Revisiting model complexity in the wake of overparameterized learning

1 Motivation and summary of contributions

Modern machine learning involves fitting a large number of parameters relative to the number of observations. Such overparameterized models are typically trained to (nearly) interpolate noisy in-sample data, and yet generalize reasonable well on out-of-sample data in many settings [27]. A series of recent work has investigated this surprising phenomenon for different models, including linear regression [3, 11, 21, 1], random features regression [19], sparse regression [16], kernel regression [17], linear classification [7, 20], boosting [18], among several others; see [2, 6] for more examples. An interesting feature of overparameterized models is the so-called "double descent" (or even "multiple descent") behavior in the generalization error curve when plotted against the raw number of model parameters or some analogous notion of model complexity. This leads us to ask the following motivating questions in this paper: (1) Is there a better and more principled measure of model complexity in general for overparameterized models? (2) More specifically, how do we compare complexity of different (near) interpolating models? We address these questions through the lens of degrees of freedom, by borrowing and extending classical ideas from optimism theory. In particular, we propose two measures of model complexity, namely *emergent and intrinsic random-X degrees of freedom*. We show the utility of our proposed complexity measures through examples of linear smoothers and interpolators, and illustrate how our proposed measures may help "reconcile" the surprising "multiple descent" generalization behaviors in modern machine learning with the "single descent" bias-variance tradeoff in classical statistical learning. In what follows, we fist summarize our proposals in Section 2, and then provide illustrative examples in Section 3.

2 New proposal for random-X degrees of freedom

Consider the standard regression setup with i.i.d. observations $(x_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, $i = 1, \dots, n$, such that $y_i = f(x_i) + \varepsilon_i$, where $f : \mathbb{R}^p \to \mathbb{R}$ is the regression function, and ε_i has mean 0 and variance σ^2. Denote by $X \in \mathbb{R}^{n \times p}$ the corresponding feature matrix and by $y \in \mathbb{R}^n$ the associated response vector. Let $\mathcal{A}$ be any fitting algorithm that maps $(X, y) \overset{\mathcal{A}}{\mapsto} \widehat{f}$, where $\widehat{f} : \mathbb{R}^p \to \mathbb{R}$ is the resulting fitted predictor. Associated with $\widehat{f}$ are three error metrics: (a) the training error, $\mathrm{ErrT}(\widehat{f}) = n^{-1} \sum_{i=1}^n (y_i - \widehat{f}(x_i))^2$, (b) the fixed-X prediction error, $\mathrm{ErrF}(\widehat{f}) = n^{-1} \sum_{i=1}^n \mathbb{E}[(\widetilde{y}_i - \widehat{f}(x_i))^2 | X, y]$, where $\widetilde{y} \in \mathbb{R}^n$ is an independent copy of y at the training points X, and (c) the random-X prediction error, $\mathrm{ErrR}(\widehat{f}) = \mathbb{E}[(y_0 - \widehat{f}(x_0))^2 | X, y]$, where (x_0, y_0) is a test observation sampled independently from the same distribution as the training data.

The training error underestimates both the fixed-X and random-X prediction error in general. In classical statistics, such downward bias is referred to as training *optimism* [13]. Define the fixed-X optimism, $\mathrm{OptF}(\widehat{f}) = \mathbb{E}[\mathrm{ErrF}(\widehat{f}) - \mathrm{ErrT}(\widehat{f}) | X]$, and the random-X optimism, $\mathrm{OptR}(\widehat{f}) = \mathbb{E}[\mathrm{ErrR}(\widehat{f}) - \mathrm{ErrT}(\widehat{f}) | X]$ [22]. The fixed-X optimism has been studied extensively and leads to the definition of *fixed-X degrees of freedom* as $\mathrm{DofF}(\widehat{f}) = \sum_{i=1}^n \mathrm{Cov}(y_i, \widehat{f}(x_i) | X)$ [8, 9, 12, 10], which under certain regularity conditions, is the same as as $\mathrm{DofF}(\widehat{f}) = \sum_{i=1}^n \mathbb{E}[\partial \widehat{f}(x_i) / \partial y_i | X]$ [26, 23]. In some cases, $\mathrm{DofF}(\widehat{f})$ can be computed explicitly: e.g., for linear smoothers $\widehat{f}(X) = L(X) y$, it is given by $\mathrm{tr}[L(X)]$ [4, 5]; for lasso, it is given by the expected number of non-zero coefficients in the fitted estimator [28, 25]; see [15, 14, 24] for various other generalizations. In classical statistics, DofF is a widely agreed-upon qualitative measure of complexity and is algorithm-specific, however it is only defined for the fixed-X setup. Despite 50+ years of work on DofF, there is no notion of random-X degrees of freedom that we know of. The goal of this paper is to propose a definition for random-X degrees of freedom, denoted by DofR, suitable for the random-X setup underlying most predictive problems.

Towards defining DofR, we first cast the classical definition of the fixed-X degrees of freedom from a different perspective. For a fitting procedure $\widehat{f} = \mathcal{A}(X, y)$, $\mathrm{DofF}(\widehat{f})$ can be shown to be equal to the value of k that satisfy the following relation: $\mathrm{OptF}(\mathcal{A}(X, y)) = \mathrm{OptF}(\mathcal{A}^{\mathrm{ref}}(U_{n \times k}, v))$, where $\mathcal{A}^{\mathrm{ref}}$ is the least squares reference algorithm, and $U_k \in \mathbb{R}^{n \times k}$ is a certain design matrix consisting n observations and $k \le n$ features, and $v \in \mathbb{R}^n$ is a noise vector with mean 0_n and covariance I_n (see Theorem 1 for more details). We then extend the same analogy and use the least squares as the reference algorithm and "match" random-X optimisms. We thus define the random-X degrees of freedom, $\mathrm{DofR}(\widehat{f})$, of any predictor $\widehat{f} = \mathcal{A}(X, y)$, as the value of k (we can show that such k always exists and is unique assuming $k \le n$; see the remarks after Theorem 1) for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, y)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)). \tag{DofR, emergent}$$

This measure, $\mathrm{DofR}(\widehat{f})$, depends of both the the predictor $\widehat{f}$ and the underlying regression function f. We call it *emergent random-X degrees of freedom*. We also define *intrinsic random-X degrees of freedom*, denoted by DofR^i, as the k (which again exists and is unique assuming $k \leq n$) for which the following relation holds:

$$\mathrm{OptR}(\mathcal{A}(X, v)) = \mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v)). \qquad \text{(DofR, intrinsic)}$$

Apart from analogy with fixed-X degrees of freedom, another reason for choosing the least squares reference algorithm to match optimisms is the following invariance property of OptR that we can show for least squares:

Theorem 1. *Let $U_k = Z_k \Sigma_k^{1/2}$, where Z_k contains i.i.d. entries of mean 0, variance 1, and bounded moment of order $4 + \mu$ for some $\mu > 0$ and $\Sigma_{k \times k}$ is a positive definite matrix whose minimum and maximum eigenvalues are uniformly bounded away from 0 and ∞. Let v contain i.i.d. entries of mean 0, variance σ^2, and bounded moment of order $4 + \nu$ for some $\nu > 0$. Denote the normalized random-X optimisms of $\widehat{f}$ by $\phi := \mathrm{OptR}(\mathcal{A}(X, y))/\sigma^2$ and $\psi := \mathrm{OptR}(\mathcal{A}(X, v))/\sigma^2$ Then, as $n, k \to \infty$ and $k/n \to \xi \in (0, 1)$, we have*

$$\frac{\mathrm{OptR}(\mathcal{A}^{\mathrm{ref}}(U_k, v))}{\sigma^2} \to \frac{1 - (1-\xi)^2}{1 - \xi}, \quad \mathrm{DofR}(\widehat{f}) \to 1 + \frac{\phi}{2} - \sqrt{1 + \frac{\phi^2}{4}}, \quad \mathrm{DofR}^i(\widehat{f}) \to 1 + \frac{\psi}{2} - \sqrt{1 + \frac{\psi^2}{4}}.$$

Remarks: There is remarkable universality in above limits: (1) They do not depend on the exact form of the distributions of U_k and v. (2) They are also independent of Σ_k. This further justifies the choice of the least squares reference algorithm for matching random-X optimisms. We can show an immediate interesting property of the random-X degrees of freedom: There is a unique number that satisfies the desired relations between $[0, n]$. We find this to be a very interpretable range for random-X degrees of freedom. The least complex predictor has DofR of 0, and the most complex predictor has DofR of n, as if the saturated model.

3 Explicit and numerical illustrative examples

In general, the random-X degrees of freedom depend of the exact form of the algorithm, but as with DofF, for linear smoothers, DofR^i takes a special interpretable form. It also shows how DofR^i is related to DofF.

Proposition 2. *Suppose $\widehat{f}$ is a linear smoother such that $\widehat{f}(X) = L(X)y$ and $\widehat{f}(x_0) = \ell(x_0)^\top y$ for some smoothing matrix $L \in \mathbb{R}^{n \times n}$ and smoothing weight function $\ell : \mathbb{R}^p \to \mathbb{R}^n$. Then, we have $\mathrm{DofR}^i(\widehat{f}) = \mathrm{tr}[L(X)] + n/2(\mathbb{E}[\ell(x_0)^\top \ell(x_0)] - \mathrm{tr}[L(X)^\top L(X)]/n) = \mathrm{DofF}(\widehat{f}) + n/2(\mathbb{E}[\ell(x_0)^\top \ell(x_0)] - \mathrm{tr}[L(X)^\top L(X)]/n)$.*

Remarks: Some special cases of interest are: (1) Interpolating models for which $L(X) = I_n$. In this case, DofR^i simplifies to $n/2 + n/2\mathbb{E}[\ell(x_0)^\top \ell(x_0)]$. Note that this number differs between different interpolating models as opposed to DofF which is always $\mathrm{tr}[L(X)] = n$ for any interpolating model. (2) In the special case of min ℓ_2-norm interpolator, we can prove the following interesting property: in the underparameterized regime when $p \leq n$, we have DofR^i/n strictly increasing from $[0, 1]$ as expected, while in the overparameterized regime when $p > n$, DofR^i/n is strictly decreasing from $(1, 0)$, so DofR^i is maximized at $p = n$. This result holds for any feature covariance Σ and shows overparameterization indeed reduces the intrinsic complexity.

Beyond linear smoothers, properties of DofR and DofR^i depend on the specific fitting procedure. Below we compare min ℓ_2-norm interpolator with min ℓ_1-norm interpolator, abbreviated mn2ls and mn1ls, whose risks are recently shown to exhibit double [11] and multiple descents [16], respectively. Note that latter is a non-linear procedure. We observe from Figure 1 that our proposed notion of intrinsic degrees-of-freedom reconciles the "bias-variance" tradeoff and turns modern "double descents" into classical "single descents".

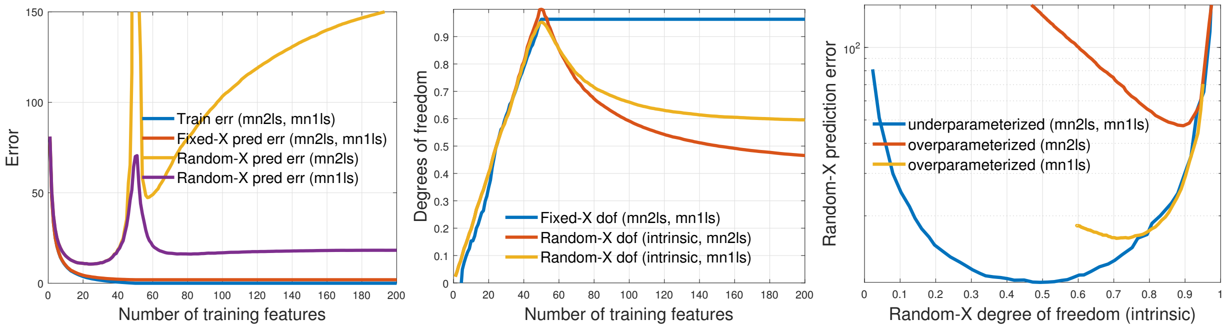

Figure 1: We consider a fixed data generating model with $n = 200$ and response non-linear in $p = 200$ feature, and consider training estimators with varying number of features. This model is similar to that used in [3].

References

[1] P. L. Bartlett, P. M. Long, G. Lugosi, and A. Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[2] P. L. Bartlett, A. Montanari, and A. Rakhlin. Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*, 2021.

[3] M. Belkin, D. Hsu, S. Ma, and S. 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.

[4] P. Craven and G. Wahba. Smoothing noisy data with spline functions. *Numerische mathematik*, 31(4):377–403, 1978.

[5] P. Craven and G. Wahba. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1979.

[6] Y. Dar, V. Muthukumar, and R. G. Baraniuk. A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning. *arXiv preprint arXiv:2109.02355*, 2021.

[7] Z. Deng, A. Kammoun, and C. Thrampoulidis. A model of double descent for high-dimensional binary linear classification. *arXiv preprint arXiv:1911.05822*, 2019.

[8] B. Efron. Estimating the error rate of a prediction rule: improvement on cross-validation. *Journal of the American statistical association*, 78(382):316–331, 1983.

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

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

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

[12] T. Hastie and R. Tibshirani. *Generalized Additive Models.* Chapman & Hall, 1990.

[13] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning.* Springer Series in Statistics, 2009. Second edition.

[14] L. Janson, W. Fithian, and T. J. Hastie. Effective degrees of freedom: a flawed metaphor. *Biometrika*, 102(2):479–485, 2015.

[15] S. Kaufman and S. Rosset. When does more regularization imply fewer degrees of freedom? sufficient conditions and counterexamples. *Biometrika*, 101(4):771–784, 2014.

[16] Y. Li and Y. Wei. Minimum ℓ_1-norm interpolators: Precise asymptotics and multiple descent. *arXiv preprint arXiv:2110.09502*, 2021.

[17] T. Liang and A. Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

[18] T. Liang and P. Sur. A precise high-dimensional asymptotic theory for boosting and minimum-l1-norm interpolated classifiers. *arXiv preprint arXiv:2002.01586*, 2020.

[19] S. Mei and A. Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.

[20] A. Montanari, F. Ruan, Y. Sohn, and J. Yan. The generalization error of max-margin linear classifiers: High-dimensional asymptotics in the overparametrized regime. *arXiv preprint arXiv:1911.01544*, 2019.

[21] V. Muthukumar, K. Vodrahalli, V. Subramanian, and A. Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1):67–83, 2020.

[22] S. Rosset and R. J. Tibshirani. From fixed-x to random-x regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, 2019.

[23] C. M. Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

[24] R. J. Tibshirani. Degrees of freedom and model search. *Statistica Sinica*, pages 1265–1296, 2015.

[25] R. J. Tibshirani and J. Taylor. Degrees of freedom in lasso problems. *The Annals of Statistics*, 40(2):1198–1232, 2012.

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

[27] C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.

[28] H. Zou, T. Hastie, and R. Tibshirani. On the “degrees of freedom” of the lasso. *The Annals of Statistics*, 35(5):2173–2192, 2007.