Bias/Variance is not the same as Approximation/Estimation

Anonymous authors
Paper under double-blind review

Abstract

We study the relation between two classical results: the bias-variance decomposition, and the approximation-estimation decomposition. Both are important conceptual tools in Machine Learning, helping us describe the nature of model fitting. It is commonly stated that they are "closely related", or "similar in spirit". However, sometimes it is said they are equivalent (spoiler: no, they're not). In reality, they have subtle connections, cutting across learning theory and classical statistics, that (very surprisingly) have not been previously observed. In this work we uncover the connections, building a bridge between these two seminal results.

1 Introduction

Geman et al. (1992) introduced the bias-variance decomposition to the Machine Learning community, and Vapnik & Chervonenkis (1974) introduced the approximation-estimation decomposition, founding the field of statistical learning theory. Both decompositions help us understand model fitting: referring to model size, and some kind of trade-off. The terms are often used interchangeably. And yet, they are different things. Given their fundamental nature and similar purposes, it is surprising that explicit connections are not widely known—perhaps due to differing notations and conventions of their respective communities. Our goal is to uncover these connections and build a bridge between these two seminal results.

The approximation-estimation decomposition refers to models drawn from some function class $\mathcal{F}$, and considers an *excess risk*—that is, the risk above that of the Bayes model—breaking it into two components:

$$\textbf{excess risk} \;=\; \textbf{approximation error} \;+\; \textbf{estimation error}. \tag{1}$$

We might choose to increase the size of our function class, perhaps by adding more parameters to our model. In this situation it is commonly understood that the approximation error will decrease, and the estimation error will increase (Von Luxburg & Schölkopf, 2011), beyond a certain point resulting in over-fitting of the model. In contrast to the abstract notion of a "function class", the *bias-variance* decomposition considers the risk of real, trained models, in expectation over possible training sets. Assuming there is a unique correct response for each given input $\boldsymbol{x}$ (i.e., no noise) it breaks the expected risk into two components:

$$\textbf{expected risk} \;=\; \textbf{bias} \;+\; \textbf{variance}. \tag{2}$$

As we increase model size: the bias tends to decrease, and the variance tends to increase, again determining the degree of over-fitting. Recently, it has become apparent that this trade-off is not always simple, e.g. with over-parameterised models; but, the decomposition still holds even if a simple trade-off does not. We note that this decomposition, as used in the Machine Learning literature, concerns inference of the response/target variable—and not of the parameters, as is more common in classical statistics.

We therefore have two decompositions: both referring to model size, with some kind of trade-off between their terms, and with bearing on the nature of over-fitting. It is easy, and common, to conflate these. From online discussion forums, to the lecture notes of esteemed institutions and well-cited research articles, one can observe innocent (but imprecise) statements such as "the two are closely related", but also the more extreme (and incorrect) "the approximation-estimation trade-off is also known as the bias-variance trade-off". To the best of our knowledge, this is the first work to discuss their connection explicitly. We consider a range of loss functions: including Bregman divergences and the 0/1 loss. We study properties of the decompositions, observing where they coincide and where they do not.

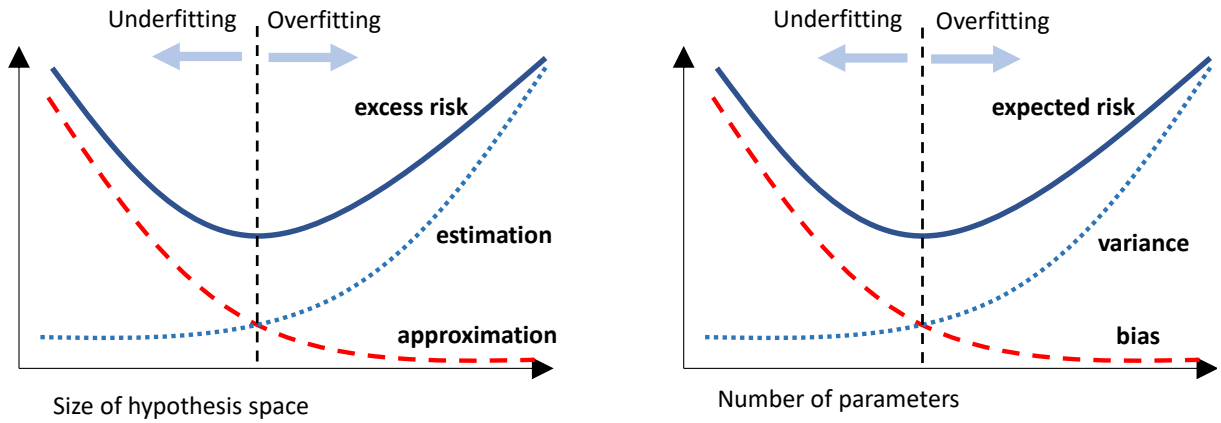

Figure 1: Two diagrams (same on left/right is intentional) illustrating how the approximation/estimation and bias/variance trade-offs are commonly described, and easily confused.

2 Background

We introduce notation and review the two decompositions. We introduce these ideas in an intentionally didactic/comprehensive manner, to avoid any possibility of confusion in terminology.

2.1 Preliminaries

Consider a standard supervised learning setup, where the task is to map from an input $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^d$ to an output $\boldsymbol{y} \in \mathcal{Y} \subseteq \mathbb{R}^k$, and assume there exists an unknown distribution $P(\boldsymbol{x}, \boldsymbol{y})$. This is achieved by learning the parameters of a model $f : \boldsymbol{x} \to \boldsymbol{y}$, which can also be seen as selecting a function f from a set $\mathcal{F} \subset \mathcal{F}_{all}$, a subset of all measurable functions. The discrepancy of $f(\boldsymbol{x})$ from the true $\boldsymbol{y}$ is quantified with a loss function $\ell(\boldsymbol{y}, f(\boldsymbol{x}))$, which may or may not be symmetric. Using this, we define the *risk* of a given model f,

$$R(f) := \mathbb{E}_{\boldsymbol{xy}}[\ell(\boldsymbol{y}, f(\boldsymbol{x}))] = \int \ell(\boldsymbol{y}, f(\boldsymbol{x}))\, dP(\boldsymbol{x}, \boldsymbol{y}). \tag{3}$$

The *Bayes* model $\boldsymbol{y}^*$ is the hypothetical function which minimizes this quantity at each $\boldsymbol{x}$, i.e.

$$\boldsymbol{y}^* := \arg\inf_{f \in \mathcal{F}_{all}} R(f), \tag{4}$$

where we acknowledge a slight abuse of notation, using $\boldsymbol{y}^*$ as a function in $\mathcal{F}_{all}$ or a vector in $\mathbb{R}^k$ as needed—the intention will always be made clear from context. Given that we picked a restricted family $\mathcal{F} \subset \mathcal{F}_{all}$, we have no guarantee that it contains $\boldsymbol{y}^*$. The best-in-family model f^* is defined

$$f^* := \arg\inf_{f \in \mathcal{F}} R(f). \tag{5}$$

Both of these are defined in terms of the true distribution $P(\boldsymbol{x}, \boldsymbol{y})$. In practice, we only have a sample $\mathcal{S} = \{(\boldsymbol{x}_i, \boldsymbol{y}_i)\}_{i=1}^n$, drawn from a random variable $D \sim P(\boldsymbol{x}, \boldsymbol{y})^n$. We write the *empirical risk* as:

$$R_{emp}(f; \mathcal{S}) := \frac{1}{n} \sum_{i=1}^{n} \ell(\boldsymbol{y}_i, f(\boldsymbol{x}_i)), \tag{6}$$

and a model in $\mathcal{F}$ that minimizes this, known as an *empirical risk minimizer*, is defined:

$$\hat{f}_{erm} := \arg\inf_{f \in \mathcal{F}} R_{emp}(f; \mathcal{S}). \tag{7}$$

We use a 'hat' notation to emphasize that the ERM is dependent on D, the random variable over training data samples. For some models/losses an ERM is achievable—e.g. the closed-form solution for linear models under squared loss. However in general, a training procedure will not necessarily result in an ERM. We can now cover the specifics for the two decompositions.

2.2 The Approximation-Estimation decomposition

The approximation-estimation decomposition is a seminal observation from the 1970s work of Vapnik and Chervonenkis, reviewed in Vapnik (1999). An excellent historical account can be found in Bottou (2013). The result deals with the *excess* risk $R(\hat{f}_{erm}) - R(\boldsymbol{y}^*)$, i.e. the risk of $\hat{f}_{erm}$ above that of the Bayes model, $\boldsymbol{y}^*$. The approximation-estimation decomposition, applicable for any loss ℓ, breaks this into two terms:

$$\underbrace{R(\hat{f}_{erm}) - R(\boldsymbol{y}^*)}_{\textbf{excess risk}} = \underbrace{R(\hat{f}_{erm}) - R(f^*)}_{\textbf{estimation error}} + \underbrace{R(f^*) - R(\boldsymbol{y}^*)}_{\textbf{approximation error}}. \tag{8}$$

The approximation error is the additional risk due to using a restricted family $\mathcal{F}$, rather than the space of all functions $\mathcal{F}_{all}$. This is a systematic quantity, not dependent on any particular data sample. The estimation error is the additional risk due to our finite training data, when trying to find $f^* \in \mathcal{F}$. This a random variable, dependent on the particular data sample used to obtain $\hat{f}_{erm}$. There is a natural trade-off (see Figure 1, left) as we change the size of $\mathcal{F}$, keeping data size fixed. As we increase $|\mathcal{F}|$, approximation error will likely decrease (potentially to zero, if $\boldsymbol{y}^* \in \mathcal{F}$), but estimation error will increase, as it becomes harder to find f^* in the larger space. The reason behind this is, in effect, the classical *multiple hypothesis testing* problem—we cannot reliably distinguish many hypotheses when our dataset is small. Bottou & Bousquet (2007) extended Equation 8, recognising that it is often intractable to find $\hat{f}_{erm}$, and we can only have a sub-optimal model $\hat{f}$. An additional risk component then emerges, and the excess risk of $\hat{f}$ now decomposes into a sum of *optimisation* error, estimation error, and approximation error:

$$\underbrace{R(\hat{f}) - R(\boldsymbol{y}^*)}_{\textbf{excess risk of } \hat{f}} = \underbrace{R(\hat{f}) - R(\hat{f}_{erm})}_{\textbf{optimisation error}} + \underbrace{R(\hat{f}_{erm}) - R(f^*)}_{\textbf{estimation error}} + \underbrace{R(f^*) - R(\boldsymbol{y}^*)}_{\textbf{approximation error}}. \tag{9}$$

These three terms describe the learning process in abstract form: accounting respectively for the choice of learning algorithm, the quality/amount of data, and the capacity of the model family.

2.3 The Bias-Variance decomposition

A bias-variance decomposition involves the *expected* risk of a trained model $\hat{f}$, where the expectation $\mathbb{E}_D$ is over the random variable $D \sim P(\boldsymbol{x}, y)^n$, i.e., all possible training sets of a fixed size n. Focusing on a squared loss, and $y \in \mathbb{R}$, Geman et al. (1992) showed:

$$\underbrace{\mathbb{E}_D\left[\mathbb{E}_{\boldsymbol{x}y}[(y - \hat{f}(\boldsymbol{x}))^2]\right]}_{\textbf{expected risk}} = \underbrace{\mathbb{E}_{\boldsymbol{x}y}\left[(y - y^*)^2\right]}_{\textbf{noise}} + \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\left(y^* - \mathbb{E}_D[\hat{f}(\boldsymbol{x})]\right)^2\right]}_{\textbf{bias}} + \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_D[(\hat{f}(\boldsymbol{x}) - \mathbb{E}_D[\hat{f}(\boldsymbol{x})])^2]\right]}_{\textbf{variance}}. \tag{10}$$

where $y^* = \mathbb{E}_{y|\boldsymbol{x}}[y]$ is the Bayes-optimal prediction at each point $\boldsymbol{x}$. The bias is a systematic component, independent of any particular training sample, and commonly regarded as measuring the 'strength' of a model. The variance measures the sensitivity of $\hat{f}$ to changes in the training sample, independent of the true label y. The noise is a constant, independent of any model parameters. There is again a perceived trade-off with these terms (see Figure 1, right). As the size of the (un-regularised) model increases: bias *tends* to decrease, and variance *tends* to increase. However, the trade-off can be more complex (e.g. with over-parameterized models) and the exact dynamics are an open research issue.

Bias-Variance decompositions hold for more than just squared loss. In fact, the same form holds for any *Bregman divergence* (Bregman, 1967; Pfau, 2013). For a domain $\mathcal{Y} \subseteq \mathbb{R}^k$, define $\ell : \mathcal{Y} \times \mathrm{ri}(\mathcal{Y}) \to \mathbb{R}_+$ as an arbitrary Bregman divergence, parameterised by a strictly convex *generator*[1] function ϕ, then:

$$\underbrace{\mathbb{E}_D\left[\mathbb{E}_{\boldsymbol{x}\boldsymbol{y}}[\ell(\boldsymbol{y}, \hat{f}(\boldsymbol{x}))]\right]}_{\textbf{expected risk}} = \underbrace{\mathbb{E}_{\boldsymbol{x}\boldsymbol{y}}\left[\ell(\boldsymbol{y}, \boldsymbol{y}^*)\right]}_{\textbf{noise}} + \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\ell(\boldsymbol{y}^*, \mathring{f}_\phi(\boldsymbol{x}))\right]}_{\textbf{bias}} + \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_D[\ell(\mathring{f}_\phi(\boldsymbol{x}), \hat{f}(\boldsymbol{x}))]\right]}_{\textbf{variance}}. \tag{11}$$

where $\mathring{f}_\phi(\boldsymbol{x}) := \arg\min_{\boldsymbol{z} \in \mathcal{Y}} \mathbb{E}_D[\ell(\boldsymbol{z}, \hat{f}(\boldsymbol{x}))]$, and $\boldsymbol{y}^* := \arg\min_{\boldsymbol{z} \in \mathrm{ri}(\mathcal{Y})} \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}[\ell(\boldsymbol{y}, \boldsymbol{z})] = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}[\boldsymbol{y}]$ is again the Bayes model.

[1] We refer the reader to Banerjee et al. (2005b) for an excellent tutorial on Bregman divergences.

This general Bregman form covers many well-known losses as special cases, e.g. squared and Poisson loss. For a squared loss, we have $\mathring{f}_\phi(\boldsymbol{x}) = \mathbb{E}_D[\hat{f}(\boldsymbol{x})]$, but this is not always the case. In general, $\mathring{f}_\phi(\boldsymbol{x})$ is a generalised measure of centrality, which is known (in the information geometry community) as a *Bregman centroid* (Nielsen & Nock, 2009). As this will be important in the coming sections, we define it formally.

Definition 1 (Centroid prediction at $\boldsymbol{x}$) *Assume an arbitrary loss $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$. For a single input $\boldsymbol{x}$, given a distribution of predictions from a model $\hat{f}$, induced by a random variable D, the centroid prediction is the point in $\mathcal{Y}$ closest on average to all others: $\mathring{f}(\boldsymbol{x}) := \arg\min_{\boldsymbol{z}\in\mathcal{Y}} \mathbb{E}_D[\ell(\boldsymbol{z}, \hat{f}(\boldsymbol{x}))]$.*

For the special case of ℓ as a Bregman divergence, this is available in closed-form (Nielsen & Nock, 2009).

Definition 2 (Bregman centroid prediction at $\boldsymbol{x}$) *If ℓ is a Bregman divergence $B_\phi : \mathcal{Y} \times \mathrm{ri}(\mathcal{Y}) \to \mathbb{R}_+$ defined by a strictly convex function $\phi : \mathcal{Y} \to \mathbb{R}$, then the centroid prediction at $\boldsymbol{x}$ is:*

$$\mathring{f}_\phi(\boldsymbol{x}) := \arg\min_{\boldsymbol{z}\in\mathcal{Y}} \mathbb{E}_D\left[B_\phi(\boldsymbol{z}, \hat{f}(\boldsymbol{x}))\right] = [\nabla\phi]^{-1}\left(\mathbb{E}_D\left[\nabla\phi(\hat{f}(\boldsymbol{x}))\right]\right). \tag{12}$$

This closed-form for the Bregman centroid[2] prediction will turn out to be very useful. Examples of Bregman centroids can be found in the table below.

Table 1: Examples of losses which admit a bias-variance decomposition, with corresponding centroids.

Name	**Domain**	$B_\phi(\boldsymbol{y}, \hat{f}(\boldsymbol{x}))$	**Centroid** $\mathring{f}_\phi(\boldsymbol{x})$
Squared	$y \in \mathbb{R}$	$(y - \hat{f}(\boldsymbol{x}))^2$	$\mathbb{E}_D[\hat{f}(\boldsymbol{x})]$
Poisson	$y \in \{0, 1, 2, ...\}$	$y \ln \frac{y}{\hat{f}} - (y - \hat{f})$	$\exp(\mathbb{E}_D[\ln \hat{f}])$
KL-divergence	$\boldsymbol{y} \in \mathbb{R}^k, s.t. \sum_c y_c = 1$	$D_{KL}(\boldsymbol{y} \,\|\|\, \hat{f}(\boldsymbol{x}))$	$Z^{-1}\exp(\mathbb{E}_D[\ln \hat{f}(\boldsymbol{x})])$
Ikatura-Saito	$y \in [0, \infty)$	$\frac{y}{\hat{f}(\boldsymbol{x})} - \ln \frac{y}{\hat{f}(\boldsymbol{x})} - 1$	$1/\mathbb{E}_D[\hat{f}(\boldsymbol{x})^{-1}]$

The bias/variance terms take *different functional forms* for each Bregman divergence. Note that the KL-divergence example implies a decomposition for the cross-entropy loss, since the two differ only by a constant. It is interesting to note that generalised decompositions only appeared in the ML community with Heskes (1998), but the idea seems to be known much earlier in statistics, e.g., Hastie & Tibshirani (1986, Eq. 19).

Bias-Variance decompositions do not hold for all losses. The approximation-estimation decomposition, Equation 9, applies for *any* loss. This is not so for the bias-variance decomposition. As is well-documented (Geurts, 2002), the form of Equation 11 does not hold for the 0/1 loss, and we have an *inequality*:

$$\underbrace{\mathbb{E}_D\left[\mathbb{E}_{\boldsymbol{x}y}[\ell_{0/1}(y, \hat{f}(\boldsymbol{x}))]\right]}_{\textbf{expected 0/1 risk}} \neq \underbrace{\mathbb{E}_{\boldsymbol{x}y}\left[\ell_{0/1}(y, y^*)\right]}_{\textbf{noise}} + \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\ell_{0/1}(y^*, \mathring{f}(\boldsymbol{x}))\right]}_{\textbf{bias}} + \underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_D[\ell_{0/1}(\mathring{f}(\boldsymbol{x}), \hat{f}(\boldsymbol{x}))]\right]}_{\textbf{variance}}, \tag{13}$$

where $\mathring{f}(\boldsymbol{x})$ is the *modal* value. Several authors have proposed alternative decompositions, following sets of axioms to define what constitutes a 'bias-variance' decomposition (Wolpert, 1997; James & Hastie, 1997; Heskes, 1998). The necessary and sufficient conditions for such a decomposition are an open research question.

2.4 Summary

These decompositions are *conceptual* tools to describe the nature of model fitting. They are by no means perfect reflections of the process, most especially in the context of over-parameterized models (Nagarajan & Kolter, 2019; Zhang et al., 2021). However, it is *extremely* common to see papers making the incorrect assumption/claim that the two are equivalent, or that one is a special case of the other. Our purpose with this work is to correct these false assumptions, identifying *precisely* how the two connect.

[2]Note that this is a minimization over the first (left-hand) argument, so it is technically a *left* centroid. The *right* centroid can be similarly defined, turning out to be simply $\mathbb{E}_D[\hat{f}(\boldsymbol{x})]$ for any valid ϕ (Banerjee et al., 2005a), which explains why the Bayes-optimal prediction is $\boldsymbol{y}^* = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}[\boldsymbol{y}]$ for any Bregman divergence.

3 Bias/Variance is not the same as Approximation/Estimation

By now it should be evident that these decompositions are related, but are not quite the same thing. Perhaps the most obvious difference is that they are on different quantities—the excess risk of an ERM, versus the expected risk of an arbitrary trained model. We now build a bridge between the two. We first define a concept that we will refer to repeatedly in the coming sections: the 'centroid model'.

Definition 3 (Centroid model) *For a model $\hat{f}$ dependent on a random variable D, the centroid model $\mathring{f}$ is the aggregate formed by taking the centroid prediction at each possible $\boldsymbol{x}$. Note that whilst by definition $\mathring{f} \in \mathcal{F}_{all}$, there is no guarantee that $\mathring{f} \in \mathcal{F}$.*

Note that this definition is general to any loss, not just Bregman divergences. We now observe that the estimation error involves $R(\hat{f}_{erm})$, making it a random variable dependent on D. We take the expectation with respect to D, and separate it into two components using $R(\mathring{f})$, the risk of the centroid model.

Proposition 1 (Decomposing the Expected Estimation Error) *For an arbitrary loss ℓ, the expected estimation error decomposes as follows:*

$$\underbrace{\mathbb{E}_D\left[R(\hat{f}_{erm}) - R(f^*)\right]}_{\textbf{expected estimation error}} = \underbrace{\mathbb{E}_D\left[R(\hat{f}_{erm}) - R(\mathring{f})\right]}_{\textbf{estimation variance}} + \underbrace{R(\mathring{f}) - R(f^*)}_{\textbf{estimation bias}}. \tag{14}$$

The *estimation variance*, $\mathcal{E}_{est(v)}$, measures the *random* variations of $\hat{f}_{erm}$ around the centroid model. The *estimation bias*, $\mathcal{E}_{est(b)}$, measures the *systematic* difference between the centroid model and the best-in-family model. Using these concepts, we can present the relation between the two decompositions.

Theorem 1 (Bias-Variance in terms of Approximation-Estimation) *For any Bregman divergence, $\ell(\boldsymbol{y}, f(\boldsymbol{x})) = B_\phi(\boldsymbol{y}, f(\boldsymbol{x}))$, the following decomposition of the bias and variance applies.*

$$\underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\ell(\boldsymbol{y}^*, \mathring{f}_\phi(\boldsymbol{x}))\right]}_{\textbf{bias}} = \underbrace{R(f^*) - R(\boldsymbol{y}^*)}_{\textbf{approximation error}} + \underbrace{R(\mathring{f}_\phi) - R(f^*)}_{\textbf{estimation bias}} \tag{15}$$

$$\underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_D[\ell(\mathring{f}_\phi(\boldsymbol{x}), \hat{f}(\boldsymbol{x}))]\right]}_{\textbf{variance}} = \underbrace{\mathbb{E}_D\left[R(\hat{f}) - R(\hat{f}_{erm})\right]}_{\textbf{optimisation error}} + \underbrace{\mathbb{E}_D\left[R(\hat{f}_{erm}) - R(\mathring{f}_\phi)\right]}_{\textbf{estimation variance}} \tag{16}$$

This confirms the premise of our work. Bias is *not* approximation error, and variance is *not* estimation error. It is not even the case that one is a special case of the other, as is sometimes stated. The true relation is more subtle. The approximation error is in fact just *one component of the bias*, and, the estimation error *contributes to both bias and variance.* The theorem above is illustrated in Figure 2.

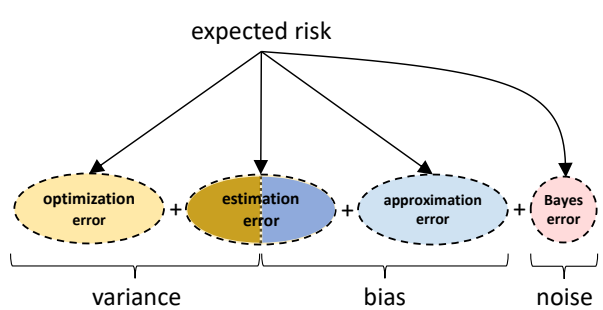

Figure 2: Illustration of Theorem 1. The bias is only partly determined by approximation error (i.e. choice of **model**), while the rest is due to expected estimation error (i.e. choice of **data**). Similarly, variation in data accounts for only part of the variance, and the rest is due to optimisation error (i.e. choice of **algorithm**).

4 Discussion

A simplistic description of bias and variance would say they are the error *'due to the model'* (bias) and the error *'due to the data'* (variance). Theorem 1 shows there is more nuance to understand. We now discuss the subtleties and implications of these results.

4.1 The bias is a flawed proxy for model capacity.

It is common to assume the bias is an indication of how simple/complex a model is—expected to be lower if the model has higher 'capacity'. But what is model 'capacity'? If we take it to be the ability to minimize the population risk, then the *ultimate* measure of model capacity is the *approximation error*. We see from Equation 15 that the bias contains exactly this, but also the additional *estimation bias* term, which gives it some surprising dynamics.

For a squared loss with a linear model, Equation 15 has been noted[3] before (Hastie et al., 2017, Eq 7.14). Our result generalises it to a broader range of losses and arbitrary *non-linear* models. For tractability, their analyses were restricted to linear models. However tractable, they were unable to observe a critical fact—that in the general case, estimation bias $R(\mathring{f}) - R(f^*)$, can take *negative values,* i.e.

$$\textbf{bias} \quad = \quad \underbrace{\left[\begin{matrix}\textit{approximation}\\ \textit{error}\end{matrix}\right]}_{\textbf{always} \geq \textbf{0}} \quad + \quad \underbrace{\left[\begin{matrix}\textit{estimation}\\ \textit{bias}\end{matrix}\right]}_{\textbf{can be negative}} \tag{17}$$

To understand how this can be, we must accept the somewhat non-intuitive idea that the centroid model can be outside the hypothesis class $\mathcal{F}$, and thus we can have $R(\mathring{f}) < R(f^*)$. This was described in Definition 3, and can be trivially illustrated, even with a simple regression stump evaluated by squared loss.

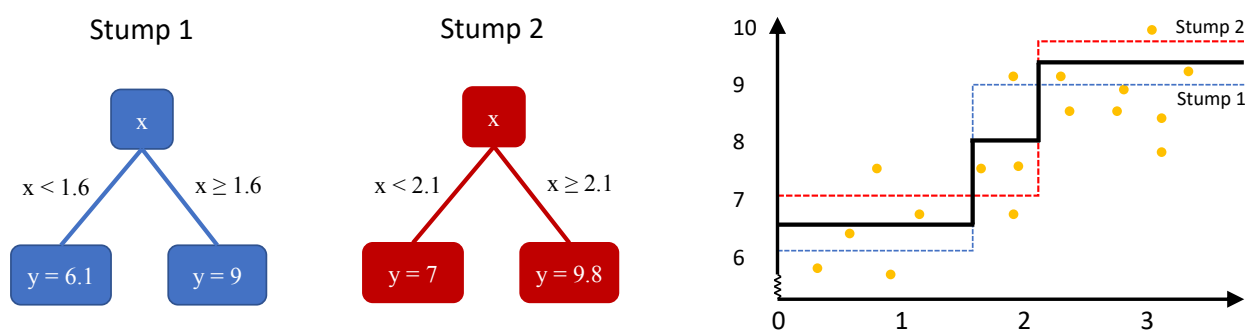

Figure 3: Two regression stumps (red/blue lines), and their centroid model (black line, arithmetic mean). Notice the centroid model is *outside* the hypothesis class, i.e. it cannot be represented as a single binary stump. As a result, the centroid model fits the data better than any $f \in \mathcal{F}$, and $\mathcal{E}_{est(b)}$ is negative.

The possibility of negative values here has significant implications. There are two ways in which bias can be zero. If $\mathcal{F}$ contains the Bayes model, then we might have $\mathcal{E}_{app} = \mathcal{E}_{est(b)} = 0$. But, there is another way. For some $\epsilon > 0$, we might have $\mathcal{E}_{app} = \epsilon$, and $\mathcal{E}_{est(b)} = -\epsilon$. In this case, the model family does *not* have sufficient capacity, since $\mathcal{E}_{app} > 0$. And yet, the bias is zero. Hence, the bias is a flawed proxy for the true model capacity.

To illustrate the point, we show experiments on a synthetic regression problem. Details in Appendix C.

[3]Hastie et al. described the first term on the right as *"the error between the best-fitting linear approximation and the true function"*. This is exactly the definition of approximation error for the linear model. We provide a proof of this relation and further discussion in the appendix.

Figure 4 shows results increasing the depth of a decision tree. The left panel shows excess risk, and the bias/variance components. We observe the classical bias/variance trade-off, including overfitting, as the depth increases beyond a certain point. It is notable that *the bias decreases to zero*, after depth 6. *Does this imply the model is 'unbiased', in the sense that it has sufficient capacity to capture the full data distribution?*

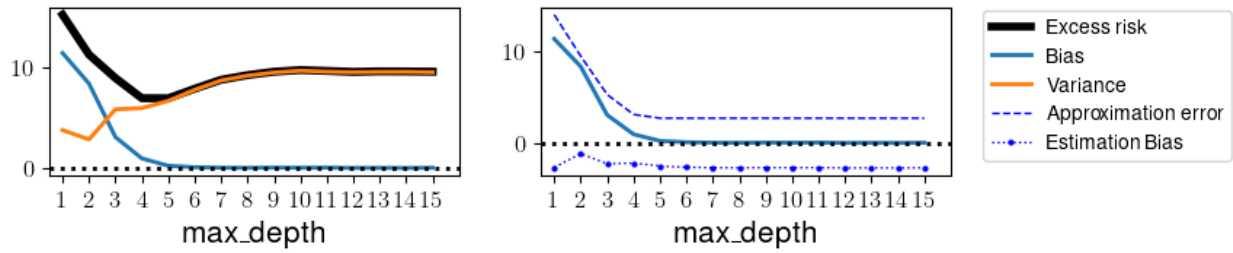

Figure 4: Risk components as we increase the depth of a regression tree.

The answer is no. A decomposition of the bias into two components (right panel) shows that the approximation error is non-zero, i.e. the best possible model *cannot* achieve zero testing error. The cause of the bias going to zero is that the estimation bias is negative, hence the bias is not a good proxy for the true model capacity. Very similar results are obtained with a k-nearest neighbour regression (Figure 5), where increasing complexity is obtained by *decreasing* the value of k.

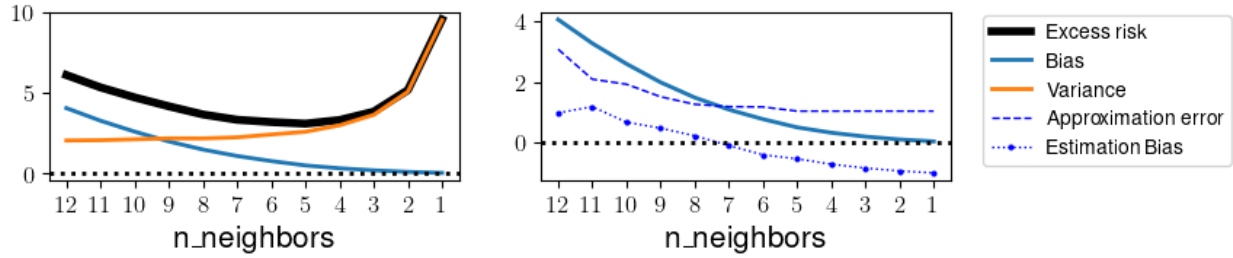

Figure 5: Risk components as we decrease the number of neighbours in a k-nn.

We can formally characterise this phenomenon, by studying the geometry of the hypothesis class $\mathcal{F}$. In particular, if the set $\mathcal{F}$ is *dual-convex* (Amari, 2008) with respect to ϕ, then $\mathring{f} \in \mathcal{F}$, and hence estimation bias is guaranteed to be non-negative.

Theorem 2 (Sufficient condition for a non-negative estimation bias.) *If the hypothesis class $\mathcal{F}$ is dual-convex then the estimation bias is non-negative.*

A simple example of a non-dual convex set is the class of regression stumps evaluated by squared loss, where $\mathring{f}(\boldsymbol{x}) = \mathbb{E}_D[\hat{f}(\boldsymbol{x})]$, illustrated in Figure 3. A simple example of a dual-convex set is the class of Generalized Linear Models evaluated by their corresponding deviance measure.

Theorem 3 (GLMs have non-negative estimation bias.) *For a Bregman divergence with generator function ϕ, define $\mathcal{F}$ as the set of all GLMs with inverse link $[\nabla\phi]^{-1}$ and natural parameters $\boldsymbol{\theta} \in \mathbb{R}^d$. Then, the estimation bias is non-negative.*

An example of this would be a logistic regression, $\hat{f}(\boldsymbol{x}) = [\nabla\phi]^{-1}(\hat{\boldsymbol{\theta}}^T\boldsymbol{x}) = 1/(1+\exp(-\hat{\boldsymbol{\theta}}^T\boldsymbol{x}))$, which results from $\phi(f) = f\ln f + (1-f)\ln(1-f)$ and the binary KL as the deviance.

4.2 New insights into the bias/variance trade-off.

In the age of deep learning, the relevance (and even existence) of a *trade-off* between bias and variance has been debated, with voices both against (Neal et al., 2018; Dar et al., 2021) and in favour (Witten, 2020). Proposition 1 places a constraint between $\mathcal{E}_{est(b)}$ and $\mathcal{E}_{est(v)}$, the estimation bias and the estimation variance. When the estimation bias is negative (e.g. Figure 3), it obviously *reduces* the bias. However, it simultaneously *increases* the variance, since $\mathcal{E}_{est(v)}$ has an imposed lower bound, satisfying the constraint. Therefore, for every single reduction in bias attributable to a negative estimation bias, *the same quantity will be lost* in the increased variance. This is an *unavoidable* trade-off. Obviously other components of the bias/variance may mask this behaviour, making the trade-off less visible.

4.3 The estimation variance plays a role in double descent.

In many models, an increasing degree of over-parameterisation has been associated with a 'peaking' trend in the variance (Nakkiran, 2019; Yang et al., 2020), ultimately causing a *double descent* in the risk.

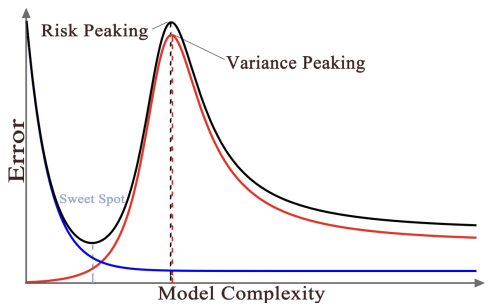

Figure 6: Illustration of double descent, caused by a 'peaking' variance (red line) and monotonically decreasing bias (blue line). Image credit Yang et al. (2020).

Such models often fit their training data perfectly (Belkin et al., 2019; Zhang et al., 2021), i.e. they *interpolate* the data. If we consider this in the context of Equation 16, we see that:

$$\textbf{variance} \;=\; \begin{bmatrix} \textit{estimation} \\ \textit{variance} \end{bmatrix} \;+\; \underbrace{\begin{bmatrix} \textit{optimisation} \\ \textit{error} \end{bmatrix}}_{\approx\, \mathbf{0}\ \textbf{for interpolating models}}.$$

i.e., the optimization error is close to zero. This observed 'peaking' variance must therefore be primarily due to the *estimation variance*, $\mathcal{E}_{est(v)}$. Furthermore, very deep models are likely to be able to fit any function, i.e. their approximation error is zero. In these scenarios, the *only terms* remaining in the expected risk are $\mathcal{E}_{est(b)}$ and $\mathcal{E}_{est(v)}$. *Why* such models can push training error to zero, even on random labels, and still generalise well, remains an open question for modern machine learning (Zhang et al., 2021). Overall, we believe this warrants further study in the context of deep models.

4.4 What if a bias-variance decomposition doesn't hold?

Our goal has been to build a 'bridge' between the bias-variance decomposition, and the approximation-estimation decomposition. So far, we have considered this for a restricted class of losses, Bregman divergences, where we know a bias-variance decomposition holds. However, as mentioned, a bias-variance decomposition in the form of Equation 11 does not hold for all losses. One side of our 'bridge' seems to be missing.

James & Hastie (1997) present an alternative decomposition, for *any loss*, which links neatly with our results. In the special case of a loss with a bias-variance decomposition, their decomposition is equivalent. The key observation is to distinguish the *measurement* of variance, from its *effect* on the expected risk. Similarly, they distinguish two terms for the bias: the *measurement*, and its *effect* on the expected risk. The measurement

and the effect of each are not necessarily the same numerical quantity. In their own words: *"This double role of both bias and variance is so automatic that we often fail to consider it"*. The *measurement* is considered to be the 'natural' form for the terms, as in Equation 13. They then proceed to define the *effect* of the terms: bias-*effect* and variance-*effect*. When averaged over $P(\boldsymbol{x}, y)$, for *any* loss, these are:

$$\textbf{bias-effect} := R(\mathring{f}) - R(\boldsymbol{y}^*), \tag{18}$$

$$\textbf{variance-effect} := \mathbb{E}_D\Big[R(\hat{f}) - R(\mathring{f})\Big]. \tag{19}$$

These quantify the *effect on the risk* of using one predictor versus another. The *bias-effect* is the change in risk, for the centroid model versus the Bayes model. A link to our results is apparent—the bias-effect term is simply the excess risk of the centroid model. The *variance-effect* is defined similarly: the change in risk for a model $\hat{f}$ versus the centroid model, averaged over the distribution of D. *For losses where a bias-variance decomposition holds, the measurement is equal to the effect.* For example, with squared loss, the bias-effect is equal to the bias, $\mathbb{E}_{\boldsymbol{x}y}[(\mathbb{E}_D[\hat{f}] - y^*)^2]$. For the 0/1 loss, this is not the case. However, they observe that with these definitions, *for any loss*, we have the decomposition:

$$\underbrace{\mathbb{E}_D\Big[R(\hat{f})\Big]}_{\textbf{expected risk}} = \underbrace{R(\boldsymbol{y}^*)}_{\textbf{noise}} + \underbrace{R(\mathring{f}) - R(\boldsymbol{y}^*)}_{\textbf{bias-effect}} + \underbrace{\mathbb{E}_D[R(\hat{f}) - R(\mathring{f})]}_{\textbf{variance-effect}}. \tag{20}$$

If the loss is a Bregman divergence, Equation 20 reduces to Equation 11. We can relate the terms above to the approximation-estimation decomposition, using the same overall strategy as before.

Proposition 2 (Bias/Variance Effects, in terms of Approximation-Estimation) *For any loss ℓ, we have the following decomposition of the bias-effect and variance-effect.*

$$\underbrace{R(\mathring{f}) - R(\boldsymbol{y}^*)}_{\textbf{bias-effect}} = \underbrace{R(f^*) - R(\boldsymbol{y}^*)}_{\textbf{approximation error}} + \underbrace{R(\mathring{f}) - R(f^*)}_{\textbf{estimation bias}}, \tag{21}$$

$$\underbrace{\mathbb{E}_D[R(\hat{f}) - R(\mathring{f})]}_{\textbf{variance-effect}} = \underbrace{\mathbb{E}_D\left[R(\hat{f}) - R(\hat{f}_{erm})\right]}_{\textbf{optimisation error}} + \underbrace{\mathbb{E}_D\left[R(\hat{f}_{erm}) - R(\mathring{f})\right]}_{\textbf{estimation variance}}. \tag{22}$$

And the full relation to our earlier observations can be illustrated as follows.

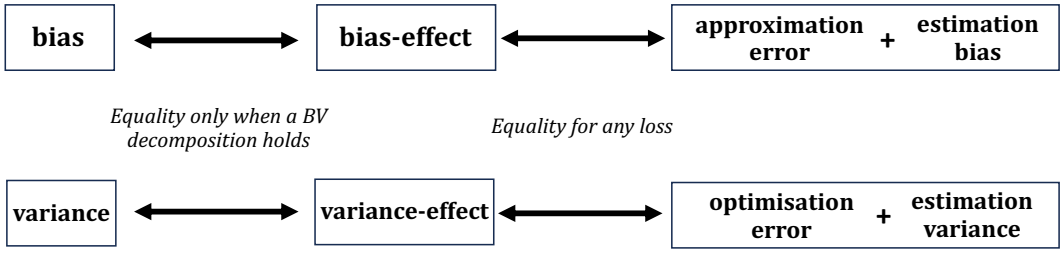

Figure 7: Relations between several decompositions that we have considered in this work.

It is notable that James & Hastie (1997) also use the concept of the centroid for an arbitrary loss, i.e. Definition 1. For 0/1 loss, the centroid prediction is the *modal* value of the predictions. Taking the mode is equivalent to a plurality/majority vote across the distribution of predictions from $\hat{f}$. Weighted voting classifiers are well-studied, e.g., Boosting (Schapire, 2003). In this context it is well-appreciated that a voted combination of weak (half-plane linear) models results a *non-linear* decision boundary. This implies $\mathring{f} \notin \mathcal{F}$, and thus again it is possible for estimation bias to be negative. Further characterisation of the terms in Figure 7, for the general case of any loss, would therefore be desirable.

5 Conclusions

We analysed the precise connections between two seminal results: the bias-variance decomposition, and the approximation-estimation decomposition. Perhaps the most surprising aspect of this work was that it had not been explored before—two such foundational ideas, not previously connected. On a literature review, we found numerous sources incorrectly stating the two were equivalent, or related as a special case / general case. This is false. The true relation, given by Theorem 1, is more intricate, and yielded interesting novel observations that we detailed, including links to the phenomenon of double descent in deep learning. We focused on Bregman divergences, but also briefly considered the case of general losses, where a bias-variance decomposition does not hold, e.g., 0/1 loss. In this case the geometry of such losses is not well-understood, leaving several open issues. In all cases, the *centroid model* (Definition 3), turned out to be a key mathematical object in bridging the decompositions. We conjecture that further study of this object, and its role in generalisation, may yield yet deeper and interesting insights.

References

Shun-ichi Amari. Information geometry and its applications: Convex function and dually flat manifold. In *LIX Fall Colloquium on Emerging Trends in Visual Computing*, pp. 75–102. Springer, 2008.

Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005a.

Arindam Banerjee, Srujana Merugu, Inderjit S Dhillon, Joydeep Ghosh, and John Lafferty. Clustering with bregman divergences. *Journal of machine learning research*, 6(10), 2005b.

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

Léon Bottou. *In Hindsight: Doklady Akademii Nauk SSSR, 181(4), 1968*, pp. 3–5. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013. ISBN 978-3-642-41136-6. doi: 10.1007/978-3-642-41136-6_1. URL https://doi.org/10.1007/978-3-642-41136-6_1.

Léon Bottou and Olivier Bousquet. The tradeoffs of large scale learning. *Advances in neural information processing systems*, 20, 2007.

Lev M Bregman. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR Comp. Mathematics*, 7(3):200–217, 1967.

Yehuda Dar, Vidya Muthukumar, and Richard G Baraniuk. A farewell to the bias-variance tradeoff? an overview of the theory of overparameterized machine learning. *arXiv preprint arXiv:2109.02355*, 2021.

Stuart Geman, Elie Bienenstock, and René Doursat. Neural networks and the bias/variance dilemma. *Neural computation*, 4(1):1–58, 1992.

Pierre Geurts. *Contributions to decision tree induction: bias/variance tradeoff and time series classification*. PhD thesis, University of Liège Belgium, 2002.

Trevor Hastie and Robert Tibshirani. Generalized Additive Models. *Statistical Science*, 1(3):297 – 310, 1986. doi: 10.1214/ss/1177013604. URL https://doi.org/10.1214/ss/1177013604.

Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, volume 12th printing, January 13th 2017. Springer, 2017.

Tom Heskes. Bias/variance decompositions for likelihood-based estimators. *Neural Computation*, 10(6): 1425–1433, 1998.

Gareth James and Trevor Hastie. Generalizations of the bias/variance decomposition for prediction error. Technical report, Dept. Statistics, Stanford Univ., Stanford, CA, 1997.

Vaishnavh Nagarajan and J Zico Kolter. Uniform convergence may be unable to explain generalization in deep learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Preetum Nakkiran. More data can hurt for linear regression: Sample-wise double descent. *arXiv preprint arXiv:1912.07242*, 2019.

Brady Neal, Sarthak Mittal, Aristide Baratin, Vinayak Tantia, Matthew Scicluna, Simon Lacoste-Julien, and Ioannis Mitliagkas. A modern take on the bias-variance tradeoff in neural networks. *arXiv preprint arXiv:1810.08591*, 2018.

Frank Nielsen and Richard Nock. Sided and symmetrized bregman centroids. *IEEE transactions on Information Theory*, 55(6):2882–2904, 2009.

David Pfau. A Generalized Bias-Variance Decomposition for Bregman Divergences. Technical report, Columbia University, 2013.

Robert E Schapire. The boosting approach to machine learning: An overview. *Nonlinear estimation and classification*, pp. 149–171, 2003.

Vladimir Vapnik. *The nature of statistical learning theory*. Springer science & business media, 1999.

Vladimir Vapnik and Alexey Chervonenkis. *Theory of pattern recognition*. Nauka, Moscow, 1974.

Ulrike Von Luxburg and Bernhard Schölkopf. Statistical learning theory: Models, concepts, and results. In *Handbook of the History of Logic*, volume 10, pp. 651–706. Elsevier, 2011.

Daniela Witten. Twitter thread: *The Bias-Variance Trade-Off & "DOUBLE DESCENT"*, 2020. URL https://x.com/daniela_witten/status/1292293102103748609. Posted 3.54am, 9th August, 2020.

David H Wolpert. On bias plus variance. *Neural Computation*, 9(6):1211–1243, 1997.

Zitong Yang, Yaodong Yu, Chong You, Jacob Steinhardt, and Yi Ma. Rethinking bias-variance trade-off for generalization of neural networks. In *International Conf. on Machine Learning*, 2020.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Appendix

A Proofs of Theorems

A.1 Proof of Theorem 1 (Bias-Variance in terms of Approximation-Estimation).

We wish to prove the following statements:

$$\underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\ell(\boldsymbol{y}^*, \mathring{f}_\phi(\boldsymbol{x}))\right]}_{\textbf{bias}} = \underbrace{R(f^*) - R(\boldsymbol{y}^*)}_{\textbf{approximation error}} + \underbrace{R(\mathring{f}_\phi) - R(f^*)}_{\textbf{estimation bias}} \tag{23}$$

$$\underbrace{\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_D\left[\ell(\mathring{f}_\phi(\boldsymbol{x}), \hat{f}(\boldsymbol{x}))\right]\right]}_{\textbf{variance}} = \underbrace{\mathbb{E}_D\left[R(\hat{f}) - R(\hat{f}_{erm})\right]}_{\textbf{optimisation error}} + \underbrace{\mathbb{E}_D\left[R(\hat{f}_{erm}) - R(\mathring{f}_\phi)\right]}_{\textbf{estimation variance}} \tag{24}$$

To show Equation 23, we note that the $R(f^*)$ terms cancel, so we just need to prove:

$$\mathbb{E}_{\boldsymbol{x}}\left[\ell(\boldsymbol{y}^*, \mathring{f}_\phi(\boldsymbol{x}))\right] = R(\mathring{f}_\phi) - R(\boldsymbol{y}^*). \tag{25}$$

The proof below builds on the *Bregman 3-point property* (Nielsen & Nock, 2009).

Definition (Bregman three-point identity) *The Bregman three-point property states, for any p, q, r,*

$$B_\phi(p, r) = B_\phi(p, q) + B_\phi(q, r) + \langle\ p - q\ ,\ \nabla\phi(q) - \nabla\phi(r)\ \rangle \tag{26}$$

We then have the following, where we apply the three-point property to $\boldsymbol{y}, \mathring{f}_\phi$, with $\boldsymbol{y}^*$ as the mid-point.

$$B_\phi(\boldsymbol{y}, \mathring{f}_\phi) = B_\phi(\boldsymbol{y}, \boldsymbol{y}^*) + B_\phi(\boldsymbol{y}^*, \mathring{f}_\phi) + \langle\ \boldsymbol{y} - \boldsymbol{y}^*\ ,\ \nabla\phi(\boldsymbol{y}^*) - \nabla\phi(\mathring{f}_\phi)\ \rangle \tag{27}$$

Take the expected value w/r $p(y|\boldsymbol{x})$ and the inner product term vanishes, since $\boldsymbol{y}^* = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{x}}[y]$. Rearranging terms and further taking expectation w/r $\boldsymbol{x}$, we recover:

$$R(\mathring{f}_\phi) - R(\boldsymbol{y}^*) \quad = \quad \mathbb{E}_{\boldsymbol{x}}\left[B_\phi(\boldsymbol{y}^*, \mathring{f}_\phi)\right] \tag{28}$$

which is the desired result, proving Equation 23.

To show Equation 24, we follow a similar pattern. Take the 3-point property for $\boldsymbol{y}, \hat{f}$ with $\mathring{f}_\phi$ as the mid-point.

$$B_\phi(\boldsymbol{y}, \hat{f}) = B_\phi(\boldsymbol{y}, \mathring{f}_\phi) + B_\phi(\mathring{f}_\phi, \hat{f}) + \langle\ \boldsymbol{y} - \mathring{f}_\phi\ ,\ \nabla\phi(\mathring{f}_\phi) - \nabla\phi(\hat{f})\ \rangle \tag{29}$$

Take the expected value w/r D and the inner product term vanishes, since $\nabla\phi(\mathring{f}_\phi) = \mathbb{E}_D\left[\nabla\phi(\hat{f})\right]$.

Rearranging terms and further taking expectation over $p(\boldsymbol{x})$, we recover:

$$\mathbb{E}_D\left[R(\hat{f}) - R(\mathring{f}_\phi)\right] \quad = \quad \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_D\left[B_\phi(\mathring{f}_\phi, \hat{f})\right]\right] \tag{30}$$

which is the desired result, completing the theorem.

■

Special case of Theorem 1 for squared loss. The following presents the special case of squared loss, included for didactic purposes due to its ubiquity and links to the results for linear models. We wish to prove the following statements:

$$\mathbb{E}_{\boldsymbol{x}}\left[(\mathbb{E}_D[\hat{f}(\boldsymbol{x})] - \mathbb{E}_{y|\boldsymbol{x}}[y])^2\right] = \mathcal{E}_{app} + \mathcal{E}_{est(b)} \tag{31}$$

$$\mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_D[(\hat{f}(\boldsymbol{x}) - \mathbb{E}_D[\hat{f}(\boldsymbol{x})])^2]\right] = \mathcal{E}_{opt} + \mathcal{E}_{est(v)} \tag{32}$$

$$\mathbb{E}_{\boldsymbol{x}y}[(y - \mathbb{E}_{y|\boldsymbol{x}}[y])^2] = R(y^*) \tag{33}$$

To show Equation 33, we simply note that $y^* = \mathbb{E}_{y|\boldsymbol{x}}[y]$, so the expression is true by definition.

To show Equation 31 we note, as an intermediate step, that:

$$\mathcal{E}_{app} + \mathcal{E}_{est(b)} = \Big(R(f^*) - R(y^*)\Big) + \Big(R(\mathbb{E}_D[\hat{f}]) - R(f^*)\Big) = R(\mathbb{E}_D[\hat{f}]) - R(y^*). \tag{34}$$

We then have the following, again using the definition of y^*.

$$\begin{aligned}
R(\mathbb{E}_D[\hat{f}]) - R(y^*) &= \mathbb{E}_{\boldsymbol{x}y}\left[(\mathbb{E}_D[\hat{f}] - y)^2\right] - \mathbb{E}_{\boldsymbol{x}y}\left[(y - \mathbb{E}_{y|\boldsymbol{x}}[y])^2\right]. \\
&= \mathbb{E}_{\boldsymbol{x}y}\left[\left(\mathbb{E}_D[\hat{f}]\right)^2 - 2y\mathbb{E}_D[\hat{f}] - \mathbb{E}_{y|\boldsymbol{x}}[y]^2 + 2y\mathbb{E}_{y|\boldsymbol{x}}[y]\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\left(\mathbb{E}_D[\hat{f}]\right)^2 - 2\mathbb{E}_{y|\boldsymbol{x}}[y]\mathbb{E}_D[\hat{f}(\boldsymbol{x})] - \mathbb{E}_{y|\boldsymbol{x}}[y]^2 + 2\mathbb{E}_{y|\boldsymbol{x}}[y]^2\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\left(\mathbb{E}_D[\hat{f}]\right)^2 - 2\mathbb{E}_{y|\boldsymbol{x}}[y]\mathbb{E}_D[\hat{f}(\boldsymbol{x})] + \mathbb{E}_{y|\boldsymbol{x}}[y]^2\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[(\mathbb{E}_D[\hat{f}] - \mathbb{E}_{y|\boldsymbol{x}}[y])^2\right],
\end{aligned}$$

which is the bias, and the desired result.

To show Equation 32, we follow a similar pattern. From definitions:

$$\mathcal{E}_{opt} + \mathcal{E}_{est(v)} = \mathbb{E}_D\left[R(\hat{f}) - R(\hat{f}_{erm})\right] + \mathbb{E}_D\left[R(\hat{f}_{erm}) - R(\mathbb{E}_D[\hat{f}])\right] = \mathbb{E}_D\left[R(\hat{f}) - R(\mathbb{E}_D[\hat{f}])\right]. \tag{35}$$

We then have the following.

$$\begin{aligned}
\mathbb{E}_D\left[R(\hat{f}) - R(\mathbb{E}_D[\hat{f}])\right] &= \mathbb{E}_D\left[\mathbb{E}_{\boldsymbol{x}y}\left[(\hat{f} - y)^2\right] - \mathbb{E}_{\boldsymbol{x}y}\left[(\mathbb{E}_D[\hat{f}] - y)^2\right]\right] \\
&= \mathbb{E}_D\left[\mathbb{E}_{\boldsymbol{x}y}\left[\hat{f}^2 - 2y\hat{f} - \mathbb{E}_D[\hat{f}]^2 + 2y\mathbb{E}_D[\hat{f}]\right]\right] \\
&= \mathbb{E}_{\boldsymbol{x}y}\left[\mathbb{E}_D[\hat{f}^2] - 2y\mathbb{E}_D[\hat{f}] - \mathbb{E}_D[\hat{f}]^2 + 2y\mathbb{E}_D[\hat{f}]\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_D[\hat{f}^2] - \mathbb{E}_D[\hat{f}]^2\right] \\
&= \mathbb{E}_{\boldsymbol{x}}\left[\mathbb{E}_D[(\hat{f} - \mathbb{E}_D[\hat{f}])^2]\right]
\end{aligned}$$

where the final step is the standard definition of variance, giving the desired result.
■

A.2 Proof of Theorem 2 (Sufficient condition for a non-negative estimation bias).

To prove Theorem 2, we demonstrate that under a certain condition, $\mathring{f} \in \mathcal{F}$, which implies $R(\mathring{f}) \geq R(f^*)$, and therefore $R(\mathring{f}) - R(f^*) \geq 0$. We use the following definition, due to Amari (2008).

Definition 4 (Dual convex set) *Let ϕ be a strictly convex function. A set $\mathcal{F}$ is dually convex with respect to ϕ iff, for any pair of points $f, g \in \mathcal{F}$ and for all $\lambda \in [0, 1]$*

$$\lambda \nabla\phi(f) + (1 - \lambda)\nabla\phi(g) \in \mathcal{F}$$

i.e. the set $\mathcal{F}$ is dual-convex iff it is convex in its dual coordinate representation.

An arbitrary set $\mathcal{C}$ is convex iff for any random variable X defined over elements of $\mathcal{C}$, its expectation is also in $\mathcal{C}$, i.e. $\mathbb{E}[X] \in \mathcal{C}$.

Therefore, for a dual convex set $\mathcal{F}$, we have that the point $\mathbb{E}_D[\nabla\phi(f)] \in \mathcal{F}$. The primal coordinate representation of this point, $\nabla\phi^{-1}(\mathbb{E}_D[\nabla\phi(f)])$, is also a member of $\mathcal{F}$, i.e. $\mathring{f} \in \mathcal{F}$, proving the theorem.
■

A.3 Proof of Theorem 3 (GLMs have non-negative estimation bias).

We demonstrate that $\mathcal{E}_{est(b)} \geq 0$ if $\hat{f}$ is a GLM of a particular form. We give two proofs: a direct one and one that makes use of Theorem 2.

Direct proof. The estimation bias is defined:

$$\mathcal{E}_{est(b)} = R(\mathring{f}) - R(f^*). \tag{36}$$

This involves the definition of the centroid prediction, which for a Bregman divergence is,

$$\mathring{f}_\phi(\boldsymbol{x}) := [\nabla\phi]^{-1}\left(\mathbb{E}_D\left[\nabla\phi(\hat{f}(\boldsymbol{x}))\right]\right). \tag{37}$$

Given a Bregman divergence with generator ϕ, define $\mathcal{F}$ as the class of GLMs with inverse link $[\nabla\phi]^{-1}$, parameterised by $\boldsymbol{\theta} \in \mathbb{R}^d$. In this case, each $\hat{f} \in \mathcal{F}$ takes the form:

$$\hat{f}(\boldsymbol{x}) := [\nabla\phi]^{-1}\left(\boldsymbol{\theta}^T\boldsymbol{x}\right), \tag{38}$$

where $\boldsymbol{\theta}$ are the natural parameters. Substituting this into the centroid prediction gives us,

$$\begin{aligned}\mathring{f}_\phi(\boldsymbol{x}) &= [\nabla\phi]^{-1}\left(\mathbb{E}_D\left[\nabla\phi\left([\nabla\phi]^{-1}\left(\boldsymbol{\theta}^T\boldsymbol{x}\right)\right)\right]\right), \\ &= [\nabla\phi]^{-1}\left(\mathbb{E}_D\left[\boldsymbol{\theta}\right]^T\boldsymbol{x}\right).\end{aligned} \tag{39}$$

Since $\mathbb{E}_D[\boldsymbol{\theta}]$ is within the convex hull of the distribution of $\boldsymbol{\theta}$ induced by D, the centroid prediction is the same form of GLM as $\hat{f}(\boldsymbol{x})$, for all $\boldsymbol{x}$, and therefore the centroid model $\mathring{f}_\phi \in \mathcal{F}$. Then, since by definition f^* is the risk minimizer in $\mathcal{F}$, we must have that $R(\mathring{f}_\phi) \geq R(f^*)$, and therefore Equation 36 is non-negative.
■

Proof using Theorem 2. To show that the estimation bias is non-negative, it suffices to show that the class of GLMs of a particular form is *dually-convex*. We verify that the property of dual-convexity holds. Define $\mathcal{F}$ = GLMs with inverse link $\nabla\phi^{-1}$.

By definition, if $f \in \mathcal{F}$, it is parameterised by a vector $\boldsymbol{\theta}$ as follows: $f(\boldsymbol{x}) = \nabla\phi^{-1}(\boldsymbol{\theta}^T\boldsymbol{x})$.

Let h be the function, expressed in primal coordinates, corresponding to the convex combination of two arbitrary GLMs in their dual coordinates, i.e. $h = \nabla\phi^{-1}(\lambda\nabla\phi(f) + (1-\lambda)\nabla\phi(g))$, with $\lambda \in [0,1]$, and with f and g two GLMs $f = \nabla\phi^{-1}(\boldsymbol{\theta}^T\boldsymbol{x})$ and $g = \nabla\phi^{-1}(\boldsymbol{\xi}^T\boldsymbol{x})$. We need to show that $h \in \mathcal{F}$. But

$$\begin{aligned}
h &= \nabla\phi^{-1}(\lambda\nabla\phi(f) + (1-\lambda)\nabla\phi(g))) \\
&= \nabla\phi^{-1}(\lambda\nabla\phi(\nabla\phi^{-1}(\boldsymbol{\theta}^T\boldsymbol{x})) + (1-\lambda)\nabla\phi(\nabla\phi^{-1}(\boldsymbol{\xi}^T\boldsymbol{x})))) \\
&= \nabla\phi^{-1}(\lambda\boldsymbol{\theta}^T\boldsymbol{x} + (1-\lambda)\boldsymbol{\xi}^T\boldsymbol{x}) \\
&= \nabla\phi^{-1}((\lambda\boldsymbol{\theta}^T + (1-\lambda)\boldsymbol{\xi}^T)\boldsymbol{x})
\end{aligned}$$

which is again a GLM in $\mathcal{F}$.
■

B Discussion of related work by Hastie et al, 2017

We detail related observations by Hastie et al. (2017), who assume a linear model with squared loss, i.e. $\ell(y, \hat{f}(\boldsymbol{x})) = (y - \hat{\boldsymbol{\theta}}^T\boldsymbol{x})^2$. The optimal parameters are $\boldsymbol{\theta}_* := \arg\min_{\hat{\boldsymbol{\theta}}} \mathbb{E}_{\boldsymbol{x}y}[(y - \hat{\boldsymbol{\theta}}^T\boldsymbol{x})^2] = \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^T]^{-1}\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}y^*]$, and the Bayes model is $y^* = \arg\min_z \mathbb{E}_{y|\boldsymbol{x}}[(y-z)^2] = \mathbb{E}_{y|\boldsymbol{x}}[y]$. In this case, the bias-variance decomposition is,

$$\underbrace{\mathbb{E}_D\Big[\mathbb{E}_{\boldsymbol{x}y}\Big[(y - \hat{\boldsymbol{\theta}}^T\boldsymbol{x})^2\Big]\Big]}_{expected\ risk} = \underbrace{\mathbb{E}_{\boldsymbol{x}y}\Big[(y-y^*)^2\Big]}_{noise} + \underbrace{\mathbb{E}_{\boldsymbol{x}}\Big[(y^* - \mathbb{E}_D[\hat{\boldsymbol{\theta}}^T\boldsymbol{x}])^2\Big]}_{bias} + \underbrace{\mathbb{E}_{\boldsymbol{x}}\Big[\mathbb{E}_D\Big[(\hat{\boldsymbol{\theta}}^T\boldsymbol{x} - \mathbb{E}_D[\hat{\boldsymbol{\theta}}^T\boldsymbol{x}])^2\Big]\Big]}_{variance}. \tag{40}$$

Hastie et al. (2017, Eq 7.14) show that the bias decomposes more finely:

$$\mathbb{E}_{\boldsymbol{x}}\Big[(y^* - \mathbb{E}_D[\hat{\boldsymbol{\theta}}^T\boldsymbol{x}])^2\Big] = \mathbb{E}_{\boldsymbol{x}}\Big[(y^* - \boldsymbol{\theta}_*^T\boldsymbol{x})^2\Big] + \mathbb{E}_{\boldsymbol{x}}\Big[(\boldsymbol{\theta}_*^T\boldsymbol{x} - \mathbb{E}_D[\hat{\boldsymbol{\theta}}^T\boldsymbol{x}])^2\Big]. \tag{41}$$

Hastie et al. describe this expression as:

> *"The first term on the right-hand side is the average squared model bias, the error between the best-fitting linear approximation and the true function. The second term is the average squared estimation bias, the error between the average estimate [..] and the best-fitting linear approximation."*

The first point to note is that Hastie et al. deal only with a decomposition for squared loss. When referring to bias, they use nomenclature 'squared bias', whereas in fact the square is an artefact of using squared loss, and is not present in the general case.

When they refer to the error between "the best-fitting linear approximation and the true function", we note that this is precisely a description of the *approximation error* for a linear model. This is no coincidence. To see the connection precisely, we note the following property of the approximation error.

Theorem 4 *For a loss ℓ, if a bias-variance decomposition holds, then the approximation error $R(f^*) - R(y^*)$ simplifies as follows.*

$$\begin{aligned}
\mathcal{E}_{app} &= R(f^*) - R(y^*) \\
&= \mathbb{E}_{\boldsymbol{x}y}[\ell(y^*, f^*(\boldsymbol{x})].
\end{aligned} \tag{42}$$

i.e. the difference-of-risks is equal to the divergence between the two models themselves.

Proof sketch. Use the 3-point theorem in exactly the same manner as in the proof of Theorem 1, i.e. between y, f^* with y^* as the mid-point, then take expectation successively over $P(y|\boldsymbol{x})$ then $P(\boldsymbol{x})$. ■

For the case of squared loss, this yields the term from Hastie et al, shown above in Equation 41, i.e.,

$$R(\boldsymbol{\theta}_*^T \boldsymbol{x}) - R(y^*) = \mathbb{E}_{\boldsymbol{x}}\Big[(y^* - \boldsymbol{\theta}_*^T \boldsymbol{x})^2\Big]. \tag{43}$$

Overall, this shows that Hastie et al. (2017, Eq 7.14) is the special case of our Equation 15 for squared loss/linear models.

Estimation bias: The second term on the right of equation 41 is described as the error between the expected model and the best-fitting linear approximation. This is equivalent to the standard definition of estimation bias, but for the specific case of a linear model and squared loss, i.e.

$$\underbrace{R(\mathbb{E}_D[\hat{\boldsymbol{\theta}}^T \boldsymbol{x}]) - R(\boldsymbol{\theta}_*^T \boldsymbol{x})}_{estimation\ bias} \;=\; \mathbb{E}_{\boldsymbol{x}}\Big[(\boldsymbol{\theta}_*^T \boldsymbol{x} - \mathbb{E}_D[\hat{\boldsymbol{\theta}}^T \boldsymbol{x}])^2\Big]. \tag{44}$$

However, the squared loss seems to be unique in that Equation 44 holds. This is a consequence of taking an expectation over $\boldsymbol{x}$, and the properties of the OLS solution. In the general Bregman case we have an inequality:

$$\underbrace{R(\mathring{f}_\phi) - R(f^*)}_{estimation\ bias} \;\neq\; \mathbb{E}_{\boldsymbol{x}}\Big[B_\phi(f^*(\boldsymbol{x}), \mathring{f}_\phi(\boldsymbol{x}))\Big]. \tag{45}$$

Hastie et al. observed that in unregularized linear models, the estimation bias will[4] be zero. With ridge regression, $\boldsymbol{\theta}_* := \mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}\boldsymbol{x}^T + \lambda]^{-1}\mathbb{E}_{\boldsymbol{x}}[\boldsymbol{x}y^*]$, and thus the estimation bias will be non-negative for $\lambda > 0$. Since the OLS solution is closed-form, $\mathcal{E}_{opt} = 0$, and the estimation variance is simply $Var(\hat{\boldsymbol{\theta}}^T \boldsymbol{x})$.

Hastie et al. (2017, Figure 7.2) also briefly alludes to the idea of estimation *variance*—from this we assume that Hastie *et al.* were well aware of these terms in the context of squared loss / linear models. However, the *difference-of-risks* formulation that we use generalises these ideas to any model family, and any Bregman divergence.

[4] Assuming the Gauss-Markov conditions hold.

C Experimental details

We summarise our methodology to generate the illustrative experiments shown in the paper. **For the purposes of anonymous submission, an outline is provided below. For the final submission we will supply all code for reproducible research.**

We use a synthetic 1-d problem: $x \in [0, 15]$, and the true label is $y = x + 5\sin(2x) + \epsilon$, where ϵ is Gaussian noise with zero mean and $\sigma = 3$. Training data is $n = 100$ points, illustrated below.

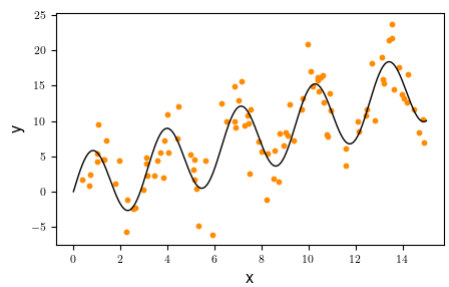

Figure 8: Synthetic problem for experiments.

Since this is a regression problem, $\ell(y, f(\boldsymbol{x})) = (y - f(\boldsymbol{x}))^2$, and $\mathring{f}(\boldsymbol{x}) := \mathbb{E}_D[\hat{f}(\boldsymbol{x})]$.

The function class $\mathcal{F}$ is defined as the set of all trained models obtained over $T = 1000$ independently sampled datasets, each of size $n = 100$. The best-in-class model is the minimum across the T trials:

$$f^* := \arg\min_D \hat{R}(\hat{f}_D) \tag{46}$$

where the risk $R(f)$ is approximated by sample of uniformly sampled points at a resolution of 0.001, giving a total of $n = 15,000$ test points. To simplify analysis, we assume $\hat{f} = \hat{f}_{erm}$.