

On the Bias-Variance Tradeoff: Textbooks Need an Update

A Modern Perspective

Posted on January 5, 2020

On the Bias-Variance Tradeoff: Textbooks Need an Update ()

TL;DR and Preamble

The Bias-Variance Tradeoff

Motivation for Revisiting this Tradeoff

Test Error Decreasing with Network Width

Red Flags in Geman et al. (1992)

Test Error Analysis is Not Sufficient

The Lack of a Tradeoff

Main Result

Why does variance decrease?

Double Descent Curve

Suggested Updates to Textbooks, Courses, and our Mental Models

Concluding Thoughts

References

TL;DR: It is not always necessary to trade bias for variance when increasing model complexity.

Preamble: the charge of the electron, Millikan's experiment, and anchoring

Richard Feynman made an insightful observation about the history of measurements of the charge of the electron: Millikan's first measurement was noticeably lower than the (now known) true value. Over the next few decades, other groups produced estimates of the charge of the electron. Interestingly, almost all of them underestimated the true value (see Stack Exchange thread (<https://hsm.stackexchange.com/questions/264/timeline-of-measurements-of-the-electrons-charge>) for more info). According to Feynman's account, the subsequent measurements were biased by Millikan's original measurement. This is a specific example of the cognitive bias known as anchoring (<https://en.wikipedia.org/wiki/Anchoring>).

Anchoring causes us to use an initial piece of information as an "anchor" that draws our interpretation of later information toward the anchor. This human tendency is not necessarily bad. It makes perfect sense from a Bayesian perspective where one tries to make the best use of finite resources (Lieder et al., 2012), but it can have negative consequences. The intuitive concept of the bias-variance tradeoff that is commonly taught has acted as a strong anchor in the machine learning community. We can see it incorrectly biasing the interpretation of experiments in Geman et al. (1992)'s landmark bias-variance tradeoff paper. It's time to cut this anchor loose.

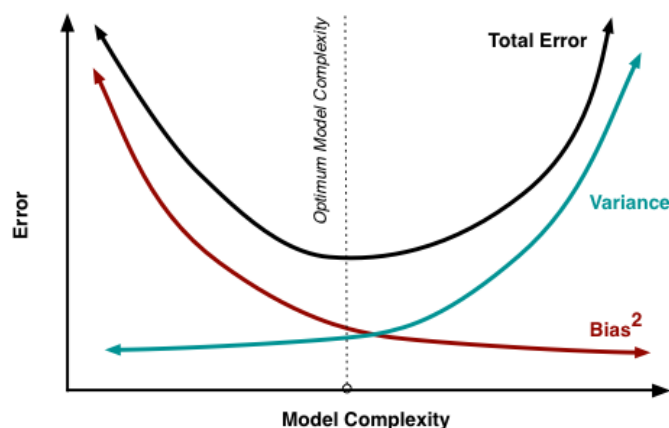
Note: Because this is a blog post, it will be light on details. For many more details, please see the featured papers: On the Bias-Variance Tradeoff: Textbooks Need an Update (MSc Thesis) (<https://arxiv.org/abs/1912.08286>) (Neal, 2019) and A Modern Take on the Bias-Variance Tradeoff in Neural Networks (<https://arxiv.org/abs/1810.08591>) (Neal et al., 2018).

The Bias-Variance Tradeoff

Stuart Geman, the man who influenced Steve Jobs' uniform clothing style (according to a shaky Wikipedia contribution (https://en.wikipedia.org/w/index.php?title=Stuart_Geman&oldid=913795070#Influences)), gave us a concise statement of the bias-variance tradeoff:

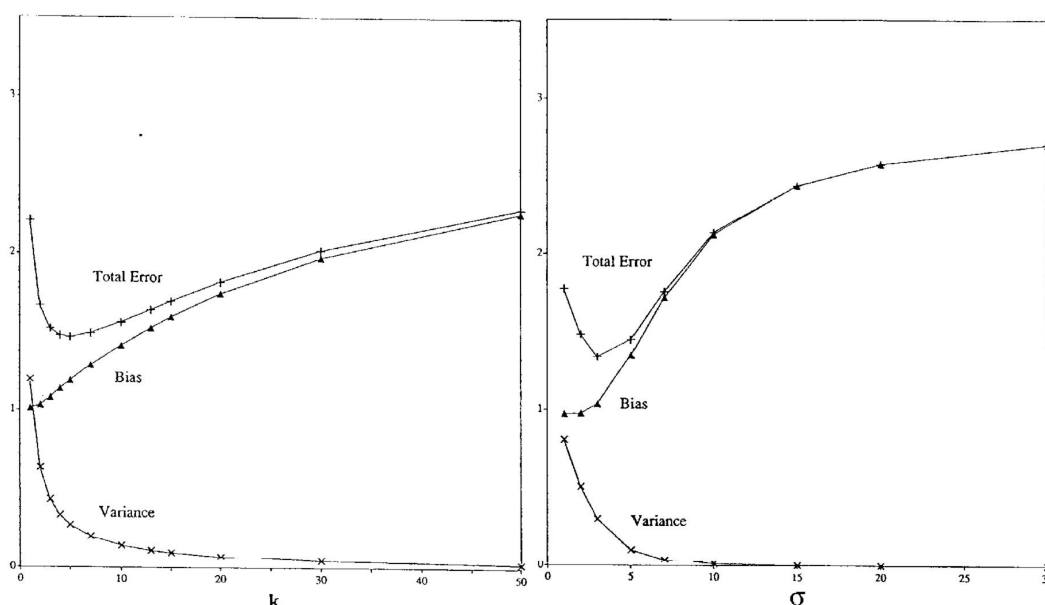
The price to pay for achieving low bias is high variance (Geman et al., 1992).

In other words, when our model is too simple, it will have high bias and low variance; in order to achieve low bias, we can increase the complexity of the model, but this will inevitably result in high variance. This concept known as the bias-variance tradeoff is commonly illustrated with this graph (Fortmann-Roe, 2012):



The U-shaped test error curve is a key consequence of the bias-variance tradeoff, as it means that there is an optimal model complexity that perfectly balances bias and variance.

There is a fair amount of **evidence for the bias-variance tradeoff** in nonparametric models. For example, there is evidence of a bias-variance tradeoff in k-nearest neighbors (when varying k) and in kernel regression (when varying kernel width σ) (Geman et al., 1992):



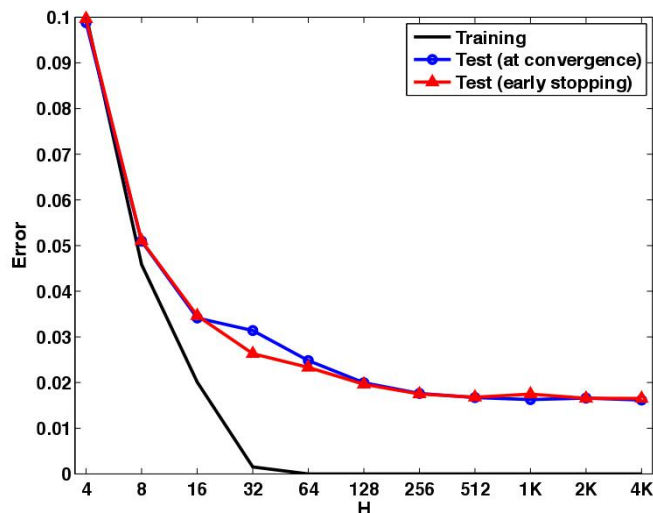
Note that model complexity decreases with increasing k and σ , so the x-axis is reversed in these two graphs compared to the usual bias-variance tradeoff graph above.

For more examples of where we see the bias-variance tradeoff and more background on why we believe the bias-variance tradeoff, see Section 3.4 “Why do we believe the Bias-Variance Tradeoff?” of my MSc thesis (<https://arxiv.org/abs/1912.08286>).

Motivation for Revisiting this Tradeoff

Test Error Decreasing with Network Width

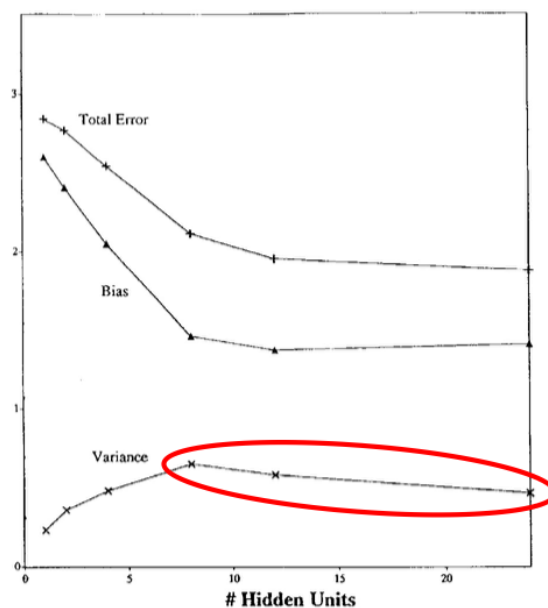
Despite the U-shape test error curve that the bias-variance tradeoff implies, Neyshabur et al. (2015) actually found that test error simply decreases with network width:



There has been a large amount of supporting evidence for this finding since 2015, with some extra detail such as the “double descent” curve, which we briefly cover further below.

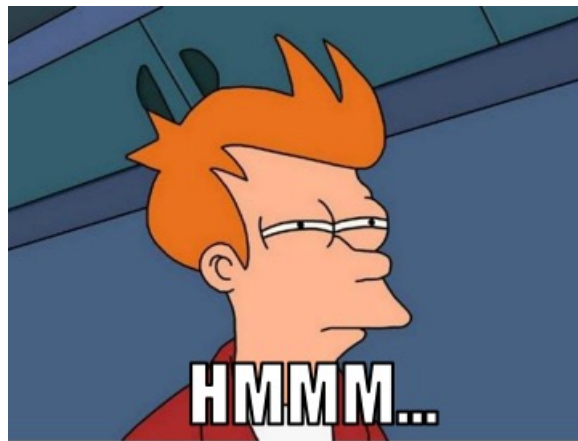
Red Flags in Geman et al. (1992)

The decreasing test error curve above can even be seen in Geman et al. (1992)’s own results. They largely focus on neural networks (their paper’s title is “Neural Networks and the Bias/Variance Dilemma (<http://www.dam.brown.edu/people/geman/Homepage/Essays%20and%20ideas%20about%20neurobiology/bias-variance.pdf>),” after all). In particular, they ran experiments, varying the number of hidden units in a single hidden layer network (network width). Here are their neural network experiments on a handwritten character recognition dataset (red emphasis is mine):



Geman et al. (1992) then maintain their claim that there is a bias-variance tradeoff in network width:

The basic trend is what we expect: bias falls and variance increases with the number of hidden units.

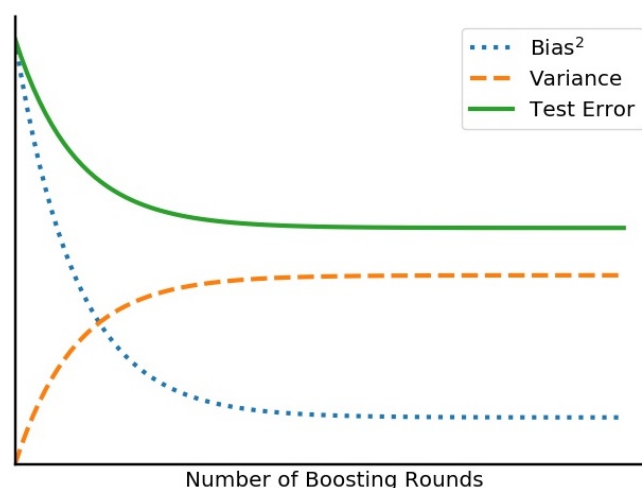


They end up noting that the “effects [of the bias-variance tradeoff] are not perfectly demonstrated” and attributing this discrepancy to convergence issues. Geman et al. (1992)’s inclusion of this result where variance decreases for the last 2/3 of the plot, despite it running counter to their claim, is great scholarship. This conflict between their result and their claim does not prevent this paper from being highly influential. This paper has over 3700 citations, over 180 of which were in 2019.

Test Error Analysis is Not Sufficient

It is time to revisit the bias-variance tradeoff in neural networks. The evidence that Geman et al. (1992) provide for their claims of a bias-variance tradeoff in neural networks is unsatisfying. The recent accumulating evidence that test error does not behave as a U-shaped curve in neural networks seems to suggest that it might not be necessary to trade bias for variance. However, decreasing test error does not imply the lack of a bias-variance tradeoff.

For a clear example where we see both decreasing test error and a tradeoff between bias and variance, consider boosting. In boosting, test error often decreases with the number of rounds (Schapire & Singer, 1999). In spite of this monotonicity in test error in boosting, Bühlmann & Yu (2003, Theorem 1) show that variance *does* grow, just at an exponentially decaying rate, calling this an “exponential bias-variance tradeoff.” We provide a visualization of this exponential bias-variance tradeoff below:

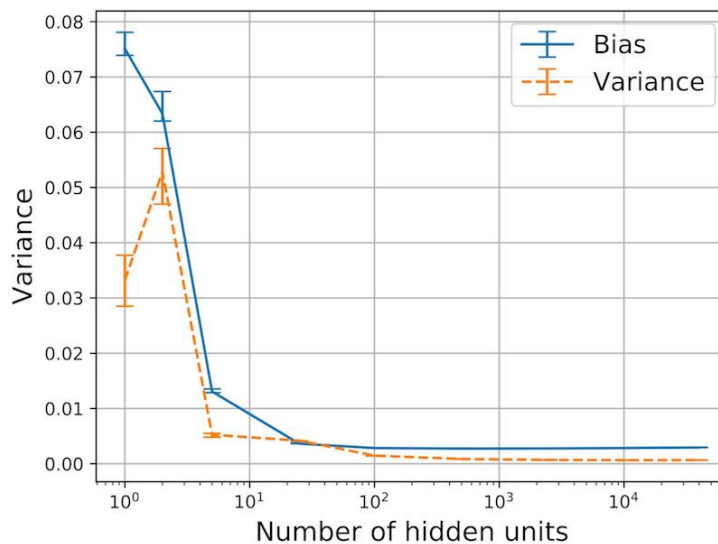


This means that a test error analysis is not enough to know if there is a bias-variance tradeoff. In the next section we address this by measuring the specific quantities of interest: bias and variance.

The Lack of a Tradeoff


Main Result

So, in 2018 we decided that it is time to revisit the bias-variance tradeoff. In contrast to traditional bias-variance tradeoff wisdom, as exemplified by Geman et al. (1992)’s claims above (1 and 2), we find that *both bias and variance* decrease with network width, in the over-parameterized regime (Neal et al., 2018):



We demonstrate these results on CIFAR10, SVHN, MNIST, small MNIST, and a sinusoid regression task. We also find these results are robust to many experimental details such as which gradient-based optimizer is used and which stopping criterion is used.

Statisticians Hate Him



Get low bias AND
LOW VARIANCE
with this one
WEIRD trick

LEARN THE TRUTH NOW

Why does variance decrease?

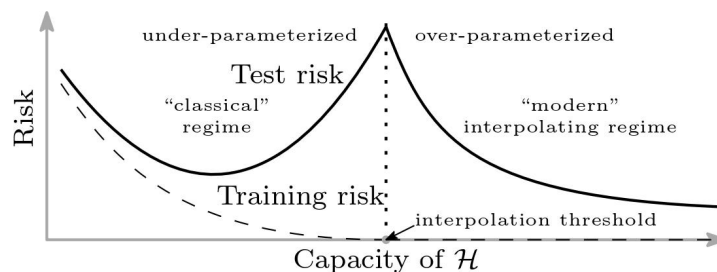
This section is an incomplete summary that is meant to give you just enough details to see if you want to read more. Please see *A Modern Take on the Bias-Variance Tradeoff in Neural Networks* (<https://arxiv.org/abs/1810.08591>) for full details.

In order to better understand why variance decreases with network width in the over-parameterized setting, we introduce a decomposition of the variance, decomposing it into *variance due to sampling* of the training set (the usual “variance” in the classic bias-variance tradeoff) and *variance due to optimization* (relevant in non-convex settings with SGD from a random initialization). We also run experiments to estimate these two terms of our variance decomposition.

We provide sufficient conditions that imply decreasing variance with number of parameters (Neal et al., 2018, Section 5). One of the main assumptions is the notion of an “intrinsic dimension” that does not grow with the number of parameters once the network is large enough (Li et al., 2018).

Double Descent Curve

There is growing evidence that test error acts as the classic U-shaped curve in the under-parameterized regime and monotonically decreases in the over-parameterized regime. Belkin et al. (2019) depict this “double descent” as follows:



Our work, concurrent to the double descent work (Spigler et al., 2018; Geiger et al., 2019; Belkin et al., 2019), focused on the over-parameterized regime. More specifically, rather than searching for a cusp in the test error, we were trying to go far enough out on the x-axis to search for an increase in variance. That said, we did observe humps in the variance, so there is probably a similar phenomenon in variance, which could be an interesting direction for future work. For more discussion on the double descent curve, please see Section 4.4 of my MSc thesis (<https://arxiv.org/abs/1912.08286>).

Suggested Updates to Textbooks, Courses, and our Mental Models

Our textbooks, classes, and the majority of the broader machine learning community's intuition still reflect the classic, rigid tradeoff, missing this newly discovered nuance. Here, we list some proposed updates to the machine learning curriculum. The benefits of these updates range from better informing practitioners on how to think about model selection to helping young machine learning theory researchers have accurate intuitions that will improve their work. These updates are taken from Section 4.5 of my MSc thesis (<https://arxiv.org/abs/1912.08286>):

1. **The bias-variance tradeoff should not be assumed to be universal.** For example, we have discussed how there is not a bias-variance tradeoff in the width of neural networks. There might be a similar phenomenon in random forests (Belkin et al., 2019).
2. **The bias-variance decomposition (https://en.wikipedia.org/wiki/Bias-variance_tradeoff#Bias-variance_decomposition_of_squared_error) does not imply a tradeoff.** If we consider a fixed test error, the decomposition does imply a tradeoff between bias and variance. However, in general, test error does not remain fixed as model complexity varies, so this implication does not hold. This lack of implication should be made explicit in textbooks and other learning resources. Currently, the decomposition is often presented in close proximity to the tradeoff as ambiguous evidence for it.
3. **Loose bounds should not be interpreted as necessarily predictive of the quantities they bound,** unless they have been shown to be. For example, upper bounds on the generalization gap, variance, or estimation error often grow with some measure of model complexity. It is not uncommon to interpret these bounds as confirming our intuition that the bounded quantities (e.g. generalization gap) will grow with model complexity, which supports the bias-variance tradeoff hypothesis. However, many of these bounds are actually *negatively* correlated with the quantities they are bounding in neural networks (Neyshabur et al., 2017). Of course, if these bounds were tight, these problems would be alleviated (Dziugaite & Roy, 2017).

Concluding Thoughts

While the concept of the bias-variance tradeoff is useful and accurate sometimes, especially in nonparametric models (see The Bias-Variance Tradeoff), thinking of the tradeoff as universal can be misleading. For example, we show that it is not necessary to trade bias for variance in neural networks (see Main Result). There are several updates that we can make to our textbooks, courses, and mental models that will help better inform practice and help us have accurate intuitions for research.

Although there is a lot of evidence for the double descent curve when increasing neural network width, even in architectures such as ResNets and transformers (Nakkiran et al., 2020), the double descent curve is also not universal. For example, it does not appear to be present in nonparametric models such as k-nearest neighbors and kernel regression. We do not comprehensively know in what models we see the classic U-shaped curve, the double

descent curve, or something else. Even within neural networks, there are ways to increase model complexity other than increasing network width; for example, increasing network depth is a natural one (see, e.g., Neal et al. (2018, Appendix C)). The test error curve may very well look different, depending on how model complexity is varied.

Furthermore, measuring the more specific quantities, bias and variance (as opposed to just test error), is important to support claims about the bias-variance tradeoff or lack thereof (see Test Error Analysis is Not Sufficient).

Acknowledgments

Thanks to Florent Krzakala (<https://florentkrzakala.com>) for pointing out the connection to the historical measurements of the charge of the electron. Thanks to Ioannis Mitliagkas (<http://mitliagkas.github.io>) for recommending I put this connection in the preamble, for recommendations on the general ordering of content, and for many edits. Thanks to Xavier Bouthillier (<https://mila.quebec/en/person/xavier-bouthillier/>) and Sherjil Ozair (<https://sherjilozair.github.io>) for helpful feedback on the blog post.

References

- Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine learning practice and the classical bias-variance trade-off. *Proceedings of the National Academy of Sciences*.
- Bühlmann, P., & Yu, B. (2003). Boosting with the L2 Loss: Regression and Classification. *Journal of the American Statistical Association*.
- Dziugaite, G. K., & Roy, D. M. (2017). Computing Nonvacuous Generalization Bounds for Deep (Stochastic) Neural Networks with Many More Parameters than Training Data. *UAI 2017*.
- Fortmann-Roe, S. (2012). Understanding the Bias-Variance Tradeoff. Retrieved from <http://scott.fortmann-roe.com/docs/BiasVariance.html>
- Geiger, M., Spigler, S., d'Ascoli, S., Sagun, L., Baity-Jesi, M., Biroli, G., & Wyart, M. (2019). The jamming transition as a paradigm to understand the loss landscape of deep neural networks. *Physical Review*. E.
- Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural Networks and the Bias/Variance Dilemma. *Neural Computation*.
- Li, C., Farkhoor, H., Liu, R., & Yosinski, J. (2018). Measuring the Intrinsic Dimension of Objective Landscapes. *ICLR 2018*.
- Lieder, F., Griffiths, T., & Goodman, N. (2012). Burn-in, bias, and the rationality of anchoring. *NeurIPS 2012*.
- Nakkiran, P., Kaplun, G., Bansal, Y., Yang, T., Barak, B., & Sutskever, I. (2020). Deep Double Descent: Where Bigger Models and More Data Hurt. *ICLR 2020*.
- Neal, B. (2019). On the Bias-Variance Tradeoff: Textbooks Need an Update (Master's thesis).
- Neal, B., Mittal, S., Baratin, A., Tania, V., Scicluna, M., Lacoste-Julien, S., & Mitliagkas, I. (2018). A Modern Take on the Bias-Variance Tradeoff in Neural Networks. *ICML 2019 Workshop on Identifying and Understanding Deep Learning Phenomena*.
- Neyshabur, B., Bhojanapalli, S., Mcallester, D., & Srebro, N. (2017). Exploring Generalization in Deep Learning. *NeurIPS 2017*.
- Neyshabur, B., Tomioka, R., & Srebro, N. (2015). In Search of the Real Inductive Bias: On the Role of Implicit Regularization in Deep Learning. *ICLR Workshop Track*.
- Schapire, R. E., & Singer, Y. (1999). Improved Boosting Algorithms Using Confidence-rated Predictions. *Machine Learning*.
- Spigler, S., Geiger, M., d'Ascoli, S., Sagun, L., Biroli, G., & Wyart, M. (2018). A jamming transition from under- to over-parametrization affects loss landscape and generalization.
- Tags: machine learning (/tags#machine learning) bias-variance tradeoff (/tags#bias-variance tradeoff)



Cite this blog post (BibTeX)

[← PREVIOUS POST \(/WHICH-CAUSAL-INFERENCE-BOOK\)](#)

[NEXT POST → \(/FUNDAMENTAL-PROBLEM-OF-CAUSAL-INFERENCE-NO-PROBLEM\)](#)

2 Comments

 Login ▼

G

Join the discussion...

LOG IN WITH

OR SIGN UP WITH DISQUS 

Name

 2 [Share](#)

[Best](#) [Newest](#) [Oldest](#)

B

[Bootstamp2](#)

3 years ago edited



This purported evidence is a little confusing/disconcerting to me. You test your thesis on datasets curated by humans, so it's possible that improvements that "aren't subject to the tradeoff" are merely researchers/engineers learning to (indirectly) exploit the structure *common to all datasets*. Here's a way to check this: train things on "bad" data (like pure noise). Would you still expect the 'Double Descent Curve'?

If anything begins to sound like "I don't need inductive bias" (and perhaps it's debatable whether what you're saying sounds like that), I have to flag it.

I should add that I really liked this post, and am thankful to the wikipedia edit that led me here.

1 0 [Reply](#) • [Share](#) ›



[Brady Neal](#) Mod [→ Bootstamp2](#)

3 years ago



Thanks for the comment and the words of appreciation!

I think this evidence is/was disconcerting to a lot of people. Note that we aren't saying that you will never see the bias-variance tradeoff. For example, if someone told me they had constructed a data distribution that leads larger and larger trained neural networks to be higher and higher variance, I would probably believe them and go check out their work. I think people are generally much less interested in these theoretical constructions than actual datasets, though.

Hmm, so with pure noise example you asked about, the Bayes error (best error you could expect from any classifier) is $1 - 1/k$, where k is the number of classes. Any randomly initialized neural network will get roughly that. So then, I would expect that it wouldn't learn much, and you'd see roughly a horizontal line in performance when increasing width. It might be more interesting to add lower levels of label noise. From a quick search, it looks like that's what they do in this paper (see Figure 4b): <https://arxiv.org/abs/1912.....>. They are able to find a double descent curve with up to 20% label noise (not sure they tested it with higher label noises).

I'm definitely not saying we "don't need inductive bias"! :) Even infinitely large neural networks have inductive bias. See the initial work that Radford Neal did on this and the recent revamping that it has received.

I think a big part of machine learning these days is people trying to find the right inductive biases that are specific enough to be useful for learning, but general enough in that they can exploit the structure that is shared throughout our world.

1 0 [Reply](#) • [Share](#) ›

[Subscribe](#) [Privacy](#) [Do Not Sell My Data](#)

6/28/23, 2:10 AM

On the Bias-Variance Tradeoff: Textbooks Need an Update



(/feed.xml)



(mailto:bradyneal1@gmail.com)



(https://github.com/bradyneal)



(https://twitter.com/CasualBrady)



(https://linkedin.com/in/bradyneal)

Brady Neal • 2021

Theme by beautiful-jekyll (<https://deanattali.com/beautiful-jekyll/>)