Prediction, Generalization, and Complexity: Revisiting the View from Classical Statistics

Ryan J. Tibshirani

Simons MPG Bootcamp, August 2024

Abstract

Classical statistical decision theory treats the relationship between the prediction error, generalization gap (which the statistics literature calls optimism), and complexity of a model from a perspective that treats the covariates X as fixed, nonrandom values. This fixed-X perspective delivers a number of insights, many of which are well-known to statisticians, but perhaps less well-known to researchers in machine learning. This talk reviews these results from the classical statistics literature, and then emphasizes the ways in which this fixed-X theory is insufficient to explain the prediction error, generalization gap, and complexity of a model in a random-X setting—which is the predominant view in machine learning—especially as a predictive model becomes flexible enough to interpolate the training data. Finally, we show how to reinterpret some of the fixed-X classical statistics concepts in order to extend them to a random-X setting.

1 Models and metrics: fixed-X and random-X

- Random-X model: i.i.d. pairs $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \ldots, n$ of features/covariates and responses.
- Fixed-X model: relates features and responses through signal plus noise with i.i.d. errors,

$$\begin{gathered} y_i = f(x_i) + \epsilon_i, \quad i = 1, \ldots, n, \\ \epsilon_i,\ i = 1, \ldots, n \text{ i.i.d., with each } \epsilon_i \sim (0, \sigma^2), \\ \text{and where } x_i \in \mathbb{R}^p,\ i = 1, \ldots, n \text{ are fixed.} \end{gathered}$$

- When can we derive the second from the first, conditional on X (feature matrix with rows x_i)? Take $f(x) = \mathbb{E}[y_i | x_i] = x$ and we get mean zero errors $\epsilon_i = y_i - f(x_i)$. But for these to be i.i.d. conditional on X, we require that each $x_i \perp\!\!\!\perp \epsilon_i$. This is a strong assumption! (Implies no omitted variables.)
- The random-X and fixed-X models call to mind distinct notions of risk/prediction error. Given $\hat{f}$, fit on the training data (x_i, y_i), $i = 1, \ldots, n$, a natural notion of risk in the random-X model is

$$\mathsf{Risk}_{\mathrm{R}}(\hat{f}) = \mathbb{E}\big[(f(x_0) - \hat{f}(x_0))^2\big],$$

  where x_0 is an independent draw from the feature distribution. Expectation is over everything that is random (i.e., x_0 and (x_i, y_i), $i = 1, \ldots, n$).
- In the fixed-X model, a natural notion of risk is

$$\mathsf{Risk}_{\mathrm{F}}(\hat{f}; X) = \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - \hat{f}(x_i))^2 \,\bigg|\, X\bigg].$$

- Note that these only differ from analogous random-X and fixed-X prediction error metrics by a constant σ^2 (not depending on $\hat{f}$):

$$\mathsf{Err}_{\mathrm{R}}(\hat{f}) = \mathbb{E}\big[(y_0 - \hat{f}(x_0))^2\big] = \mathsf{Risk}_{\mathrm{R}}(\hat{f}) + \sigma^2,$$

where (x_0, y_0) is an independent draw from the joint distribution on the features and response, and similarly

$$\mathsf{Err}_{\mathrm{F}}(\hat{f}; X) = \mathbb{E}_{\mathrm{F}}\bigg[\frac{1}{n}\sum_{i=1}^{n}(y_i^* - \hat{f}(x_i))^2 \,\bigg|\, X\bigg] = \mathsf{Risk}_{\mathrm{F}}(\hat{f}; X) + \sigma^2,$$

where each (x_i, y_i^*) is an independent copy of (x_i, y_i) (i.e., y_i^* is drawn independently from the law of $y_i \,|\, x_i$).

- Thus there is a sort of equivalence in reasoning about risk and prediction error and we'll fluidly move back and forth and use either one as convenient. Finally, a lot of what is said here extends beyond squared loss to more general loss functions (e.g., Bregman divergences) though we'll omit details for simplicity.

- Classically, statisticians seem to love studying fixed-X risk/error. A big part of this is probably that it gives a clean framework in which we can develop various powerful tools for analysis (which we will review shortly).

- To be more charitable, there may have also been more genuine historical motivations for fixed-X risk/error from experimental design. If the values x_i, $i = 1, \ldots, n$ were chosen (i.e., designed), and we ran an experiment in which we measured y_i, $i = 1, \ldots, n$, and we plan to measure y_i^*, $i = 1, \ldots, n$, again in the future at the same design values, then studying $\mathsf{Err}_{\mathrm{F}}(\hat{f}; X)$ would actually be natural.

- The fixed-X view is also more in line with metrics of interest in signal processing: imagine x_i simply encodes an index (say, of a sequence or a matrix).

- In machine learning (and many modern applications of statistical prediction), of course, random-X notions of risk/error are the standard, and seem much more natural overall.

- So this begs the question: how different can random-X and fixed-X metrics be? We will discuss this later. First, we will adopt the fixed-X perspective and review some classical results from statistical decision theory.

2 Fixed-X optimism, covariance, and degrees of freedom

2.1 Optimism and Efron's covariance formula

- The *optimism* of a predictor $\hat{f}$ is defined by

$$\mathsf{Opt}_{\mathrm{F}}(\hat{f}; X) = \mathsf{Err}_{\mathrm{F}}(\hat{f}; X) - \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \,\bigg|\, X\bigg].$$

  In other words, it is the difference between the fixed-X prediction error and the training error of $\hat{f}$. The term "optimism" appears to be due to Brad Efron (precise references below).

- A key relationship, in the fixed-X data model, is given by *Efron's covariance formula*:

$$\mathsf{Opt}_{\mathrm{F}}(\hat{f}; X) = \frac{2}{n}\sum_{i=1}^{n}\mathrm{Cov}[y_i, \hat{f}(x_i) \,|\, X].$$

  This is named in recognition of Efron's work on the topic (Efron, 1983, 1986, 2004), but some version of this fact had been known earlier (Akaike, 1973; Mallows, 1973; Stein, 1981).

- The covariance formula holds without any assumptions on $\hat{f}$, or any distributional assumptions on the individual errors ϵ_i, except $\mathbb{E}[\epsilon_i] = 0$ and $\mathrm{Var}[\epsilon_i] = \sigma^2$. A quick way to verify it: consider Bregman divergence loss $D_\phi(y_i, \hat{f}(x_i))$, where

$$D_\phi(u, v) = \phi(u) - \phi(v) - \phi'(v)(u - v).$$

Squared error is the special case in which $\phi(u) = u^2$. For such losses,

$$D_\phi(y_i^*, \hat{f}(x_i)) - D_\phi(y_i, \hat{f}(x_i)) = \phi(y_i^*) - \phi(y_i) - \phi'(\hat{f}(x_i))(y_i^* - y_i),$$

so the optimism is

$$\begin{aligned}
\mathsf{Opt}_{\text{F}}(\hat{f}; X) &= \frac{1}{n} \sum_{i=1}^n \mathbb{E}\Big[D_\phi(y_i^*, \hat{f}(x_i)) - D_\phi(y_i, \hat{f}(x_i)) \,\Big|\, X \Big] \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}\Big[\phi(y_i^*) - \phi(y_i) - \phi'(\hat{f}(x_i))(y_i^* - y_i) \,\Big|\, X \Big] \\
&= \frac{1}{n} \sum_{i=1}^n \mathbb{E}\Big[\phi'(\hat{f}(x_i))(y_i - f(x_i) + f(x_i) - y_i^*) \,\Big|\, X \Big] \\
&= \frac{1}{n} \sum_{i=1}^n \mathrm{Cov}[y_i, \phi'(\hat{f}(x_i)) \,|\, X].
\end{aligned}$$

For squared error loss $\phi'(u) = 2u$, and this recovers the claimed covariance formula. We can also see that something more general holds for Bregman divergences.

2.2 Degrees of freedom

- The (effective) *degrees of freedom* of $\hat{f}$ is defined by

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

  Some authors motivate degrees of freedom intuitively by commenting that the more complex the fitting procedure $\hat{f}$, the more "self-influence" each response y_i will have on the corresponding predicted value $\hat{f}(x_i)$ (and hence the higher the degrees of freedom in total).

- The literature on degrees of freedom is of course therefore closely tied to that on optimism and covariance penalties. It is unclear where the term "degrees of freedom" originated, but early work that uses this term includes Hastie and Tibshirani (1987, 1990); Ye (1998); Efron (2004).

- The utility of degrees of freedom is twofold:

  1. It serves as a unitless, fundamental measure of the complexity of a predictive model. This enables us to reason about the behavior of models, and to compare them, more rigorously.
  2. It provides an avenue for estimating of fixed-X prediction error: note that by Efron's covariance formula combined with the definition of degrees of freedom,

$$\mathsf{Err}_{\text{F}}(\hat{f}; X) = \mathbb{E}\bigg[\frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 \,\bigg|\, X \bigg] + \frac{2\sigma^2}{n} \mathrm{df}(\hat{f}).$$

     Thus if $\widehat{\mathrm{df}}(\hat{f})$ is an unbiased estimate of $\mathrm{df}(\hat{f})$, then

$$\widehat{\mathsf{Err}}_{\text{F}}(\hat{f}; X) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \frac{2\sigma^2}{n} \widehat{\mathrm{df}}(\hat{f}) \tag{2}$$

     is an unbiased estimate of fixed-X prediction error.

- There is a sizeable literature on deriving exact formulae or unbiased estimates of degrees of freedom for a variety of core predictors of interest. This is enabled by a seminal result due to Stein, which we will cover in the next subsection; first, we give some basic examples where degrees of freedom can be computed directly.

2.2.1 Least squares regression

- As usual let $X \in \mathbb{R}^{n \times p}$ denote the feature matrix (whose i^{th} row is x_i) and let $Y \in \mathbb{R}^n$ denote the response vector (whose i^{th} element is y_i). Assume that X has linearly independent columns, which necessarily requires that $d \leq n$. The least squares predictor of Y on X is given by

$$\hat{f}^{\text{ls}}(x) = x^\mathsf{T} \hat{\beta}^{\text{ls}}, \quad \text{where } \hat{\beta}^{\text{ls}} = (X^\mathsf{T} X)^{-1} X^\mathsf{T} Y.$$

- Abbreviating $P_X = X(X^\mathsf{T} X)^{-1} X^\mathsf{T}$, we can compute

$$\begin{aligned} \mathrm{df}(\hat{f}^{\text{ls}}) &= \frac{1}{\sigma^2} \operatorname{tr}(\operatorname{Cov}[X\hat{\beta}^{\text{ls}}, Y \,|\, X]) \\ &= \frac{1}{\sigma^2} \operatorname{tr}(\operatorname{Cov}[P_X Y, Y \,|\, X]) \\ &= \operatorname{tr}(P_X) \\ &= p, \end{aligned}$$

  where the last line uses $\operatorname{tr}(P_X) = \operatorname{tr}(X^\top X (X^\top X)^{-1}) = p$.

- In other words, the degrees of freedom of the least squares predictor is simply the number of estimated parameters, which is highly intuitive.

- In this case, the unbiased estimator from (2) is simply

$$\widehat{\mathsf{Err}}_{\text{F}}(\hat{f}; X) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \frac{2\sigma^2}{n} p,$$

  which is known as *Mallows' C_p estimator* (Mallows, 1973) of the fixed-X error of least squares.

2.2.2 k-nearest-neighbors regression

- Now let $\mathcal{N}_k(x)$ denote the indices of the k nearest points among x_i, $i = 1, \ldots, n$ to a given $x \in \mathbb{R}^p$ (as measured in ℓ_2 distance). The k-nearest-neighbors predictor is

$$\hat{f}^{\text{knn}}(x) = \frac{1}{k} \sum_{i \in \mathcal{N}_k(x)} y_i.$$

- Let $w(x) \in \mathbb{R}^n$ with $w(x)_j = 1/k$ if $j \in \mathcal{N}_k(x)$ and 0 otherwise. Then can abbreviate

$$\hat{f}^{\text{knn}}(x) = w(x)^\mathsf{T} Y.$$

- Letting $L_X \in \mathbb{R}^{n \times n}$ denote the matrix with i^{th} row $w(x_i)$, we can compute

$$\begin{aligned} \mathrm{df}(\hat{f}^{\text{knn}}) &= \frac{1}{\sigma^2} \operatorname{tr}(\operatorname{Cov}[L_X Y, Y \,|\, X]) \\ &= \operatorname{tr}(L_X) \\ &= \frac{n}{k}. \end{aligned}$$

- In other words, the degrees of freedom of k-nearest-neighbors regression is n/k. This makes intuitive sense: it equals 1 when $k = n$ (and the predictor reduces to an average), and it equals n when $k = 1$ (and the predictor reduces to nearest-neighbor interpolation).

- The second-to-last line in the previous display also shows that the degrees of freedom of an arbitrary linear smoother is the trace $\operatorname{tr}(L_X)$ of the smoother matrix. This gives use explicit degrees of freedom formulae for predictors like ridge, kernel ridge, smoothing splines, thin-plate splines, etc.

2.3 Interlude: Stein's formula

- This subsection can be skipped without interrupting the flow of main ideas. We use it as an opportunity to cover a beautiful idea which perhaps not everybody in machine learning knows about.

- *Stein's formula* (Stein, 1981) says if $\hat{f}$ is weakly differentiable as a function of Y, and we additionally assume Gaussian errors $\epsilon_i \sim N(0,\sigma)^2$ in the fixed-X data model, then

$$\mathrm{df}(\hat{f}) = \mathbb{E}\bigg[\sum_{i=1}^n \frac{\partial \hat{f}(x_i)}{\partial y_i} \,\bigg|\, X\bigg]. \tag{3}$$

  Like the covariance definition, Stein's formula for degrees of freedom is quite intuitive: the more complex the predictor $\hat{f}$, the greater the dependence of $\hat{f}(x_i)$ on y_i, and hence the steeper the derivative.

- Based on (3), we are able to form an unbiased estimate of $\mathrm{df}(\hat{f})$, namely,

$$\widehat{\mathrm{df}}(\hat{f}) = \sum_{i=1}^n \frac{\partial \hat{f}(x_i)}{\partial y_i}.$$

  Plugging this into (2) gives

$$\widehat{\mathsf{Err}}_{\mathrm{F}}(\hat{f}; X) = \frac{1}{n}\sum_{i=1}^n (y_i - \hat{f}(x_i))^2 + \frac{2\sigma^2}{n}\sum_{i=1}^n \frac{\partial \hat{f}(x_i)}{\partial y_i},$$

  which is known as *Stein's unbiased risk estimator* (SURE).

- Computing the Stein divergence $\widehat{\mathrm{df}}(\hat{f}) = \sum_{i=1}^n \partial \hat{f}(x_i)/\partial y_i$ is no trivial feat, but there is literature showing it can be done in closed-form for predictors like wavelet denoising, shape-constrained regression, quantile regression, lasso and various generalizations, and low-rank matrix factorization; see, e.g., Donoho and Johnstone (1995); 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). (The Stein divergence could probably also be done or approximated by auto-differentiation techniques.)

- At the core of (3) is *Stein's lemma*: if $h : \mathbb{R} \to \mathbb{R}$ is weakly differentiable, and $Z \sim N(0,1)$, then

$$\mathbb{E}[Zh(Z)] = \mathbb{E}[h'(Z)].$$

  The proof is simple. Letting $\phi(z) = \exp(-z^2/2)/\sqrt{2\pi}$ denote the standard normal density, observe

$$\begin{aligned}
\mathbb{E}[h'(Z)] &= \int h'(z)\phi(z)\,dz \\
&= h(z)\phi(z)\Big|_{-\infty}^{\infty} - \int h(z)\phi'(z)\,dz \\
&= \int zh(z)\phi(z)\,dz \\
&= \mathbb{E}[Zh(Z)].
\end{aligned}$$

2.4 Bootstrap, Bregman, exponential families, etc.

- Some very quick remarks before moving on to random-X. First, when Stein's formula does not apply (or cannot be computed), parametric bootstrap methods can be used estimate degrees of freedom in (1). The classical references are Breiman (1992); Ye (1998); Efron (2004). A recent improvement is given in Oliveira et al. (2021).

- Second, a lot of what has been said here carries over to Bregman divergence loss functions and data from exponential family models. A core reference is Efron (1986). There are also analogs of Stein's lemma for continuous and discrete exponential family distributions; e.g., the result for the Poisson case is called *Hudson's lemma*, due to Hudson (1978).

3 Fixed-X versus random-X views: how different are they?

- Given all the development just covered in fixed-X decision theory, one might wonder how much of this is appropriate for a random-X perspective on prediction.
- After all, tools like cross-validation (or more simply train/validation splitting) are the de facto standard for estimating random-X prediction error. Why aren't we all instead using covariance penalties, degrees of freedom, SURE, etc.?
- The applicability of these fixed-X tools hinges on whether random-X prediction error behaves similarly to fixed-X prediction error. How different are they?
- The short answer is:
  - *Usually not very different* in classical regimes: low dimensions p, and smooth functions $f, \hat{f}$.
  - *But can be very different* in modern regimes: high dimensions p, or nonsmooth functions $f, \hat{f}$.
- We won't really go into further detail on that paper, and instead we'll discuss some examples/evidence that support the "short answer" given above, across the next two subsections.

3.1 Insights from least squares regression

- Let's return to least squares predictor $\hat{f}^{\text{ls}}$ of Y on X, where we assume $X \in \mathbb{R}^{n \times p}$ has linearly independent columns. To investigate its risk properties as simply as possible, let's assume an underlying linear model

$$y_i = x_i^T \beta + \epsilon_i, \quad i = 1, \ldots, n,$$

  for some unknown coefficients $\beta \in \mathbb{R}^p$.

- To compute its fixed-X risk, first note that

$$\mathbb{E}[\hat{\beta}^{\text{ls}} \,|\, X] = (X^T X)^{-1} X^T \mathbb{E}[Y|X] = \beta.$$

  That is, the least squares coefficient estimate $\hat{\beta}^{\text{ls}}$ is itself unbiased. Thus the fixed-X squared bias, in a decomposition of fixed-X risk, is

$$\mathsf{B}_{\text{F}}(\hat{f}; X) = \frac{1}{n} \big\| \mathbb{E}[X\hat{\beta}^{\text{ls}} \,|\, X] - X\beta \big\|_2^2 = 0.$$

  Meanwhile, the fixed-X variance, in a decomposition of fixed-X risk, is

$$\begin{aligned} \mathsf{V}_{\text{F}}(\hat{f}; X) &= \frac{1}{n} \operatorname{tr}(\operatorname{Cov}[X\hat{\beta}^{\text{ls}} \,|\, X]) \\ &= \frac{1}{n} \operatorname{tr}(\operatorname{Cov}[P_X Y \,|\, X]) \\ &= \frac{\sigma^2}{n} \operatorname{tr}(P_X) \\ &= \frac{\sigma^2 p}{n}. \end{aligned}$$

  Therefore the fixed-X risk is

$$\mathsf{Risk}_{\text{F}}(\hat{f}; X) = \mathsf{B}_{\text{F}}(\hat{f}; X) + \mathsf{V}_{\text{F}}(\hat{f}; X) = \frac{\sigma^2 p}{n}.$$

- Meanwhile, for random-X risk, conditioning on both X, x_0 we can define perform a similar bias-variance decomposition with squared bias

$$\mathsf{B}_{\text{R}}(\hat{f}; X, x_0) = \big(\mathbb{E}[x_0^T \hat{\beta}^{\text{ls}} \,|\, X, x_0] - x_0^T \beta \big)^2 = 0.$$

The variance in this decomposition is

$$\begin{aligned}\mathsf{V}_{\mathrm{R}}(\hat{f}, X; x_0) &= \mathrm{Var}[x_0^\mathsf{T}\hat{\beta} \,|\, X, x_0] \\ &= x_0^\mathsf{T}(X^\mathsf{T}X)^{-1}X^\mathsf{T}\,\mathrm{Cov}[Y \,|\, X]\, X(X^\mathsf{T}X)^{-1}x_0 \\ &= \sigma^2 x_0^\mathsf{T}(X^\mathsf{T}X)^{-1}x_0 \\ &= \sigma^2\,\mathrm{tr}\left(x_0 x_0^\mathsf{T}(X^\mathsf{T}X)^{-1}\right).\end{aligned}$$

The the random-X risk, after integrating over X, x_0, is

$$\mathsf{Risk}_{\mathrm{R}}(\hat{f}) = \mathbb{E}[\mathsf{B}_{\mathrm{R}}(\hat{f}; X, x_0)] + \mathbb{E}[\mathsf{V}_{\mathrm{R}}(\hat{f}; X, x_0)] = \sigma^2\,\mathrm{tr}\left(\mathbb{E}[x_0 x_0^\mathsf{T}]\mathbb{E}[(X^\mathsf{T}X)^{-1}]\right),$$

where we have used the independence of X and x_0.

- There are various ways to attack the last formula to get more insights; one particularly simple approach is to assume that each $x_i \sim N(0, \Sigma)$ (random matrix theory will then show that a broad class of feature models lead to the same answer under proportional asymptotics). Under this assumption $\mathbb{E}[x_0 x_0^\mathsf{T}] = \Sigma$, by definition, and $(X^\mathsf{T}X)^{-1}$ has an inverse Wishart distribution, whose expectation is known to be $\Sigma/(n-p-1)$, thus

$$\mathsf{Risk}_{\mathrm{R}}(\hat{f}) = \sigma^2\,\mathrm{tr}\left(\Sigma\frac{\Sigma^{-1}}{n-p-1}\right) = \frac{\sigma^2 p}{n-p-1}.$$

- To recap, we have shown that least squares regression on p linearly independent (and jointly Gaussian) features has fixed-X and random-X risk:

$$\mathsf{Risk}_{\mathrm{F}}(\hat{f}; X) = \frac{\sigma^2 p}{n} \quad \text{and} \quad \mathsf{Risk}_{\mathrm{R}}(\hat{f}) = \frac{\sigma^2 p}{n-p-1}.$$

  These are similar when p is small relative to n, but *very different* when p approaches n. The random-X error explodes at $p = n-1$. This and related behavior have of course been studied intensely in the recent literature on double descent and benign interpolation (as has the behavior of the ridgeless least squares predictor when $p > n$). Nice reviews: Bartlett et al. (2021); Belkin (2021).

3.2 Insights from empirical process theory

- From a different perspective, note that we can also view $\mathsf{Risk}_{\mathrm{F}}(\hat{f})$ and $\mathsf{Risk}_{\mathrm{F}}(\hat{f}; X)$ in terms of $L^2(P)$ and $L^2(P_n)$ distances: defining

$$\begin{aligned}\|f - \hat{f}\|_{L^2(P)}^2 &= \int (f(x) - \hat{f}(x))^2\, dP(x), \\ \|f - \hat{f}\|_{L^2(P_n)}^2 &= \int (f(x) - \hat{f}(x))^2\, dP_n(x),\end{aligned}$$

  where P is the covariate distribution and P_n the empirical distribution of x_i, $i = 1, \ldots, n$, we have

$$\mathsf{Risk}_{\mathrm{R}}(\hat{f}) = \mathbb{E}\big[\|f - \hat{f}\|_{L^2(P)}^2\big], \quad \text{and} \quad \mathsf{Risk}_{\mathrm{F}}(\hat{f}; X) = \mathbb{E}\big[\|f - \hat{f}\|_{L^2(P_n)}^2 \,|\, X\big].$$

- Empirical process theory gives bounds on the differences between $L^2(P_n)$ and $L^2(P)$ norms. Given a function class $\mathcal{G}$, recall its *localized Rademacher complexity* is defined by

$$R_n(\delta) = \mathbb{E}\left[\sup_{g \in \mathcal{G},\, \|g\|_{L^2(P)} \le \delta} \frac{1}{n}\left|\sum_{i=1}^n \sigma_i g(x_i)\right|\right],$$

  where σ_i, $i = 1, \ldots, n$ are i.i.d. Rademacher variables (taking values ± 1 with equal probability) and x_i, $i = 1, \ldots, n$ are i.i.d. from P. The following is adapted from Theorem 14.1 of Wainwright (2019).

Theorem 1. *Let $\mathcal{G}$ be a star-shaped and b-uniformly bounded class of functions for some $b > 0$, and denote by δ_n the smallest positive solution to*

$$R_n(\delta) \leq \delta^2/b. \tag{4}$$

Then there exists universal constants $c_0, c_1, c_2 > 0$ such that

$$\Big| \|g\|_{L^2(P_n)} - \|g\|_{L^2(P)} \Big| \leq c_0 \delta_n, \quad \text{for all } g \in \mathcal{G}, \tag{5}$$

with probability at least $1 - c_1 \exp(-c_2 n \delta_n^2 / b^2)$.

- A consequence can be read off as follows. Take $g = f - \hat{f}$ and suppose that $f - \hat{f} \in \mathcal{G}$. For example, if each of $f, \hat{f}$ are Lipschitz continuous with parameter L, then $f - \hat{f}$ is Lipschitz with parameter $2L$. So we can take $\mathcal{G}$ to be the class of Lipschitz functions with parameter $2L$, and then (5) implies

$$\Big| \big\|f - \hat{f}\big\|_{L^2(P_n)} - \big\|f - \hat{f}\big\|_{L^2(P)} \Big| \leq c_0 \delta_n, \tag{6}$$

  with high probability. If the criticial radius of δ_n of $\mathcal{G}$, the smallest positive solution to (4), is small, then (6) gives us a meaningful bound. The analogous statement also holds after taking expectations over the training set, i.e., for fixed-X and random-X risk, and this is one sense in which we can prove that these metrics are close (when δ_n is small).

- We can now flip this result and talk about the contrapositive. What about when we *know* that the fixed-X and random-X risks differ substantially? An example is a generalizing interpolator: here, we would have $\mathsf{Risk}_{\mathrm{R}}(\hat{f}) \to 0$ as $n \to \infty$, but

$$\mathsf{Risk}_{\mathrm{F}}(\hat{f}; X) = \mathbb{E}\bigg[\frac{1}{n} \sum_{i=1}^n (f(x_i) - y_i)^2 \,\bigg|\, X \bigg] = \sigma^2.$$

  This represents a huge difference between the two metrics: one vanishing, and the other pinned at the noise level. But this cannot happen under (6) if δ_n vanishes. Therefore we arrive at the following "theorem of alternatives" for interpolators.

  Theorem 2. *Let $\mathcal{G}$ be a star-shaped and b-uniformly bounded class of functions, and assume that $g, h \in \mathcal{G}$ implies $g - h \in \alpha\mathcal{G}$ for some universal constant $\alpha > 0$ (this happens, e.g., if $\mathcal{G}$ is a ball defined with respect to some norm or seminorm). Assume that $f \in \mathcal{G}$, and let $\hat{f}$ interpolate the training data. If the critical radius of $\mathcal{G}$ satisfies $\delta_n \to 0$ as $n \to \infty$, then the following two statements are mutually exclusive:*

  1. *$\hat{f}$ generalizes: $\mathsf{Risk}_{\mathrm{R}}(\hat{f}) \to 0$ as $n \to \infty$;*
  2. *$\hat{f} \in \beta\mathcal{G}$ for any constant $\beta > 0$.*

  To make things more concrete we can again just take $\mathcal{G}$ to be the class of functions that are Lipschitz continuous with parameter L. This is known to have critical radius $\delta_n \asymp n^{-\frac{1}{3}}$ for $d = 1$ and $\delta_n \asymp n^{-\frac{1}{2d}}$ for $d \geq 2$. Then for an interpolator $\hat{f}$, either:

  1. $\hat{f}$ generalizes: $\mathsf{Risk}_{\mathrm{R}}(\hat{f}) \to 0$ as $n \to \infty$; or
  2. $\hat{f}$ is Lipschitz continuous with any finite Lipschitz parameter.

4 Bias and variance, revisited

- The comparison between fixed-X and random-X errors is the central focus in Rosset and Tibshirani (2020). This paper compares the bias and variance components of these two error metrics separately, and proves that the random-X components are often (though not always) larger than their fixed-X counterparts for common various predictors, via finite-sample analysis. Unfortunately we won't have time to go into this.

5 Degrees of freedom, revisited

5.1 Limitations of classical degrees of freedom

- A critical limitation of classical degrees of freedom, as defined earlier, is straightforward to state. For any interpolator, satisfying $\hat{f}(x_i) = y_i$, $i = 1, \ldots, n$, we have the trivial answer:

$$\mathrm{df}(\hat{f}) = \frac{1}{\sigma^2} \sum_{i=1}^{n} \mathrm{Cov}[y_i, y_i \,|\, X] = n.$$

- 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, as any interpolator has zero training error and the same fixed-X prediction error.

- On the other hand, if we are to think of degrees of freedom as a general measure of model complexity, then the above display leaves a lot to be desired. As we know, some interpolators—in particular, implicitly regularized ones—are actually quite well-behaved and can generalize well to unseen data. Classical degrees of freedom lacks the ability to distinguish between well-behaved interpolators, which are smooth in between the covariate points, and wild ones, which are arbitrarily nonsmooth.

5.2 Reinterpreting classical degrees of freedom

What follows reflects ongoing work with Pratik Patil and Jin-Hong Du. A preprint will be on arXiv soon.

- Recall that if the feature matrix $X \in \mathbb{R}^{n \times p}$ has linearly independent columns, then least squares regression predictor $\hat{f}^{\mathrm{ls}}$ of Y on X has degrees of freedom p, and hence by Efron's covariance formula,

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

- Given an arbitrary predictor $\hat{f}$, we know that it still satisfies, again by Efron's covariance formula,

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

- Comparing the last two displays, we see that we may hence interpret the degrees of freedom of $\hat{f}$ 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 $\mathsf{Opt}_{\mathrm{F}}(\hat{f})$. This is simply a reformulation of the original definition (1), and the next proposition records this idea precisely.

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

$$\mathsf{Opt}_{\mathrm{F}}(\hat{f}^{\mathrm{rf}}_d) = \frac{2\sigma^2}{n} d, \quad d = 1, \ldots, n.$$

  Extend these reference values so that we may write for all nonnegative d,

$$\overline{\mathsf{Opt}}_{\mathrm{F}}(\hat{f}^{\mathrm{rf}}_d) = \frac{2\sigma^2}{n} d, \quad d \in [0, \infty].$$

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

$$\mathsf{Opt}_{\mathrm{F}}(\hat{f}) = \overline{\mathsf{Opt}}_{\mathrm{F}}(\hat{f}^{\mathrm{rf}}_d).$$

  Then $\mathrm{df}(\hat{f}) = d$.

5.3 Defining random-X degrees of freedom

- The idea behind the last proposition (reinterpretation) 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 $\hat{f}$, we require two things:

  1. a *metric* met, which we assume (without loss of generality) is negatively-oriented: the lower the value of $\mathsf{met}(\hat{f})$, the less complex we deem $\hat{f}$;
  2. a *reference class* $\{\hat{f}_d^{\mathrm{rf}} : d \in D\}$, 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 $\hat{f}$ a complexity of d where d is smallest value in D for which $\mathsf{met}(\hat{f}) \leq \mathsf{met}(\hat{f}_d^{\mathrm{rf}})$. 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 $\hat{f}$.

- Classical 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 an equality: $\mathsf{met}(\hat{f}) = \mathsf{met}(\hat{f}_d^{\mathrm{rf}})$ for some $d \geq 0$. 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 with random-X optimism, defined as

$$\mathsf{Opt}_{\mathrm{R}}(\hat{f}) = \mathsf{Err}_{\mathrm{R}}(\hat{f}) - \mathbb{E}\bigg[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2\bigg].$$

  This is sometimes called the *generalization gap* in the machine learning literature.

- Before we pursue this, it is important to note that the classical fixed-X definition, which uses least squares and fixed-X optimism in the equivalent reformulation given in the last subsection, is special for two reasons. The metric assigned to the reference model here, i.e., the fixed-X optimism 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 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 $n \geq 2$. For each fixed $d \leq n-1$, let $\widetilde{X}_d \in \mathbb{R}^{n\times d}$ have i.i.d. rows from $N(0,\Sigma)$, with $\Sigma \in \mathbb{R}^{d\times d}$ an arbitrary deterministic positive definite covariance matrix. Let*

$$\widetilde{Y}|\widetilde{X}_d \sim N(\widetilde{X}_d\beta, \sigma^2 I),$$

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

$$\mathsf{Opt}_{\mathrm{R}}(\hat{f}_d^{\mathrm{rf}}) = \sigma^2\bigg(\frac{d}{n} + \frac{d}{n-d-1}\bigg), \quad d = 1, \ldots, n-1.$$

  Extend these reference values so that we may write

$$\overline{\mathsf{Opt}}_{\mathrm{R}}(\hat{f}_d^{\mathrm{rf}}) = \sigma^2\bigg(\frac{d}{n} + \frac{d}{n-d-1}\bigg), \quad d \in [0, n-1].$$

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

$$\mathsf{Opt}_{\mathrm{R}}(\hat{f}) = \overline{\mathsf{Opt}}_{\mathrm{R}}(\hat{f}_d^{\mathrm{rf}}).$$

- Several remarks are in order. First, recall that fixed-X degrees of freedom ranges from 0 to ∞. (In fact, negative values are also allowed, but we implicitly rule this out in our reinterpretation.) That is, we cannot rule out arbitrarily large values of fixed-X degrees of freedom, a property that has been criticized by some authors. 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. Fixed-X optimism scaled linearly with d, whereas random-X optimism 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.

- The choice of Gaussian features $\widetilde{X}_d$ in facilitates the calculation of the random-X optimism of least squares regression (as we have already seen, we can leverage well-known properties of the inverse Wishart distribution). Interestingly, the result does not depend on the feature covariance Σ. By standard arguments in random matrix theory, the same formula is actually asymptotically exact (as $d/n \to \xi < 1$) for a broad class of feature models.

- The linear mean $\mathbb{E}[\widetilde{Y}|\widetilde{X}_d] = \widetilde{X}_d\beta$ is important, but the assumption of Gaussian errors is not. The calculation actually only requires 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 of the reference model 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.

5.4 An intrinsic version of model complexity

- The reference model we use in the previous definition is least squares regression on *well-specified* data, where the mean is linear in the covariates. As previously commented, 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, we are comparing $\mathsf{Opt}_{\mathrm{R}}(\hat{f})$—which is generically comprised of both bias and variance, to $\overline{\mathsf{Opt}}_{\mathrm{R}}(\hat{f}^{\mathrm{rf}}_d)$—which is made up of variance alone. This is intentional. The notion of random-X degrees of freedom given in the last subsection determines the complexity of the given predictor $\hat{f}$ 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 may want to match variance to variance in determining degrees of freedom, i.e., we may want to exclude bias effects in calculating the random-X optimism of the given model $\hat{f}$. This gives rise to a different notion of model complexity, which we define next.

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

$$\mathsf{Opt}_{\mathrm{R}}(\hat{f}(\cdot; X, v)) = \overline{\mathsf{Opt}}_{\mathrm{R}}(\hat{f}^{\mathrm{rf}}_d).$$

- The difference between the previous definition and the current one is that in the latter we are measuring the random-X optimism of $\hat{f}$ when it is being fit on "pure noise" $v \sim N(0, \sigma^2 I)$. Note that because the random-X optimism of least squares does not depend on β in the reference data model, we may set $\beta = 0$ and write the current definition as equivalently matching

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

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

- As it turns out, the emergent notion is generally larger than the intrinsic one. In short, the presence of bias generally "adds complexity". This can be proved formally for a number of prediction models (under no assumptions on the data distribution), such as least squares regression, ridge regression, kernel ridge, smoothing splines, and thin-plate splines. Experiments verify that this continues to be the case for other predictors (random forests, neural nets) in general.

5.5 Distribution shift, decomposition, and attribution

- As discussed, we can interpret $\mathrm{df}_{\mathrm{R}}^{\mathrm{e}}(\hat{f}) - \mathrm{df}_{\mathrm{R}}^{\mathrm{i}}(\hat{f})$ as the degrees of freedom "due to bias" (or perhaps more accurately, model misspecification).

- Generic decompositions are also possible. Let $e_1, \ldots, e_m$ be any list of user-chosen "sources of error". Here we are also allowing for distribution shift between training and test data sets. For example, we may be interested inpsecting the potential contributions of bias (e_1) and covariate shift (e_2).

- For any subset $S \subseteq \{e_1, \ldots, e_m\}$, we define $\mathrm{df}_{\mathrm{R}}^{S}(\hat{f}) = d$ to be the number d for which

$$\mathsf{Opt}_{\mathrm{R}}(\hat{f}(\cdot; \widetilde{X}^S, \widetilde{Y}^S)) = \overline{\mathsf{Opt}}_{\mathrm{R}}(\hat{f}_d^{\mathrm{rf}}).$$

  On the right-hand side is the least squares reference model, as usual, trained on well-specified data (with Gaussian features). On the left-hand side is the given predictor, trained on a modified data set $(\widetilde{X}^S, \widetilde{Y}^S)$ in which the sources of error indexed by S have been isolated (i.e., they are activated, and nothing else is present).

- We can then attribute d_i degrees of freedom to each error component e_i, as follows:

$$d_i = \sum_{S \subseteq \{e_1,\ldots,e_m\} \setminus \{e_i\}} \frac{|S|!(m - |S| - 1)!}{m!} (\mathrm{df}_{\mathrm{R}}^{S \cup \{e_i\}}(\hat{f}) - \mathrm{df}_{\mathrm{R}}^{S}(\hat{f})).$$

- This is an instance of a Shapley value. Therefore it obeys all of the Shapley axioms; in particular, efficiency:

$$\sum_{i=1}^{m} d_i = \mathrm{df}_{\mathrm{R}}(\hat{f}).$$

References

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

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

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

Leo Breiman. The little bootstrap and other methods for dimensionality selection in regression: X-fixed prediction error. *Journal of the American Statistical Association*, 87(419):738–754, 1992.

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.

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.

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.

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.

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.

H. Malcolm Hudson. A natural identity for exponential families with applications in multiparameter estimation. *Annals of Statistics*, 6(3):473–484, 1978.

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.

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

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.

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

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.

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

Jianming Ye. On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, 93(441):120–131, 1998.

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.