Understanding Double Descent Requires a Fine-Grained Bias-Variance Decomposition

Ben Adlam*†  Jeffrey Pennington*
Google Brain
{adlam, jpennin}@google.com

Abstract

Classical learning theory suggests that the optimal generalization performance of a machine learning model should occur at an intermediate model complexity, with simpler models exhibiting high bias and more complex models exhibiting high variance of the predictive function. However, such a simple trade-off does not adequately describe deep learning models that simultaneously attain low bias and variance in the heavily overparameterized regime. A primary obstacle in explaining this behavior is that deep learning algorithms typically involve multiple sources of randomness whose individual contributions are not visible in the total variance. To enable fine-grained analysis, we describe an interpretable, symmetric decomposition of the variance into terms associated with the randomness from sampling, initialization, and the labels. Moreover, we compute the high-dimensional asymptotic behavior of this decomposition for random feature kernel regression, and analyze the strikingly rich phenomenology that arises. We find that the bias decreases monotonically with the network width, but the variance terms exhibit non-monotonic behavior and can diverge at the interpolation boundary, even in the absence of label noise. The divergence is caused by the interaction between sampling and initialization and can therefore be eliminated by marginalizing over samples (i.e. bagging) or over the initial parameters (i.e. ensemble learning).

1 Introduction

It is undeniable that modern neural networks (NNs) are becoming larger and more complex, with many state-of-the-art models now employing billions of trainable parameters [1–3]. While parameter count may be a crude way of quantifying complexity, there is little doubt that these models have enormous capacity, often far more than is needed to perfectly fit the training data, even if the labels are pure noise [4]. Surprisingly, these same high-capacity models generalize well when trained on real data.

These observations conflict with classical generalization theory, which contends that models of intermediate complexity should generalize best, striking a balance between the bias and the variance of their predictive functions. A paradigm for understanding the observed generalization behavior of modern methods is known as double descent [5], in which the test error behaves as predicted by classical theory and follows the standard U-shaped curve until the point where the training set can be fit exactly, but after this point it begins to descend again, eventually finding its global minimum in the overparameterized regime.

While double descent has been the focus of significant research, a concrete and interpretable theoretical explanation for the phenomenon has thus far been lacking. One of the challenges in developing

*Both authors contributed equally to this work. †Work done as a member of the Google AI Residency program (https://g.co/airesidency).

while decomposing the variance has been proposed before, prior work has naively relied on the
mathematically sound, can lead to ambiguous conclusions, as we discuss in Sec. 4.

Another challenge in understanding double descent is that the classical bias-variance decomposition is
similar to ours is advanced in \cite{21}

The main mathematical tools we utilize come from random matrix theory and build on the results
of \cite{30} for studying random matrices with nonlinear dependencies. We also rely on techniques
from operator-valued free probability for computing traces of large block matrices \cite{31}. One
advantage of these tools is that they facilitate the extension of our analysis to more general settings,

### 1.1 Related Work

The idea of a trade-off between bias and variance has a long history, with theoretical and experimental
support having been well established in a variety of contexts over the years. The seminal paper of
Geman et al. \cite{10} examines a number of models, ranging from kernel regression to $k$-nearest
to neural networks, and concludes that the trade-off exists in all cases.\footnote{Interestingly, the variance of simple feed-forward neural networks was observed to eventually be a decreasing function of width, but the authors rationalized this early evidence of double descent as a quirk of the optimization.} The resulting
U-shaped test error curve was verified theoretically in a variety of classical settings, see e.g. \cite{11}.

In recent years, these conclusions have been called into question by the intriguing experimental results
of \cite{4,12}, which were later replicated in a number of settings, see e.g. \cite{13}, which showed that deep
neural networks and kernel methods can generalize well even in the interpolation regime, implying
that both the bias and the variance can decrease as the model complexity increases. A number of
theoretical results have since established this behavior in certain settings, such as interpolating nearest
neighbor schemes \cite{14} and kernel regression \cite{15,16}. These observations have given rise to the
double descent paradigm for understanding how test error depends on model complexity \cite{5}. The
influential work \cite{17} (which actually predates \cite{5}) established initial theoretical insights for linear
networks and found empirical evidence of double descent for nonlinear networks; more evidence has
followed recently in \cite{13,18}. Precise theoretical predictions soon confirmed this picture for linear
regression in various scenarios \cite{19,22}, and recently even for kernel regression \cite{23,24} with random
features related to neural networks.

The primary focus of these recent works has been on double descent in the total test error, or perhaps
the standard bias-variance decomposition with respect to label noise \cite{23}. A multivariate philosophy
similar to ours is advanced in \cite{25}, which revisited the empirical study of the bias-variance tradeoff
in neural networks from \cite{10} and showed the variance can decrease in the overparameterized regime.
However, in that work the variance is simply decomposed using the law of total variance, which,
while mathematically sound, can lead to ambiguous conclusions, as we discuss in Sec. 4.

The main mathematical tools we utilize come from random matrix theory and build on the results
of \cite{26,30} for studying random matrices with nonlinear dependencies. We also rely on techniques
from operator-valued free probability for computing traces of large block matrices \cite{31}. One
advantage of these tools is that they facilitate the extension of our analysis to more general settings,
including the case of kernel regression with respect to the Neural Tangent Kernel (NTK) \cite{NTK}. To ease the exposition we have deferred the discussion of the NTK and all proofs to the Supplementary Material (SM).

While finalizing this manuscript, we became aware of several concurrent works that examine similar questions. Yang et al. \cite{Yang2019} define the total bias and variance similarly to \cite{Bottou2018}, but they do not attempt a decomposition of the variance. Their results can be derived as a special case of our fine-grained decomposition by summing the variance terms in Thm. 1. Jacot et al. \cite{Jacot2018} study the relationship between the random feature model and the nonparametric Gaussian process which it approximates. The bias-variance decomposition considered in that paper is again univariate and is with respect to the randomness in the random features (the expressions are subsequently averaged over the training data).

Closest to our work is \cite{Lan2020}, which also studies a multivariate decomposition of the random feature model in the high-dimensional limit. Unlike our approach, their decomposition is not symmetric with respect to the underlying random variables, and the results depend on the chosen order of conditioning. Their particular choice, and indeed all possible choices, arise as special cases of our general result. See Sec. S8 for a detailed discussion.

**1.2 Our Contributions**

1. We develop a symmetric, interpretable variance decomposition suitable for modern deep learning algorithms
2. We compute this decomposition analytically for random feature kernel regression in the high-dimensional asymptotic regime
3. We prove that the bias is monotonically decreasing as the width increases and that it is finite at the interpolation threshold
4. We clarify the relationship between label noise and double descent: while the test loss can diverge at the interpolation threshold without label noise, the divergence is exacerbated by it
5. We provide a quantitative description of how both ensemble and bagging methods can eliminate double descent, since the divergence is caused by variance terms due to the interactions between sampling and initialization

**2 Bias-Variance Decomposition**

In this section, we trace through the evolution of several ways to analyze the bias-variance trade-off. By analyzing their shortcomings, we motivate our fine-grained analysis that follows.

**2.1 Classical Bias-Variance Decomposition**

The bias-variance trade-off has long served as a useful paradigm for understanding the generalization of machine learning algorithms. For a given test point $x$, it decomposes the expected error as

$$
E[(\hat{y}(x) - y(x))^2] = (E[\hat{y}(x) - E[y(x)])^2 + V[\hat{y}(x)] + V[y(x)],
$$

and subsequently averages over the test point to obtain a decomposition of the test error in which the first term is the bias, the second term is the variance, and the third term is the irreducible noise. In classical settings, the randomness of the predictive function is usually regarded as coming from randomness in the training data, i.e. sampling noise. This leads to two common conventions, where the expectations in eqn. (1) are over both $X$ and $y$ or are conditional on $X$ and only over the label noise in $y$. For concreteness and to simplify the exposition, in this subsection we adopt the latter convention and make the common modelling assumption that the sampling noise is an additive term $\varepsilon$ on the training labels but is zero on the test labels $y(x)$. Using $E_x$ to denote expectation over the test point, we have

$$
E_{\text{test}} := E_xE_\varepsilon [(\hat{y}(x) - y(x))^2 = E_x(E_\varepsilon[\hat{y}(x)] - y(x))^2 + E_xV_\varepsilon [\hat{y}(x)]].
$$

We refer to eqn. (2) as the classical bias-variance decomposition.

**2.2 Bias-Variance Decompositions for Modern Learning Methods**

Modern methods for training neural networks often utilize additional sources of randomness, such as the initial parameter values, minibatch selection, etc., which we collectively denote by $\Theta$. One is
We can interpret the variance as a measure of erroneous assumptions in the model. Another is that the predictive function \( \hat{y}(x) \) appearing in the bias \( B \) is not conditional on any random variables. As we discuss in Sec. 4.2, different choices yield different terms and can lead to ambiguous interpretations. The downside of this perspective is that the variance \( V \) no longer admits a simple interpretation since it contains contributions from multiple random variables. This problem can be remedied by further decomposing the variance.

2.2.1 Symmetric Decomposition of the Variance

To gain further insight into the structure of the total variance \( V \) and how individual random variables contribute to it, it can be useful to write \( V \) as a sum of individual terms, each with an unambiguous meaning.

One path forward is to rely on the law of total variance: \( \mathbb{V}[Y] = \mathbb{E}[\mathbb{V}[Y|X]] + \mathbb{V}[\mathbb{E}[Y|X]] \), where the terms represent the variance of \( Y \) unexplained and explained by \( X \) respectively. However, one is immediately confronted by the question of which source of randomness to condition on. As we discuss in Sec. 4.2, different choices yield different terms and can lead to ambiguous interpretations.

To avoid this ambiguity, we introduce a fully-symmetric decomposition, which turns out to be unique if we additionally require self-consistency under marginalization with respect to all variables.

**Proposition 1.** Let \( X_1, \ldots, X_K \), and \( Y \) be random variables and \( X := \{X_1, \ldots, X_K\} \). We define a variance decomposition of \( Y \) to be a multiset \( \{V_1, \ldots, V_N\} \) of nonnegative real numbers such that \( \mathbb{V}[Y] = \sum_i V_i \). Then there exists a unique variance decomposition \( V := \{V_s : s \subseteq X\} \) such that \( V \) is invariant under permutations of \( X \), and such that for all \( S \subseteq X \) the marginal variances satisfy the subset-sum relation,

\[
\mathbb{V}[Y|X] = \sum_{S \subseteq X} V_S \, .
\]

**Example 1.** Consider the case of two random variables, the parameters \( P \) and the data \( D \). Then \( X = \{P, D\} \) and the decomposition satisfying Prop. 1 is given by

\[
V_P := \mathbb{E}_X \mathbb{V}[\hat{y}|P] \, ,
\]

\[
V_D := \mathbb{E}_X \mathbb{V}[\hat{y}|D] \, ,
\]

\[
V_{PD} := \mathbb{E}_X \mathbb{V}[\hat{y}|P, D] - \mathbb{E}_X \mathbb{V}[\hat{y}|P] - \mathbb{E}_X \mathbb{V}[\hat{y}|D] \, .
\]

We can interpret \( V_{PD} \) as the variance explained by the parameters and data together beyond what they explain individually.
We consider predictive functions

where we dropped the outer expectation over $Y_i$.

and the model’s predictions are given by

For this model, the variance of the predictions is

Specifically, we define the random features on the training set

network with random weights. Specifically, we define the random features on the training set

high-dimensional asymptotics, the much more general setting in which the labels are produced by

SNR $\sigma/\sqrt{n}$ are generated by a linear function parameterized by $\beta \in \mathbb{R}^{n_0}$, whose entries are drawn independently from $\mathcal{N}(0, 1)$. Concretely, we let

\[ y(x_i) = \beta^T x_i / \sqrt{n_0} + \epsilon_i, \]

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is additive label noise on the training points, yielding a signal-to-noise ratio SNR $= \sigma^{-2}$. Although this may seem like a simple data distribution, it turns out that, in these high-dimensional asymptotics, the much more general setting in which the labels are produced by a non-linear teacher neural network can be exactly modeled with a linear teacher of this form (see Sec. 2.2).

We consider predictive functions $\hat{y}$ defined by approximate kernel ridge regression using the random feature model of $[36, 37]$, for which the random features are given by a single-layer neural network with random weights. Specifically, we define the random features on the training set $X = \{x_1, \ldots, x_m\}$ and test point $x$ to be

\[ f := \sigma(W_1 x / \sqrt{n_0}), \]

for a weight matrix $W_1 \in \mathbb{R}^{n_1 \times n_0}$ with iid entries $[W_1]_{ij} \sim \mathcal{N}(0, 1)$. The kernel induced by these random features is

\[ K(x_1, x_2) := \frac{1}{n_1} \sigma(W_1 x_1 / \sqrt{n_0})^T \sigma(W_1 x_2 / \sqrt{n_0}), \]

and the model’s predictions are given by

\[ \hat{y}(x) = Y K^{-1} K_x, \]

where $Y := [y(x_1), \ldots, y(x_m)]$, $K := K(X, X) + \gamma I_m$, $K_x := K(X, x)$, and $\gamma$ is a ridge regularization constant. For this model, $W_1$ plays the role of $\theta$ from Sec. 2.2.

Altogether, the test loss can be written as

\[ E_{\text{test}} = \mathbb{E}_\beta \mathbb{E}_x (y(x) - \hat{y}(x))^2 = \mathbb{E}_x (\beta^T x / \sqrt{n_0} - Y K^{-1} K_x)^2, \]

where we dropped the outer expectation over $\beta$ because the distribution concentrates around its mean (see the SM).

3 Asymptotic Variance Decomposition for Random Feature Regression

Problem setup and notation. Following prior work modeling double descent $[21, 23, 24]$, we perform our analysis in the high-dimensional asymptotic scaling limit in which the dataset size $m$, feature dimensionality $n_0$, and hidden layer size $n_1$ all tend to infinity at the same rate, with $\phi := n_0 / m$ and $\psi := n_0 / n_1$ held constant.

We consider the task of learning an unknown function from $m$ independent samples $(x_i, y_i) \in \mathbb{R}^{n_0} \times \mathbb{R}$, $i = 1, \ldots, m$, where the datapoints are standard Gaussian, $x_i \sim \mathcal{N}(0, I_{n_0})$, and the labels are generated by a linear function parameterized by $\beta \in \mathbb{R}^{n_0}$, whose entries are drawn independently from $\mathcal{N}(0, 1)$. Concretely, we let

\[ y(x_i) = \beta^T x_i / \sqrt{n_0} + \epsilon_i, \]

where $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is additive label noise on the training points, yielding a signal-to-noise ratio SNR $= \sigma^{-2}$. Although this may seem like a simple data distribution, it turns out that, in these high-dimensional asymptotics, the much more general setting in which the labels are produced by a non-linear teacher neural network can be exactly modeled with a linear teacher of this form (see Sec. 2.2).

We consider predictive functions $\hat{y}$ defined by approximate kernel ridge regression using the random feature model of $[36, 37]$, for which the random features are given by a single-layer neural network with random weights. Specifically, we define the random features on the training set $X = \{x_1, \ldots, x_m\}$ and test point $x$ to be

\[ f := \sigma(W_1 x / \sqrt{n_0}), \]

for a weight matrix $W_1 \in \mathbb{R}^{n_1 \times n_0}$ with iid entries $[W_1]_{ij} \sim \mathcal{N}(0, 1)$. The kernel induced by these random features is

\[ K(x_1, x_2) := \frac{1}{n_1} \sigma(W_1 x_1 / \sqrt{n_0})^T \sigma(W_1 x_2 / \sqrt{n_0}), \]

and the model’s predictions are given by

\[ \hat{y}(x) = Y K^{-1} K_x, \]

where $Y := [y(x_1), \ldots, y(x_m)]$, $K := K(X, X) + \gamma I_m$, $K_x := K(X, x)$, and $\gamma$ is a ridge regularization constant. For this model, $W_1$ plays the role of $\theta$ from Sec. 2.2.

Altogether, the test loss can be written as

\[ E_{\text{test}} = \mathbb{E}_\beta \mathbb{E}_x (y(x) - \hat{y}(x))^2 = \mathbb{E}_x (\beta^T x / \sqrt{n_0} - Y K^{-1} K_x)^2, \]

where we dropped the outer expectation over $\beta$ because the distribution concentrates around its mean (see the SM).

\[ ^3 \text{See the SM for an extension to the Neural Tangent Kernel of a single-hidden-layer neural network [32].} \]

\[ ^4 \text{Any non-zero variance $\sigma_{W_1}^2$ can be absorbed into a redefinition of $\sigma$.} \]
A bivariate symmetric decomposition of the variance resolves this ambiguity and shows that the

\[ \text{Corollary 1.} \]

such that

\[ \text{are given by the unique solutions to the coupled polynomial equations,} \]

\[ \text{noise exacerbates but does not cause double descent. Since } V_x = V_{Pe} = 0, \text{ they are not shown in (j).} \]

3.1 Main Result: Exact Asymptotics for the Fine-Grained Variance Decomposition

**Lemma 1.** Let \( \eta := \mathbb{E}[\sigma(g)^2] \) and \( \zeta := (\mathbb{E}[g\sigma(g)])^2 \) for \( g \sim \mathcal{N}(0, 1) \). Then, in the high-dimensional asymptotics defined above, the traces \( \tau_1(\gamma) := \frac{1}{m} \mathbb{E} \text{tr}(K^{-1}) \) and \( \tau_2(\gamma) := \frac{1}{m} \mathbb{E} \text{tr}(\frac{1}{\sigma_n} X^\top X K^{-1}) \)

are given by the unique solutions to the coupled polynomial equations,

\[
\zeta \tau_1 (1 - \gamma \tau_1) = \phi/\psi \left( \zeta \tau_2 + \phi(\tau_2 - \tau_1) \right) = (\tau_1 - \tau_2) \phi \left( (\eta - \zeta) \tau_1 + \zeta \tau_2 \right), \tag{21}
\]

such that \( \tau_1, \tau_2 \in \mathbb{C}^+ \) for \( \gamma \in \mathbb{C}^- \).

**Theorem 1.** Let \( \tau_1 \) and \( \tau_2 \) be defined as in Lemma 1 and use the prime symbol to denote their derivatives with respect to \( \gamma \). Then, as \( \Im(\gamma) \to 0^+ \), the asymptotic bias and variance terms of eqns. (9), (15) are given by

\[
\begin{align*}
B &= \tau_2^2/\tau_1^2, & V_{PX} &= -\tau_2^2/\tau_1^2 - B - V_P - V_X, \\
V_P &= \tau_2^2/\tau_1^2 - B, & V_{Pe} &= 0, \\
V_X &= \phi B(\tau_1 - \tau_2)^2/(\tau_2^2 - \phi(\tau_1 - \tau_2)^2), & V_{Xe} &= \sigma_x^2 V_X, \\
V_x &= 0, & V_{PXe} &= \sigma_x^2 \left( -\tau_1^2/\tau_1^2 - 1 \right) - V_{Xe}. \tag{22}
\end{align*}
\]

**Corollary 1.** In the ridgeless setting, the bias \( B \) is a non-increasing function of the overparameterization ratio \( n_1/m = \phi/\psi \). Furthermore, at the interpolation boundary \( \psi = \phi \), \( V_{PX} \) and \( V_{PXe} \) are divergent while the remaining terms are bounded.

4 Fine-Grained Analysis of Double Descent

The fine-grained variance decomposition given in Thm. 1 provides a powerful tool for understanding the origins of double descent. In this section, we use this tool to reinterpret several counterintuitive observations made in prior work and to provide a clear and unambiguous characterization of the source of double descent.

4.1 Semi-classical Approach: The Bias Diverges

In [21, 23], double descent in random feature kernel regression was analyzed through the lens of the semi-classical bias-variance decomposition introduced in eqn. (3). In our setting,

\[
E_{\text{test}} = B_{SC} + V_{SC}, \tag{23}
\]
where,

$$B_{SC} = \mathbb{E}_X \mathbb{E}_{P,X} \left( \mathbb{E}_x[y(j|x)|P,X] - y(x) \right)^2 , \quad \text{and} \quad V_{SC} = \mathbb{E}_X \mathbb{E}_{P,X} [V_e[y|P,X]|x].$$  \hspace{1cm} (24)

To gain further insight into this decomposition, we can express $B_{SC}$ and $V_{SC}$ in terms of the variables in Thm. 1:

$$B_{SC} = B + V_P + V_X + V_{PX} , \quad \text{and} \quad V_{SC} = V_e + V_{P\epsilon} + V_{X\epsilon} + V_{PX\epsilon} .$$  \hspace{1cm} (25)

Using the correspondence between the variance terms and areas mentioned in Remark 1, we can adopt the multivariate approach outlined in Sec. 2.2 and decompose the test loss in terms of two sources of randomness, the optimization/initial parameters $P$ and data sampling $D$. The total variance is additionally decomposed according to the law of total variance:

$$V = \mathbb{E}_X \mathbb{E}_D [\mathbb{E}_P[y(D)|X]] + \mathbb{E}_X \mathbb{E}_D [V_P[y(D)|X]], \hspace{1cm} (26)$$

where Neal et al. \cite{25} suggest an interpretation for the two terms as “variance due to sampling” and “variance due to optimization,” respectively. While the expressions in eqn. \cite{26} are themselves unambiguous, we will see that attributing such an interpretation to them can be somewhat misleading.

Some simple algebra allows us to express $V_D^\epsilon$ in terms of the variables in Thm. 1 as

$$V_D^\epsilon = V_P + V_{PX} + V_{P\epsilon} + V_{PX\epsilon} .$$  \hspace{1cm} (27)

Because eqn. \cite{27} contains $V_{PX}$ and $V_{PX\epsilon}$, Corollary 1 implies that $V_D^\epsilon$ diverges at the interpolation threshold, and indeed we observe that in Fig. \cite{25}(g). From the above interpretation of the meaning of $V_D^\epsilon$, we might therefore conclude that the “variance due to optimization” is the source of double descent.

On the other hand, we could have equally well decided to decompose the variance by conditioning on $P$ instead of $D$, yielding:

$$V = \mathbb{E}_X \mathbb{E}_P [\mathbb{E}_D[y|P]|X] + \mathbb{E}_X \mathbb{E}_P [V_D[y|P]|X] .$$  \hspace{1cm} (28)

The corresponding interpretations of these terms would then be “variance due to optimization” and “variance due to sampling,” respectively. As above, it is straightforward to express $V_D^\epsilon$ as,

$$V_P^\epsilon = V_X + V_{PX} + V_{X\epsilon} + V_{PX\epsilon} .$$  \hspace{1cm} (29)

In this case, Corollary 1 implies that $V_D^\epsilon$ diverges at the interpolation threshold, as Fig. \cite{25}(h) confirms. In this case, we might therefore conclude that the “variance due to sampling” is the source of double descent.

The above analysis reveals conflicting explanations for the source double descent, depending on which source of randomness is conditioned on when applying the law of total variance. We believe this ambiguity is undesirable and provides further motivation for the symmetric variance decomposition in Prop. 1.
We find that while the other two variance terms we plot the optimal ratio. The understanding we have developed for the sources of variance enables explicit prediction of the variable setting, the symmetric decomposition can be written as (see Example 1),

\[
V = V_p + V_D + V_{PD}.
\]

See Fig. 1(d) for an illustration of this decomposition. This figure shows that \(V_{PD}\) inhabits the ambiguous overlap region that was responsible for the inconsistent interpretations arising from a naive application of the law of total variance. From the theoretical results shown in Fig. 1(i), it is clear that neither the variance explained by the parameters, \(V_p\), nor the variance explained by the data, \(V_D\), can be responsible for double descent; instead it must be \(V_{PD}\) that is causing the divergence.

Recalling the definition of \(V_{PD}\) in Ex. 1 we conclude that the divergence at the interpolation boundary is caused by the “variance explained by the parameters and training data together beyond what they explain individually.”

One implication of this interpretation is that if we had a way of removing either the variance from the parameters or the variance from the data, then the divergence would be eliminated. We examine this phenomenon from the perspective of ensemble and bagging methods in Sec. 5 and confirm empirically that this is indeed the case. See Fig. 2.

**Bivariate Symmetric Decomposition: \(V_{PD}\) is the Source of Divergence.** In the previous two-variable setting, the symmetric decomposition can be written as (see Example 1).

\[
V = V_p + V_D + V_{PD}.
\]

Trivariate Symmetric Decomposition: Divergence Persists in Absence of Label Noise. Returning to the full model from Sec. 3 with three sources of randomness, we know from Thm. 1 that

\[
V = V_p + V_X + V_{PX} + V_{X\varepsilon} + V_{PX\varepsilon},
\]

while the other two variance terms \(V_\varepsilon\) and \(V_{PE}\) vanish. The seven variance terms are illustrated in Fig. 1(e). The dependence of the five non-zero terms on the model’s capacity is plotted in Fig. 1(j)). We find that \(V_{PX}\) and \(V_{PX\varepsilon}\) both diverge at the interpolation threshold while the other terms remain finite. This result helps explain recent empirical results that have found that label noise amplifies the double descent phenomena [14]; because \(V_{PX}\) itself diverges, there is double descent even without label noise, but because \(V_{PX\varepsilon}\) also diverges, label noise can exacerbate the effect.

## 5 Ensemble Learning

The understanding we have developed for the sources of variance enables explicit prediction of the effectiveness of ensemble and bagging techniques. We consider averaging the predictive functions of several independently initialized base learners as well as bagging the predictions from models with independent samples of training data. Specifically, we consider \(k_P\) independent samples of the parameters, \(P_j\), and \(k_D\) independent samples of the training data, \(X_j\) and \(\varepsilon_j\). Then our predictive function on a test point \(x\) is

\[
\hat{y}^*(x) := \frac{1}{k_P k_D} \sum_{i,j} \hat{y}_{ij}(x),
\]

where the indicies of \(\hat{y}\) indicate the specific sample of parameters and training data used to construct the predictor. A simple calculation gives the variance decomposition of \(\hat{y}^*\) as

\[
V_{PD} = \frac{V_p}{k_P}, \quad V_X = \frac{V_X}{k_D}, \quad V_\varepsilon = \frac{V_\varepsilon}{k_D}, \quad V_{X\varepsilon} = \frac{V_{X\varepsilon}}{k_D}.
\]
\[ V^\epsilon_{P_k} = \frac{V_P}{k_P k_D}, \quad V_{P_X} = \frac{V_{PX}}{k_P k_D}, \quad \text{and} \quad V^\epsilon_{P_X} = \frac{V_{PX\epsilon}}{k_P k_D}, \] (34)

while the bias remains the same. We illustrate these results empirically in Fig. 2 and show that ensembles of base learners and bagging are both able to independently reduce the divergence around the interpolation threshold, as they reduce the divergent terms \( V_{P_X} \) and \( V_{PX\epsilon} \).

As the computation of eqn. (32) requires evaluating \( k_P k_D \) base learners, it is natural to try to characterize the optimal combination of ensembles and bagging given a fixed computational budget. We find the optimal ratio is given as

\[ \frac{k_D}{k_P}_{\text{optimal}} = \frac{V_X + V_\epsilon + V_{X\epsilon}}{V_P}. \] (35)

See Fig. 2 which shows that, for the kernel regression problem studied here, ensembles are typically more efficient at small width and bagging is more efficient at large width.

6 Conclusion

We analyzed the bias and variance trade-off in the modern setting, where the difference to the classical picture of under- and overfitting is marked. We argued that understanding the behavior of the bias and variance in learning algorithms that depend on large sources of randomness requires rethinking the classical definitions to encompass these sources.

We presented a bias-variance decomposition that is suitable for these settings, and showed how it can help attribute components of the loss to their causes, while avoiding counterintuitive or ambiguous conclusions. For random feature kernel regression, we gave exact predictions for all of the terms in the decomposition and proved that the bias is monotonically decreasing and identified the source of divergence at the interpolation threshold to be the interaction between the noise from sampling and initialization. We showed that while label noise does not cause the divergence, it can exacerbate the effect. Finally, we made exact predictions for ensemble learning and bagging and provided the computationally optimal strategy to combine them.

Broader Impact

While it is hard to envision all future applications of this research, the authors do not believe this theoretical work will raise any ethical concerns or will generate any adverse future societal consequences.

Acknowledgments and Disclosure of Funding

We are grateful to Boris Hanin, Jaehoon Lee, Mihai Nica, D. Sculley, Jasper Snoek, and Lechao Xiao for valuable feedback on an earlier version of the paper. We also thank the anonymous reviewers for pointing us to many related works, including the connection to ANOVA.

Funding in direct support of this work came from Google. No third party funding was used.

References


