The Implicit Regularization of Ordinary Least Squares Ensembles

Daniel LeJeune
Rice University

Hamid Javadi
Rice University

Richard G. Baraniuk
Rice University

Abstract

Ensemble methods that average over a collection of independent predictors that are each limited to a subsampling of both the examples and features of the training data command a significant presence in machine learning, such as the ever-popular random forest, yet the nature of the subsampling effect, particularly of the features, is not well understood. We study the case of an *ensemble of linear predictors*, where each individual predictor is fit using ordinary least squares on a random submatrix of the data matrix. We show that, under standard Gaussianity assumptions, when the number of features selected for each predictor is optimally tuned, the asymptotic risk of a large ensemble is equal to the asymptotic *ridge regression* risk, which is known to be optimal among linear predictors in this setting. In addition to eliciting this *implicit regularization* that results from subsampling, we also connect this ensemble to the dropout technique used in training deep (neural) networks, another strategy that has been shown to have a ridge-like regularizing effect.

1 INTRODUCTION

Ensemble methods (Breiman, 1996; Amit and Geman, 1997; Josse and Wager, 2016) are an oft-used strategy employed successfully in a broad range of problems in machine learning and statistics, in which one combines a number of *weak predictors* together to obtain one powerful predictor. This is accomplished by giving each weak learner a different *view* of the training data. Various strategies for changing this training data view exist, among which many are simple sampling-based techniques in which each predictor is (independently) given access to a subsampling of the rows (examples) and columns (features) of the training data matrix, such as *bagging* (Breiman, 1996; Bühlmann and Yu, 2002). Another noteworthy technique is *boosting* (Freund and Schapire, 1997; Breiman, 1998), in which the training data examples are reweighted adaptively according to how badly they have been misclassified while building the ensemble. In this work, we consider the former class of techniques—those that train each weak predictor using an independent subsampling of the training data.

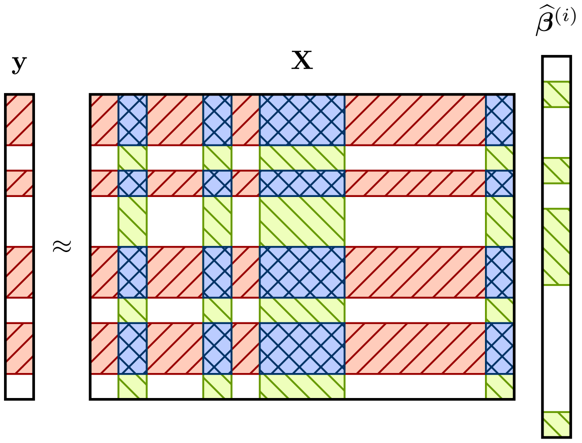

Figure 1: Example (rows) and feature (columns) subsampling of the training data used in the ordinary least squares fit for one member of the ensemble. The i-th member of the ensemble is only allowed to predict using its subset of the features (green). It must learn its parameters $\widehat{\boldsymbol{\beta}}^{(i)}$ by performing ordinary least squares using the subsampled examples of $\mathbf{y}$ (red) and the subsampled examples (rows) and features (columns) of the data matrix $\mathbf{X}$ (blue, crosshatched).

Ensemble methods based on independent example and feature subsampling are attractive for two reasons. First, they are computationally appealing in that they are massively parallelizable, and since each member

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. Copyright 2020 by the author(s).

of the ensemble uses only part of the data, they are able to overcome memory limitations faced by other methods (Louppe and Geurts, 2012). Second, ensemble methods are known to achieve lower risk due to the fact that combining several different predictors reduces variance (Bühlmann and Yu, 2002; Wager et al., 2014; Scornet et al., 2015), and empirically they have been found to perform very well. *Random forests* (Breiman, 2001; Athey et al., 2019; Friedberg et al., 2018), for example, ensemble methods that combine example and feature subsampling with decision trees by choosing the most useful feature from a random subset of the features at each branch of the tree, remain among the best-performing off-the-shelf machine learning methods available (Cutler and Zhao, 2001; Fernández-Delgado et al., 2014; Wyner et al., 2017).

Let $\mathbf{X} \in \mathbb{R}^{n\times p}$ be the training data matrix consisting of n examples of data points each having p features. While there exist theoretical results on the benefits of *example (row) subsampling* (Bühlmann and Yu, 2002), the exact nature of the effect of *feature (column) subsampling* on ensemble performance remains poorly understood. In this paper, we study the prototypical form of this problem in the context of linear regression. That is, given the data matrix $\mathbf{X}$ and target variables $\mathbf{y} \in \mathbb{R}^n$, we study the ensemble $\widehat{\boldsymbol{\beta}}^{\text{ens}} = \frac{1}{k}\sum_{i=1}^{k} \widehat{\boldsymbol{\beta}}^{(i)}$, where each $\widehat{\boldsymbol{\beta}}^{(i)}$ is learned using ordinary least squares on an independent random subsampling of both the examples and features of the training data. This subsampling is illustrated in Figure 1. We show that under such a scheme, the resulting predictor of this ensemble performs as well as the *ridge regression* (Hoerl and Kennard, 1970; Friedman et al., 2001) predictor fit using the entire training data, which is known to be the optimal linear predictor under the data assumptions that we consider. Further, the asymptotic risk of the ensemble depends *only on the amount of feature subsampling* and not on the amount of example subsampling, provided each individual ordinary least squares problem is underdetermined. Our main result in Theorem 3.6 can be summarized as follows:

Theorem 3.6 (informal statement)**.** *When the features and underlying model weights both follow i.i.d. Gaussian distributions, the optimal asymptotic risk for an ensemble of ordinary least squares predictors is equal to the optimal asymptotic ridge regression risk.*

We can interpret this result as an example of *implicit regularization* (Mahoney, 2012; Neyshabur et al., 2014; Gunasekar et al., 2017; Arora et al., 2019). That is, while the individual ordinary least squares subproblems are completely unregularized, the ensemble behaves as if it had been regularized using a ridge regression penalty. Recently, there has been much interest in investigating the implicit regularization effects of commonly used heuristic methods, particularly in cases where they enable the training of highly *overparameterized* models that generalize well to test data despite having the capacity to overfit the training data (Zhang et al., 2017; Belkin et al., 2018). Examples of heuristic techniques that have been shown to have implicit regularization effects include stochastic gradient descent (Hardt et al., 2016) and *dropout* (Srivastava et al., 2014). Incidentally, we show a strong connection between the ensemble of ordinary least squares predictors and dropout, which is known to have a ridge-like regularizing effect (Wager et al., 2013), and we make this link via stochastic gradient descent.

Contributions We summarize our contributions as follows: **[C1]** We prove that when the amount of feature subsampling is optimized to minimize risk, an ensemble of ordinary least squares predictors achieves the same risk as the optimal ridge regression predictor asymptotically as $n, p \to \infty$ (see Section 3). **[C2]** We demonstrate the converge of the ensemble risk to the optimal ridge regression risk via simulation (see Section 4.1). **[C3]** We reveal a connection between the ordinary least squares ensemble and the popular *dropout* technique used in deep (neural) network training (see Section 4.3) and from the insight gained from this connection develop a recipe for mitigating excess risk under suboptimal feature subsampling via simple output scaling (see Section 4.4).

2 ENSEMBLES OF ORDINARY LEAST SQUARES PREDICTORS

We consider the familiar setting of *linear regression*, where there exists a linear relationship between the target variable $y \in \mathbb{R}$ and the feature variables $\mathbf{x} \in \mathbb{R}^p$—i.e., $y = \langle \mathbf{x}, \boldsymbol{\beta} \rangle$, where $\boldsymbol{\beta} \in \mathbb{R}^p$ is the model parameter vector. The goal of a machine learning algorithm is to estimate these parameters given n i.i.d. noisy samples $\left\{\mathbf{x}^{(i)}, y^{(i)}\right\}_{i=1}^{n}$. The noise relationship is given by

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \sigma\mathbf{z}, \tag{1}$$

where $[\mathbf{X}]_{ij} = [\mathbf{x}^{(i)}]_j$, $[\mathbf{y}]_i = y^{(i)}$, and $[\mathbf{z}]_i = z^{(i)}$, where $z^{(i)}$ are i.i.d. zero-mean random variables with unit variance independent of $\mathbf{X}$. We assume a Gaussian $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$ distribution on $\mathbf{x}$, and for the results in this paper, we assume $\boldsymbol{\Sigma} = \mathbf{I}_p$.

Our ensemble consists of k linear predictors each fit using ordinary least squares on a submatrix of $\mathbf{X}$, and the resulting prediction is the average of the outputs. Equivalently, our ensemble is defined by its estimate of the parameters

$$\widehat{\boldsymbol{\beta}}^{\text{ens}} \triangleq \frac{1}{k} \sum_{i=1}^{k} \widehat{\boldsymbol{\beta}}^{(i)}, \tag{2}$$

where $\widehat{\boldsymbol{\beta}}^{(i)}$ is the parameter estimate of the i-th member of the ensemble. To characterize the estimates $\widehat{\boldsymbol{\beta}}^{(i)}$, we first introduce some notation. Let the *selection matrix* $\mathbf{S}$ corresponding to a subset of indices $S \subseteq [p]$, where $[p] = \{1, \ldots, p\}$, denote the the $p \times |S|$ matrix obtained by selecting from $\mathbf{I}_p$ the columns corresponding to the indices in S, where $\mathbf{I}_p$ denotes the $p \times p$ identity matrix. With this definition of selection matrices, for $S \subseteq [p]$ and $T \subseteq [n]$, we have that $\mathbf{T}^\top \mathbf{X} \mathbf{S}$ is the matrix of size $|T| \times |S|$ obtained from $\mathbf{X}$ by selecting (subsampling) the rows and columns indicated by sets T and S. Returning to the ensemble, let $\mathcal{S} \triangleq (S_i)_{i=1}^k$ and $\mathcal{T} \triangleq (T_i)_{i=1}^k$ denote the collection of *feature subsets* and *example subsets*, respectively, where each $S_i \subseteq [p]$ and each $T_i \subseteq [n]$. Then, assuming $|S_i| < |T_i|$, for each member of the ensemble we let

$$\widehat{\boldsymbol{\beta}}^{(i)}_{S_i} = \underset{\boldsymbol{\beta}'}{\arg\min} \left\| \mathbf{T}_i^\top \left(\mathbf{X} \mathbf{S}_i \boldsymbol{\beta}' - \mathbf{y} \right) \right\|_2, \tag{3}$$

$$\widehat{\boldsymbol{\beta}}^{(i)}_{S_i^c} = \mathbf{0}, \tag{4}$$

where $S_i^c = [p] \setminus S_i$ denotes the complement of the set S_i. This can alternatively be written in closed form as

$$\widehat{\boldsymbol{\beta}}^{(i)} = \mathbf{S}_i \left(\mathbf{T}_i^\top \mathbf{X} \mathbf{S}_i \right)^\dagger \mathbf{T}_i^\top \mathbf{y}, \tag{5}$$

where $(\cdot)^\dagger$ denotes the Moore–Penrose pseudoinverse. Thus, the closed-form expression for the ensemble parameter estimate is given by

$$\widehat{\boldsymbol{\beta}}^{\text{ens}} = \frac{1}{k} \sum_{i=1}^{k} \mathbf{S}_i \left(\mathbf{T}_i^\top \mathbf{X} \mathbf{S}_i \right)^\dagger \mathbf{T}_i^\top \mathbf{y}. \tag{6}$$

3 ENSEMBLE RISK

We define the *risk* of a linear predictor as the expected squared error of a prediction of the target variable on an independent data point $\mathbf{x}$:

$$\begin{aligned} R(\boldsymbol{\beta}') &\triangleq \mathbb{E}_{\mathbf{x}} \left[\langle \mathbf{x}, \boldsymbol{\beta} - \boldsymbol{\beta}' \rangle^2 \right] \\ &= \langle \boldsymbol{\beta} - \boldsymbol{\beta}', \boldsymbol{\Sigma} (\boldsymbol{\beta} - \boldsymbol{\beta}') \rangle . \end{aligned} \tag{7}$$

For any predictor of the form $\boldsymbol{\beta}' = f(\mathbf{X})\mathbf{y}$, for some $f : \mathbb{R}^{n \times p} \to \mathbb{R}^{p \times n}$, we can rewrite parameter estimation error as

$$\boldsymbol{\beta} - \boldsymbol{\beta}' = (\mathbf{I}_p - f(\mathbf{X})\mathbf{X})\boldsymbol{\beta} - \sigma f(\mathbf{X})\mathbf{z}. \tag{8}$$

Then by the independence of $\mathbf{X}$ and $\mathbf{z}$ and some algebra, we can decompose the risk into the so-called "bias" and "variance" components

$$\begin{aligned} \mathbb{E}_{\mathbf{z}} \left[R(\boldsymbol{\beta}') \right] = {} & \underbrace{\left\langle \boldsymbol{\beta}\boldsymbol{\beta}^\top, (\mathbf{I}_p - f(\mathbf{X})\mathbf{X})^\top \boldsymbol{\Sigma} (\mathbf{I}_p - f(\mathbf{X})\mathbf{X}) \right\rangle}_{\text{bias}(\boldsymbol{\beta}')} \\ & + \underbrace{\sigma^2 \left\langle f(\mathbf{X}), \boldsymbol{\Sigma} f(\mathbf{X}) \right\rangle}_{\text{variance}(\boldsymbol{\beta}')} . \end{aligned} \tag{9}$$

For the ensemble, we obtain for the bias and variance

$$\text{bias}(\widehat{\boldsymbol{\beta}}^{\text{ens}}) = \frac{1}{k^2} \sum_{i,j=1}^{k} \text{bias}_{ij}(\widehat{\boldsymbol{\beta}}^{\text{ens}}) \tag{10}$$

$$\text{variance}(\widehat{\boldsymbol{\beta}}^{\text{ens}}) = \frac{1}{k^2} \sum_{i,j=1}^{k} \text{variance}_{ij}(\widehat{\boldsymbol{\beta}}^{\text{ens}}), \tag{11}$$

where

$$\begin{aligned} \text{bias}_{ij}(\widehat{\boldsymbol{\beta}}^{\text{ens}}) = {} & \Big\langle \boldsymbol{\beta}\boldsymbol{\beta}^\top, \Big(\mathbf{I}_p - \mathbf{S}_i \left(\mathbf{T}_i^\top \mathbf{X} \mathbf{S}_i \right)^\dagger \mathbf{T}_i^\top \mathbf{X} \Big)^\top \\ & \times \boldsymbol{\Sigma} \Big(\mathbf{I}_p - \mathbf{S}_j \left(\mathbf{T}_j^\top \mathbf{X} \mathbf{S}_j \right)^\dagger \mathbf{T}_j^\top \mathbf{X} \Big) \Big\rangle, \end{aligned} \tag{12}$$

$$\begin{aligned} \text{variance}_{ij}(\widehat{\boldsymbol{\beta}}^{\text{ens}}) = \sigma^2 \Big\langle & \mathbf{S}_i \left(\mathbf{T}_i^\top \mathbf{X} \mathbf{S}_i \right)^\dagger \mathbf{T}_i^\top, \\ & \boldsymbol{\Sigma} \mathbf{S}_j \left(\mathbf{T}_j^\top \mathbf{X} \mathbf{S}_j \right)^\dagger \mathbf{T}_j^\top \Big\rangle. \end{aligned} \tag{13}$$

Thus, evaluating the risk of the ensemble is a matter of evaluating these pairwise interaction terms.

To begin evaluating the above terms, we need to introduce additional assumptions. Specifically, we assume that the subsets are independent and that all indices are equally likely to be included in each subset.

Assumption 3.1 (finite subsampling)**.** The subsets in the collections $\mathcal{S}$ and $\mathcal{T}$ are selected at random such that $|S_i| < |T_i| - 1$ and that the following hold:

- $\Pr(j \in S_i) = |S_i|/p$ for all $j \in [p]$,
- $\Pr(m \in T_i) = |T_i|/n$ for all $m \in [n]$,
- The subsets $S_1, S_2, \ldots, S_k, T_1, T_2, \ldots, T_k$ are conditionally independent given the example subset sizes $(|T_i|)_{i=1}^k$.

A simple sampling strategy that satisfies these assumptions is to fix $|S_i|$ and $|T_i|$ such that $|S_i| < |T_i| - 1$ and select subsets uniformly at random of the given sizes. Another strategy is to construct the subsets by flipping a coin for each index, rejecting any resulting subsets that fail to satisfy $|S_i| < |T_i| - 1$.

With Assumption 3.1, we are now equipped to evaluate the pairwise interaction terms. The following two lemmas enable us to characterize the bias and variance components of the risk in the finite-dimensional setting. The proofs of these lemmas are exercises in linear algebra and conditional expectations and can be found in the Appendix.

With some slight abuse of notation, we allow $\mathbb{E}_{\mathcal{S},\mathcal{T}}$ to denote the expectation taken with respect to the choice of indices in the subsets, but not their sizes. In other words, $\mathbb{E}_{\mathcal{S},\mathcal{T}}$ indicates the conditional expectation over $\mathcal{S}$ and $\mathcal{T}$, conditioned on the subset sizes indicated by the context.

Lemma 3.2 (bias)**.** *Assume that $\mathbf{\Sigma} = \mathbf{I}_p$ and that Assumption 3.1 holds. Then*

$$\mathbb{E}_{\mathbf{X},\mathcal{S},\mathcal{T}}\left[\text{bias}_{ij}(\widehat{\boldsymbol{\beta}}^{\text{ens}})\right] = \begin{cases} \frac{|S_i^c \cap S_j^c|}{p}\left(1 + \frac{|S_i \cap S_j|}{n - |S_i \cap S_j| - 1}\right)\|\boldsymbol{\beta}\|_2^2 & \text{if } i \neq j, \\ \frac{|S_i^c|}{p}\left(1 + \frac{|S_i|}{|T_i| - |S_i| - 1}\right)\|\boldsymbol{\beta}\|_2^2 & \text{if } i = j. \end{cases} \tag{14}$$

Lemma 3.3 (variance)**.** *Assume that $\mathbf{\Sigma} = \mathbf{I}_p$ and that Assumption 3.1 holds. Then*

$$\mathbb{E}_{\mathbf{X},\mathcal{S},\mathcal{T}}\left[\text{variance}_{ij}(\widehat{\boldsymbol{\beta}}^{\text{ens}})\right] = \begin{cases} \frac{\sigma^2 |S_i \cap S_j|}{n - |S_i \cap S_j| - 1} & \text{if } i \neq j, \\ \frac{\sigma^2 |S_i|}{|T_i| - |S_i| - 1} & \text{if } i = j. \end{cases} \tag{15}$$

One observation that we can make already from these results is that the example subsampling only affects the terms where $i = j$. Assuming that the subsampling procedure is the same for each i, so that for large k the $i \neq j$ terms are sure to dominate the sum, this means that in the limit as $k \to \infty$, the effects of example subsampling are non-existent. We note that this is a result of the assumption that $|S_i| < |T_i|$, and that if we were to have $|S_i| > |T_i|$, then we would observe effects of example subsampling when $i \neq j$, which we discuss further in Section 5.2.

We now turn our attention to the setting where $n, p \to \infty$ in order to better reason about the results contained in these lemmas. We introduce the following additional assumption.

Assumption 3.4 (asymptotic subsampling)**.** For some $\alpha, \eta \in [0, 1]$, the subsets in the collections $\mathcal{S}$ and $\mathcal{T}$ are selected randomly such that $|S_i|/p \xrightarrow{\text{a.s.}} \alpha$ as $p \to \infty$ and $|T_i|/n \xrightarrow{\text{a.s.}} \eta$ as $n \to \infty$ for all $i \in [k]$.

This assumption is easily satisfied. For example, in the sampling strategy where we fix $|S_i|$ and $|T_i|$, we can choose $|S_i| = \lfloor \alpha p \rfloor$ and $|T_i| = \lfloor \eta n \rfloor$. For the coin-flipping strategy, we can select feature subsets with a coin of probability α and example subsets with a coin of probability η.

Under this assumption, and additionally assuming without loss of generality that $\|\boldsymbol{\beta}\|_2 = 1$, if $n, p \to \infty$ such that $p/n \to \gamma$ and $\eta > \alpha\gamma$, the quantities in (14) and (15) converge almost surely as follows:

$$\mathbb{E}_{\mathbf{X},\mathcal{S},\mathcal{T}}\left[\text{bias}_{ij}(\widehat{\boldsymbol{\beta}}^{\text{ens}})\right] \xrightarrow{a.s.} \begin{cases} (1-\alpha)^2\left(1 + \frac{\alpha^2\gamma}{1 - \alpha^2\gamma}\right) & \text{if } i \neq j, \\ (1-\alpha)\left(1 + \frac{\alpha\gamma}{\eta - \alpha\gamma}\right) & \text{if } i = j, \end{cases} \tag{16}$$

and

$$\mathbb{E}_{\mathbf{X},\mathcal{S},\mathcal{T}}\left[\text{variance}_{ij}(\widehat{\boldsymbol{\beta}}^{\text{ens}})\right] \xrightarrow{a.s.} \begin{cases} \frac{\sigma^2\alpha^2\gamma}{1 - \alpha^2\gamma} & \text{if } i \neq j, \\ \frac{\sigma^2\alpha\gamma}{\eta - \alpha\gamma} & \text{if } i = j. \end{cases} \tag{17}$$

We are now equipped to state our asymptotic risk result for the ensemble of ordinary least squares predictors. Denote for an ensemble satisfying Assumptions 3.1 and 3.4 with parameters α, η, and k the *limiting risk*

$$R^{\text{ens}}_{\alpha,\eta,k} \triangleq \lim_{n,p\to\infty} \mathbb{E}_{\mathbf{X},\mathbf{z},\mathcal{S},\mathcal{T}}\left[R(\widehat{\boldsymbol{\beta}}^{\text{ens}})\right]. \tag{18}$$

From (10) and (11), we know that both the bias and variance components of the limiting risk are the averages of k^2 terms, and from (16) and (17), we know that the $k(k-1)$ terms where $i \neq j$ will take one value and the remaining k terms where $i = j$ will take another. Thus we have the *limiting bias*

$$\lim_{n,p\to\infty} \mathbb{E}_{\mathbf{X},\mathcal{S},\mathcal{T}}\left[\text{bias}(\widehat{\boldsymbol{\beta}}^{\text{ens}})\right] = \frac{k-1}{k}\left(\frac{(1-\alpha)^2}{1-\alpha^2\gamma}\right) + \frac{1}{k}\left(\frac{\eta(1-\alpha)}{\eta - \alpha\gamma}\right) \tag{19}$$

and *limiting variance*

$$\lim_{n,p\to\infty} \mathbb{E}_{\mathbf{X},\mathcal{S},\mathcal{T}}\left[\text{variance}(\widehat{\boldsymbol{\beta}}^{\text{ens}})\right] = \frac{k-1}{k}\left(\frac{\sigma^2\alpha^2\gamma}{1-\alpha^2\gamma}\right) + \frac{1}{k}\left(\frac{\sigma^2\alpha\gamma}{\eta - \alpha\gamma}\right). \tag{20}$$

Upon careful examination of these quantities, we observe that in fact both the limiting bias *and* the limiting variance are *decreasing* in k, and thus the ensemble serves not only as a means to reduce variance (as is well understood), *but also* to reduce bias. We defer further discussion to Section 4.2. Adding the limiting bias and variance together yields the following result.

Theorem 3.5 (limiting risk)**.** *Assume that $\mathbf{\Sigma} = \mathbf{I}_p$ and $\|\boldsymbol{\beta}\|_2 = 1$ and that Assumptions 3.1 and 3.4 hold. Then in the limit as $n, p \to \infty$ with $p/n \to \gamma$, for $\eta > \alpha\gamma$, we have almost surely that*

$$R^{\text{ens}}_{\alpha,\eta,k} = \frac{k-1}{k}\left(\frac{(1-\alpha)^2 + \sigma^2\alpha^2\gamma}{1-\alpha^2\gamma}\right) + \frac{1}{k}\left(\frac{\eta(1-\alpha) + \sigma^2\alpha\gamma}{\eta - \alpha\gamma}\right). \tag{21}$$

Here we see again more explicitly that for large k, the effect of example subsampling vanishes. This leaves us with the *large-ensemble* risk

$$\begin{aligned} R_\alpha^{\text{ens}} &\triangleq \lim_{k\to\infty} R_{\alpha,\eta,k}^{\text{ens}} \\ &= \frac{(1-\alpha)^2 + \sigma^2\alpha^2\gamma}{1-\alpha^2\gamma}. \end{aligned} \tag{22}$$

We note that while the large-ensemble risk depends only upon α, we cannot realize this risk with an ensemble if $\eta \le \alpha\gamma$. Our remaining results concern the large-ensemble risk and therefore assume that $\eta = 1$ for simplicity, but we caution the reader that some of these results may not be valid for some smaller values of η, depending on σ and γ.

Because α is an algorithmic hyperparameter, it can be tuned to minimize the risk. If we do so, then what we obtain is the perhaps surprising result that the optimal large-ensemble risk of the ordinary least squares predictor is equal to the limiting risk of the *ridge regression* predictor under our assumptions. The ridge regression predictor with parameter λ is defined as

$$\begin{aligned} \widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}} &\triangleq \arg\min_{\boldsymbol{\beta}'} \|\mathbf{X}\boldsymbol{\beta}' - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}'\|_2^2 \\ &= \left(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_p\right)^{-1}\mathbf{X}^\top\mathbf{y}. \end{aligned} \tag{23}$$

We formally state this result in the following theorem, which leverages the recent analysis of the limiting risk of ridge regression by Dobriban and Wager (2018).[1] The proof is found in the Appendix.

Theorem 3.6. *Assume that $\boldsymbol{\Sigma} = \mathbf{I}_p$ and $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, p^{-1}\mathbf{I}_p)$ and that Assumptions 3.1 and 3.4 hold with $\eta = 1$. Then in the limit as $n, p \to \infty$ with $p/n \to \gamma$, we have almost surely that*

$$\inf_{\alpha<\gamma^{-1}} R_\alpha^{\text{ens}} = \inf_\lambda R\left(\widehat{\boldsymbol{\beta}}_\lambda^{\text{ridge}}\right). \tag{24}$$

The implication of Theorem 3.6 is quite strong. Under the assumption of the theorem that true parameters $\boldsymbol{\beta}$ have a Gaussian distribution with covariance $p^{-1}\mathbf{I}_p$, the ridge regression predictor (the maximum a posteriori estimator for this setting) is the predictor with the lowest expected risk of all predictors of the form $\boldsymbol{\beta}' = f(\mathbf{X})\mathbf{y}$. To see this, note that if we take the expectation of (9) with respect to $\boldsymbol{\beta}$, we find that the optimal $f(\mathbf{X})$ must satisfy the first order optimality condition

$$\boldsymbol{\Sigma} f(\mathbf{X})(\mathbf{X}\mathbf{X}^\top + p\sigma^2\mathbf{I}_p) = \boldsymbol{\Sigma}\mathbf{X}^\top, \tag{25}$$

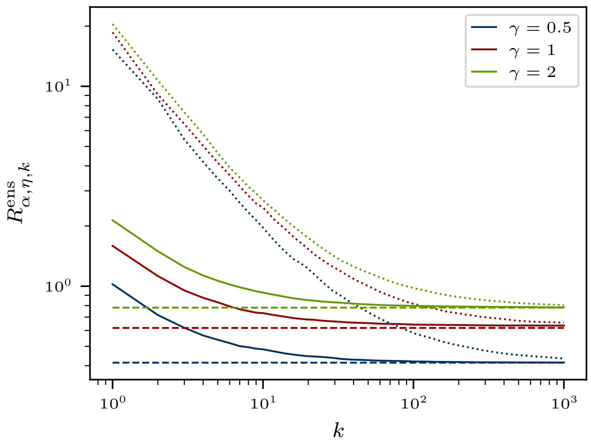

Figure 2: Approximate limiting risk (averaged over 50 trials with $n = 200, \sigma = 1$) when using $\eta = 1$ (solid) and $\eta = 1.1 \times \alpha\gamma$ (dotted). For each value of γ, both ensembles converge to the theoretical optimal ridge regression risk (dashed).

which for invertible $\boldsymbol{\Sigma}$ yields the optimally tuned ridge regression predictor. Thus, in the $\boldsymbol{\Sigma} = \mathbf{I}_p$ setting, the optimally tuned ensemble achieves the optimal risk for any linear predictor.

A curious result obtained during the proof of this theorem is the following corollary relating the optimal large ensemble risk to the optimal choice of the hyperparameter α.

Corollary 3.7. *Assume that $\boldsymbol{\Sigma} = \mathbf{I}_p$ and $\|\boldsymbol{\beta}\|_2 = 1$ and that Assumptions 3.1 and 3.4 hold with $\eta = 1$. Then in the limit as $n, p \to \infty$ with $p/n \to \gamma$, we have almost surely that*

$$R_{\alpha_*}^{\text{ens}} = 1 - \alpha_*, \tag{26}$$

where $\alpha_* = \arg\min_{\alpha<\gamma^{-1}} R_\alpha^{\text{ens}}$.

4 DISCUSSION

4.1 Convergence

In practice, any ensemble will have only a finite number of members. Therefore, it is important to understand the rates at which the risk of the ensemble converges to large-ensemble risk in (22). From Theorem 3.5, it is clear that as a function of k, the limiting risk converges to the large-ensemble risk at a rate $O(1/k)$. However, as the choice of η approaches $\alpha\gamma$, this rate becomes slower. In Figure 2, we plot[2] the convergence in k of the limiting risk to the large-ensemble risk for $\eta = 1$ (using all examples) and for

[1] We note that results on MMSE estimation error from the wireless communication community (see, e.g., Tulino and Verdú, 2004) predate the more general result of Dobriban and Wager (2018), and that these apply to the $\boldsymbol{\Sigma} = \mathbf{I}_p$ setting we consider, where risk is equal to estimation error.

[2] See `https://github.com/dlej/ensemble-ols`.

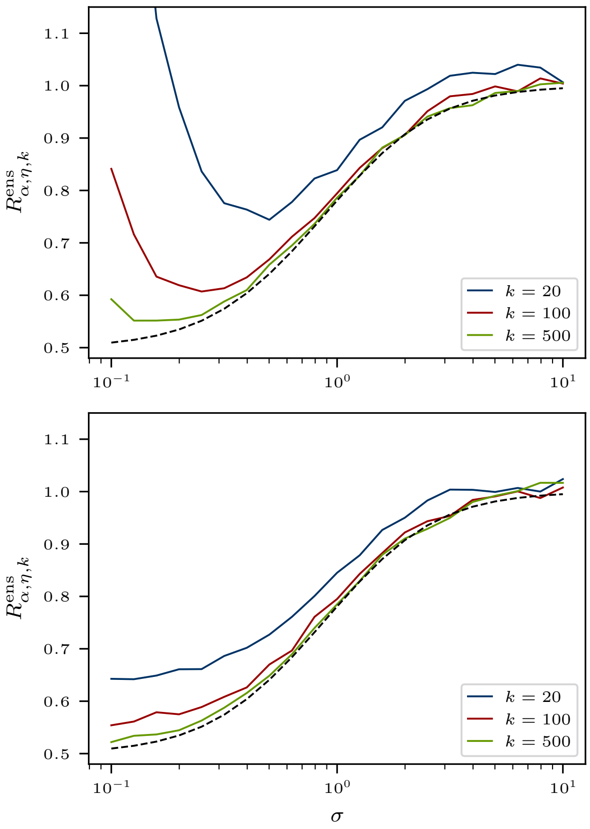

Figure 3: Approximate limiting risk (averaged over 100 trials with $n = 200, p = 400$) when using $\alpha = \alpha_*$ (top) and $\alpha = \arg\min_{\alpha'} R^{\text{ens}}_{\alpha',\eta,k}$ (bottom). As k increases, in both cases the risk converges to the theoretical optimal ridge regression risk (black dashed).

$\eta = 1.1 \times (\alpha\gamma)$ (near to as small as possible while still having $|S_i| < |T_i|$). We plot these curves for $\sigma = 1$ and for three different values of γ, using $n = 200$, which is sufficient to realize the convergence in n and p. We choose $\alpha = \alpha_*$, the minimizer of the large-ensemble risk. What we observe is that, indeed, for both choices of η, the risks converge to the optimal ridge risk. As expected, however, with the smaller choice of η the risk converges nearly an order of magnitude more slowly.

While the choice of $\alpha = \alpha_*$ will result in optimal risk for large enough ensembles, for finite k this choice can in some cases be undesirable. For instance, consider the setting where $\eta = 1$ and $\gamma > 1$. Then as $\sigma \to 0$, $\alpha_* \to \gamma^{-1}$ (see expressions for α_* in the Appendix). This obviously yields the optimal large-ensemble risk, by definition, but for any finite k, the limiting risk tends to infinity for this choice of α. However, if we know what the size of our ensemble will be, we can tune α to the limiting risk for finite k instead of the large ensemble risk. In general, this means choosing an α smaller than α_*. In Figure 3, we demonstrate the convergence in k to the large-ensemble risk as a function of σ for $\alpha = \alpha_*$ and for $\alpha = \arg\min_{\alpha'} R^{\text{ens}}_{\alpha',\eta,k}$. We plot these curves for $\gamma = 2$ and $\sigma \in [0.1, 10]$, using $n = 200$. While for both choices of α we see convergence in k for each σ, as $\sigma \to 0$, the risk is very large for $\alpha = \alpha_*$. For α adapted to the choice of k, however, this effect is mitigated.

4.2 Bias and Variance Decrease with Ensemble Size

We return here to the observation made in Section 3 that the limiting bias and variance are both *decreasing* in k. Thus, although there is a bias–variance tradeoff in α, there is no such tradeoff with k. This can be seen by comparing the $i = j$ and $i \neq j$ terms in each case. In the case of bias, for the bias to be decreasing, it must be that

$$\frac{(1-\alpha)^2}{1-\alpha^2\gamma} < \frac{\eta(1-\alpha)}{\eta-\alpha\gamma}. \tag{27}$$

Since $\alpha^2\gamma < 1$ and $\eta > \alpha\gamma$, after some algebra, this reduces to

$$\gamma(\alpha - 1) < \eta(1 - \alpha\gamma). \tag{28}$$

Because $\alpha \leq 1$, the left-hand side is non-positive, and since $\alpha < \gamma^{-1}$, the right-hand side is strictly positive. Thus this inequality always holds, and the bias is decreasing.

In the case of variance, for the variance to be decreasing, we must have

$$\frac{\alpha^2\gamma}{1-\alpha^2\gamma} < \frac{\alpha\gamma}{\eta-\alpha\gamma}. \tag{29}$$

Again since $\alpha^2\gamma < 1$ and $\eta > \alpha\gamma$, this reduces to

$$\alpha\eta < 1. \tag{30}$$

So, unless both $\alpha = 1$ and $\eta = 1$, in which case every member of the ensemble is the ordinary least squares predictor fit using the entire training data, the variance is decreasing.

4.3 Dropout and Ridge Regression

There is an interesting connection between the ordinary least squares ensemble with $\eta = 1$ and the popular *dropout* technique (Srivastava et al., 2014) used in deep (neural) network training, which consists of randomly masking the features at each iteration of (stochastic) gradient descent. To draw this connection, define

$$\ell_i(\boldsymbol{\beta}') = \left\| \mathbf{X}\mathbf{S}_i\mathbf{S}_i^\top \boldsymbol{\beta}' - \mathbf{y} \right\|_2^2. \tag{31}$$

Then our ensemble member parameter estimates are minimizers of this loss function.

$$\widehat{\boldsymbol{\beta}}^{(i)} = \arg\min_{\boldsymbol{\beta}'} \ell_i(\boldsymbol{\beta}') \text{ s.t. } \boldsymbol{\beta}'_{S_i^c} = \mathbf{0}. \tag{32}$$

For each i, the i-th member of the ensemble is able to solve its subproblem independently of the other members. As a result, we can consider the ensemble to be a model with $\sum_{i=1}^k |S_i|$ parameters that are eventually averaged to reduce them down to p parameters. If we were to instead constrain ourselves so that we were allowed to use *only* p parameters, such that we could not optimize each member of the ensemble independently, we might try to optimize them jointly by minimizing the average loss. That is,

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}'} \frac{1}{k} \sum_{i=1}^{k} \ell_i(\boldsymbol{\beta}'). \tag{33}$$

If we go a step further and let $k \to \infty$ and optimize this loss using stochastic gradient descent where at each iteration we use the gradient of an individual ℓ_i selected at random, then our ensemble becomes equivalent to the predictor learned using dropout. It is well-known that dropout with linear regression has a very strong connection to ridge regression (Srivastava et al., 2014); specifically, we find that

$$\widehat{\boldsymbol{\beta}} = \frac{1}{\alpha}\left(\mathbf{X}^\top \mathbf{X} + \frac{1-\alpha}{\alpha}\operatorname{diag}(\mathbf{X}^\top\mathbf{X})\right)^{-1} \mathbf{X}^\top \mathbf{y}. \tag{34}$$

In the case of $\boldsymbol{\Sigma} = \mathbf{I}_p$, $n^{-1}\operatorname{diag}(\mathbf{X}^\top\mathbf{X})$ will converge to $\mathbf{I}_p$ as $n, p \to \infty$, in which case dropout and ridge regression are equivalent up to a rescaling. We discuss the case where $\boldsymbol{\Sigma} \neq \mathbf{I}_p$ in Section 5.1.

4.4 Scaled Ensembles

Our ensemble combines the individual predictors by simple averaging. However, in light of the fact that dropout is only equivalent to ridge regression up to a rescaling of the output, it is worth considering the effect of using an equally-weighted linear combination but using different weights from $1/k$ in constructing the ensemble predictor. That is, we consider the risk of the μ-*scaled* predictor $\widehat{\boldsymbol{\beta}}^{\text{ens}}_\mu = (\mu/k)\sum_{i=1}^k \widehat{\boldsymbol{\beta}}^{(i)}$. A simple calculation, proved in the Appendix, shows that under the assumptions of Theorem 3.5 the large-ensemble risk of the μ-scaled predictor is given by

$$R^{\text{ens}}_{\alpha,\mu} = \mu^2 R^{\text{ens}}_\alpha + (1-\mu)^2 + 2\mu(1-\mu)(1-\alpha). \tag{35}$$

Hence, it is possible to minimize the risk of $\widehat{\boldsymbol{\beta}}^{\text{ens}}_\mu$ over the choice of parameter μ. This results in

$$\mu_* = \frac{\alpha}{R^{\text{ens}}_\alpha + 2\alpha - 1} \tag{36}$$

as the optimal choice for μ and

$$R^{\text{ens}}_{\alpha,\mu_*} = 1 - \frac{\alpha^2}{2\alpha - 1 + R^{\text{ens}}_\alpha} \tag{37}$$

as the achieved risk for the optimally-scaled ensemble. Note that as a result of Corollary 3.7, $R^{\text{ens}}_{\alpha_*} = 1 - \alpha_*$. Therefore, for ensembles with optimally-tuned $\alpha = \alpha_*$ we have $\mu_* = 1$, and any scaling in constructing the ensemble predictor will not further improve the achieved risk. However, it is easy to see that when $\alpha > \alpha_*$ (the ensemble members select *more* features than is optimal), $\mu_* < 1$, and the risk is improved by adding extra shrinkage to the ensemble predictor. Similarly, if $\alpha < \alpha_*$, (the ensemble members select *less* features than is optimal), $\mu_* > 1$, and the risk is improved by inflating the ensemble predictor. We illustrate the improvement in risk to be had in Figure 4, where we plot the risk with ($\mu = \mu_*$) and without ($\mu = 1$) optimal scaling for two choices of α—one where we always select half as many features as optimal ($\alpha = \alpha_*/2$), and one where we always use half of the available features ($\alpha = 1/2$).

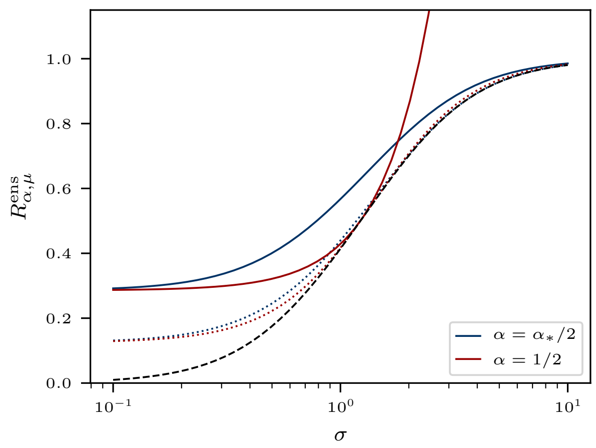

Figure 4: μ-scaled large-ensemble risk (theoretical, $\gamma = 0.5$) when using $\mu = 1$ (solid) and $\mu = \mu_*$ (dotted). For both the setting where we use fewer features than optimal with $\alpha = \alpha_*/2$ (blue) and the fixed $\alpha = 1/2$ setting (red), we see significantly improved risk by scaling.

5 FUTURE DIRECTIONS

5.1 Non-Identity Covariance

Of course, it is important to understand the behavior of the ordinary least squares ensemble in the case where $\boldsymbol{\Sigma} \neq \mathbf{I}_p$ when considering applications of the method to real data. As discussed in Section 3, provided $\mathbf{\Sigma}$ is invertible, ridge regression remains the optimal linear predictor, and whether the ensemble (or extensions thereto) still achieves the optimal risk in this setting remains an open question.

By inspection of the closed-form solution of dropout in (34), we see that it is no longer equivalent (as $n, p \to \infty$) to ridge regression in this setting and is therefore no longer optimal. We believe that this is likely the case for the ensemble as well. However, if we extend the coin-flipping strategy for feature subset selection to one where we have a collection of coin with probabilities $\boldsymbol{\alpha} \in [0,1]^p$, one for each feature, we can extend the result in (34) to obtain the closed-form dropout solution

$$\widehat{\boldsymbol{\beta}} = \mathbf{A}^{-1}\left(\mathbf{X}^\top\mathbf{X} + (\mathbf{I}_p - \mathbf{A})\,\mathbf{A}^{-1}\mathrm{diag}(\mathbf{X}^\top\mathbf{X})\right)^{-1}\mathbf{X}^\top\mathbf{y}, \tag{38}$$

where $\mathbf{A} = \mathrm{diag}(\boldsymbol{\alpha})$. We prove this result in the Appendix. Thus, if $\boldsymbol{\alpha}$ is chosen such that

$$\frac{1-\alpha_j}{\alpha_j} = \frac{\lambda}{n[\mathbf{\Sigma}]_{jj}}, \tag{39}$$

then the corrected dropout estimator

$$\widetilde{\boldsymbol{\beta}} = \mathbf{A}\widehat{\boldsymbol{\beta}} \tag{40}$$

is equivalent to ridge regression with parameter λ as $n, p \to \infty$. This leads us to believe that the optimal ensemble in the $\mathbf{\Sigma} \neq \mathbf{I}_p$ setting should also use non-uniform feature sampling, and extending our analysis to this case is an interesting area for future work.

5.2 Beyond Ordinary Least Squares: Ensembles of Interpolators

Throughout this work we have assumed that the members of the ensemble solve their subproblems using ordinary least squares, which yields the unique solution that minimizes the squared error given $|T_i|$ observations of $|S_i|$ variables, and this uniqueness requires that $|T_i|$ be no less than $|S_i|$. In the case where $|T_i| < |S_i|$, there are infinitely many solutions that minimize the squared error. However, we could in this case opt to *regularize* the solution to solve this problem. While analysis of the effect of regularizing the solution of the subproblems in the ensemble is beyond the scope of this work, we comment briefly on what would happen if we were to simply use the same solution presented in (5)—i.e., use the pseudoinverse solution, which has the smallest ℓ^2 norm of all solutions to the least squares problem. In this case, when $\eta = 1$, the learned predictor would be an *interpolator* (Belkin et al., 2018; Hastie et al., 2019) of the training data, and such methods have recently become increasingly of interest given the ability of deep (neural) network methods to have extremely good test performance while having (nearly) zero training error (Zhang et al., 2017; Belkin et al., 2019).

Specifically, it becomes immediately clear that in this setting, the effect of the choice of η does not vanish as $k \to \infty$. Lemma 3.3 can easily be extended to this setting, since the roles of S_i and T_i in (13) can simply be reversed, and as $n, p \to \infty$, we obtain

$$\mathbb{E}_{\mathbf{X},\mathcal{S},\mathcal{T}}\left[\mathrm{variance}_{ij}(\widehat{\boldsymbol{\beta}}^{\mathrm{ens}})\right] \xrightarrow{a.s.} \begin{cases} \frac{\sigma^2\eta^2}{\gamma-\eta^2} & \text{if } i \neq j, \\ \frac{\sigma^2\eta}{\alpha\gamma-\eta} & \text{if } i = j. \end{cases} \tag{41}$$

Thus, the variance component of the large-ensemble risk in this setting is equal to $\sigma^2\eta^2/(\gamma-\eta^2)$ and does not depend upon α. In future work, we plan to extend our analysis for the bias component of the large-ensemble risk to this setting, and we expect that in this case the bias will depend on *both* α and η.

5.3 Optimal Ensemble Mixing

In the ordinary least squares ensemble, we have used equal weighting when taking the average of our predictors. Instead, we could extend the idea presented in Section 4.4 to consider unequal weighting parameterized by $\boldsymbol{\mu} \in \mathbb{R}^k$, giving us the ensemble parameter estimate $\widehat{\boldsymbol{\beta}}^{\mathrm{ens}}_{\boldsymbol{\mu}} = \sum_{i=1}^k \mu_i \widehat{\boldsymbol{\beta}}^{(i)}$. While equal weighting gives us optimal risk in the setting where $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, p^{-1}\mathbf{I}_p)$, where ridge regression is optimal, under other distributional assumptions on $\boldsymbol{\beta}$, such as sparsity, where ridge regression is not optimal, unequal weighting has the potential to yield better ensembles.

Using the sparsity example, consider $\boldsymbol{\beta}$ such that $\|\boldsymbol{\beta}\|_0 = s \ll p$, and suppose that for some i, $S_i = S_{\boldsymbol{\beta}}$, where $S_{\boldsymbol{\beta}} = \{j : \beta_j \neq 0\}$. For simplicity, assume that $\eta = 1$, so that $T_i = [n]$ for all i. In this case, any predictor that uses the remaining $p - s$ features injects noise into its predictions, so the best predictor uses only the s features in $S_{\boldsymbol{\beta}}$. Under the i.i.d. Gaussian noise assumption, the predictor with lowest risk is in fact

$$\widehat{\boldsymbol{\beta}} = \operatorname*{arg\,min}_{\boldsymbol{\beta}' : \boldsymbol{\beta}'_{S^c_{\boldsymbol{\beta}}} = \mathbf{0}} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}'\|_2 = \widehat{\boldsymbol{\beta}}^{(i)}, \tag{42}$$

where i is such that $S_i = S_{\boldsymbol{\beta}}$. Thus an optimal weighting $\boldsymbol{\mu}$ is given by

$$\mu_i = \begin{cases} \frac{1}{C} & \text{if } S_i = S_{\boldsymbol{\beta}}, \\ 0 & \text{otherwise,} \end{cases} \tag{43}$$

where $C = |\{i : S_i = S_{\boldsymbol{\beta}}\}|$. This optimal weighting is decidedly non-uniform, and this raises the question of what schemes could be employed, either adaptively or non-adaptively, to minimize risk, and how they would fit into this analysis framework.

Acknowledgements

We would like to thank Ryan Tibshirani for helpful discussions and the anonymous reviewers for their helpful feedback. This work was supported by NSF grants CCF-1911094, IIS-1838177, and IIS-1730574; ONR grants N00014-18-12571 and N00014-17-1-2551; AFOSR grant FA9550-18-1-0478; DARPA grant G001534-7500; and a Vannevar Bush Faculty Fellowship, ONR grant N00014-18-1-2047.

References

Y. Amit and D. Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, 1997.

S. Arora, N. Cohen, W. Hu, and Y. Luo. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems 32*, pages 7413–7424. 2019.

S. Athey, J. Tibshirani, and S. Wager. Generalized random forests. *The Annals of Statistics*, 47(2): 1148–1178, Apr. 2019.

M. Belkin, D. J. Hsu, and P. Mitra. Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate. In *Advances in Neural Information Processing Systems 31*, pages 2300–2311. 2018.

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.

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, Aug. 1996.

L. Breiman. Arcing classifier (with discussion and a rejoinder by the author). *The Annals of Statistics*, 26(3):801–849, June 1998.

L. Breiman. Random forests. *Machine Learning*, 45 (1):5–32, Oct. 2001.

P. Bühlmann and B. Yu. Analyzing bagging. *The Annals of Statistics*, 30(4):927–961, Aug. 2002.

A. Cutler and G. Zhao. PERT - perfect random tree ensembles. *Computing Science and Statistics*, page 497, 2001.

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

M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim. Do we need hundreds of classifiers to solve real world classification problems? *Journal of Machine Learning Research*, 15:3133–3181, 2014.

Y. Freund and R. E. Schapire. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55(1):119–139, 1997.

R. Friedberg, J. Tibshirani, S. Athey, and S. Wager. Local linear forests. *arXiv preprint arXiv:1807.11408*, 2018.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning.* Springer Series in Statistics, 2001.

S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems 30*, pages 6151–6159. 2017.

M. Hardt, B. Recht, and Y. Singer. Train faster, generalize better: Stability of stochastic gradient descent. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48, pages 1225–1234, June 2016.

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

A. E. Hoerl and R. W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.

J. Josse and S. Wager. Bootstrap-based regularization for low-rank matrix estimation. *Journal of Machine Learning Research*, 17(1):4227–4255, Jan. 2016.

G. Louppe and P. Geurts. Ensembles on random patches. In *Machine Learning and Knowledge Discovery in Databases*, pages 346–361, Berlin, Heidelberg, 2012.

M. W. Mahoney. Approximate computation and implicit regularization for very large-scale data analysis. In *Proceedings of the 31st ACM SIGMOD-SIGACT-SIGAI Symposium on Principles of Database Systems*, PODS 12, pages 143–154, 2012.

B. Neyshabur, R. Tomioka, and N. Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

E. Scornet, G. Biau, and J.-P. Vert. Consistency of random forests. *The Annals of Statistics*, 43(4): 1716–1741, Aug. 2015.

N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958, Jan. 2014.

A. M. Tulino and S. Verdú. Random matrix theory and wireless communications. *Foundations and Trends in Communications and Information Theory*, 1(1): 1–182, 2004.

S. Wager, S. Wang, and P. S. Liang. Dropout training as adaptive regularization. In *Advances in Neural Information Processing Systems 26*, pages 351–359. 2013.

S. Wager, T. Hastie, and B. Efron. Confidence intervals for random forests: The jackknife and the infinitesimal jackknife. *Journal of Machine Learning Research*, 15:1625–1651, 2014.

A. J. Wyner, M. Olson, J. Bleich, and D. Mease. Explaining the success of AdaBoost and random forests as interpolating classifiers. *Journal of Machine Learning Research*, 18(48):1–33, 2017.

C. Zhang, S. Bengio, M. Hardt, B. Recht, and O. Vinyals. Understanding deep learning requires rethinking generalization. In *5th International Conference on Learning Representations, ICLR 2017*, 2017.