beta}^{\mathrm{lasso}}_0 = \lim_{\lambda \to 0^+} \widehat{\beta}^{\mathrm{lasso}}_{\lambda},$$

and correspondingly define the predictor $\widehat{f}^{\mathrm{lasso}}_0(x) = x^\top \widehat{\beta}^{\mathrm{lasso}}_0$. When $p \leqslant n$ (and $\mathrm{rank}(X) = p$), this is no different from the ordinary least squares estimator. Meanwhile, when $p > n$ (and $\mathrm{rank}(X) = n$), the lassoless estimator as defined above is an interpolator with minimum ℓ_1 norm:

$$\widehat{\beta}^{\mathrm{ridge}}_0 \in \mathop{\arg\min}_{b \in \mathbb{R}^p} \{\|b\|_1 : y = Xb\}.$$

We note that even when the lasso solution is not unique, it is still possible to construct a sequence of lasso solutions converging to a minimum ℓ_1 norm interpolator; see Tibshirani (2013) for details.

We will derive the asymptotics of random-X degrees of freedom for the lassoless predictor, in the proportional asymptotics model from the previous subsection. In preparation for this, for $\gamma > 1$, let $(\tau, \mu) \in \mathbb{R}^2$ be the unique solution to the nonlinear system:

$$\tau^2 = \sigma^2 + \gamma \mathbb{E}[(\mathsf{soft}(B + \tau H; \mu) - B)^2], \tag{54}$$

$$1 = \gamma \mathbb{E}[\mathsf{soft}'(B + \tau H; \mu)], \tag{55}$$

where again $B \sim F$ and $H \sim \mathcal{N}(0,1)$ are independent. This system is studied in Li and Wei (2021) (we modify its presentation to suit our purposes), and is the limit of (49), (50) as $\lambda \to 0^+$. Similar to our earlier convention, let (τ_0, μ_0) denote the solution to the above system when $B = 0$.

Theorem 11. *For the lassoless predictor $\widehat{f}_0^{\text{lasso}}$, under the same conditions as Theorem 9, we have the following asymptotic equivalences:*

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}_0^{\text{lasso}})/n \simeq \begin{cases} \gamma & \text{for } \gamma \leqslant 1 \\ 1 & \text{for } \gamma > 1, \end{cases} \tag{56}$$

$$\mathsf{df}_{\mathrm{R}}^{\mathrm{i}}(\widehat{f}_0^{\text{lasso}}) \simeq \begin{cases} \gamma & \text{for } \gamma \leqslant 1 \\ \omega(\tau_0^2/\sigma^2) & \text{for } \gamma > 1, \end{cases} \tag{57}$$

$$\mathsf{df}_{\mathrm{R}}(\widehat{f}_0^{\text{lasso}}) \simeq \begin{cases} \gamma & \text{for } \gamma \leqslant 1 \\ \omega(\tau^2/\sigma^2) & \text{for } \gamma > 1. \end{cases} \tag{58}$$

The proof of Theorem 11 is given in Appendix B.6, and numerical verification of the results can be found in Appendix C.6.

As before, in the ridgeless setting, we can leverage the asymptotics above to develop monotonicity properties for intrinsic and emergent random-X degrees for freedom in lassoless regression.

Proposition 12. *The following properties hold for the asymptotic limits from Theorem 11.*

1. *The asymptotic limit of intrinsic random-X degrees of freedom $\mathsf{df}_{\mathrm{R}}^{\mathrm{i}}(\widehat{f}_0^{\text{lasso}})/n$ in (57) is increasing in γ on $(0,1)$, maximized at $\gamma = 1$, and decreasing in γ on $(1, \infty)$.*
2. *The asymptotic limit of emergent random-X degrees of freedom $\mathsf{df}_{\mathrm{R}}(\widehat{f}_0^{\text{lasso}})/n$ in (58) is increasing in γ on $(0,1)$, and maximized at $\gamma = 1$.*

Note: recall that when $\gamma \leqslant 1$, the lassoless solution is just least squares, as is the ridgeless solution. Therefore the underparameterized statements in the lassoless results above are duplicates of those in Proposition 8. However, for ease of interpretation, we leave the underparameterized cases in the presentation of Proposition 12.

The proof of Proposition 12 is given in Appendix B.7, and numerical illustrations can be found in Appendix C.6. Our next and last subsection generalizes the study of ridge and lasso estimators.

5.5 Convex regularized least squares

Given a proper closed convex function $\mathsf{reg} : \mathbb{R} \to [0, \infty]$, consider defining a regularized least squares estimator by

$$\widehat{\beta}_\lambda^{\text{convex}} \in \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \frac{1}{2} \sum_{i=1}^n (y_i - x_i^\top b)^2 + \lambda \sum_{i=1}^p \mathsf{reg}(b_i). \tag{59}$$

for a tuning parameter $\lambda > 0$. The corresponding predictor is defined as $\widehat{f}_\lambda^{\text{convex}} = x^\top \widehat{\beta}_\lambda^{\text{convex}}$. Note that the element notation above emphasizes the fact that the solution in (59) need not be unique; the theory below applies to any one of its solutions.

We will extend the degrees of freedom analysis in Section 5.3 to the regularized estimator in (59). To introduce some relevant notation, recall that the proximal operator reg is defined as

$$\mathsf{prox}_{\mathsf{reg}}(x;t) = \operatorname*{arg\,min}_{z\in\mathbb{R}} \frac{1}{2t}(x-z)^2 + \mathsf{reg}(z),$$

for a parameter $t > 0$. Still working under Assumption B for our asymptotic analysis, we will now describe the nonlinear system which generalizes (49), (50): define $(\tau,\mu) \in \mathbb{R}^2$ to solve

$$\tau^2 = \sigma^2 + \gamma \mathbb{E}[(\mathsf{prox}_{\mathsf{reg}}(B+\tau H;\mu) - B)^2], \tag{60}$$

$$\mu = \lambda + \gamma\mu \mathbb{E}[\mathsf{prox}'_{\mathsf{reg}}(B+\tau H;\mu)], \tag{61}$$

where $B \sim F$ and $H \sim \mathcal{N}(0,1)$ are independent. This system is adapted from Thrampoulidis et al. (2018), who show that its solution determines the limiting behavior of the regularized estimator in (59). (We modify the form of the system, with the full details given in the proof of our next result.) Moreover, we use (τ_0,μ_0) to denote the solution in (49), (50) when we replace F by a point mass at 0 (i.e., set $B = 0$). We are ready to state our asymptotic results.

Theorem 13. *Consider the convex regularized predictor $\widehat{f}_\lambda^{\text{convex}}$ with tuning parameter $\lambda > 0$. Under Assumption B, the following asymptotic equivalences hold, as $n,p \to \infty$ such that $p/n \to \gamma \in (0,\infty)$:*

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}_\lambda^{\text{convex}})/n \simeq 1 - \lambda/\mu, \tag{62}$$

$$\mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}_\lambda^{\text{convex}})/n \simeq \omega\big((1-\lambda^2/\mu^2)\tau_0^2/\sigma^2\big), \tag{63}$$

$$\mathsf{df}_{\mathrm{R}}(\widehat{f}_\lambda^{\text{convex}})/n \simeq \omega\big((1-\lambda^2/\mu^2)\tau^2/\sigma^2\big). \tag{64}$$

Here we use $a_n \simeq b_n$ to mean $|a_n - b_n| \to 0$ as $n \to \infty$ (in probability, if a_n, b_n are random).

Note that Theorem 9 is a special case of Theorem 13 when the regularizer is the ℓ_1 norm, $\mathsf{reg} = \|\cdot\|_1$, and the proximal operator is soft-thresholding, $\mathsf{prox}_{\|\cdot\|_1} = \mathsf{soft}$. The proof of Theorem 13 is given in Appendix B.8.

We now give an interpretation of the asymptotic limits (63), (64) for random-X degrees of freedom by drawing an analogy to generalized cross-validation (GCV). Initially designed for linear smoothers, GCV scales the training error by a factor that involves the trace of the smoothing matrix. This can be understood more broadly, beyond linear smoothers, as a fixed-X degrees of freedom adjustment. This leads to the following approximation, writing $\mathsf{err}_{\mathrm{T}}(\widehat{f}_\lambda^{\text{convex}})$ for the training error of $\widehat{f}_\lambda^{\text{convex}}$:

$$\mathsf{err}_{\mathrm{R}}(\widehat{f}_\lambda^{\text{convex}}) \approx \frac{\mathsf{err}_{\mathrm{T}}(\widehat{f}_\lambda^{\text{convex}})}{(1-\mathsf{df}_{\mathrm{F}}(\widehat{f}_\lambda^{\text{convex}})/n)^2}. \tag{65}$$

For ridge regression (Patil et al., 2021), and regularized least squares with convex penalties (Bellec, 2023) more broadly, this approximation is exact under proportional asymptotics. To connect this to the results in Theorem 13, consider

$$\begin{aligned}\mathsf{opt}_{\mathrm{R}}(\widehat{f}_\lambda^{\text{convex}}) &= \mathsf{err}_{\mathrm{R}}(\widehat{f}_\lambda^{\text{convex}}) - \mathsf{err}_{\mathrm{T}}(\widehat{f}_\lambda^{\text{convex}}) \\ &\approx \mathsf{err}_{\mathrm{R}}(\widehat{f}_\lambda^{\text{convex}}) - (1-\mathsf{df}_{\mathrm{F}}(\widehat{f}_\lambda^{\text{convex}})/n)^2 \cdot \mathsf{err}_{\mathrm{R}}(\widehat{f}_\lambda^{\text{convex}})\end{aligned}$$

$$\approx \mathsf{err}_{\mathrm{R}}(\widehat{f}_\lambda^{\text{convex}}) - (\lambda^2/\mu^2) \cdot \mathsf{err}_{\mathrm{R}}(\widehat{f}_\lambda^{\text{convex}}),$$

where the second line uses (65), and the third line uses (62). Since the parameters τ_0 and τ are the limiting random-X prediction errors in the intrinsic and emergent settings, respectively, the last line provides a way to understand (63), (64) (after applying ω to map optimism to degrees of freedom).

6 Case studies: experiments

We continue our study of degrees of freedom, now via numerical experiments. Python code to reproduce our experiments is available at: `https://github.com/jaydu1/model-complexity`; additional results for k-nearest-neighbor (kNN) and random features regression are given in Appendix D.

6.1 Lasso regression

Returning to the lasso, whose degrees of freedom we studied asymptotically in the previous section, we now empirically compare its random-X degrees of freedom to the (average) number of nonzero coefficients in the lasso solution. The latter is known to be its fixed-X degrees of freedom (Zou et al., 2007; Tibshirani and Taylor, 2012), in general.

We simulate data according to a sparse linear model $y_i = x_i^\top \beta + \varepsilon_i$, $i \in [n]$. The entries of $x_i \in \mathbb{R}^p$ and ε_i are all i.i.d. standard normal. The first s entries of β are equal to α, while the remaining are equal to zero. We choose α so that the signal-to-noise ratio (SNR) is 1, and compute all quantities by averaging over 500 repetitions (500 times drawing the simulated data set; and in each repetition, we compute prediction errors using an independent test set of 1000 samples).

Figures 4 and 5 show the results for underparameterized and overparameterized cases, respectively. The underparameterized case uses $n = 200$, $p = 30$, and $s = 10$, while the overparameterized case uses $n = 200$, $p = 300$, and $s = 100$. As expected by the theory, the fixed-X degrees of freedom of the lasso (middle panel) equals the average number of nonzero coefficients in its solution, over the full path of λ values. Interestingly, in the underparameterized case, this also appears to be true of the intrinsic random-X degrees of freedom: it also coincides with the average number of nonzero lasso coefficients.

By comparison, in this same setting, the emergent random-X degrees of freedom is initially quite a bit larger, and then eventually settles back down to coincide with the number of nonzero estimated coefficients, at higher levels of estimated sparsity. As shown in the right panel, once it has estimated a little more than $s = 10$ nonzero coefficients, it soon achieves near-perfect support recovery in our simulations. Though this underparameterized problem setup with large n, small p, and uncorrelated features is somewhat idealistic (the lasso is able to identify a near-perfect support), it is nonetheless interesting to observe the behavior the degrees of freedom "due to bias", emergent minus intrinsic degrees of freedom, here: it is initially quite large, for lower levels of estimated sparsity, and then it vanishes, at higher levels of estimated sparsity.

In the overparameterized case, with $p = 300$ features, the lasso is only able to recover about half of the true support once it has estimated $s = 100$ nonzero coefficients, as the right panel of Figure 5 shows. From the middle panel of the figure, we see that the intrinsic random-X degrees of freedom grows increasingly smaller than the fixed-X degrees of freedom, as number of nonzero coefficients increases. Meanwhile, the emergent random-X degrees of freedom exceeds fixed-X degrees of freedom up until the point at which the lasso exhibits roughly 100 nonzero coefficients, when it drops below

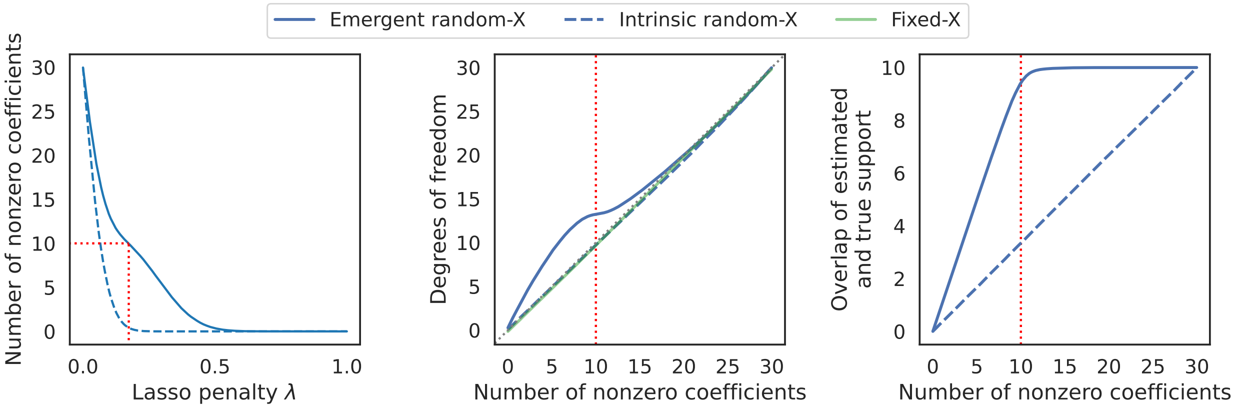

Figure 4: Degrees of freedom of lasso predictors, parameterized by the average number of nonzero coefficients, in a problem setting with $n = 200$, $p = 30$, and sparsity level $s = 10$.

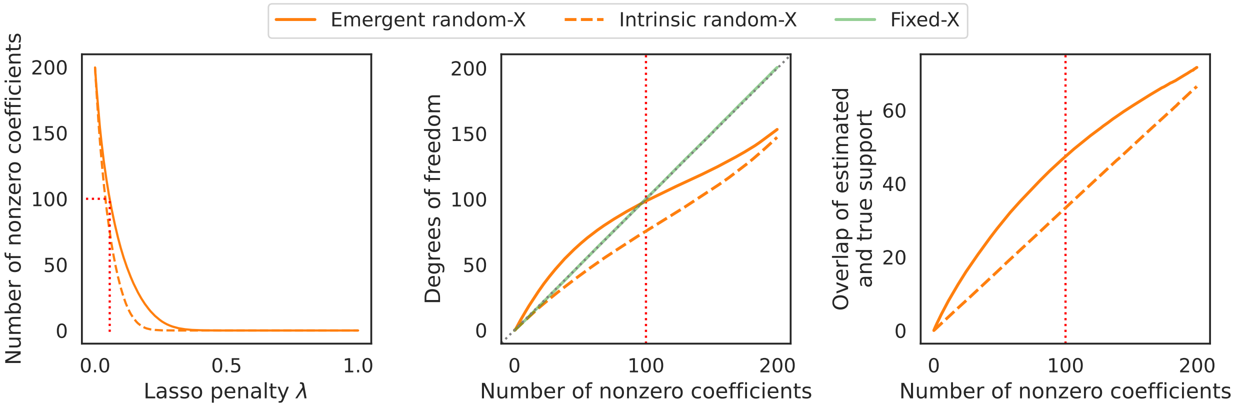

Figure 5: Degrees of freedom of lasso predictors, parameterized by the average number of nonzero coefficients, in a problem setting with $n = 200$, $p = 300$, and sparsity level $s = 100$.

fixed-X degrees of freedom. Lastly, the degrees of freedom "due to bias", the different in emergent and intrinsic degrees of freedom, is fairly large throughout.

6.2 Random forests

Next, we study random forests. Following the experimental setup in Belkin et al. (2019), we use a single tree (rather than an average of trees over subsamples of training data) up until the point of interpolation, increasing the maximum number of leaves $N_{\text{leaf}}^{\max}$ allowed in the tree until we reach zero training error. After interpolation, we keep $N_{\text{leaf}}^{\max}$ fixed and increase the number of trees N_{tree}.

We draw data according to a linear model $y_i = x_i^\top \beta + \varepsilon_i$, $i \in [n]$, where each $x_i \sim \mathcal{N}(0, \Sigma_{\text{AR1},\rho=0.25})$, $\varepsilon_i \sim \mathcal{N}(0, 0.5^2)$, and β is drawn uniformly from the unit sphere in $\mathbb{R}^d$. Here the covariance matrix $\Sigma_{\text{AR1},\rho}$ has entries $\rho^{|i-j|}$. The SNR in this setup is 4. As before, we average all quantities over 500 repetitions (and in each one, compute prediction errors over an independent test set of size 1000).

Figure 6 shows the results for $n = 2000$ and $p = 50$. As we can see in the left panel, the random-X prediction error initially decreases, then it increases again as the number of leaves approaches the interpolation threshold. After this point, it decreases as we increase the total number of leaves by including more trees. As expected, the fixed-X degrees of freedom increases before the interpolation threshold, while remaining constant beyond the point, as shown in the middle panel. On the other hand, both the emergent and intrinsic random-X degrees of freedom decrease after this threshold, and generally remain much smaller than the trivial saturation value (of n degrees of freedom). The degrees of freedom "due to bias", emergent minus intrinsic, is also consistently large throughout.

6.3 Degrees of freedom comparisons

So far we have mostly examined the behavior of degrees of freedom for individual classes of models, and have drawn comparisons between members of one class. Now, we shift our focus to comparing *across* model classes. Figure 7 studies such comparisons across ridge regression, kNN, and random forest predictors. The top row shows results for data simulated from a linear model with $n = 200$ and $p = 100$, a setting that favors ridge regression; the bottom row shows the results for data simulated with the same n, p, but in a way that favors random forests. (This uses the `make_classification` function from `scikit-learn` v1.2.2; see Pedregosa et al. (2011)). In the top row, we can see that optimally-tuned ridge regression achieves the best random-X prediction error (rightmost panel), but interestingly, does so at a much larger emergent random-X degrees of freedom than optimally-tuned kNN. In the bottom row, the optimally-tuned random forest achieves the best prediction error, and does so at a much larger degrees of freedom than either ridge regression or kNN.

As a follow-up on the comparisons just discussed, one may naturally ask: can a model achieve both the best prediction error and emergent degrees of freedom, simultaneously? In a sense, this would put the model on the "Pareto frontier" traced out by predictive accuracy versus complexity. Both ridge regression and random forests achieve the best predictive accuracy (in top and bottom rows, respectively) but fail to do so at the lowest complexity, in Figure 7. Yet, this is only a snapshot of their performance at a given sample size. In Figure 8, we examine the optimally-tuned ridge, kNN, and random forest predictors as we vary the sample size n from 100 to 1000. A each n, we measure the excess random-X prediction error (the random-X prediction error minus the Bayes error) and the normalized emergent random-X degrees of freedom (scaled by n) of each optimally-tuned predictor. For the linear model simulation (corresponding to the left panel of Figure 8), ridge quickly becomes "Pareto optimal" as n increases, eventually demonstrating a lower emergent degrees of freedom than kNN. For the simulation designed to favor random forests (right panel), random forests fail to be "Pareto optimal" at any n, as ridge and kNN each offer a nontrivial tradeoff in balancing predictive accuracy versus complexity. (As a side note, it is interesting to note that the dynamic range of the emergent degrees of freedom of optimally-tuned kNN is very small, in both settings.)

7 Degrees of freedom decomposition

In previous sections, we spoke frequently of the degrees of freedom "due to bias", which refers to the difference in emergent and intrinsic random-X degrees of freedom. Here we describe how this general idea—decomposing degrees of freedom by attributing complexity to different sources of errors—can be extended to problems involving distribution shift.

We demonstrate the idea in the context of covariate shift. Given a predictor $\widehat{f}$, we consider four scenarios (Table 1 gives a summary). In all cases, the reference model remains the least squares predictor on well-specified data, as in Definition 1 or Definition 2, but the left-hand side of the matching equations (18) or (19) changes.

1. The *total emergent model* (both signal and covariate shift): for the left-hand side in (18), the optimism is computed using random-X prediction error under covariate shift, and the result is

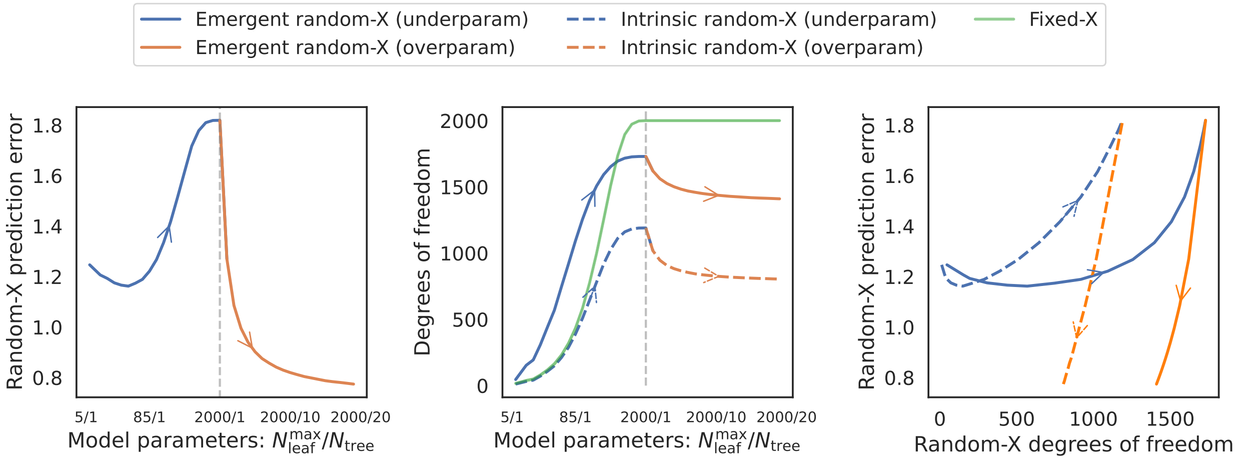

Figure 6: Prediction error and degrees of freedom of random forest predictors, as we vary the number of trees N_{tree} and the maximum number of leaves for each tree $N_{\text{leaf}}^{\text{max}}$, in a problem with $n = 2000$, $p = 50$.

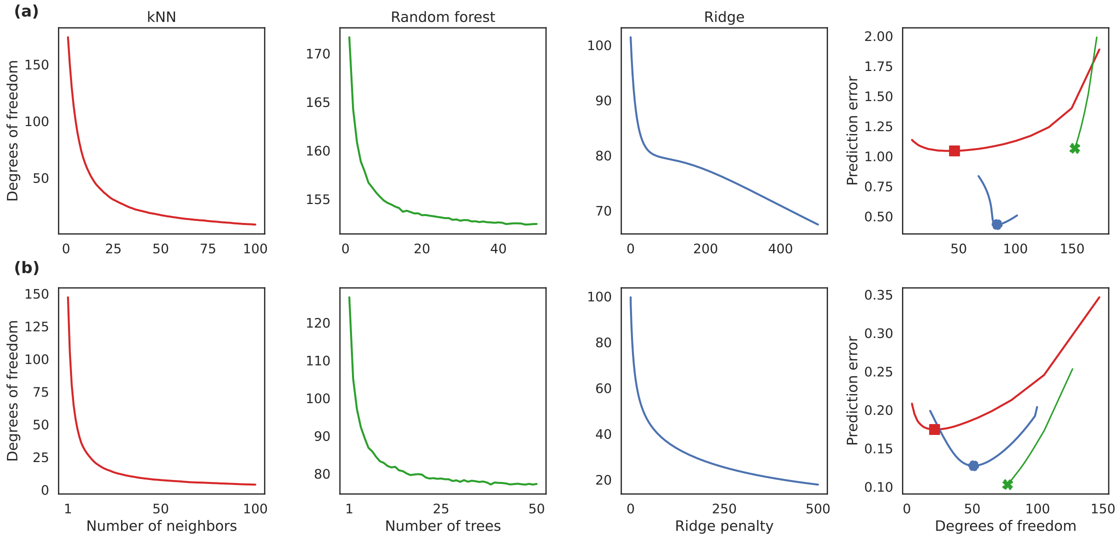

Figure 7: Prediction error and degrees of freedom for ridge regression, kNN, and random forests. In both rows, $n = 200$ and $p = 100$. The top row displays data drawn from a linear model, which favors ridge. The bottom displays data drawn from a model that favors random forests.

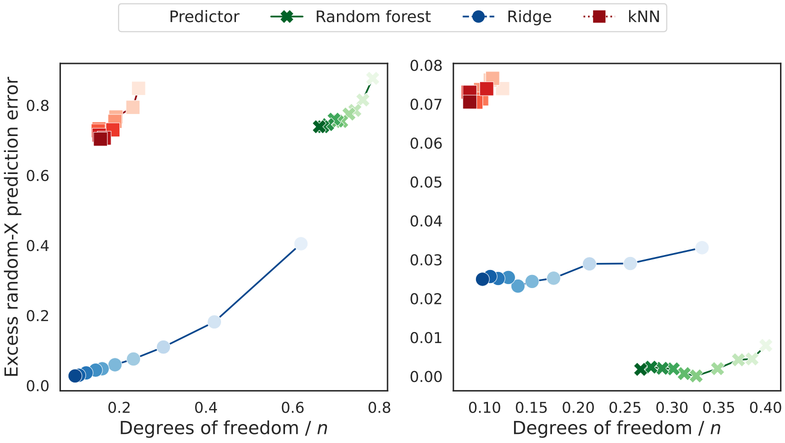

Figure 8: Comparisons of excess prediction error and degrees of freedom for optimally-tuned ridge regression, kNN, and random forests, as the sample size n varies. Lighter, more transparent colors indicate smaller n, whereas darker, more opaque colors indicate larger n. The left panel uses data simulated as in the top row of Figure 7, and the right panel uses data simulated as in the bottom row of Figure 7.

Signal presence	Covariate shift	
	✗	✓
✗	$\mathsf{df}^{00}(\widehat{f})$	$\mathsf{df}^{01}(\widehat{f})$
✓	$\mathsf{df}^{10}(\widehat{f})$	$\mathsf{df}^{11}(\widehat{f})$

Table 1: Scenarios for decomposing degrees of freedom due to bias and covariate shift.

denoted $\mathsf{df}^{11}(\widehat{f})$, which we simply call emergent degrees of freedom $\mathsf{df}_{\mathrm{R}}(\widehat{f})$.

2. The *partial emergent model* (with signal but no covariate shift): for the left-hand side in (18), the optimism is computed using random-X prediction error without covariate shift, and the result is denoted $\mathsf{df}^{10}(\widehat{f})$.
3. The *partial intrinsic model* (with no signal but with covariate shift): for the left-hand side in (19), the optimism is computed using random-X prediction error without signal yet still with covariate shift, and the result is denoted $\mathsf{df}^{01}(\widehat{f})$.
4. The *intrinsic model* (with no signal and no covariate shift): for the left-hand side in (19), the optimism is computed using random-X prediction error without signal or covariate shift, and the result is denoted $\mathsf{df}^{00}(\widehat{f})$, which we simply call the intrinsic degrees of freedom $\mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f})$.

In order to attribute an amount of degrees of freedom to each source of error—bias and covariate shift—we use a definition akin to Shapley values (Shapley, 1953):

$$\phi^{\mathrm{sig}}(\widehat{f}) = \frac{1}{2}(\mathsf{df}_{\mathrm{R}}(\widehat{f}) - \mathsf{df}^{01}(\widehat{f})) + \frac{1}{2}(\mathsf{df}^{10}(\widehat{f}) - \mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f})),$$
$$\phi^{\mathrm{cov}}(\widehat{f}) = \frac{1}{2}(\mathsf{df}_{\mathrm{R}}(\widehat{f}) - \mathsf{df}^{10}(\widehat{f})) + \frac{1}{2}(\mathsf{df}^{01}(\widehat{f}) - \mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f})).$$

Note that by construction (which is also a Shapley axiom called "efficiency"), we have:

$$\mathsf{df}_{\mathrm{R}}(\widehat{f}) = \mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}) + \phi^{\mathrm{sig}}(\widehat{f}) + \phi^{\mathrm{cov}}(\widehat{f}).$$

In other words, we have created a bona fide decomposition of the total emergent degrees of freedom into constituent parts—attributed to variance, bias, and covariate shift (first three terms above, respectively). The same idea can be extended to an arbitrary number of sources of error.

As a simple example, we revisit the setup used in the first row of Figure 7, but introduce covariate shift by drawing test features from a scaled and shifted version of the training feature distribution. Figure 9 displays the degrees of freedom of ridge, kNN, and random forest predictors broken down into components due to variance, bias, and covariate shift. Figure 10 shows the same quantities, but restricted to the optimally-tuned model within each class (which minimizes out-of-distribution prediction error). We can see that the kNN predictor exhibits the smallest intrinsic complexity; the ridge predictor exhibits the smallest complexity due to bias (recall, the true model here is linear); and quite interestingly, random forests display by the smallest complexity due to covariate shift.

8 Discussion

A high-level summary of our proposal is as follows. In order to define the complexity of an arbitrary prediction model, we consider two critical components: a metric and a reference model. The metric

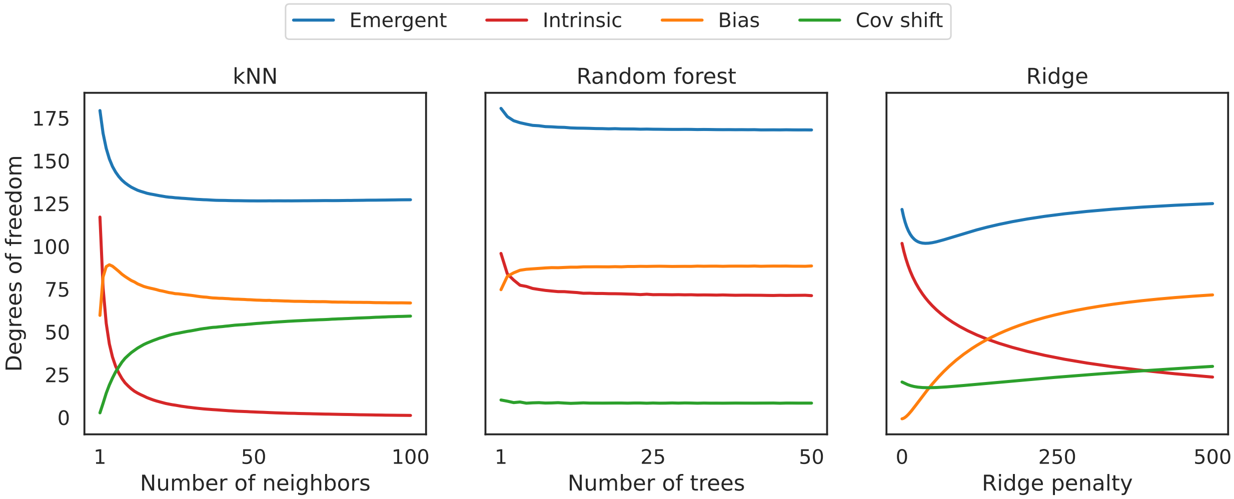

Figure 9: Decomposing degrees of freedom for ridge, kNN, and random forest predictors. The setup is as in the top row of Figure 7 but with covariate shift.

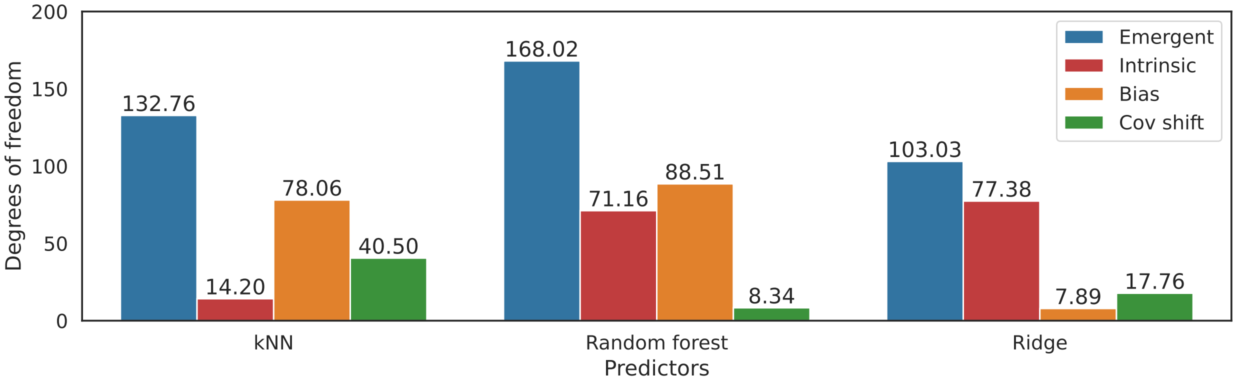

Figure 10: Decomposition for the optimally-tuned model within each class in Figure 9.

quantifies complexity, while the reference model provides context, which is analogous adding units to a measurement. Precisely, we define the complexity of a given model as the number of parameters in the reference model we require in order to obtain an identical value of the metric.

Each choice of a metric and reference model gives rise to a different notion of model complexity. In fact, if we take the metric to be fixed-X optimism (the difference between random-X prediction error and training error) as the metric, and least squares as the reference model, then this formulation reproduces the classical notion of (effective) degrees of freedom. With this motivation in mind, we focused on examining random-X optimism (the difference between random-X prediction error and training error) as the metric, and least squares as the reference model, which allowed us to define a new random-X notion of degrees of freedom.

By changing the metric—to measure the random-X optimism when the given model is run on pure noise, we can isolate the degrees of freedom due to variance, which we call the intrinsic random-X degrees of freedom. Then, taking the difference between the original random-X degrees of freedom and this intrinsic version allows us to isolate the degrees of freedom due to bias. A similar idea can be used to isolate the degrees of freedom due to other sources of error in settings with distribution shift, such as covariate shift.

Of course, our choice to focus on regression, and metrics based on squared error, does not reflect a fundamental restriction. An interesting direction for follow-up work would be to use the framework we proposed in this paper in order to study model complexity in classification.

Acknowledgements

Preliminary versions of our work were presented at a workshop on the Theory of Overparameterized Machine Learning (TOPML) in 2022, and the Joint Statistical Meetings (JSM) in 2023. We thank the participants for their comments and feedback.

PP and JHD thank Pierre Bellec and Takuya Koriyama for useful discussions about the degrees of freedom of convex regularized M-estimators. RJT thanks Trevor Hastie, Saharon Rosset, and Rob Tibshirani for many insightful conversations about optimism and model complexity over the years.

References

Hirotogu Akaike. Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory*, pages 267–281, 1973.

Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024.

Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010.

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

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

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: A statistical viewpoint. *Acta Numerica*, 30:87–201, 2021.

Mohsen Bayati and Andrea Montanari. The lasso risk for Gaussian matrices. *IEEE Transactions on Information Theory*, 58(4):1997–2017, 2011.

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

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

Misha Belkin. Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation. *Acta Numerica*, 30:203–248, 2021.

Pierre C. Bellec. Out-of-sample error estimate for robust M-estimators with convex penalty. *Information and Inference*, 12(4):2782–2817, 2023.

Pierre C. Bellec and Alexandre Tsybakov. Bounds on the prediction error of penalized least squares estimators with convex penalty. In *Modern Problems of Stochastic Analysis and Statistics*, 2017.

Blake Bordelon, Abdulkadir Canatar, and Cengiz Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. In *International Conference on Machine Learning*, pages 1024–1034, 2020.

Leo Breiman and David Freedman. How many variables should be entered in a regression equation? *Journal of the American Statistical Association*, 78(381):131–136, 1983.

Leo Breiman and Philip Spector. Submodel selection and evaluation in regression. The x-random case. *International Statistical Review*, 60(3):291–319, 1992.

Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *Annals of Statistics*, 17(2):453–510, 1989.

T. Tony Cai. Adaptive wavelet estimation: A block thresholding and oracle inequality approach. *Annals of Statistics*, 27(3):898–924, 1999.

Emmanuel J. Candès, Carlos M. Sing-Long, and Joshua D. Trzasko. Unbiased risk estimates for singular value thresholding and spectral estimators. *IEEE Transactions on Signal Processing*, 61(19):4643–4657, 2013.

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

Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general gaussian designs with applications to hypothesis testing. *The Annals of Statistics*, 51(5):2194–2220, 2023.

Xi Chen, Qihang Lin, and Bodhisattva Sen. On degrees of freedom of projection estimators with applications to multivariate nonparametric regression. *Journal of the American Statistical Association*, 115(529): 173–186, 2020.

Chen Cheng and Andrea Montanari. Dimension free ridge regression. *arXiv preprint arXiv:2210.08571*, 2022.

Peter Craven and Grace Wahba. Smoothing noisy data with spline functions. *Numerische Mathematik*, 31 (4):377–403, 1978.

Alicia Curth, Alan Jeffares, and Mihaela van der Schaar. A u-turn on double descent: Rethinking parameter counting in statistical learning. *arXiv preprint arXiv:2310.18988*, 2023.

Lee H. Dicker. Optimal equivariant prediction for high-dimensional linear models with arbitrary predictor covariance. *Electronic Journal of Statistics*, 7:1806–1834, 2013.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *Annals of Statistics*, 46(1):247–279, 2018.

David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

Jin-Hong Du, Pratik Patil, and Arun Kumar Kuchibhotla. Subsample ridge ensembles: Equivalences and generalized cross-validation. In *International Conference on Machine Learning*, 2023.

Rick Durrett. *Probability: Theory and Examples.* Cambridge University Press, 2010. Fourth edition.

Bradley Efron. Estimating the error rate of a prediction rule: Improvement on cross-validation. *Journal of the American Statistical Association*, 78(382):316–331, 1983.

Bradley Efron. How biased is the apparent error rate of a prediction rule? *Journal of the American Statistical Association*, 81(394):461–470, 1986.

Bradley Efron. The estimation of prediction error: Covariance penalties and cross-validation. *Journal of the American Statistical Association*, 99(467):619–632, 2004.

Vyacheslav L. Girko. *Theory of Random Determinants.* Kluwer Academic Publishers, 1990.

Vyacheslav L. Girko. *Statistical Analysis of Observations of Increasing Dimension.* Kluwer Academic Publishers, 1995.

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

Trevor Hastie and Robert Tibshirani. Generalized additive models: Some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

Trevor Hastie and Robert Tibshirani. *Generalized Additive Models.* Chapman & Hall, 1990.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2):949–986, 2022.

Ronald R. Hocking. The analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49, 1976.

Daniel J. Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14:569–600, 2014.

Lucas Janson, William Fithian, and Trevor Hastie. Effective degrees of freedom: A flawed metaphor. *Biometrika*, 102(2):479–485, 2015.

Andrey Kolmogorov. On tables of random numbers. *Sankhya: The Indian Journal of Statistics, Series A*, 25 (4):369–375, 1963.

Takuya Koriyama, Pratik Patil, Jin-Hong Du, Kai Tan, and Pierre C. Bellec. Precise asymptotics of bagging regularized M-estimators. *arXiv preprint arXiv:2409.15252*, 2024.

Daniel LeJeune, Pratik Patil, Hamid Javadi, Richard G. Baraniuk, and Ryan J. Tibshirani. Asymptotics of the sketched pseudoinverse. *SIAM Journal on Mathematics of Data Science*, 6(1):199–225, 2024.

Yue Li and Yuting Wei. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*, 2021.

Bo Luan, Yoonkyung Lee, and Yunzhang Zhu. Predictive model degrees of freedom in linear regression. *arXiv preprint arXiv:2106.15682*, 2021.

Bo Luan, Yoonkyung Lee, and Yunzhang Zhu. On measuring model complexity in heteroscedastic linear regression. *arXiv preprint arXiv:2204.07021*, 2022.

Neil Mallinar, James B. Simon, Amirhesam Abedsoltan, Parthe Pandit, Mikhail Belkin, and Preetum Nakkiran. Benign, tempered, or catastrophic: A taxonomy of overfitting. *arXiv preprint arXiv:2207.06569*, 2022.

Colin Mallows. Some comments on C_p. *Technometrics*, 15(4):661–675, 1973.

Mary Meyer and Michael Woodroofe. On the degrees of freedom in shape-restricted regression. *Annals of Statistics*, 28(4):1083–1104, 2000.

Frederik Riis Mikkelsen and Niels Richard Hansen. Degrees of freedom for piecewise Lipschitz estimators. *Annales de l'Institut Henri Poincaré Probabilités et Statistiques*, 54(2):819–841, 2018.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of Machine Learning.* MIT Press, 2018.

Natalia L. Oliveira, Jing Lei, and Ryan J. Tibshirani. Unbiased risk estimation in the normal means problem via coupled bootstrap techniques. *arXiv preprint arXiv:2111.09447*, 2021.

Natalia L. Oliveira, Jing Lei, and Ryan J. Tibshirani. Unbiased test error estimation in the poisson means problem via coupled bootstrap techniques. *arXiv preprint arXiv:2212.01943*, 2022.

Pratik Patil and Jin-Hong Du. Generalized equivalences between subsampling and ridge regularization. In *Neural Information Processing Systems*, 2023.

Pratik Patil, Yuting Wei, Alessandro Rinaldo, and Ryan J. Tibshirani. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3178–3186. PMLR, 2021.

Pratik Patil, Arun Kumar Kuchibhotla, Yuting Wei, and Alessandro Rinaldo. Mitigating multiple descents: A model-agnostic framework for risk monotonization. *arXiv preprint arXiv:2205.12937*, 2022.

Pratik Patil, Jin-Hong Du, and Ryan J. Tibshirani. Optimal ridge regularization for out-of-distribution prediction. In *International Conference on Machine Learning*, 2024.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot, and Édouard Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

Saharon Rosset and Ryan J. Tibshirani. From fixed-X to random-X regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, 15(529):138–151, 2020.

Gideon Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.

Stanley L. Sclove. On criteria for choosing a regression equation for prediction. Technical report, University of Illinois, Chicago, 1969.

A. V. Serdobolskii. Solution of empirical SLAE unimprovable in the mean. *Review of Applied and Industrial Mathematics*, 8(1):321–326, 2001.

A. V. Serdobolskii. Unimprovable solution to systems of empirical linear algebraic equations. *Statistics and Probability Letters*, 60(1):1–6, 2002.

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

Lloyd S. Shapley. *A value for n-person games*. Princeton University Press, 1953.

Peter Sollich. Gaussian process regression with mismatched models. In *Neural Information Processing Systems*, 2001.

Charles Stein. Multiple regression. In *Contributions to Probability and Statistics: Essays in Honor of Harold Hotelling*, pages 424–443. Stanford University Press, 1960.

Charles Stein. Estimation of the mean of a multivariate normal distribution. *Annals of Statistics*, 9(6):1135–1151, 1981.

Mary L. Thompson. Selection of variables in multiple regression: Part I. A review and evaluation. *International Statistical Review*, 46(1):1–19, 1978a.

Mary L. Thompson. Selection of variables in multiple regression: Part II. Chosen procedures, computations and examples. *International Statistical Review*, 46(2):129–146, 1978b.

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

Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288, 1996.

Ryan J. Tibshirani. The lasso problem and uniqueness. *Electronic Journal of Statistics*, 7:1456–1490, 2013.

Ryan J. Tibshirani. Degrees of freedom and model search. *Statistica Sinica*, 25(3):1265–1296, 2015.

Ryan J. Tibshirani and Jonathan Taylor. Degrees of freedom in lasso problems. *Annals of Statistics*, 40(2): 1198–1232, 2012.

David N. C. Tse and Stephen V. Hanly. Linear multiuser receivers: effective interference, effective bandwidth and user capacity. *IEEE Transactions on Information Theory*, 45(2):641–657, 1999.

David N. C. Tse and Ofer Zeitouni. Linear multiuser receivers in random environments. *IEEE Transactions on Information Theory*, 46(1):171–188, 2000.

John W. Tukey. Discussion of "Topics in the investigation of linear relations fitted by the method of least squares". *Journal of the Royal Statistical Society: Series B*, 29(1):2–52, 1967.

Vladimir N. Vapnik and Alexey Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications*, 16(2):264–280, 1971.

Sergio Verdu. *Multiuser Detection.* Cambridge University Press, 1998.

Sergio Verdu and Slomo Shamai. Multiuser detection with random spreading and error-correction codes: Fundamental limits. *Conference on Communications, Control, and Computing*, 1997.

Chris S. Wallace and D. M Boulton. An information measure for classification. *The Computer Journal*, 11 (2):185–194, 1968.

Shuaiwen Wang, Haolei Weng, and Arian Maleki. Which bridge estimator is the best for variable selection? *Annals of Statistics*, 48(5):2791–2823, 2020.

Haolei Weng, Arian Maleki, and Le Zheng. Overcoming the limitations of phase transition by higher order analysis of regularization techniques. *Annals of Statistics*, 46(6):3099–3129, 2018.

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

Tong Zhang. Learning bounds for kernel regression using effective data dimensionality. *Neural Computation*, 17(9):2077–2098, 2005.

Lijia Zhou, James B. Simon, Gal Vardi, and Nathan Srebro. An agnostic view on the cost of overfitting in (kernel) ridge regression. *arXiv preprint arXiv:2306.13185*, 2023.

Hui Zou and Ming Yuan. Regularized simultaneous model selection in multiple quantiles regression. *Computational Statistics and Data Analysis*, 52(12):5296–5304, 2008.

Hui Zou, Trevor Hastie, and Robert Tibshirani. On the "degrees of freedom" of the lasso. *Annals of Statistics*, 35(5):2173–2192, 2007.

A Proofs for Sections 3 and 4

A.1 Derivation of right-hand side in (16)

Recall the least squares regression estimator on $(\widetilde{X}_d, \widetilde{y})$ is given by

$$\widehat{\beta}^{\mathrm{ls}} = (\widetilde{X}_d^\top \widetilde{X}_d)^{-1} \widetilde{X}_d^\top \widetilde{y}.$$

The predicted values at the design points are

$$\widehat{f}^{\mathrm{ls}}(\widetilde{X}_d) = \widetilde{X}_d \widehat{\beta}^{\mathrm{ls}} = L_d \widetilde{y},$$

where $L_d = \widetilde{X}_d(\widetilde{X}_d^\top \widetilde{X}_d)^{-1}\widetilde{X}_d^\top \in \mathbb{R}^{n \times n}$ is the smoothing matrix for the least squares estimator. The training error is thus

$$\begin{aligned}
\mathbb{E}\bigg[\frac{1}{n}\|\widetilde{y} - L_d \widetilde{y}\|_2^2\bigg] &= \frac{\sigma^2}{n}\mathbb{E}[\mathrm{tr}(I - L_d)^2] \\
&= \frac{\sigma^2}{n}\mathbb{E}[\mathrm{tr}(I - L_d)] \\
&= \sigma^2(1 - d/n).
\end{aligned}$$

In the second equality above, we used the fact that the matrix $I - L_d$ is idempotent. Now let $(\widetilde{x}_0, \widetilde{y}_0)$ denote a sample which is i.i.d. to the training data $(\widetilde{x}_i, \widetilde{y}_i)$, $i \in [n]$. Conditional on $\widetilde{X}_d, \widetilde{x}_0$, we can decompose the random-X prediction error of the least squares predictor $\widehat{f}^{\mathrm{ls}}$ into irreducible error squared bias plus variance, as usual:

$$\begin{aligned}
\mathbb{E}\big[\big(\widetilde{y}_0 - \widehat{f}^{\mathrm{ls}}(\widetilde{x}_0)\big)^2 \,|\, \widetilde{X}_d, \widetilde{x}_0\big] &= \sigma^2 + \big(\mathbb{E}[\widetilde{x}_0^\top \widehat{\beta}^{\mathrm{ls}} \,|\, \widetilde{X}_d, \widetilde{x}_0] - \widetilde{x}_0^\top \beta\big)^2 + \mathrm{Var}[\widetilde{x}_0^\top \widehat{\beta}^{\mathrm{ls}} \,|\, \widetilde{X}_d, \widetilde{x}_0] \\
&= \sigma^2 + \big(\widetilde{x}_0^\top (\widetilde{X}_d^\top \widetilde{X}_d)^{-1}\widetilde{X}_d^\top \widetilde{X}_d \beta - \widetilde{x}_0^\top \beta\big)^2 + \sigma^2 \widetilde{x}_0^\top (\widetilde{X}_d^\top \widetilde{X}_d)^{-1}\widetilde{x}_0 \\
&= \sigma^2\big(1 + \widetilde{x}_0^\top (\widetilde{X}_d^\top \widetilde{X}_d)^{-1}\widetilde{x}_0\big).
\end{aligned}$$

where in the second line we used $\widetilde{y} = \widetilde{X}_d \beta + v$, where $\mathbb{E}[v|\widetilde{X}_d] = 0$ and $\mathrm{Cov}[v|\widetilde{X}_d] = \sigma^2 I$. Taking an expectation over $\widetilde{x}_0$, which is independent of $\widetilde{X}_d$, gives

$$\mathbb{E}\big[\big(\widetilde{y}_0 - \widehat{f}^{\mathrm{ls}}(\widetilde{x}_0)\big)^2 \,|\, \widetilde{X}_d\big] = \sigma^2\big(1 + \mathrm{tr}[\Sigma(\widetilde{X}_d^\top \widetilde{X}_d)^{-1}]\big).$$

Finally, taking an expectation over $\widetilde{X}_d$, and using the fact that $(\widetilde{X}_d^\top \widetilde{X}_d)^{-1} \sim W^{-1}(\Sigma^{-1}, n)$ (inverse Wishart distributed),

$$\begin{aligned}
\mathbb{E}\big[\big(\widetilde{y}_0 - \widehat{f}^{\mathrm{ls}}(\widetilde{x}_0)\big)^2\big] &= \sigma^2\left(1 + \mathrm{tr}\left[\Sigma \frac{\Sigma^{-1}}{n - d - 1}\right]\right) \\
&= \sigma^2\left(1 + \frac{d}{n - d - 1}\right).
\end{aligned}$$

The random-X optimism is therefore given by

$$\begin{aligned}
\mathsf{opt}_{\mathrm{R}}(\widehat{f}^{\mathrm{ls}}) &= \sigma^2\left(1 + \frac{d}{n - d - 1}\right) - \sigma^2\left(1 - \frac{d}{n}\right) \\
&= \sigma^2\left(\frac{d}{n} + \frac{d}{n - d - 1}\right),
\end{aligned}$$

which completes the derivation.

A.2 Proof of approximation result in (24)

Let $z = d/n$, and rewrite

$$x = \frac{d}{n} + \frac{d}{n-d-1}$$

as

$$x = z + \frac{z}{1-z-1/n} \iff z^2 - (x+1-1/n)z + (1-1/n)x = 0.$$

Note that we can write $\omega_n(x)/n$ as a solution of the above quadratic equation in z,

$$\omega_n(x)/n = \frac{b_n - \sqrt{b_n^2 - 4c_n}}{2},$$

where we define

$$b_n = x + 1 - 1/n \quad \text{and} \quad c_n = 1 - 1/n.$$

Meanwhile, one can check that $\omega(x)$ solves

$$x = z + \frac{z}{1-z} \iff z^2 - (x+1)z + x = 0,$$

and indeed we can write

$$\omega(x) = \frac{b - \sqrt{b^2 - 4c}}{2},$$

where $b = x+1$ and $c = 1$. The desired fact (24) therefore follows using $b_n \to b$, $c_n \to c$, and using continuity.

B Proofs for Section 5

B.1 Proof of Proposition 5

Due to the monotonicity of ω_n in (22), it suffices to show that the intrinsic random-X optimism $\mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_\lambda)$ is decreasing in λ. From (27), recall, this is

$$\mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_\lambda) = \sigma^2 \mathbb{E}\bigg[\frac{2}{n}\operatorname{tr}[L_X(X)] + \mathbb{E}[L_X(x_0)^\top L_X(x_0) \,|\, X] - \frac{1}{n}\operatorname{tr}[L_X(X)^\top L_X(X)]\bigg],$$

where recall for ridge, we have $L_X(x) = X(\widehat{\Sigma} + \lambda I)^{-1}x$, with $\widehat{\Sigma} = X^\top X/n$. From Proposition 2 of Rosset and Tibshirani (2020), we know that the middle term (which is $V + V^+$ in their notation) is decreasing in λ. For the first and last term, writing $s_i \geqslant 0$, $i \in [p]$ for the eigenvalues of $\widehat{\Sigma}$, observe

$$\begin{aligned}\frac{2}{n}\operatorname{tr}[L_X(X)] - \frac{1}{n}\operatorname{tr}[L_X(X)^\top L_X(X)] &= \sum_{i=1}^p \bigg(\frac{2s_i}{s_i+\lambda} - \frac{s_i^2}{(s_i+\lambda)^2}\bigg) \\ &= \sum_{i=1}^p \frac{2s_i^2 + 2\lambda s_i - s_i^2}{(s_i+\lambda)^2} \\ &= \sum_{i=1}^p \bigg(1 - \frac{\lambda^2}{(s_i+\lambda)^2}\bigg).\end{aligned}$$

Each summand here is decreasing in λ, which means that their sum is, and hence this remains true after taking an expectation with respect to X. This completes the proof.

B.2 Proof of Theorem 6

Throughout the proof, we will use the language of asymptotic equivalents. For sequences $\{A_p\}_{p\geqslant 1}$ and $\{B_p\}_{p\geqslant 1}$ of (random or deterministic) matrices of growing dimension, we say that A_p and B_p are asymptotically equivalent, and write this as $A_p \simeq B_p$, provided $\lim_{p\to\infty} |\operatorname{tr}[C_p(A_p - B_p)]| = 0$ almost surely for any sequence $\{C_p\}_{p\geqslant 1}$ of matrices with bounded trace norm, $\limsup \|C_p\|_{\operatorname{tr}} < \infty$ as $p \to \infty$. The notion of asymptotic equivalence satisfies various calculus rules that we will use in our proofs. We refer readers to Lemma E.3 of Patil and Du (2023) for a list of these rules.

We collect below three equivalences that we will use in the proofs. These are standard and we refer readers to Section S.6.5 of Patil et al. (2022) for more details.

Lemma 14. *Under Assumption A.1, as $n, p \to \infty$ with $0 < \liminf_{n\to\infty} \gamma_n \leqslant \limsup_{n\to\infty} \gamma_n < \infty$, the following asymptotic equivalences hold for any $\lambda > 0$:*

1. *First-order basic equivalence:*

$$\lambda(\widehat{\Sigma} + \lambda I)^{-1} \simeq (v(\lambda; \gamma_n)\Sigma + I)^{-1}, \tag{66}$$

where $v_n = v(\lambda; \gamma_n) > 0$ is the unique solution to the fixed point equation:

$$v_n^{-1} = \lambda + \gamma_n \,\overline{\operatorname{tr}}[\Sigma(v_n\Sigma + I)^{-1}], \tag{67}$$

2. *Second-order variance-type equivalence:*

$$(\widehat{\Sigma} + \lambda I)^{-1}\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-1} \simeq \widetilde{v}_v(\lambda; \gamma_n)(v(\lambda; \gamma_n)\Sigma + I)^{-1}\Sigma(v(\lambda; \gamma_n)\Sigma + I)^{-1}, \tag{68}$$

where $\widetilde{v}_v(\lambda; \gamma_n)$ is defined through $v(\lambda; \gamma_n)$ via the following equation:

$$\widetilde{v}_v(\lambda; \gamma_n) = \frac{1}{v(\lambda; \gamma_n)^{-2} - \gamma_n \,\overline{\operatorname{tr}}[\Sigma^2(v(\lambda; \gamma_n)\Sigma + I)^{-2}]}.$$

3. *Second-order bias-type equivalence:*

$$\lambda^2(\widehat{\Sigma} + \lambda I)^{-1}A(\widehat{\Sigma} + \lambda I)^{-1} \simeq (v(\lambda; \gamma_n)\Sigma + I)^{-1}(\widetilde{v}_b(\lambda; \gamma_n, A)\Sigma + A)(v(\lambda; \gamma_n)\Sigma + I)^{-1}, \tag{69}$$

for any matrix $A \in \mathbb{R}^{p\times p}$ with bounded operator norm which is independent of $\widehat{\Sigma}$, and where $\widetilde{v}_b(\lambda; \gamma_n, A)$ is defined through $v(\lambda; \gamma_n)$ by the following equation:

$$\widetilde{v}_b(\lambda; \gamma_n, A) = \frac{\gamma_n \,\overline{\operatorname{tr}}[A\Sigma(v(\lambda; \gamma_n)\Sigma + I)^{-2}]}{v(\lambda; \gamma_n)^{-2} - \gamma_n \,\overline{\operatorname{tr}}[\Sigma^2(v(\lambda; \gamma_n)\Sigma + I)^{-2}]}.$$

With this background, we are now ready to derive the asymptotic equivalents for the fixed-X and random-X degrees of freedom of the ridge predictor below. Note that v_n in (67) is the reciprocal of μ_n in (36).

Fixed-X degrees of freedom. Recall that the fixed-X degrees of freedom of ridge regression is:

$$\begin{aligned} \mathsf{df}_{\mathrm{F}}(\widehat{f}_\lambda^{\mathrm{ridge}})/n &= \operatorname{tr}[L_X(X)]/n \\ &= \gamma_n \,\overline{\operatorname{tr}}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-1}] \\ &= \gamma_n \,\overline{\operatorname{tr}}[I - \lambda(\widehat{\Sigma} + \lambda I)^{-1}] \end{aligned}$$

$$= \gamma_n - \gamma_n \lambda \,\overline{\mathrm{tr}}[(\widehat{\Sigma} + \lambda I)^{-1}].$$

Thus, using (66), we have the following asymptotic equivalence:

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}_\lambda^{\mathrm{ridge}})/n \simeq \gamma_n - \gamma_n \,\overline{\mathrm{tr}}[(v_n \Sigma + I)^{-1}] = \gamma_n \,\overline{\mathrm{tr}}[v_n \Sigma (v_n \Sigma + I)^{-1}]. \tag{70}$$

Now, multiplying the fixed point equation (67) by v_n, note that that the final expression in (70) is simply $1 - \lambda v_n$. In addition, substituting $\mu_n = v_n^{-1}$ yields the final expression in (37), as desired.

Intrinsic random-X degrees of freedom. Recall from (27) that the intrinsic random-X optimism of ridge regression is:

$$\mathsf{opt}_{\mathrm{R}}^{\mathrm{i}}(\widehat{f}_\lambda^{\mathrm{ridge}} \mid X)/\sigma^2 = 2 \,\mathrm{tr}[L_X(X)]/n + \mathbb{E}_{x_0}[L_X(x_0)^\top L_X(x_0)] - \mathrm{tr}[L_X(X)^\top L_X(X)]/n. \tag{71}$$

We now rewrite the three terms in (71) to make them amenable for applications of asymptotic equivalents described in the background above.

On one hand, note that:

$$\begin{aligned}
&2 \,\mathrm{tr}[L_X(X)]/n - \mathrm{tr}[L_X(X)^\top L_X(X)]/n \\
&= -\,\mathrm{tr}[(I - L_X(X))^2]/n + 1 \\
&= -\,\mathrm{tr}[(I - L_X(X))]/n + \mathrm{tr}[L_X(X)(I - L_X(X))]/n + 1 \\
&= -1 + \mathrm{tr}[L_X(X)]/n + \mathrm{tr}[L_X(X)(I - L_X(X))]/n + 1 \\
&= -1 + \gamma_n - \lambda \,\mathrm{tr}[(\widehat{\Sigma} + I)^{-1}]/n + \lambda \,\mathrm{tr}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2}]/n + 1 \\
&= \gamma_n - \lambda \,\mathrm{tr}[(\widehat{\Sigma} + I)^{-1}]/n + \lambda \,\mathrm{tr}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2}]/n.
\end{aligned} \tag{72}$$

On the other hand, note that:

$$\begin{aligned}
\mathbb{E}[\mathrm{tr}[L_X(x_0)^\top L_X(x_0)]] &= \mathrm{tr}[\Sigma \widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2}]/n \\
&= \mathrm{tr}[\Sigma(\widehat{\Sigma} + \lambda I)^{-1}]/n - \lambda \,\mathrm{tr}[\Sigma(\widehat{\Sigma} + \lambda)^{-2}]/n.
\end{aligned} \tag{73}$$

Substituting (72), (73) into (71), our goal is reduced to obtaining an asymptotic equivalent for:

$$\begin{aligned}
\mathsf{opt}_{\mathrm{R}}^{\mathrm{i}}(\widehat{f}_\lambda^{\mathrm{ridge}})/\sigma^2 &= \mathrm{tr}[\Sigma(\widehat{\Sigma} + \lambda I)^{-1}]/n - \lambda \,\mathrm{tr}[\Sigma(\widehat{\Sigma} + \lambda)^{-2}]/n \\
&\quad + \gamma_n - \lambda \,\mathrm{tr}[(\widehat{\Sigma} + I)^{-1}]/n + \lambda \,\mathrm{tr}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2}]/n \\
&= (\mathrm{tr}[\Sigma(\widehat{\Sigma} + \lambda I)^{-1}]/n + 1) - (1 - \gamma_n + \lambda \,\mathrm{tr}[(\widehat{\Sigma} + \lambda I)^{-1}]/n) \\
&\quad - (\lambda \,\mathrm{tr}[\Sigma(\widehat{\Sigma} + \lambda)^{-2}]/n - \lambda \,\mathrm{tr}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2}]/n).
\end{aligned} \tag{74}$$

Now observe that for the first line in (74):

$$\begin{aligned}
1 - \gamma_n + \lambda \,\mathrm{tr}[(\widehat{\Sigma} + \lambda I)^{-1}]/n &\simeq 1 - \gamma_n + \gamma_n \,\overline{\mathrm{tr}}[(v_n \Sigma + I)^{-1}] \\
&= 1 - \gamma_n + \lambda v_n + \gamma_n - 1 \\
&= \lambda v_n (1 - \lambda v_n) + \lambda^2 v_n^2 \\
&= \lambda v_n^2 \gamma_n \,\overline{\mathrm{tr}}[\Sigma(v_n \Sigma + I)^{-1}] + \lambda^2 v_n^2 \\
&\simeq \lambda^2 v_n^2 (\mathrm{tr}[\Sigma(\widehat{\Sigma} + \lambda I)^{-1}]/n + 1).
\end{aligned} \tag{75}$$

Similarly, observe that for the second line in (74):

$$
\begin{aligned}
\operatorname{tr}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2}]/n &= \gamma_n \,\overline{\operatorname{tr}}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2}] \\
&\simeq \frac{1}{v_n^{-2} - \gamma_n \,\overline{\operatorname{tr}}[\Sigma^2 (v_n \Sigma + I)^{-2}]} \cdot \gamma_n \,\overline{\operatorname{tr}}[\Sigma (v_n \Sigma + I)^{-2}] \\
&= v_n^2 \left(\frac{\gamma_n \,\overline{\operatorname{tr}}[\Sigma^2 (v_n \Sigma + I)^{-2}]}{v_n^{-2} - \gamma_n \,\overline{\operatorname{tr}}[\Sigma^2 (v_n \Sigma + I)^{-2}]} + 1 \right) \cdot \gamma_n \,\overline{\operatorname{tr}}[\Sigma (v_n \Sigma + I)^{-2}] \\
&\simeq \lambda^2 v_n^2 \operatorname{tr}[\Sigma(\widehat{\Sigma} + \lambda I)^{-2}]/n. \qquad (76)
\end{aligned}
$$

Hence, substituting (75) and (76) into (74), we have

$$
\begin{aligned}
\mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_{\lambda} \,|\, X)/\sigma^2 &\simeq (1 - \lambda^2 v_n^2)(\operatorname{tr}[\Sigma(\widehat{\Sigma} + \lambda I)^{-1}]/n + 1 - \lambda \operatorname{tr}[\Sigma(\widehat{\Sigma} + \lambda)^{-2}]/n) \\
&= (1 - \lambda^2 v_n^2)(\operatorname{tr}[\Sigma \widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-2}]/n + 1) \\
&= (1 - \lambda^2 v_n^2)\left(\frac{1}{v_n^{-2} - \gamma_n \,\overline{\operatorname{tr}}[\Sigma^2 (v_n \Sigma + I)^{-2}]} \gamma_n \,\overline{\operatorname{tr}}[\Sigma^2 (v_n \Sigma + I)^{-2}] + 1 \right) \\
&\simeq (1 - \lambda^2 v_n^2)\left(\frac{\gamma_n \,\overline{\operatorname{tr}}[\Sigma^2 (\Sigma + \mu_n I)^{-2}]}{1 - \gamma_n \,\overline{\operatorname{tr}}[\Sigma^2 (\Sigma + \mu_n I)^{-2}]} + 1 \right) \\
&= (1 - \lambda^2 v_n^2)\left(\frac{V_n}{D_n} + 1 \right),
\end{aligned}
$$

where in the second-to-last step above, we used $\mu_n = v_n^{-1}$ to simplify the expressions. Now applying the mapping ω to bring on the degrees of freedom scale, we have that

$$
\omega(\mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_{\lambda} \,|\, X)/\sigma^2) \simeq \omega((1 - \lambda^2/\mu^2)(V_n/D_n + 1)). \qquad (77)
$$

Note that the right-hand side of (77) is always bounded by 1 (by construction), thus we can apply the dominated convergence theorem to conclude that same asymptotic equivalence (77) holds after we take an expectation with respect to X. This yields the result in (38).

Emergent random-X degrees of freedom. Using (32), the only additional quantity we need to deal with is the excess bias, whose asymptotic equivalent we will derive next.

Recalling (35), let us abbreviate $f_{\mathrm{LI}}(x) = x^\top \beta$ and thus $f(x) = f_{\mathrm{LI}}(x) + f_{\mathrm{NL}}(x)$. We can decompose the excess bias $B^+ = B^+(\widehat{f}^{\mathrm{ridge}}_{\lambda})$ into linear, nonlinear, and cross components as follows:

$$
\begin{aligned}
B^+ &= \mathbb{E}_{x_0}[(f(x_0) - L_X(x_0)^\top f(X))^2] - \|(I - L_X(X))f(X)\|_2^2/n \\
&= \mathbb{E}_{x_0}[(f_{\mathrm{LI}}(x_0) + f_{\mathrm{NL}}(x_0) - L_X(x_0)^\top (f_{\mathrm{LI}}(X) + f_{\mathrm{NL}}(X)))^2] \\
&\quad - \|(I - L_X(X))(f_{\mathrm{LI}}(X) + f_{\mathrm{NL}}(X))\|_2^2/n \\
&= B^+_{\mathrm{LI}} + B^+_{\mathrm{NL}} + C^+, \qquad (78)
\end{aligned}
$$

where B^+_{LI}, B^+_{NL}, and C^+ are defined as:

$$
\begin{aligned}
B^+_{\mathrm{LI}} &= \mathbb{E}_{x_0}[(f_{\mathrm{LI}}(x_0) - L_X(x_0)^\top f_{\mathrm{LI}}(X))^2] - \|(I - L_X(X))f_{\mathrm{LI}}(X)\|_2^2/n, \\
B^+_{\mathrm{NL}} &= \mathbb{E}_{x_0}[(f_{\mathrm{NL}}(x_0) - L_X(x_0)^\top f_{\mathrm{NL}}(X))^2] - \|(I - L_X(X))f_{\mathrm{NL}}(X)\|_2^2/n, \\
C^+ &= 2\mathbb{E}_{x_0}[(f_{\mathrm{LI}}(x_0) - L_X(x_0)^\top f_{\mathrm{LI}}(X))(f_{\mathrm{NL}}(x_0) - L_X(x_0)^\top f_{\mathrm{NL}}(X))] \\
&\quad - 2 f_{\mathrm{LI}}(X)^\top (I - L_X(X))^2 f_{\mathrm{NL}}(X).
\end{aligned}
$$

We will obtain the asymptotic equivalents for B^+_{LI}, B^+_{NL}, and C^+ separately below.

Asymptotic equivalent for C^+. For the first term in C^+, we have:

$$\begin{aligned}
&\mathbb{E}_{x_0}[(f_{\mathrm{LI}}(x_0) - L_X(x_0)^\top f_{\mathrm{LI}}(X))(f_{\mathrm{NL}}(x_0) - L_X(x_0)^\top f_{\mathrm{NL}}(X))] \\
&= -\mathbb{E}_{x_0}[f_{\mathrm{LI}}(x_0) L_X(x_0)^\top f_{\mathrm{NL}}(X)] + \mathbb{E}_{x_0}[f_{\mathrm{LI}}(X)^\top L_X(x_0) L_X(x_0)^\top f_{\mathrm{NL}}(X)].
\end{aligned}$$

Here we used the fact that $\mathbb{E}_{x_0}[f_{\mathrm{NL}}(x_0) f_{\mathrm{LI}}(x_0)] = 0$ and $\mathbb{E}_{x_0}[f_{\mathrm{NL}}(x_0) L_X(x_0)^\top f_{\mathrm{LI}}(X)] = 0$ because $\mathbb{E}_{x_0}[f_{\mathrm{NL}}(x_0) x_0] = 0$. For the remaining two terms in C^+, observe that:

$$\mathbb{E}_{x_0}[f_{\mathrm{LI}}(x_0) L_X(x_0)^\top f_{\mathrm{NL}}(X)] = \beta^\top \Sigma (\widehat{\Sigma} + \lambda I)^{-1} X^\top f_{\mathrm{NL}}(X)/n, \tag{79}$$

$$\mathbb{E}_{x_0}[f_{\mathrm{LI}}(X)^\top L_X(x_0) L_X(x_0)^\top f_{\mathrm{NL}}(X)] = \beta^\top X^\top/n (\widehat{\Sigma} + \lambda I)^{-1} \Sigma (\widehat{\Sigma} + \lambda I)^{-1} X^\top f_{\mathrm{NL}}(X)/n. \tag{80}$$

Invoking Lemma A.3 of Patil and Du (2023), we conclude that the right-hand sides of both (79) and (80) almost surely vanish. In a similar way, we can show that the second term of C^+ vanishes almost surely. Thus, we have $C^+ \simeq 0$.

Asymptotic equivalent for B^+_{LI}. For the linear component of excess bias, we have

$$\begin{aligned}
B^+_{\mathrm{LI}} &= \mathbb{E}_{x_0}[(f_{\mathrm{LI}}(x_0) - L_X(x_0)^\top f_{\mathrm{LI}}(X))^2] - \|(I - L_X(X)) f_{\mathrm{LI}}(X)\|_2^2/n \\
&= \mathbb{E}_{x_0}[(\beta^\top x_0 - \beta^\top X^\top L_X(x_0))^2] - \|(I - L_X(X)) X\beta\|_2^2/n \\
&= \mathbb{E}_{x_0}[(\beta^\top (I - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I)^{-1}) x_0)^2] - \|(I - X(\widehat{\Sigma} + \lambda I)^{-1} X^\top/n) X\beta\|_2^2/n \\
&= \lambda^2 \beta^\top (\widehat{\Sigma} + \lambda I)^{-1} \Sigma (\widehat{\Sigma} + \lambda I)^{-1} \beta - \|X(I - (\widehat{\Sigma} + \lambda I)^{-1} \widehat{\Sigma})\beta\|_2^2/n \\
&= \lambda^2 \beta^\top (\widehat{\Sigma} + \lambda I)^{-1} \Sigma (\widehat{\Sigma} + \lambda I)^{-1} \beta - \lambda^2 \beta^\top (\widehat{\Sigma} + \lambda I)^{-1} \widehat{\Sigma} (\widehat{\Sigma} + \lambda I)^{-1} \beta.
\end{aligned} \tag{81}$$

From (68) and (69), we have the following equivalence:

$$\begin{aligned}
B^+_{\mathrm{LI}}/\sigma^2 &\simeq (1 + \widetilde{v}_b(\lambda; \gamma_n, \Sigma)) \beta^\top (v_n \Sigma + I)^{-1} \Sigma (v_n \Sigma + I)^{-1} \beta \\
&\quad - \lambda^2 \widetilde{v}_v(\lambda; \gamma_n) \beta^\top (v(\lambda; \gamma_n)\Sigma + I)^{-1} \Sigma (v(\lambda; \gamma_n)\Sigma + I)^{-1} \beta \\
&\simeq (1 + \widetilde{v}_b(\lambda; \gamma_n, \Sigma)) \beta^\top (v_n \Sigma + I)^{-1} \Sigma (v_n \Sigma + I)^{-1} \beta \\
&\quad - \lambda^2 v_n^2 (1 + \widetilde{v}_b(\lambda; \gamma_n, \Sigma)) \beta^\top (v_n \Sigma + I)^{-1} \Sigma (v_n \Sigma + I)^{-1} \beta \\
&\simeq (1 - \lambda^2 v_n^2)(1 + \widetilde{v}_b(\lambda; \gamma_n, \Sigma)) \beta^\top (v_n \Sigma + I)^{-1} \Sigma (v_n \Sigma + I)^{-1} \beta/\sigma^2 \\
&= \frac{(1 - \lambda^2 v_n^2) \beta^\top (v_n \Sigma + I)^{-1} \Sigma (v_n \Sigma + I)^{-1} \beta/\sigma^2}{v_n^{-2} - \gamma_n \, \overline{\mathrm{tr}}[\Sigma^2 (v_n \Sigma + I)^{-2}]} \\
&= \frac{(1 - \lambda^2 v_n^2) \mu^2 \beta^\top (\Sigma + \mu_n I)^{-1} \Sigma (\Sigma + \mu_n I)^{-1} \beta/\sigma^2}{1 - \gamma_n \, \overline{\mathrm{tr}}[\Sigma^2 (\Sigma + \mu_n I)^{-2}]} \\
&= (1 - \lambda^2 v_n^2) \frac{B_n}{D_n},
\end{aligned}$$

where we again used the parameterization $\mu_n = v_n^{-1}$ to simplify the expression.

Asymptotic equivalent for B^+_{NL}. For the nonlinear component of excess bias, we have

$$\begin{aligned}
B^+_{\mathrm{NL}} &= \mathbb{E}_{x_0}[(f_{\mathrm{NL}}(x_0) - L_X(x_0)^\top f_{\mathrm{NL}}(X))^2] - \|(I - L_X(X)) f_{\mathrm{NL}}(X)\|_2^2/n \\
&= \mathbb{E}_{x_0}[f_{\mathrm{NL}}(X)^\top L_X(x_0) L_X(x_0)^\top f_{\mathrm{NL}}(X)] + \sigma_{\mathrm{NL}}^2 - \|(I - L_X(X)) f_{\mathrm{NL}}(X)\|_2^2/n \\
&= f_{\mathrm{NL}}(X)^\top (X(\widehat{\Sigma} + \lambda I)^{-1} \Sigma (\widehat{\Sigma} + \lambda I)^{-1} X^\top/n) f_{\mathrm{NL}}(X)/n + \sigma_{\mathrm{NL}}^2
\end{aligned}$$

$$
\begin{aligned}
&\quad - f_{\mathrm{NL}}(X)^\top (I - X(\widehat{\Sigma} + \lambda I)^{-1} X^\top / n)^2 f_{\mathrm{NL}} / n \\
&= f_{\mathrm{NL}}(X)^\top ((\widehat{\Sigma} + \lambda I)^{-1} X^\top / n)^\top \Sigma ((\widehat{\Sigma} + \lambda I)^{-1} X^\top / n) f_{\mathrm{NL}}(X) + \sigma_{\mathrm{NL}}^2 \qquad (82) \\
&\quad - f_{\mathrm{NL}}(X)^\top (X(\widehat{\Sigma} + \lambda I)^{-1} X^\top / n - I)^\top (X(\widehat{\Sigma} + \lambda I)^{-1} X^\top / n - I) f_{\mathrm{NL}}(X) / n. \qquad (83)
\end{aligned}
$$

In the second equality above, we used the facts that $\mathbb{E}_{x_0}[f_{\mathrm{NL}}(x_0) x_0] = 0$, and $\mathbb{E}_{x_0}[f_{\mathrm{NL}}(x_0)^2] = \sigma_{\mathrm{NL}}^2$. We will now use Part (2) of Lemma A.2 of Patil and Du (2023) get asymptotic equivalents for the first term in (82) and (83). We have

$$
f_{\mathrm{NL}}(X)^\top ((\widehat{\Sigma} + \lambda I)^{-1} X^\top / n)^\top \Sigma ((\widehat{\Sigma} + \lambda I)^{-1} X^\top / n) f_{\mathrm{NL}}(X) \simeq \frac{V_n}{D_n} \sigma_{\mathrm{NL}}^2
$$

and

$$
f_{\mathrm{NL}}(X)^\top (X(\widehat{\Sigma} + \lambda I)^{-1} X^\top / n - I)^\top (X(\widehat{\Sigma} + \lambda I)^{-1} X^\top / n - I) f_{\mathrm{NL}}(X) / n \simeq \lambda^2 v_n^2 \Big(\frac{V_n}{D_n} \sigma_{\mathrm{NL}}^2 + \sigma_{\mathrm{NL}}^2 \Big).
$$

Thus, we obtain the following asymptotic equivalent for B_{NL}^+:

$$
B_{\mathrm{NL}}^+ / \sigma^2 \simeq (1 - \lambda^2 v_n^2) \Big(\frac{V_n}{D_n} + 1 \Big) \frac{\sigma_{\mathrm{NL}}^2}{\sigma^2}.
$$

And hence, we have the overall asymptotic equivalent for B^+:

$$
B^+ / \sigma^2 \simeq (1 - \lambda^2 v_n^2) \bigg(\frac{B_n}{D_n} + \Big(\frac{V_n}{D_n} + 1 \Big) \frac{\sigma_{\mathrm{NL}}^2}{\sigma^2} \bigg).
$$

Combining this with the calculation above for the intrinsic random-X optimism then applying the mapping ω, followed by the dominated convergence theorem to convert this to an expectation over X, yields the desired equivalent in (39) and finishes the proof.

B.3 Proof of Theorem 7

The proof is similar to that in Appendix B.2. We will make use of various asymptotic equivalences for ridgeless regression from Section S.6.5 of Patil et al. (2022), collected in the lemma below.

Lemma 15. *Under Assumption A.1, as $n, p \to \infty$ with $0 < \liminf_{n \to \infty} \gamma_n \leqslant \limsup_{n \to \infty} \gamma_n < \infty$, the following asymptotic equivalences hold:*

1. *First-order basic equivalence:*

$$
I - \widehat{\Sigma} \widehat{\Sigma}^\dagger \simeq \begin{cases} 0 & \gamma_n \leqslant 1 \\ (v(0; \gamma_n) \Sigma + I)^{-1} & \gamma_n > 1, \end{cases} \tag{84}
$$

where $v_n = v(0; \gamma_n) > 0$ is the unique solution to the fixed-point equation:

$$
\frac{1}{v_n} = \gamma_n \, \overline{\mathrm{tr}}[\Sigma (v_n \Sigma + I)^{-1}]. \tag{85}
$$

2. *Second-order variance-type equivalence:*

$$
\widehat{\Sigma}^\dagger \widehat{\Sigma} \widehat{\Sigma}^\dagger \simeq \begin{cases} \dfrac{\Sigma^{-1}}{1 - \gamma_n} & \gamma_n \leqslant 1 \\ \widetilde{v}_v(0; \gamma_n) (v(0; \gamma_n) \Sigma + I_p)^{-1} \Sigma (v(0; \gamma_n) \Sigma + I_p)^{-1} & \gamma_n > 1, \end{cases} \tag{86}
$$

where $\widetilde{v}_v(0;\gamma)$ is defined through $v(0;\gamma)$ via

$$\widetilde{v}_v(0;\gamma) = \frac{1}{v(0;\gamma)^{-2} - \gamma\,\overline{\operatorname{tr}}[\Sigma^2(v(0;\gamma)\Sigma + I_p)^{-2}]}.$$

3. *Second-order bias-type equivalence:*

$$\begin{aligned}
&(I_p - \widehat{\Sigma}^{+}\widehat{\Sigma})\Sigma(I_p - \widehat{\Sigma}^{+}\widehat{\Sigma}) \\
&\simeq \begin{cases} 0 & \gamma_n \leqslant 1 \\ (1 + \widetilde{v}_b(0;\gamma_n))(v(0;\gamma_n)\Sigma + I_p)^{-1}\Sigma(v(0;\gamma_n)\Sigma + I_p)^{-1} & \gamma_n > 1, \end{cases}
\end{aligned} \tag{87}$$

where $v(0;\gamma)$ is as defined in (85), and $\widetilde{v}_b(0;\gamma)$ is defined via $v(0;\gamma)$ by

$$\widetilde{v}_b(0;\gamma) = \frac{\gamma\,\overline{\operatorname{tr}}[\Sigma^2(v(0;\gamma)\Sigma + I_p)^{-2}]}{v(0;\gamma)^{-2} - \gamma\,\overline{\operatorname{tr}}[\Sigma^2(v(0;\gamma)\Sigma + I_p)^{-2}]}.$$

Note: in (84) and (87) above we use 0 to denote the all-zero matrix in $\mathbb{R}^{p\times p}$. Also, v_n in (85) is the reciprocal of μ_n in (44).

We are now ready to obtain the asymptotic equivalents for the fixed-X and random-X degrees of freedom of the ridgeless predictor below.

Fixed-X degrees of freedom. Note that the smoother matrix for ridgeless regression can be written as $L_X(X) = X\widehat{\Sigma}^{\dagger}X^{\top}/n$. The fixed-X degrees of freedom is thus:

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}_0^{\mathrm{ridge}})/n = \operatorname{tr}[L_X(X)]/n = \operatorname{tr}[\widehat{\Sigma}\widehat{\Sigma}^{\dagger}]/n.$$

Now using (84), we have

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}_0^{\mathrm{ridge}})/n \simeq \begin{cases} \gamma_n\,\overline{\operatorname{tr}}[I] = \gamma_n & \gamma_n < 1 \\ \gamma_n(1 - \overline{\operatorname{tr}}[(v_{\Sigma} + I)^{-1}]) = \gamma_n\gamma_n\,\overline{\operatorname{tr}}[v_n\Sigma(v_n\Sigma + I)^{-1}] = 1 & \gamma_n > 1, \end{cases}$$

as desired.

Intrinsic random-X degrees of freedom. Since $L_X(x_0) = x_0^{\top}\widehat{\Sigma}^{\dagger}X^{\top}/n$ for ridgeless regression, the intrinsic random-X optimism of ridgeless regression is:

$$\begin{aligned}
\mathsf{opt}_{\mathrm{R}}^{\mathrm{i}}(\widehat{f}_0^{\mathrm{ridge}} \mid X)/\sigma^2 &= 2\operatorname{tr}[\widehat{\Sigma}\widehat{\Sigma}^{\dagger}]/n - \operatorname{tr}[(\widehat{\Sigma}\widehat{\Sigma}^{\dagger})^2]/n + \operatorname{tr}[\Sigma\widehat{\Sigma}^{\dagger}\widehat{\Sigma}\widehat{\Sigma}^{\dagger}]/n \\
&= \operatorname{tr}[\widehat{\Sigma}\widehat{\Sigma}^{\dagger}]/n + \operatorname{tr}[\Sigma\widehat{\Sigma}^{\dagger}\widehat{\Sigma}\widehat{\Sigma}^{\dagger}]/n,
\end{aligned} \tag{88}$$

where we used the fact that $\widehat{\Sigma}^{\dagger}\widehat{\Sigma}\widehat{\Sigma}^{\dagger} = \widehat{\Sigma}^{\dagger}$ (a property of the Moore-Penrose pseudoinverse). We now use (84) and (86) to obtain the asymptotic equivalent for (88). We will do the underparameterized and overparameterized cases separately below.

Underparameterized regime. We have

$$\mathsf{opt}_{\mathrm{R}}^{\mathrm{i}}(\widehat{f}_0^{\mathrm{ridge}} \mid X)/\sigma^2 \simeq 2\gamma_n - \gamma_n + \frac{\gamma_n\,\overline{\operatorname{tr}}[\Sigma\Sigma^{-1}]}{1 - \gamma_n} = \gamma_n + \frac{\gamma_n}{1 - \gamma_n}.$$

Applying the mapping ω, followed by the dominated convergence theorem, yields the result.

Overparameterized regime. We have

$$
\begin{aligned}
\mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}_0^{\mathrm{ridge}} \mid X)/\sigma^2 &\simeq \gamma_n(1 - \overline{\mathrm{tr}}[(v_n\Sigma + I)^{-1}]) + \gamma_n \frac{1}{v_n^{-2} - \gamma_n \,\overline{\mathrm{tr}}[\Sigma^2(v_n\Sigma + I_p)^{-2}]}\,\overline{\mathrm{tr}}[\Sigma^2(v_n\Sigma + I)^{-2}] \\
&= \gamma_n - \gamma_n\mu_n\,\overline{\mathrm{tr}}[(\Sigma + \mu_n I)^{-1}] + \frac{\gamma_n\,\overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu_n I)^{-2}]}{1 - \gamma_n\,\overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu_n I)^{-2}]} \\
&= \gamma_n\,\overline{\mathrm{tr}}[\Sigma(\Sigma + \mu_n I)^{-1}] + \frac{\gamma_n\,\overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu_n I)^{-2}]}{1 - \gamma_n\,\overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu_n I)^{-2}]} \\
&= \frac{\gamma_n\,\overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu_n I)^{-2}]}{1 - \gamma_n\,\overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu_n I)^{-2}]} + 1 \\
&= \frac{V_n}{D_n} + 1,
\end{aligned}
$$

where we used the parameterization $\mu_n = v_n^{-1}$ to simplify the expressions on the second line and (44) on the last line. Applying the mapping ω gives the desired result.

Emergent random-X degrees of freedom. As with ridge, we will derive an asymptotic equivalent for the excess bias $B^+ = B^+(\widehat{f}_0^{\mathrm{ridge}})$, and then use the decomposition (32) to obtain the final equivalent. Let us write $B^+ = B^+_{\mathrm{LI}} + B^+_{\mathrm{NL}} + C^+$, as in (78) in the ridge proof. By similar arguments, we have $C^+ \simeq 0$. It thus suffices to obtain asymptotic equivalents for B^+_{LI} and B^+_{NL}.

Asymptotic equivalent for B^+_{LI}. For the linear component of excess bias, we have

$$
\begin{aligned}
B^+_{\mathrm{LI}} &= \mathbb{E}_{x_0}[(f(x_0) - L_X(x_0)^\top f(X))^2] - \|(I - L_X(X))f(X)\|_2^2/n \\
&= \mathbb{E}_{x_0}[(\beta^\top(I - \widehat{\Sigma}\widehat{\Sigma}^\dagger)x_0)^2] - \|(I - X\widehat{\Sigma}^\dagger X^\top/n)X\beta\|_2^2/n \\
&= \beta^\top(I - \widehat{\Sigma}\widehat{\Sigma}^\dagger)\Sigma(I - \widehat{\Sigma}\widehat{\Sigma}^\dagger)\beta - \|X(I - \widehat{\Sigma}\widehat{\Sigma}^\dagger)\beta\|_2^2/n \\
&= \beta^\top(I - \widehat{\Sigma}\widehat{\Sigma}^\dagger)\Sigma(I - \widehat{\Sigma}\widehat{\Sigma}^\dagger)\beta - \beta^\top(I - \widehat{\Sigma}\widehat{\Sigma}^\dagger)\widehat{\Sigma}(I - \widehat{\Sigma}\widehat{\Sigma}^\dagger)\beta \\
&= \beta^\top(I - \widehat{\Sigma}\widehat{\Sigma}^\dagger)\Sigma(I - \widehat{\Sigma}\widehat{\Sigma}^\dagger)\beta, \qquad (89)
\end{aligned}
$$

where we use the fact that $\widehat{\Sigma}\widehat{\Sigma}^\dagger\widehat{\Sigma} = I$. Now using (87), we can obtain the asymptotic equivalent for (89) as follows:

$$
B^+_{\mathrm{LI}} \simeq \begin{cases} 0 & \gamma_n \leqslant 1 \\ (1 + \widetilde{v}_b)(v_n\Sigma + I)^{-1}\Sigma(v_n\Sigma + I)^{-1} = \dfrac{\mu_n^2(\Sigma + \mu_n I)^{-1}\Sigma(\Sigma + \mu_n I)^{-1}}{1 - \gamma_n\,\overline{\mathrm{tr}}[\Sigma^2(\Sigma + \mu_n I)^{-2}]} = \dfrac{B_n}{D_n} & \gamma_n > 1, \end{cases}
$$

where in the last line we use the parameterization $\mu_n = v_n^{-1}$.

Asymptotic equivalent for B^+_{NL}. For the nonlinear component of excess bias, we have

$$
\begin{aligned}
B^+_{\mathrm{NL}} &= \mathbb{E}_{x_0}[(f_{\mathrm{NL}}(x_0) - L_X(x_0)^\top f_{\mathrm{NL}}(X))^2] - \|(I - L_X(X))f_{\mathrm{NL}}(X)\|_2^2/n \\
&= \mathbb{E}_{x_0}[f_{\mathrm{NL}}(X)^\top L_X(x_0)L_X(x_0)^\top f_{\mathrm{NL}}(X)] + \sigma^2_{\mathrm{NL}} - \|(I - L_X(X))f_{\mathrm{NL}}(X)\|_2^2/n \\
&= f_{\mathrm{NL}}(X)^\top(X\widehat{\Sigma}^\dagger\Sigma\widehat{\Sigma}^\dagger X^\top/n)f_{\mathrm{NL}}(X)/n + \sigma^2_{\mathrm{NL}} - f_{\mathrm{NL}}(X)^\top(I - X\widehat{\Sigma}^\dagger X^\top/n)^2 f_{\mathrm{NL}}/n \\
&= f_{\mathrm{NL}}(X)^\top(\widehat{\Sigma}^\dagger X^\top/n)^\top\Sigma(\widehat{\Sigma}^\dagger X^\top/n)f_{\mathrm{NL}}(X) + \sigma^2_{\mathrm{NL}} \qquad (90) \\
&\quad - f_{\mathrm{NL}}(X)^\top(X\widehat{\Sigma}^\dagger X^\top/n - I)^\top(X\widehat{\Sigma}^\dagger X^\top/n - I)f_{\mathrm{NL}}(X)/n. \qquad (91)
\end{aligned}
$$

As with ridge regression, in the second equality above, we used the fact that $\mathbb{E}_{x_0}[f_{\rm NL}(x_0)x_0] = 0$ and $\mathbb{E}_{x_0}[f_{\rm NL}(x_0)^2] = \sigma_{\rm NL}^2$. As shown in the proof of Theorem 1 of Patil and Du (2023), the two quadratic forms in (90) and (91) concentrate around the traces. Thus, for (90), we have

$$\begin{aligned} f_{\rm NL}(X)^\top (\widehat{\Sigma}^\dagger X^\top/n)^\top \Sigma (\widehat{\Sigma}^\dagger X^\top/n) f_{\rm NL}(X) + \sigma_{\rm NL}^2 &\simeq \mathrm{tr}[(\widehat{\Sigma}^\dagger X^\top/n)^\top \Sigma (\widehat{\Sigma}^\dagger X^\top/n)]\sigma_{\rm NL}^2 + \sigma_{\rm NL}^2 \\ &= (\mathrm{tr}[\widehat{\Sigma}^\dagger \Sigma \widehat{\Sigma}^\dagger \widehat{\Sigma}]/n + 1)\sigma_{\rm NL}^2 \\ &= (\mathrm{tr}[\widehat{\Sigma}^\dagger \Sigma]/n + 1)\sigma_{\rm NL}^2, \end{aligned}$$

where we used the fact that $\widehat{\Sigma}^\dagger \widehat{\Sigma} \widehat{\Sigma}^\dagger = \widehat{\Sigma}^\dagger$ in the third line. Similarly, for (91), we have

$$\begin{aligned} f_{\rm NL}(X)^\top (X\widehat{\Sigma}^\dagger X^\top/n - I)^\top (X\widehat{\Sigma}^\dagger X^\top/n - I) f_{\rm NL}(X)/n &\simeq \mathrm{tr}[(X\widehat{\Sigma}^\dagger X^\top/n - I)^2]/n \cdot \sigma_{\rm NL}^2 \\ &= \mathrm{tr}[X\widehat{\Sigma}^\dagger X^\top/n - I]/n \cdot \sigma_{\rm NL}^2 \\ &= (\mathrm{tr}[\widehat{\Sigma}^\dagger \widehat{\Sigma}]/n - 1)\sigma_{\rm NL}^2, \end{aligned}$$

where we used the fact that $X\widehat{\Sigma}^\dagger X^\top/n - I$ is an idempotent matrix in the second line. Combining the two asymptotic equivalents, we thus have

$$B_{\rm NL}^+/\sigma^2 \simeq \mathrm{tr}[\widehat{\Sigma}^\dagger \Sigma]/n \cdot \frac{\sigma_{\rm NL}^2}{\sigma^2} + \mathrm{tr}[\widehat{\Sigma}^\dagger \widehat{\Sigma}]/n \cdot \frac{\sigma_{\rm NL}^2}{\sigma^2}.$$

Similar to the intrinsic analysis, we obtain the following asymptotic equivalent for $B_{\rm NL}^+$:

$$B_{\rm NL}^+/\sigma^2 \simeq \begin{cases} \left(\gamma_n + \dfrac{\gamma_n}{1-\gamma_n}\right)\dfrac{\sigma_{\rm NL}^2}{\sigma^2} & \gamma_n \leqslant 1 \\ \left(\dfrac{V_n}{D_n} + 1\right)\dfrac{\sigma_{\rm NL}^2}{\sigma^2} & \gamma_n > 1. \end{cases}$$

Combining this with the results for the intrinsic random-X optimism, and passing through ω and subsequent application of the dominated convergence theorem, completes the proof.

B.4 Proof of Proposition 8

Because ω is strictly increasing, in order to analyze the monotonicity of the asymptotic equivalents for normalized degrees of freedom in γ_n, it suffices to analyze the monotonicity of the equivalents for random-X optimism in γ_n, respectively. We do this for the intrinsic and emergent cases below.

Intrinsic random-X optimism. There are two regimes to examine.

Underparameterized regime. When $\gamma_n < 1$, from the proof of Theorem 7, we have that

$$\mathsf{opt}_{\rm R}^{\rm i}(\widehat{f}_0^{\rm ridge} \,|\, X)/\sigma^2 \simeq \gamma_n + \frac{\gamma_n}{1-\gamma_n},$$

which is a strictly increasing function in $\gamma_n \in (0,1)$, with the following boundary limits:

$$\lim_{\gamma_n \to 0^+} \mathsf{opt}_{\rm R}^{\rm i}(\widehat{f}_0^{\rm ridge} \,|\, X)/\sigma^2 = 0, \quad \text{and} \quad \lim_{\gamma_n \to 1^-} \mathsf{opt}_{\rm R}^{\rm i}(\widehat{f}_0^{\rm ridge} \,|\, X)/\sigma^2 = \infty,$$

Consequently, $\mathsf{df}_{\rm R}^{\rm i}(\widehat{f}_0^{\rm ridge})/n$ is increasing from 0 to 1 in $\gamma_n \in (0,1)$.

Overparameterized regime. When $\gamma_n > 1$, by Lemma F.11 in Du et al. (2023), the solution $v(0;\gamma_n)$ to the fixed point equation (85) is finite. Then, it follows from the proof of Theorem 7 that

$$\mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_0 \,|\, X)/\sigma^2 \simeq \widetilde{v}_v(0;\gamma_n).$$

Next, we study the monotonicity of $\widetilde{v}$. Taking the derivative with respect to γ_n yields

$$\begin{aligned}
&\frac{\partial \widetilde{v}_v(0;\gamma_n)}{\partial \gamma_n} \\
&= \frac{\overline{\mathrm{tr}}[\Sigma^2(v(0;\gamma_n)\Sigma + I)^{-2}]}{\big(v(0;\gamma_n)^{-2} - \gamma_n \,\overline{\mathrm{tr}}[\Sigma^2(v(0;\gamma_n)\Sigma + I)^{-2}]\big)^3} \\
&\qquad \cdot \Big[\big(v(0;\gamma_n)^{-2} - \gamma_n \,\overline{\mathrm{tr}}[\Sigma^2(v(0;\gamma_n)\Sigma + I)^{-2}]\big)^3 - 2\gamma_n v(0;\gamma_n)^{-3}\,\overline{\mathrm{tr}}[\Sigma(v(0;\gamma_n)\Sigma + I)^{-1}]\Big] \\
&= \frac{\overline{\mathrm{tr}}[\Sigma^2(v(0;\gamma_n)\Sigma + I)^{-2}]}{\big(v(0;\gamma_n)^{-2} - \gamma_n \,\overline{\mathrm{tr}}[\Sigma^2(v(0;\gamma_n)\Sigma + I)^{-2}]\big)^3} \\
&\qquad \cdot \Big[\big(v(0;\gamma_n)^{-2} - \gamma_n \,\overline{\mathrm{tr}}[\Sigma^2(v(0;\gamma_n)\Sigma + I)^{-2}]\big)^3 - 2v(0;\gamma_n)^{-4}\Big] \\
&= \frac{\overline{\mathrm{tr}}[\Sigma^2(v(0;\gamma_n)\Sigma + I)^{-2}]}{\big(v(0;\gamma_n)^{-2} - \gamma_n \,\overline{\mathrm{tr}}[\Sigma^2(v(0;\gamma_n)\Sigma + I)^{-2}]\big)^3} \\
&\qquad \cdot \Big[- v(0;\gamma_n)^{-4} - v(0;\gamma_n)^{-2} - \big(v(0;\gamma_n)^{-2} - \gamma_n \,\overline{\mathrm{tr}}[\Sigma^2(v(0;\gamma_n)\Sigma + I)^{-2}]\big) \\
&\qquad\qquad \cdot \gamma_n \,\overline{\mathrm{tr}}[\Sigma^2(v(0;\gamma_n)\Sigma + I)^{-2}]\Big] \\
&\leqslant 0.
\end{aligned}$$

Here, we use the fact from Lemma F.11 (3) in Du et al. (2023) that

$$\frac{1}{v(0;\gamma)^2} - \gamma \,\overline{\mathrm{tr}}[\Sigma^2(v(0;\gamma)\Sigma + I)^{-2}] \geqslant 0,$$

with equality obtained only when $\gamma = \infty$. This indicates that $\mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_0 \,|\, X)$ is strictly increasing in γ_n for $\gamma_n \in (1,\infty)$, with

$$\lim_{\gamma_n \to 1^+} \mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_0 \,|\, X)/\sigma^2 = \infty, \quad \text{and} \quad \lim_{\gamma_n \to \infty} \mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_0 \,|\, X)/\sigma^2 = 0.$$

Consequently, $\mathsf{df}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_0)/n$ is decreasing from 1 to 0 in $\gamma_n \in (1,\infty)$.

Emergent random-X optimism. From the proof of Theorem 7, when $\gamma_n < 1$, we have

$$\mathsf{opt}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_0 \,|\, X)/\sigma^2 \simeq (\gamma_n + \gamma_n/(1-\gamma_n))(1 + \sigma^2_{\mathrm{NL}}/\sigma^2),$$

which is strictly increasing in $\gamma_n \in (0,1)$ with the following boundary limit:

$$\lim_{\gamma_n \to 1^-} \mathsf{opt}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_0 \,|\, X)/\sigma^2 = \infty.$$

Consequently, $\mathsf{df}_{\mathrm{R}}(\widehat{f}^{\mathrm{ridge}}_0)/n$ is increasing from 0 to 1 on $\gamma_n \in (0,1)$ and maximized at $\gamma_n = 1$. This finishes the proof.

B.5 Proof of Proposition 10

We will first parameterize the nonlinear system in (49) and (50) slightly differently by introducing a new variable $a = \mu/\tau$. Namely, we let (τ, a) solve:

$$\tau^2 = \sigma^2 + \gamma \mathbb{E}[(\mathsf{soft}(\tau H + B; a\tau) - B)^2], \tag{92}$$

$$\lambda = a\tau(1 - \gamma \mathbb{E}[\mathsf{soft}'(\tau H + B; a\tau)]). \tag{93}$$

The nonlinear system in (92) and (93) is similar to the one in Bayati and Montanari (2011). When $B = 0$ (almost surely), we denote its solution by (τ_0, a_0). Before we start the proof, we will collect the following two properties of soft-thresholding (the proximal operator for the ℓ_1 norm) for $a > 0$:

$$\mathsf{soft}(x; \kappa) = \tfrac{1}{a}\mathsf{soft}(ax; a\kappa), \tag{94}$$

$$\mathsf{soft}'(x; \kappa) = \mathsf{soft}'(ax; a\kappa). \tag{95}$$

These are straightforward to check (see, e.g., Lemma B.2 in Wang et al. (2020)). We will split the proof below into two parts, following the two statements in the proposition. As before, since ω is strictly increasing, it suffices to show the desired properties on the optimism scale.

Monotonicity of intrinsic random-X optimism. Combining (51) and (52), we can write

$$\mathsf{opt}^{\mathsf{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{lasso}}_\lambda \mid X, y) \simeq (1 - (1 - \mathsf{df}^{\mathsf{i}}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_\lambda)/n))^2 \tau_0^2.$$

Below, we will argue that each of τ_0^2 and $\mathsf{df}^{\mathsf{i}}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_\lambda)/n$ are monotonic in λ, with limits $\tau_0^2 \to \sigma^2$ and $\mathsf{df}^{\mathsf{i}}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_\lambda)/n \to 0$ as $\lambda \to \infty$.

Monotonicity of τ_0^2. We first argue below that τ_0^2 is monotonically nonincreasing in λ. We have

$$\tau_0^2 = \sigma^2 + \gamma \mathbb{E}[(\mathsf{soft}(\tau H; a_0\tau_0))^2] = \sigma^2 + \gamma\tau_0^2 \mathbb{E}[(\mathsf{soft}(H; a_0))^2],$$

where we used (94) in the second equality above. Rearranging, we get that

$$\tau_0^2 = \frac{\sigma^2}{1 - \gamma \mathbb{E}[(\mathsf{soft}(H; a_0))^2]}.$$

Now, observe that the right-hand side is monotonically nonincreasing in a_0, which follows because $x \mapsto |\mathsf{soft}(u; x)|$ is noncreasing in x for fixed u, and a_0 is nondecreasing in λ from Corollary 1.7 of Bayati and Montanari (2011). This implies that τ_0^2 is monotonically nonincreasing in λ. Lastly, by Corollary 1.7 of Bayati and Montanari (2011) once again, we have $a_0 \to \infty$ as $\lambda \to \infty$, and hence $\mathbb{E}[\mathsf{soft}(H; a_0)^2] \to 0$, and $\tau_0^2 \to \sigma^2$ as $\lambda \to \infty$.

Monotonicity of $\mathsf{df}^{\mathsf{i}}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_\lambda)/n$. To show that $\mathsf{df}^{\mathsf{i}}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_\lambda)/n$ is decreasing in λ, observe that

$$\frac{\mathsf{df}^{\mathsf{i}}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_\lambda)}{n} \simeq \gamma \mathbb{E}[\mathsf{soft}'(\tau_0 H; a_0\tau_0)].$$

To see this, note from (50), after replacing μ_0 with $a_0\tau_0$, that

$$1 - \lambda/\mu_0 = \gamma \mathbb{E}[\mathsf{soft}'(\tau_0 H; a_0\tau_0)]$$

Using (95), we have $\gamma \mathbb{E}[\mathsf{soft}'(\tau_0 H; a_0\tau_0)] = \gamma \mathbb{E}[\mathsf{soft}'(H; a_0)]$. Also, $\mathbb{E}[\mathsf{soft}'(H; a_0)] = \mathbb{P}(|H| > a_0)$, which is nonincreasing in a_0. Using the monotonically nondecreasing behavior of a_0 in λ from Corollary 1.7 of Bayati and Montanari (2011), we then have the desired monotonicity. Lastly, that $\mathsf{df}^{\mathsf{i}}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_\lambda)/n \to 0$ as $\lambda \to \infty$ follows from $\lambda \to \mu_0 \to 1$, which can be checked from (50).

Nonnegativity of emergent minus intrinsic optimism. From (51), we can write:

$$\mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{lasso}}_{\lambda} \,|\, X, y) \simeq \tau_0^2(1 - (1 - \mathsf{df}^{\mathrm{i}}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_{\lambda})/n)))^2.$$

Similarly, we can write the emergent optimism as:

$$\mathsf{opt}_{\mathrm{R}}(\widehat{f}^{\mathrm{lasso}}_{\lambda} \,|\, X, y) \simeq \tau^2(1 - (1 - \mathsf{df}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_{\lambda})/n))^2.$$

To show the asymptotic equivalent for $\mathsf{opt}_{\mathrm{R}}(\widehat{f}^{\mathrm{lasso}}_{\lambda} \,|\, X, y)$ is no less than that for $\mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\mathrm{lasso}}_{\lambda} \,|\, X, y)$, we will argue that $\tau^2 \geqslant \tau_0^2$ and $\mathsf{df}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_{\lambda}) \geqslant \mathsf{df}^{\mathrm{i}}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_{\lambda})$, below.

Nonnegativity of $\mathsf{df}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_{\lambda}) - \mathsf{df}^{\mathrm{i}}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_{\lambda})$. The two quantities we need to compare are:

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_{\lambda})/n \simeq \gamma\mathbb{E}[\mathsf{soft}'(\tau H + B; a\tau)] \quad \text{and} \quad \mathsf{df}^{\mathrm{i}}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_{\lambda})/n \simeq \gamma\mathbb{E}[\mathsf{soft}'(\tau_0 H; a_0\tau_0)].$$

Using (95), we first rewrite the asymptotic equivalents in the display above as:

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_{\lambda})/n \simeq \gamma\mathbb{E}[\mathsf{soft}'(H + B/\tau; a)] \quad \text{and} \quad \mathsf{df}^{\mathrm{i}}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_{\lambda})/n \simeq \gamma\mathbb{E}[\mathsf{soft}'(H; a_0)].$$

Observe now that for $\tau \geqslant 0$, assuming $a \leqslant a_0$, we have

$$\mathbb{E}[\mathsf{soft}'(H + B/\tau; a)] \geqslant \mathbb{E}[\mathsf{soft}'(H; a)] \geqslant \mathbb{E}[\mathsf{soft}'(H; a_0)],$$

The first inequality can be explained as follows: $\mathbb{E}[\mathsf{soft}'(H + b; a) \geqslant \mathbb{P}(|H| > a) = \mathbb{E}[\mathsf{soft}'(H; a)]$ for any fixed b and hence $\mathbb{E}[\mathsf{soft}'(H + B/\tau; a)] \geqslant \mathbb{E}[\mathsf{soft}'(H; a)]$ by conditioning on the random variable B which is independent of H. Thus we get the desired claim that $\mathsf{df}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_{\lambda}) \geqslant \mathsf{df}^{\mathrm{i}}_{\mathrm{F}}(\widehat{f}^{\mathrm{lasso}}_{\lambda})$ assuming $a \leqslant a_0$, which we will show in the next part, along with $\tau \geqslant \tau_0$.

Nonnegativity of $\tau^2 - \tau_0^2$. We consider solving the system for emergent parameters (92), (93). We will solve these using the fixed point iteration algorithm initialized at the solution τ_0, a_0 of the system with intrinsic parameters.

Namely, we will start with $a^{(0)} = a_0$ and $\tau^{(0)} = \tau_0$. If a_0 and τ_0 solve the emergent system, then we are done. Suppose they do not. Then, we first solve (92) with fixing $a = a^{(0)}$ and solving for τ. Call this solution $\tau^{(1)}$. We claim that $\tau^{(1)} \geqslant \tau^{(0)} = \tau_0$. Suppose in order to achieve a contradiction that $\tau^{(1)} < \tau^{(0)}$. Rewrite (92) after normalizing with respect to τ^2:

$$1 = \frac{\sigma^2}{\tau^2} + \gamma\mathbb{E}\left[\left(\mathsf{soft}\left(H + \frac{B}{\tau}; a\right) - \frac{B}{\tau}\right)^2\right].$$

From Lemma 12 in Weng et al. (2018), we know that the function that multiplies γ in the display above is a decreasing function of τ. Note that the function $h : x \mapsto \mathbb{E}[(\mathsf{soft}(H + xB; a) - xB)^2]$ is an even function, as $h(x) = \mathbb{E}[(\mathsf{soft}(-H + xB; a) - xB)^2] = \mathbb{E}[(\mathsf{soft}(H + (-x)B; a) - (-x)B)^2] = h(-x)$. From Lemma 6 in Weng et al. (2018), we have that $h(x)$ is increasing in x. Thus, the same function in the above display has a larger value when $B \neq 0$. Thus, if $\tau^{(1)} < \tau^{(0)}$, then both of the terms on the right-hand side of the display above increase. But we already know that $a = a^{(0)}$ satisfies the equation with $\tau^{(0)}$. This supplies the desired contradiction.

Now, fix this $\tau^{(1)}$, and solve (93) for a. Call this solution $a^{(1)}$. As before, we claim $a^{(1)} \leqslant a^{(0)} = a_0$. This follows again from a contradiction-based argument because if $a^{(1)} > a^{(0)}$, then both the terms on the right-hand side of (93) go up because the term multiplying γ is decreasing in a (since we can eliminate τ) and has a larger value when $B \neq 0$.

Iterating the above argument, we obtain two monotonic nonnegative sequences $a^{(m)}$, $\tau^{(m)}$. When $\tau^{(m)} = \infty$ one has $a^{(m)} = 0$, and when $a^{(m)} = 0$, one has $\tau^{(m+1)} = \infty$. Thus, we have $a \geqslant 0$ and $\tau \leqslant \infty$, which indicates that the process terminates as $m \to \infty$, and $\tau \geqslant \tau_0$, $a \leqslant a_0$.

B.6 Proof of Theorem 11

In the underparameterized regime (when $\gamma \leqslant 1$), the statements follow from Theorem 7 since both predictors are simply least squares in this regime. In the overparameterized regime (when $\gamma > 1$), the results follow by sending $\lambda \to 0^+$ in the results of Theorem 9. The validity of this limit, along with the existence and uniqueness of the solution to the nonlinear system (54) and (55) is shown by Li and Wei (2021).

B.7 Proof of Proposition 12

For $\gamma \in (1, \infty)$, the parameters (τ_0, a_0) solve the system:

$$\begin{aligned} \tau_0^2 &= \sigma^2 + \gamma \mathbb{E}\Big[\big(\mathsf{soft}\big(\tau_0 H; a_0\tau_0\big)\big)^2\Big] \\ 1 &= \gamma \mathbb{E}\Big[\mathsf{soft}'\big(\tau_0 H; a_0\tau_0\big)\Big]. \end{aligned}$$

Here recall that $\mathsf{soft}'(x; y)$ is the derivative of $\mathsf{soft}(x; y)$ in x. This can be simplified to:

$$\begin{aligned} 1 &= \sigma^2/\tau_0^2 + \gamma \mathbb{E}\Big[\big(\mathsf{soft}\big(H; a_0\big)\big)^2\Big] \\ 1 &= \gamma \mathbb{E}\Big[\mathsf{soft}'\big(H; a_0\big)\Big]. \end{aligned}$$

This leads to:

$$\tau_0^2 = \frac{\sigma^2}{1 - \gamma \mathbb{E}[(\mathsf{soft}(H; a_0))^2]},$$

where a_0 solves:

$$1 = \gamma \mathbb{E}[\mathsf{soft}'(H; a_0)].$$

We first conclude that a_0 is monotonically increasing in $\gamma \in (1, \infty)$ and ranges from 0 to ∞. This follows because the function $x \mapsto |\mathsf{soft}(u; x)|$ is decreasing in x, for fixed u. In particular,

$$\mathbb{E}[\mathsf{soft}'(H; a_0)] = \mathbb{P}(|H| > a_0) = 2(1 - \Phi(a_0)) = \frac{1}{\gamma}.$$

This leads to

$$a_0 = \Phi^{-1}\Big(\frac{2\gamma - 1}{2\gamma}\Big).$$

Since both the functions Φ^{-1} and $\frac{2\gamma-1}{2\gamma}$ are monotonically increasing γ, we have that the composition is monotonically increasing in γ. When $\gamma = 1$, we have $a_0 = 0$ and when $\gamma = \infty$, we have $a_0 = \infty$.

Next we will argue that $\gamma \mapsto \gamma \mathbb{E}[(\mathsf{soft}(H; a))^2]$ decreases in $\gamma \in (1, \infty)$ and ranges from 1 to 0. We do by first substituting for γ as $\frac{1}{\mathbb{E}[\mathsf{soft}'(H;a)]}$. The goal then reduces to arguing that the function

$$\gamma \mapsto \frac{\mathbb{E}[(\mathsf{soft}(H; a))^2]}{\mathbb{E}[\mathsf{soft}'(H; a)]}$$

is decreasing in γ. Since a is increasing in γ, it suffices to argue that the function

$$y \mapsto \frac{\mathbb{E}[(\mathsf{soft}(H; y))^2]}{\mathbb{E}[\mathsf{soft}'(H; y)]}$$

is decreasing in y. This follows from Lemma 16 below, and finishes the proof.

Lemma 16. *For $H \sim \mathcal{N}(0,1)$, the function*

$$y \mapsto \frac{\mathbb{E}[\mathsf{soft}(H;y)^2]}{\mathbb{E}[\mathsf{soft}'(H;y)]}$$

is monotonically decreasing in y. Here, recall, the derivative of soft *is understood to be with respect to its first argument.*

Proof. Denote the numerator and the denominator by

$$\begin{aligned} f(y) &= \mathbb{E}[\mathsf{soft}(H;y)^2] = 2\mathbb{E}[(H-y)^2\,\mathbb{1}\{H>y\}] \\ g(y) &= \mathbb{E}[\mathsf{soft}'(H;y)] = \mathbb{E}[H\mathsf{soft}(H;y)] = f(y) + 2y\mathbb{E}[(H-y)\,\mathbb{1}\{H>y\}]. \end{aligned}$$

Here, in the second equality of the second row, we use Stein's lemma.

Recall for $X \sim \mathcal{N}(0,1)$, the truncated normal distribution admits

$$\begin{aligned} \mathbb{E}[X \mid X > a] &= \varphi(a)/(1-\Phi(a)) \\ \mathrm{Var}(X \mid X > a) &= 1 + a\varphi(a)/(1-\Phi(a)) - (\varphi(a)/(1-\Phi(a)))^2 \\ \mathbb{E}[X^2 \mid X > a] &= \mathrm{Var}(X \mid X > a) + \mathbb{E}[X \mid X > a]^2 \\ &= 1 + a\varphi(a)/(1-\Phi(a)) - (\varphi(a)/(1-\Phi(a)))^2 + (\varphi(a)/(1-\Phi(a)))^2 \\ &= 1 + a\varphi(a)/(1-\Phi(a)). \end{aligned}$$

Then we have

$$\begin{aligned} \mathbb{E}[(H-y)\,\mathbb{1}\{H>y\}] &= \varphi(y) - y(1-\Phi(y)) \\ f(y) &= 2(\mathbb{E}[H^2 \mid H > y]\mathbb{P}(H > y) - 2y\mathbb{E}[H \mid H > y]\mathbb{P}(H > y) + y^2(1-\Phi(y)) \\ &= 2[(1-\Phi(y)) + y\varphi(y) - 2y\varphi(y) + y^2(1-\Phi(y)] \\ &= 2[-y\varphi(y) + (1+y^2)(1-\Phi(y))] \\ g(y) &= 2\mathbb{E}[(H-y)^2\,\mathbb{1}\{H>y\}] + 2y\mathbb{E}[(H-y)\,\mathbb{1}\{H>y\}] \\ &= f(y) + 2y\mathbb{E}[(H-y)\,\mathbb{1}\{H>y\}] \\ &= f(y) + \underbrace{2y(\varphi(y) - y(1-\Phi(y)))}_{h(y)}. \end{aligned}$$

Because $\Phi(y) = \varphi(y)$ and $\varphi'(y) = -y\varphi(y)$, we further have

$$\begin{aligned} f'(y)g(y) - f(y)g'(y) &= f'(y)[f(y) + h(y)] - f(y)[f'(y) + h'(y)] \\ &= f'(y)h(y) - f(y)h'(y) \\ &= 2[-\varphi(y) + y^2\varphi(y) + 2y(1-\Phi(y)) - (1+y^2)\varphi(y)]h(y) \\ &\qquad - f(y)2[\varphi(y) - y^2\varphi(y) - 2y(1-\Phi(y)) + y^2\varphi(y)] \\ &= 4[y(1-\Phi(y)) - \varphi(y)]h(y) + 4f(y)[y(1-\Phi(y)) - \varphi(y)] \\ &= 4[\underbrace{y(1-\Phi(y)) - \varphi(y)}_{c(y)}][h(y) + f(y)]. \end{aligned}$$

Now $c'(y) = (1 - \Phi(y) - y\varphi(y) + y\varphi(y) = 1 - \Phi(y) \geqslant 0$ and $\lim_{y\to\infty} c(y) = 0$, thus we have $c(y) \leqslant 0$ and hence

$$\frac{\partial}{\partial y}\frac{f(y)}{g(y)} = \frac{f'(y)g(y) - f(y)g'(y)}{g(y)^2} \leqslant 0,$$

which finishes the proof. □

B.8 Proof of Theorem 13

We will use results from Thrampoulidis et al. (2018), which use a slightly different scaling for the feature matrix. In particular, they use a variance scaling of $1/p$ for the entries of the feature vector x_i, whereas recall (from Assumption B), we consider a variance scaling of $1/n$. We can thus rewrite the estimator of interest from (59) (after dividing by γ) as:

$$\widehat{\beta}_{\lambda}^{\text{convex}} \in \operatorname*{arg\,min}_{b \in \mathbb{R}^p} \frac{1}{2} \sum_{i=1}^{n} (\widetilde{y}_i - \widetilde{x}_i^\top b)^2 + \widetilde{\lambda} \sum_{i=1}^{p} \mathsf{reg}(b_i), \tag{96}$$

where the transformed variables are:

$$\widetilde{x}_i = \gamma^{-1/2} x_i, \quad \widetilde{y}_i = \gamma^{-1/2} y_i, \quad \widetilde{\lambda} = \lambda/\gamma. \tag{97}$$

This transformation follows since the minimizers do not change up to positive scaling (which in our case is by γ) of the objective function. Since x_i has i.i.d. entries with variance $1/n$ and $y_i = x_i^\top \beta + \varepsilon_i$ in Assumption B, in the transformed formulation (96), the feature vectors $\widetilde{x}_i$ have i.i.d. entries with variance $1/p$, and the response variables follow the linear model $\widetilde{y}_i = \widetilde{x}_i \beta + \widetilde{\varepsilon}_i$, with the transformed noise defined as $\widetilde{\varepsilon}_i = \gamma^{-1/2} \varepsilon_i$.

With the transformation in (96), we now apply the master theorem of Thrampoulidis et al. (2018). Define the following nonlinear system of equations in four scalar variables $(\alpha, \zeta, \kappa, \nu)$:

$$\alpha^2 = \mathbb{E}\Big[\Big(\tfrac{\gamma\lambda}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\big(\tfrac{\zeta}{\nu} H + B; \tfrac{\gamma\lambda}{\nu}\big) - \tfrac{\zeta}{\nu} H\Big)^2\Big] = \mathbb{E}\Big[\Big(\mathsf{prox}_{\mathsf{reg}}\big(\tfrac{\zeta}{\nu} H + B; \tfrac{\gamma\lambda}{\nu}\big) - B\Big)^2\Big] \tag{98}$$

$$\gamma\zeta^2 = \frac{\alpha^2 + \sigma^2/\gamma}{(1+\kappa)^2} \tag{99}$$

$$\kappa\zeta = \mathbb{E}\Big[\Big(\tfrac{\gamma\lambda}{\nu} \cdot \mathsf{env}'_{\mathsf{reg}}\big(\tfrac{\zeta}{\nu} H + B; \tfrac{\gamma\lambda}{\nu}\big) - \tfrac{\zeta}{\nu} H\Big) \cdot (-H)\Big] = \mathbb{E}\Big[\Big(\mathsf{prox}_{\mathsf{reg}}\big(\tfrac{\zeta}{\nu} H + B; \tfrac{\lambda}{\nu}\big) - B\Big) \cdot H\Big] \tag{100}$$

$$\gamma\nu = \frac{1}{1+\kappa} \tag{101}$$

where $H \sim \mathcal{N}(0,1)$, and$B \sim F$ independently of H. As usual, when $B = 0$ (almost surely), we denote the solution by $(\alpha_0, \zeta_0, \kappa_0, \nu_0)$.

The parameters from (98)–(101) encode information regarding the asymptotics of various stochastic quantities that we will need in our derivation. Before we do that, we will reformulate the system above to better align with the results for the ridge and lasso predictors.

Reformulation of (98)–(101). Consider the following change of variables:

$$a = \frac{\gamma\lambda}{\zeta} \quad \text{and} \quad \tau = \frac{\zeta}{\nu}. \tag{102}$$

We will first reformulate (98)–(101) using (τ, a), yielding the following equivalent system:

$$\tau^2 = \sigma^2 + \gamma\mathbb{E}[(\mathsf{prox}_{\mathsf{reg}}(\tau H + B; a\tau) - B)^2], \tag{103}$$

$$\lambda = a\tau(1 - \gamma\mathbb{E}[\mathsf{prox}'_{\mathsf{reg}}(\tau H + B; a\tau)]), \tag{104}$$

where $H \sim \mathcal{N}(0,1)$ and $\Theta \sim F$ independent of H. The validity of this reformulation is proved later on. Letting $\mu = a\tau$, the system in (103), (104) is exactly the same (after rearranging) as (60), (61).

We are finally ready to obtain the asymptotics of the various notions of degrees of freedom, which we present in separate parts in what follows.

Fixed-X degrees of freedom. We first note that for the estimator $\widehat{\beta}_\lambda^{\mathrm{convex}}$ as defined in (59), the map $y \mapsto X\widehat{\beta}_\lambda^{\mathrm{convex}}$ is 1-Lipschitz on $\mathbb{R}^n$ (see, e.g, Proposition 3 of Bellec and Tsybakov (2017)) and has symmetric positive semidefinite Jacobian. Thus it is weakly differentiable and Stein's formula can be applied, which shows that its fixed-X degrees of freedom are then given by:

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}_\lambda^{\mathrm{convex}}) = \mathbb{E}\big[\operatorname{tr}[(\partial/\partial y)X\widehat{\beta}_\lambda^{\mathrm{convex}}] \,|\, X\big].$$

Now, observe that

$$(\partial/\partial\widetilde{y})\widetilde{X}\widehat{\beta}_\lambda^{\mathrm{convex}} = (\partial/\partial\widetilde{y})\gamma^{-1/2}X\widehat{\beta}_\lambda^{\mathrm{convex}} = (\partial/\partial y)(\partial y/\partial\widetilde{y})\gamma^{-1/2}X\widehat{\beta}_\lambda^{\mathrm{convex}} = (\partial/\partial y)X\widehat{\beta}_\lambda^{\mathrm{convex}}.$$

Thus, fixed-X degrees of freedom is unchanged under the transformation of the data in (97):

$$\mathsf{df}_{\mathrm{F}}(\widehat{f}_\lambda^{\mathrm{convex}}) = \mathbb{E}\big[\operatorname{tr}[(\partial/\partial\widetilde{y})\widetilde{X}\widehat{\beta}_\lambda^{\mathrm{convex}}] \,|\, X\big]. \tag{105}$$

In what follows, we first obtain limit in probability of the trace functional $\operatorname{tr}[(\partial/\partial\widetilde{y})\widetilde{X}\widehat{\beta}_\lambda^{\mathrm{convex}}]/p$, and then convert this convergence to obtain the desired limit of $\mathsf{df}_{\mathrm{F}}(\widehat{f}_\lambda^{\mathrm{convex}})/n$ in (105).

Define the matrix $V_\lambda = I - (\partial/\partial\widetilde{y})\widetilde{X}\widehat{\beta}_\lambda^{\mathrm{convex}} \in \mathbb{R}^{n\times n}$. By Corollary 3.2 in Bellec (2023), as $n, p \to \infty$ with $p/n \to \gamma \in (0,\infty)$, we have

$$\operatorname{tr}[V_\lambda]/p \xrightarrow{\mathrm{p}} \nu. \tag{106}$$

We mention in the passing here that the trace convergence result (106) in the special case of lasso follows from Theorem 8 of Celentano et al. (2023) and in the more general case of convex regularized M-estimators follows from Appendix A.4 of Koriyama et al. (2024). Now rearranging (106), we get

$$\operatorname{tr}[(\partial/\partial\widetilde{y})\widetilde{X}\widehat{\beta}_\lambda^{\mathrm{convex}}]/n \xrightarrow{\mathrm{p}} 1 - \gamma\nu = 1 - \gamma\mathbb{E}\big[\mathsf{prox}'_{\mathsf{reg}}(B + \tau H; a\tau)\big] = 1 - \lambda/\mu, \tag{107}$$

where the second-to-last equality follows from (121), and the last equality follows from (104) (after the change of variables $\mu = a\tau$). Finally, noting that $\operatorname{tr}[V_\lambda]/n$ ranges between $[0,1]$ for almost every y (see, e.g., Proposition 2.2 of Bellec (2023)), invoking the dominated convergence theorem (to be clear, a variant that handles convergence in probability by passing to a subsequence; see, e.g., Exercise 2.3.7 of Durrett (2010)) to convert (107) to a statement about convergence in expectation, we have that $\mathsf{df}_{\mathrm{F}}(\widehat{f}_\lambda^{\mathrm{convex}})/n$ converges to the same limit. This finishes the proof of (62).

Emergent random-X degrees of freedom. Next we consider emergent random-X optimism. Under the scaled (by n) isotropic features and linear model in Assumption B, observe that

$$\mathsf{err}_{\mathrm{R}}(\widehat{f}_\lambda^{\mathrm{convex}} \,|\, X, y) = \sigma^2 + \|\widehat{\beta}_\lambda^{\mathrm{convex}} - \beta\|_2^2/n, \tag{108}$$

$$\mathsf{err}_{\mathrm{T}}(\widehat{f}_\lambda^{\mathrm{convex}} \,|\, X, y) = \|y - X\widehat{\beta}_\lambda^{\mathrm{convex}}\|_2^2/n = \gamma\|\widetilde{y} - \widetilde{X}\widehat{\beta}^{\mathrm{convex}}\|_2^2/n. \tag{109}$$

From Theorem 4.1 in Thrampoulidis et al. (2018), for the problem (96), we note that

$$\|\widehat{\beta}_\lambda^{\mathrm{convex}} - \beta\|_2^2/p \xrightarrow{\mathrm{p}} \alpha^2 \quad \text{and} \quad \|\widetilde{y} - \widetilde{X}\widehat{\beta}_\lambda^{\mathrm{convex}}\|_2^2/p \xrightarrow{\mathrm{p}} \zeta^2, \tag{110}$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0,\infty)$. Combining (108), (109) and (110), we have

$$\mathsf{opt}_{\mathrm{R}}(\widehat{f}_\lambda^{\mathrm{convex}} \,|\, X, y) \xrightarrow{\mathrm{p}} \gamma\alpha^2 + \sigma^2 - \gamma^2\zeta^2 = \tau^2 - \gamma^2\zeta^2, \tag{111}$$

where the last line follows from combining (98) and (103). Also, note from (99), (101), and (107), we have

$$\frac{\gamma^2\zeta^2}{\gamma\alpha^2 + \sigma^2} = \frac{1}{(1+\kappa)^2} = \gamma^2\nu^2 \simeq \lambda^2/\mu^2, \tag{112}$$

where the last equivalence follows from (107). Again, using (98) and (103), note that we can rewrite (112) as

$$\gamma^2\zeta^2 \simeq \lambda^2/\mu^2 \cdot \tau^2. \tag{113}$$

Substituting (113) into (111) and applying ω and dominated convergence finishes the proof of (64).

Intrinsic random-X degrees of freedom. The proof for the intrinsic case follows similarly. When the signal is absent, we have

$$\mathsf{opt}^{\mathrm{i}}_{\mathrm{R}}(\widehat{f}^{\,\mathrm{convex}}_{\lambda} \,|\, X, y) \to \sigma^2 + \gamma\alpha_0^2 - \gamma^2\zeta_0^2 = \gamma\tau_0^2 - \lambda^2/\mu_0^2 \cdot \tau_0^2, \tag{114}$$

where we replaced α with α_0, ζ with ζ_0, and μ with μ_0 in (111) and (113). Applying ω to (114) and invoking the dominated convergence theorem finishes the proof of (63).

Derivation of the reformulation (98)–(101). Using (102), along with (60) and (101), note that (99) becomes

$$\tau^2 = \sigma^2 + \gamma\mathbb{E}[(\mathsf{prox}_{\mathsf{reg}}(\tau H + B; a\tau) - B)^2].$$

This supplies us with (103). Now, define the Moreau envelope by

$$\mathsf{env}_{\mathsf{reg}}(x; t) = \min_{z\in\mathbb{R}} \frac{1}{2t}(x - z)^2 + \mathsf{reg}(z).$$

We recall a key relationship between the proximal operator and Moreau envelope.

$$\mathsf{env}'_{\mathsf{reg}}(x; \tau) = \frac{1}{\tau}(x - \mathsf{prox}_{\mathsf{reg}}(x; \tau)). \tag{115}$$

Towards obtaining (104), from Stein's lemma, observe that

$$\mathbb{E}[\mathsf{env}'_{\mathsf{reg}}(B + \tau H; \kappa) \cdot H] = \tau\mathbb{E}[\mathsf{env}''_{\mathsf{reg}}(B + \tau H; \kappa)]. \tag{116}$$

Taking the derivative of the relation (115), we also have

$$\kappa\mathsf{env}''_{\mathsf{reg}}(B + \tau H; \kappa) = 1 - \mathsf{prox}'_{\mathsf{reg}}(B + \tau H; \kappa). \tag{117}$$

Combining (116) and (117), we obtain

$$\mathbb{E}\big[\mathsf{env}'_q(\tfrac{\zeta}{\nu}H + B; \tfrac{\lambda}{\nu}) \cdot H\big] = \tfrac{\zeta}{\nu}\mathbb{E}\big[\mathsf{env}''_q(\tfrac{\zeta}{\nu}H + B; \tfrac{\lambda}{\nu})\big] = \tfrac{\zeta}{\nu}\tfrac{\nu}{\lambda}\mathbb{E}\big[1 - \mathsf{prox}'_{\mathsf{reg}}(\tfrac{\zeta}{\nu}H + B; \tfrac{\lambda}{\nu})\big]. \tag{118}$$

Using (118), we can rewrite (61) as:

$$\kappa\zeta = \tfrac{\zeta}{\nu} - \tfrac{\lambda}{\nu}\tfrac{\zeta}{\lambda}\mathbb{E}\big[1 - \mathsf{prox}'_{\mathsf{reg}}(\tfrac{\zeta}{\nu}H + B; \tfrac{\lambda}{\nu})\big] = \tfrac{\zeta}{\nu}\big(1 - \mathbb{E}\big[1 - \mathsf{prox}'_{\mathsf{reg}}(\tfrac{\zeta}{\nu}H + B; \tfrac{\lambda}{\nu})\big]\big). \tag{119}$$

Now, using (102), we can express (119) as:

$$\kappa\nu = 1 - \mathbb{E}\big[1 - \mathsf{prox}'_{\mathsf{reg}}(B + \tau H; a\tau)\big]. \tag{120}$$

Rearranging and using (101) yields

$$\gamma\nu = 1 - \gamma\mathbb{E}\big[\mathsf{prox}'_{\mathsf{reg}}(B + \tau H; a\tau)\big]. \tag{121}$$

Multiplying both sides of (121) by $a\tau$ and using (102), we then arrive at:

$$\lambda = a\tau(1 - \gamma\mathbb{E}[\mathsf{prox}'_{\mathsf{reg}}(B + \tau H; a\tau)]). \tag{122}$$

This supplies us with (104), completing the reformulation.

C Numerical experiments for Section 5

C.1 Data models

For the simulations in Appendices C.3 and C.4, as well as that behind Figure 1, we generate data according to a nonlinear model

$$y_i = x_i^\top \beta + (\|x_i\|_2^2/d - 1) + \varepsilon_i, \quad i \in [n],$$

where each $x_i \sim \mathcal{N}(0, \Sigma_{\text{AR1},\rho=0.25})$, $\varepsilon_i \sim \mathcal{N}(0, 0.4^2)$, and β is drawn uniformly from the unit sphere in $\mathbb{R}^p$. Here, we use $\Sigma_{\text{AR1},\rho}$ to denote a covariance matrix with $\rho^{|i-j|}$. The "linearized" SNR in this setup is $\text{Var}[x_i^\top \beta]/\sigma^2 = 6.25$.

For the simulation behind Figure 1 only (i.e., not in Appendices C.3 and C.4), we sample $P = 300$ features total according to the above model, sort them in order of deceasing magnitude of $|\beta_j|$ (the linear part of the signal), and use the first p for least squares (if $p \leqslant n$), or ridgeless regression (if $p > n$), as p varies from 1 to 300. All quantities in this figure are empirical estimates computed over 500 repetitions (500 times drawing the simulated data sets), and in each repetition, the empirical prediction errors are computed based on a test set of 1000 samples.

For the simulations in Appendices C.5 and C.6, we generate data according to a linear model

$$y_i = x_i^\top \beta + \varepsilon_i, \quad i \in [n],$$

where each $x_i \sim \mathcal{N}(0, I/n)$, $\varepsilon_i \sim \mathcal{N}(0, 1)$, and we set $\beta_j = \sqrt{n/(\delta p)}$ with probability δ on, and $\beta_j = 0$ with probability $1 - \delta$, independently for $j \in [p]$. This setup has an SNR of 1.

In all figures that follow in this appendix section, Figures 11 to 14, the curves indicate theoretical quantities (asymptotic equivalents from the theorems), while the dots denote empirical estimates from averaging over 100 repetitions (100 times drawing the simulated data sets). In each repetition, empirical prediction errors are computed based on a test set of 1000 samples.

C.2 Figure formatting

For all figures in this section, we use the following formatting scheme.

- Curves in the underparameterized regime are colored blue.
- Curves in the overparameterized regime are colored orange.
- Fixed-X quantities are colored green.
- Emergent random-X quantities are denoted by solid lines (—).
- Intrinsic random-X quantities are denoted by dashed lines (- - -).

C.3 Ridge regression

Figure 11 provides empirical support for the behaviors described in Proposition 5 and Theorem 6. The top row corresponds to the underparameterized regime, while the bottom row corresponds to the overparameterized regime. Throughout, we see that the empirical estimates (dots) closely track the asymptotic equivalents (curves).

Moreover, we observe the following behaviors which align with the theory. The intrinsic random-X degrees of freedom decreases monotonically with λ in both the underparameterized and overparameterized regimes. Interestingly, the emergent random-X degrees of freedom can have nonmonotonic behavior in λ. Lastly, emergent random-X degrees of freedom is consistently higher than intrinsic random-X degrees of freedom, confirming that the presence of bias inflates degrees of freedom.

C.4 Ridgeless regression

Figure 12 provides empirical support for the behaviors described in Theorem 7 and Proposition 8. We see that the empirical estimates (dots) closely track the asymptotic equivalents (curves).

Furthermore, we observe the following behaviors which align with the theory. Both the intrinsic and emergent random-X degrees of freedom are maximized at $\gamma_n = 1$. The intrinsic random-X degrees of freedom decreases on both sides as γ_n moves away from 1. Moreover, emergent random-X degrees of freedom is always higher than intrinsic random-X degrees of freedom.

C.5 Lasso illustration

Figure 13 provides empirical support for the behaviors described in Theorem 9 and Proposition 10. The top row corresponds to the underparameterized regime, while the bottom row corresponds to the overparameterized regime. Throughout, we see that the empirical estimates (dots) closely track with the asymptotic equivalents (curves).

We also see the following behaviors which align with the theory. The intrinsic random-X degrees of freedom decreases monotonically with λ in either the underparameterized and overparameterized setting. Also, the emergent random-X degrees of freedom is always higher than intrinsic random-X degrees of freedom, confirming that the presence of bias inflates degrees of freedom.

C.6 Lassoless illustration

Figure 14 provides empirical support for the behaviors described in Theorem 11 and Proposition 12. We see that the empirical estimates (dots) closely track the asymptotic equivalents (curves).

Furthermore, we observe the following behaviors which align with the theory. Both the intrinsic and emergent random-X degrees of freedom are maximized at $\gamma_n = 1$. The intrinsic random-X degrees of freedom decreases on both sides as γ_n moves away from 1. Moreover, emergent random-X degrees of freedom is always higher than intrinsic random-X degrees of freedom.

D Additional experiments for Section 6

D.1 k-nearest neighbors regression

Here we study k-nearest neighbors (kNN) regression. Note that this is a linear smoother, hence its random-X degrees of freedom is characterized by Proposition 3, but it is not defined by a penalized least squares problem, therefore it eludes the analysis in Proposition 4 which characterizes emergent minus intrinsic degrees of freedom.

We simulate data according to the nonlinear model described in Appendix C.1. Figure 15 displays the results for an underparameterized problem with $n = 500$, $p = 300$, and Figure 16 displays the results for an overparameterized problem with $n = 200$, $p = 300$. In both cases, we can see (middle panel) that the intrinsic random-X degrees of freedom is slightly smaller than the fixed-X degrees

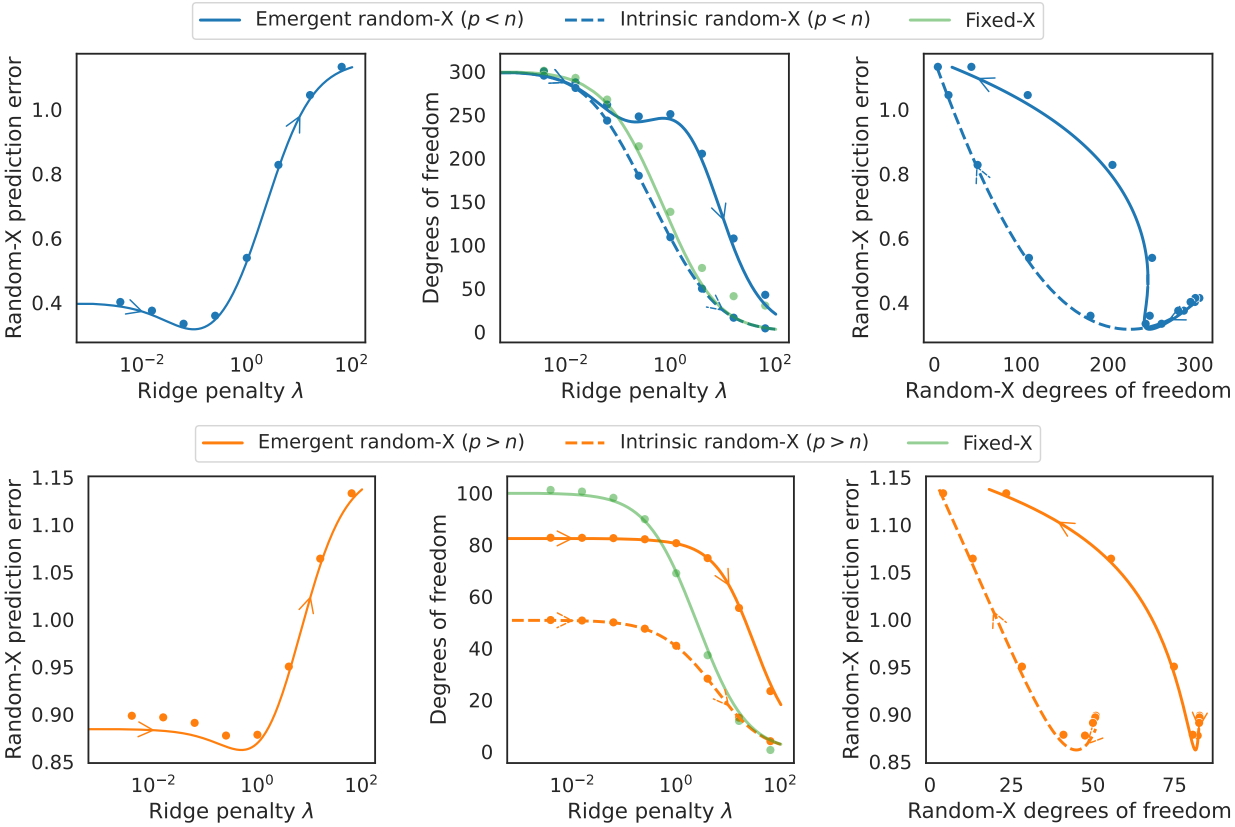

Figure 11: Prediction error and degrees of freedom of ridge predictors, over varying λ, in a problem setting with $p = 300$ features. The first row corresponds to the underparameterized regime, $n = 500$, and the second to the overparameterized regime, $n = 200$. The precise setup is as described in Appendix C.1.

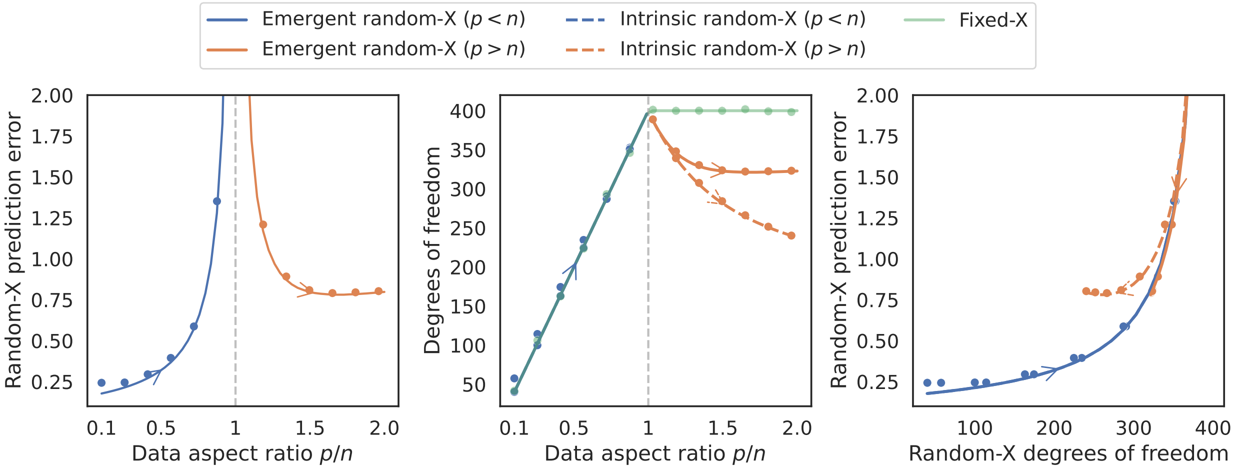

Figure 12: Prediction error and degrees of freedom of ridgeless predictors with varying aspect ratio $\gamma_n = p/n$. The number of samples is $n = 400$. The precise setup is as described in Appendix C.1.

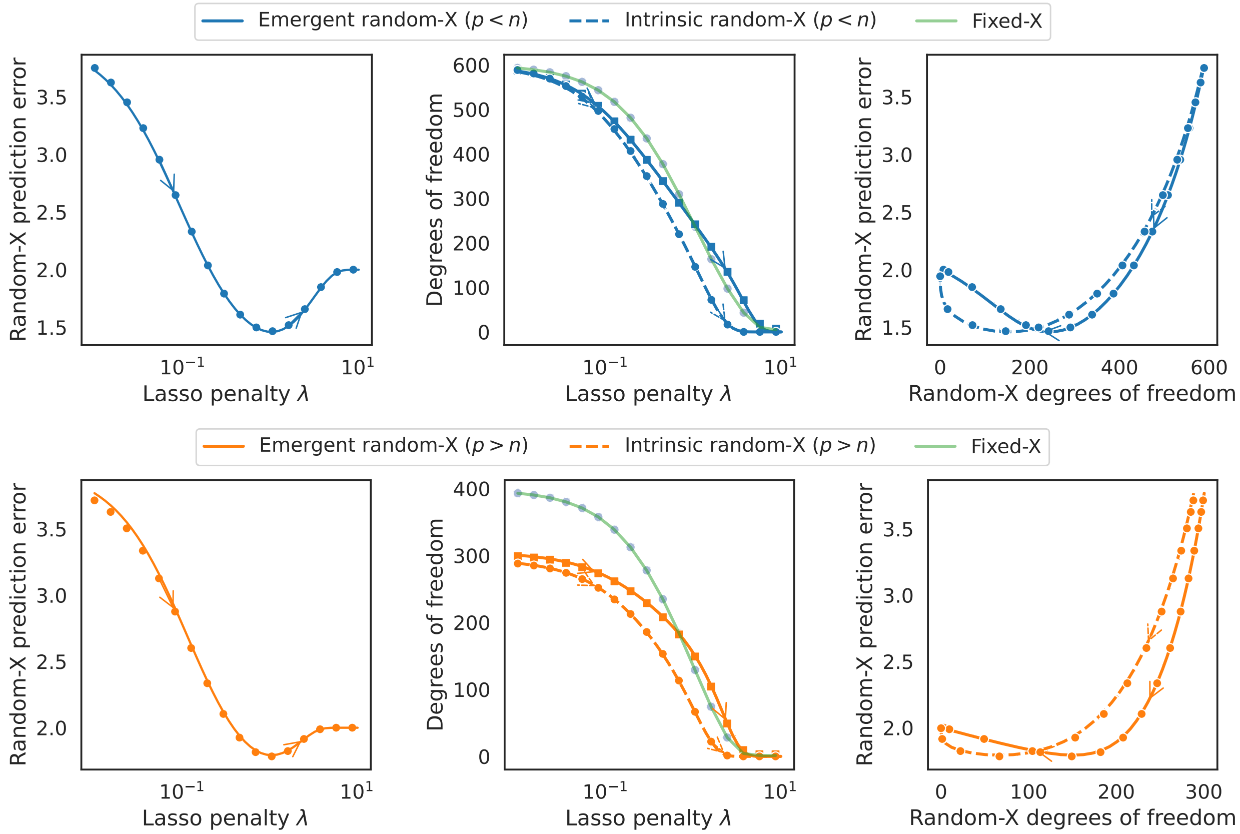

Figure 13: Prediction error and degrees of freedom of lasso predictors, over varying λ, in a problem setting with $p = 600$ features. The first row corresponds to the underparameterized regime, $n = 800$, and the second to the overparameterized regime, $n = 400$. The precise setup is as described in Appendix C.1 with $\delta = 1/6$.

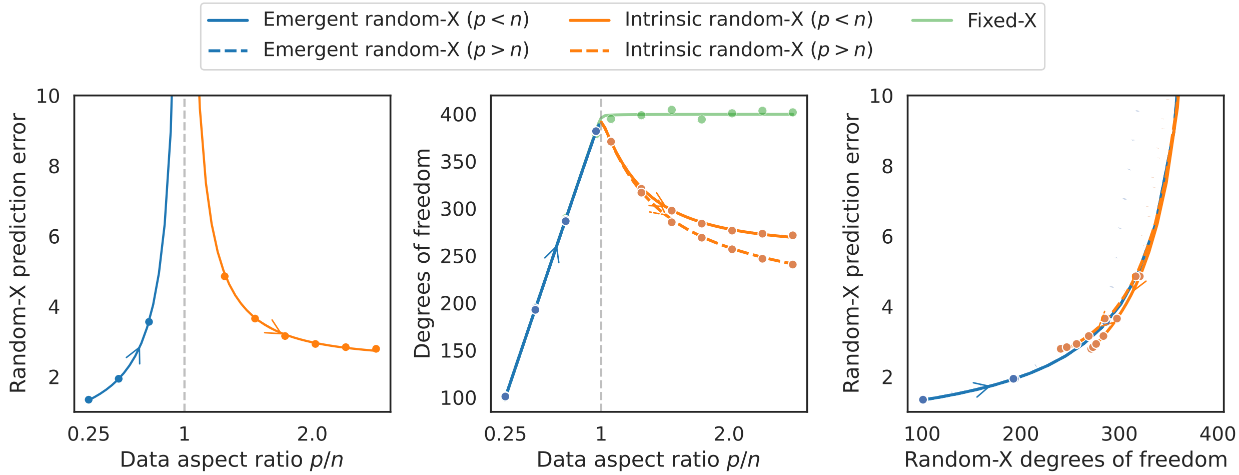

Figure 14: Prediction error and degrees of freedom of lassoless predictors with varying aspect ratio $\gamma_n = p/n$. The number of samples is $n = 400$. The precise is as described in Appendix C.1 with $\delta = 1/10$.

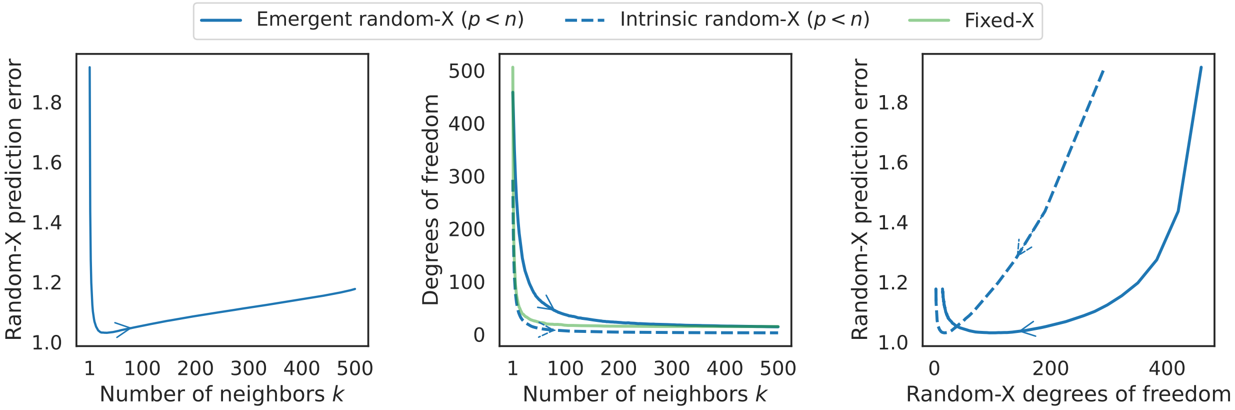

Figure 15: Prediction error and degrees of freedom of kNN predictors, in a problem with $n = 500$, $p = 300$.

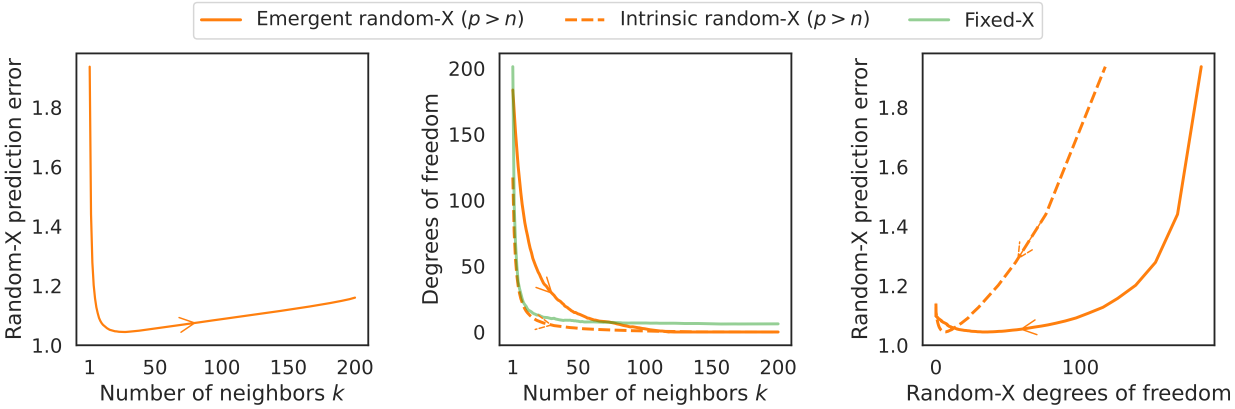

Figure 16: Prediction error and degrees of freedom of kNN predictors, in a problem with $n = 200$, $p = 300$.

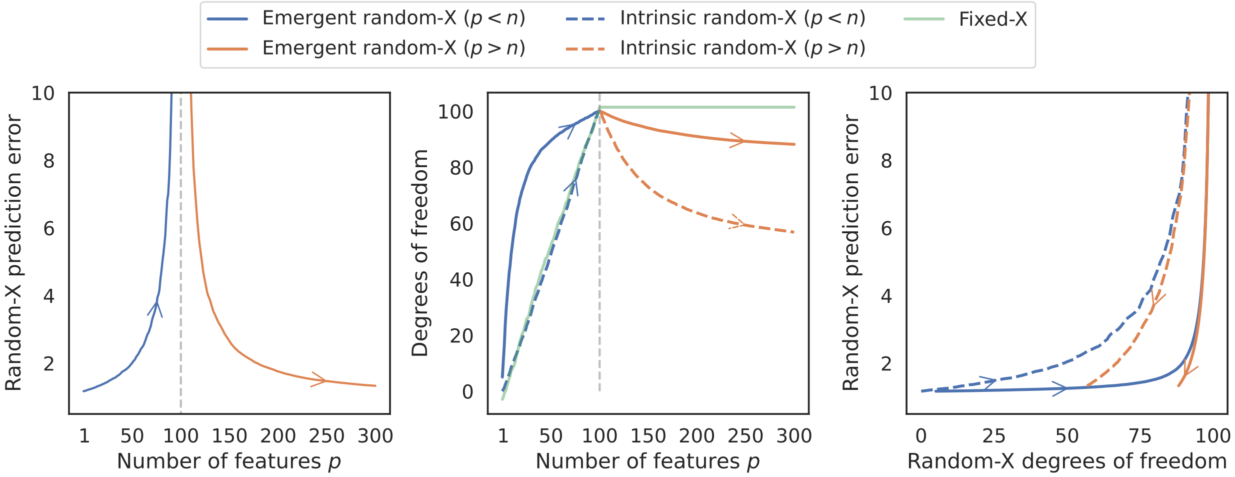

Figure 17: Prediction error and degrees of freedom of ridgeless regression on random features, in a problem where $n = 100$ and p ranges from 1 to 300.

of freedom throughout, for all k; whereas the emergent random-X degrees of freedom is somewhat larger than fixed-X degrees of freedom for small k, then it drops down for larger k. A commonality we see here, as with all other experiments, is that the degrees of freedom "due to bias" is positive. However, an interesting difference is as follows: emergent degrees of freedom is *larger* than fixed-X degrees of freedom on the less-regularized side of the model class (smaller k); with other predictors, we observe emergent degrees of freedom being smaller than fixed-X degrees of freedom on this side of the path (cf. ridge and lasso predictors for small λ in Appendices C.3 and C.5).

D.2 Random features

We examine ridgeless regression on random features. We simulate data according to the nonlinear model in Appendix C.1, with $n = 100$ samples and $P = 300$ features total, then we use features $\widetilde{x}_i = \tanh(Fx_i)$ for least squares (if $p \leqslant n$), or ridgeless regression (if $p > n$), where $F \in \mathbb{R}^{p \times P}$ has entries drawn from $\mathcal{N}(0, 1/\sqrt{P})$, and p varies from 1 to 300.

Figure 17 displays the results. These results are overall similar to Figure 1, except the emergent random-X degrees of freedom is inflated before the interpolation threshold at $p = n$.