Is Overfitting Actually Benign?

On the Consistency of Interpolating Methods

Preetum Nakkiran (UCSD)

Neil Mallinar (UCSD)

Amirhesam Abedsoltan (UCSD)

Gil Kur (MIT)

Yamini Bansal (Harvard)

Misha Belkin (UCSD)

A partial talk on partial progress. Dec 7, 2021 @ Simons

Teaser

Distribution $D_0 \coloneqq$ CIFAR-10

Noisy dist: $D_p \coloneqq$ CIFAR-10, but

w.p. p, uniformly random label

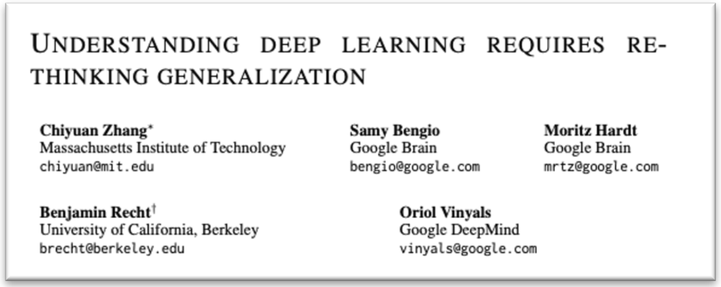
For varying $p \in [0, 1]$:

1. Interpolate N=50K iid samples from D_p
2. Evaluate test error w.r.t. D_0

“Benign overfitting”: Interpolating doesn’t hurt “too much”

…but it does hurt. Far from Bayes-optimal.

What happens as $N \to \infty$? (while still interpolating)



UNDERSTANDING DEEP LEARNING REQUIRES RETHINKING GENERALIZATION

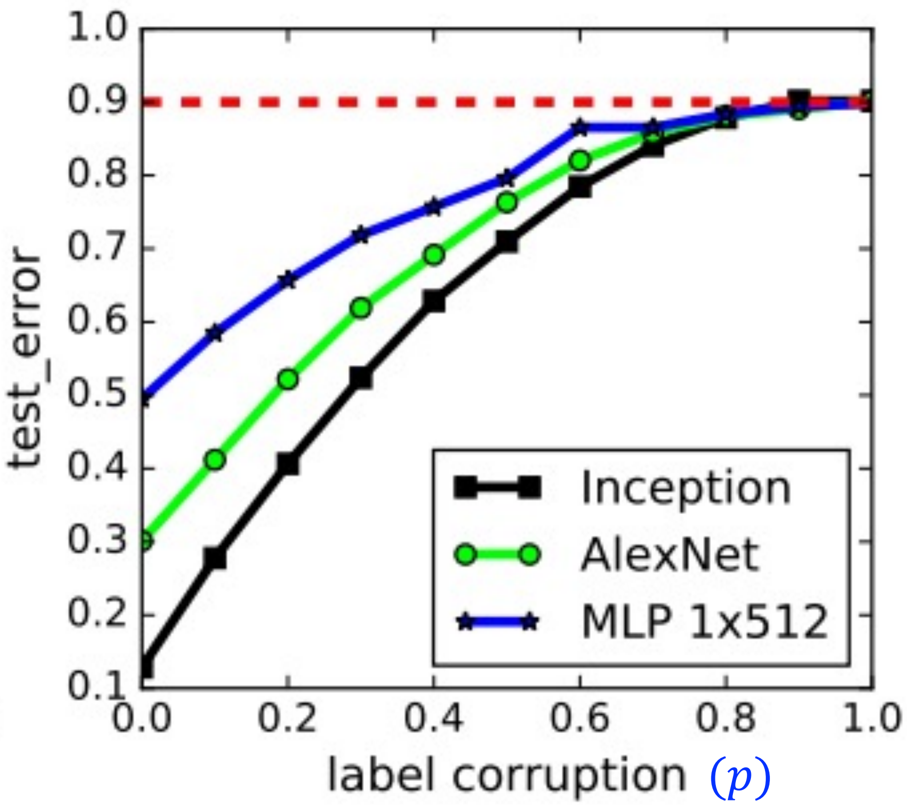
Chiyuan Zhang[*]
Massachusetts Institute of Technology
chiyuan@mit.edu

Samy Bengio
Google Brain
bengio@google.com

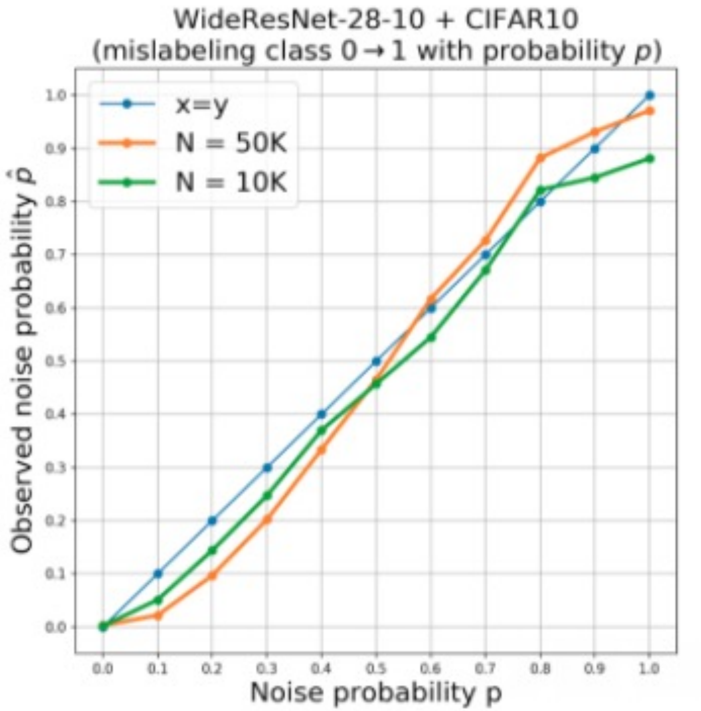
Moritz Hardt
Google Brain
mrtz@google.com

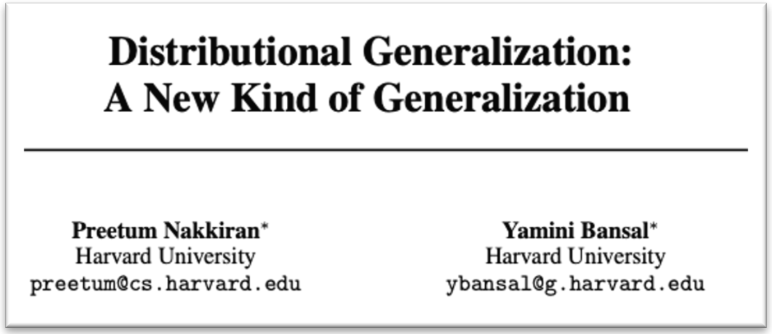
Benjamin Recht[†]
University of California, Berkeley
brecht@berkeley.edu

Oriol Vinyals
Google DeepMind
vinyals@google.com



For good interpolating networks: “p% noisy inputs → $\approx p$% noisy outputs”





Distributional Generalization:
A New Kind of Generalization

Preetum Nakkiran*
Harvard University
preetum@cs.harvard.edu

Yamini Bansal*
Harvard University
ybansal@g.harvard.edu

This project: Study *consistency implications, in simplest-possible setting*

Consistency

Setup:

Distribution $(x, y) \sim D$

Models $f : \mathcal{X} \to \mathcal{Y}$

Loss $\mathcal{L} : \mathcal{F} \to \mathbb{R}$

ex: $\mathcal{L}_D(f) = E_{(x,y)\sim D}\left[\left(y - f(x)\right)^2\right]$

ex: $\mathcal{L}_D(f) = E_{(x,y)\sim D}[\, \mathbb{1}\{y \neq f(x)\}]$

Optimal loss: $\mathcal{L}_D^* \coloneqq \inf_f \mathcal{L}_D(f)$

Learning Method: $\mathcal{A} = \{A_1, A_2, \ldots, A_n, \ldots\}$

$$A_n : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{F}$$

Def. Consistency:

A learning method $\mathcal{A}$ *is* **consistent** *on distribution* D *w.r.t. loss* $\mathcal{L}_D$ *if*

$$\mathcal{L}_D(\ A_n(D^n)\ ) \to \mathcal{L}_D^*$$

That is, if the outputs of A_n *converges to the optimal loss as* $n \to \infty$.

Q: Are modern learning methods consistent?

The Two Limits (DNNs)

Want to define sequence $\{A_1, A_2, \dots, A_n, \dots\}$
Roughly “train a neural network, of increasing size”

Inconsistent in “almost all” settings (which ones?)

Overparameterized Limit: (model >> data)

$A_n :=$ “Train a neural network of size $s(n) \gg n$, until interpolation”

Underparameterized Limit: (model << data)

Often consistent (sometimes provably so)

$A_n :=$ “Train a neural network of size $s(n) \ll n$, until convergence”

Setup:

Distribution D:
$x \sim N(0, I_d)$

$y \sim \{\pm 1\}$ independent of x $\quad (E[y] = \mu = 0.2)$

Regression: MSE loss, optimal function $f^*(x) = \mu$

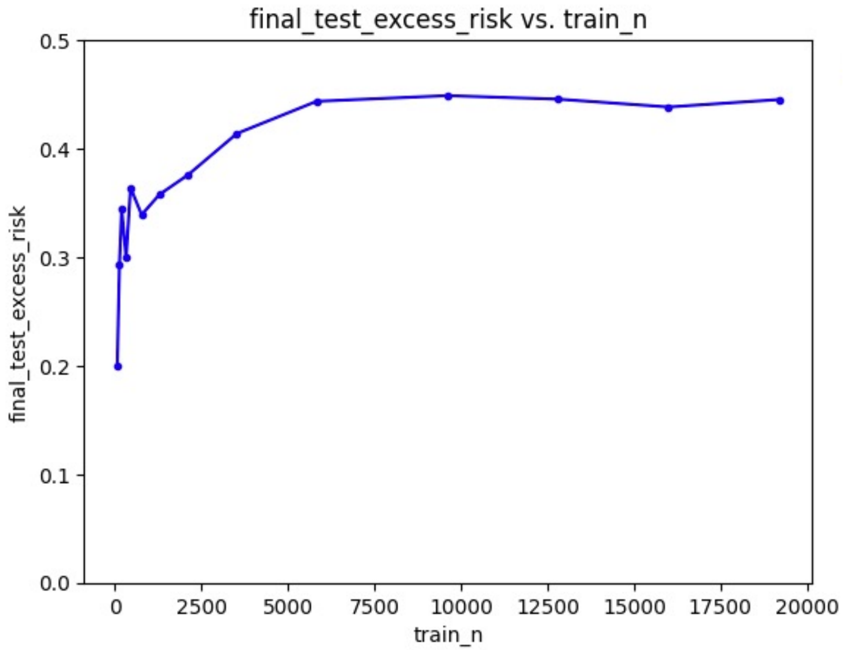
"Solve it with deep learning":

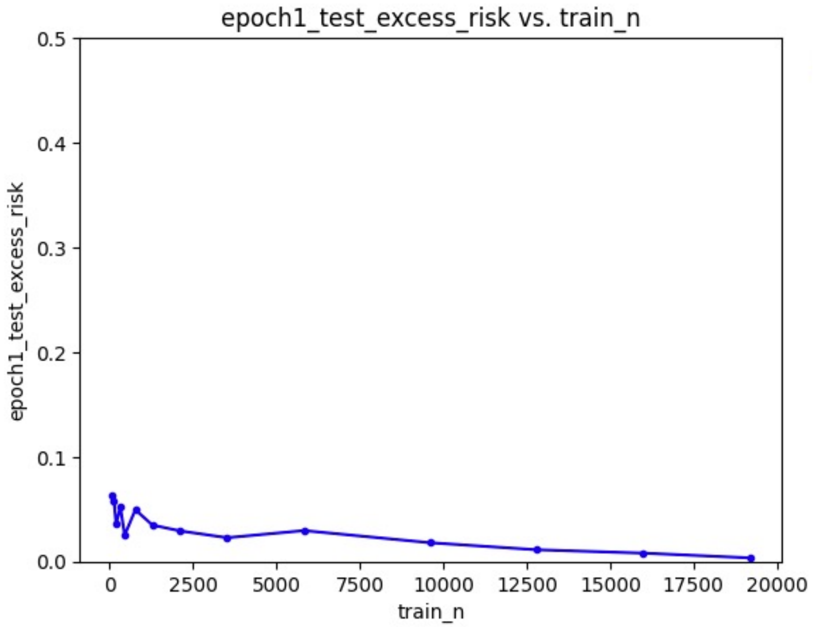
Train an interpolating MLP, with MSE loss, on samples from D.

Does it learn (close to) the constant function?

Claim: This will fail for "all reasonable hyperparameters".

That is, "almost all" interpolating-DNN learning methods are inconsistent in this setting.





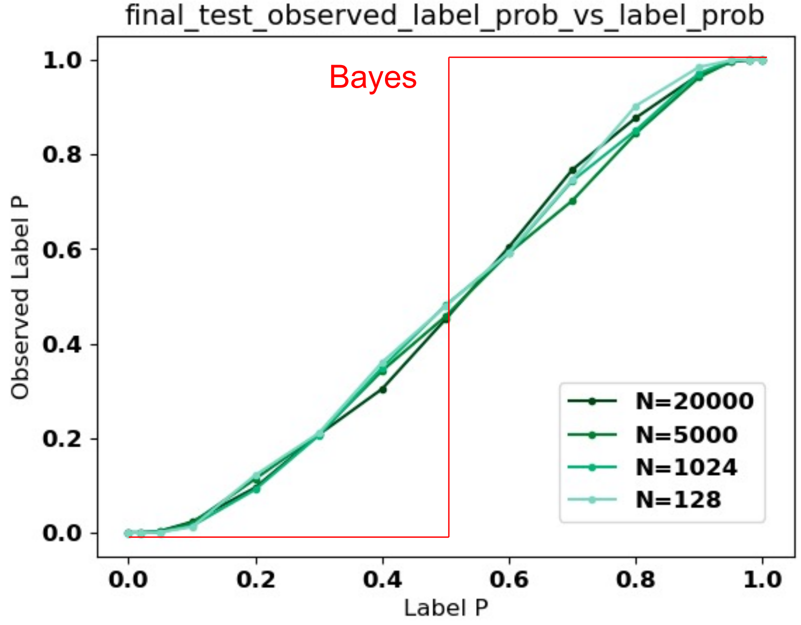
Setup:

Distribution D_p:
$x \sim N(0, I_d)$
$y \sim \{\pm 1\}$ independent of x $(\Pr[y = 1] = p)$

Classification: optimal function $f^*(x) = 1$

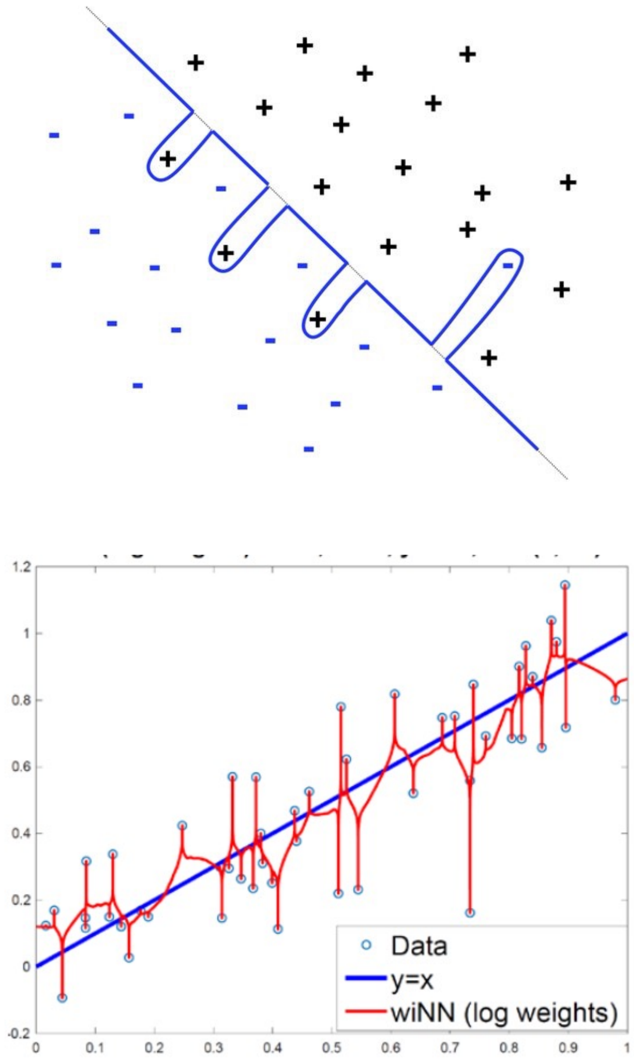
Train an interpolating MLP, with MSE loss, on samples from D.

(NB: 1-nearest-neighbors would do this)

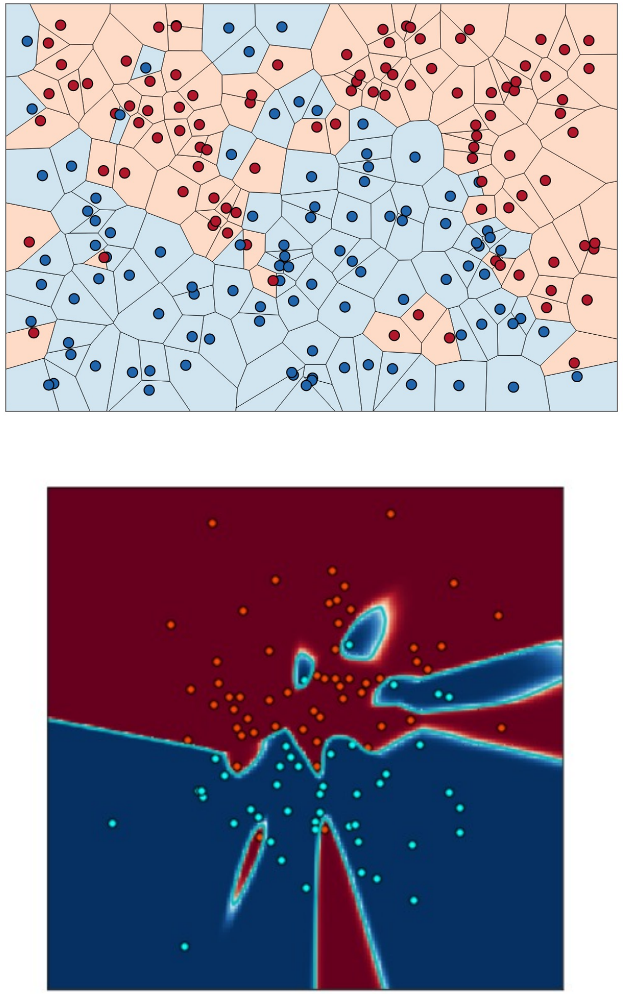


$\Pr[sign(f(x)) = 1]$

$\Pr[y = 1]$



Benign Overfitting World



Distributionally-Generalizing World

Bias + Variance

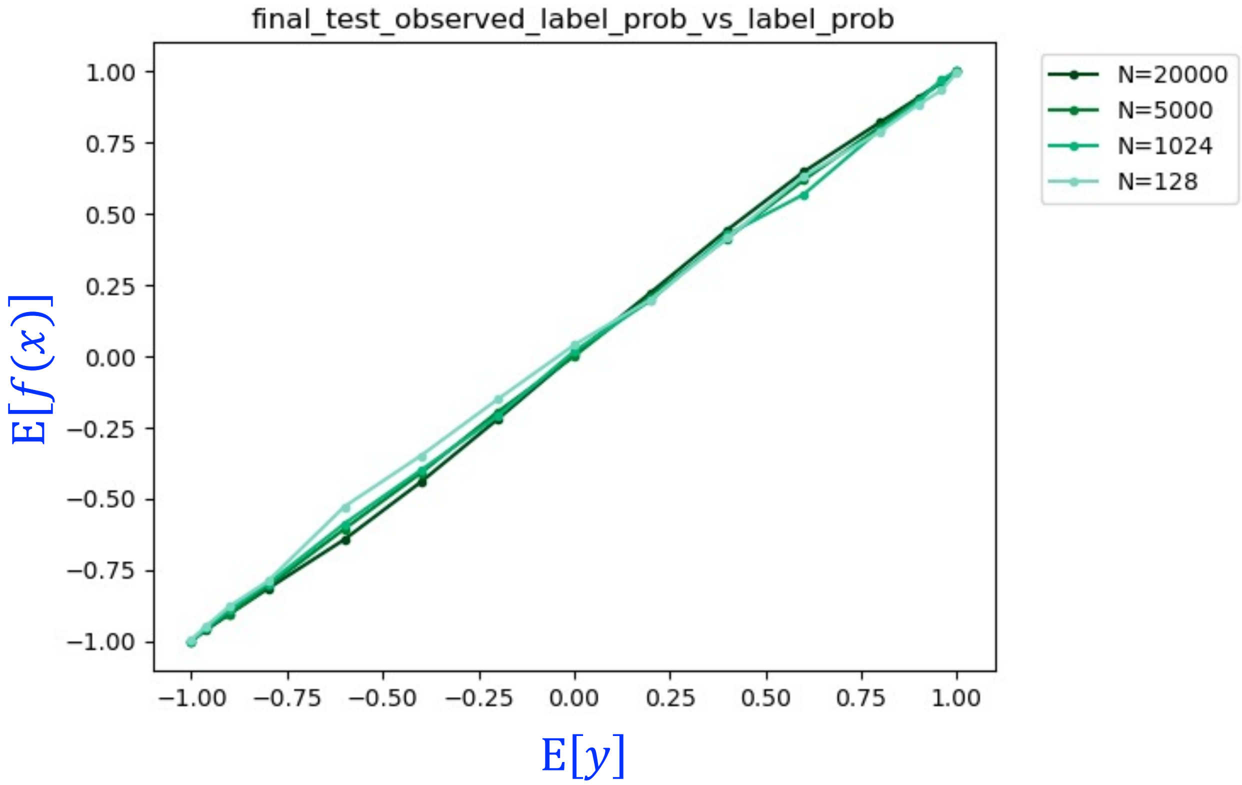
Regression:

Excess risk = bias^2 + variance

Consistency requires bias $\rightarrow 0$, variance $\rightarrow 0$.
Which one fails?

Empirical evidence:
Neural-nets are asymptotically unbiased

$$E_{f \leftarrow A_n(D^n)}[f(x)] \rightarrow E[y \mid x]$$



Problem is **variance**.

(1-nearest-neighbors would do this too)

Status / Open Questions

Observations:

1. **Negative:** “Almost all” interpolating methods are inconsistent, in “almost all” settings with non-zero Bayes risk.

(MLPs, ResNets, RBF kernels,…)

2. **Positive:** Interpolating methods appear asymptotically unbiased.

Open Questions:

1. Is there a natural definition of “almost all”?
What does consistency depend on?
(we know consistent interpolating methods exist…)

2. What separates these settings from “benign overfitting” in theory? Which assumptions are “unrealistic”?

3. Can we prove the “asymptotic unbiasedness”?

4. Is the inconsistency for some “good reason”?
(cf Distributional Generalization)

Theory Open Question

Setup (regression):

$x \sim N(0, I_d)$

$y \sim N(\mu, \sigma^2)$, independent of x

Draw n train samples.

Train unregularized RBF kernel for regression, with bandwidth $\tau(n)$.

Q: For what choices of bandwidth $\boldsymbol{\tau(n)}$ is this consistent/inconsistent?

Motivations

Differences between overparameterized & underparameterized regimes?

When do neural networks fail?

Common structure of interpolating methods, to explain:

(A) inconsistency (B) asymptotic unbiasedness

Overfitting is not benign in practice… so why is it benign in theory?

Which assumptions fail, and how should we adapt them?

Thanks!

preetum@ucsd.edu
preetum.nakkiran.org