“Benign overfitting” and the behavior of high-dimensional linear regression and classification models

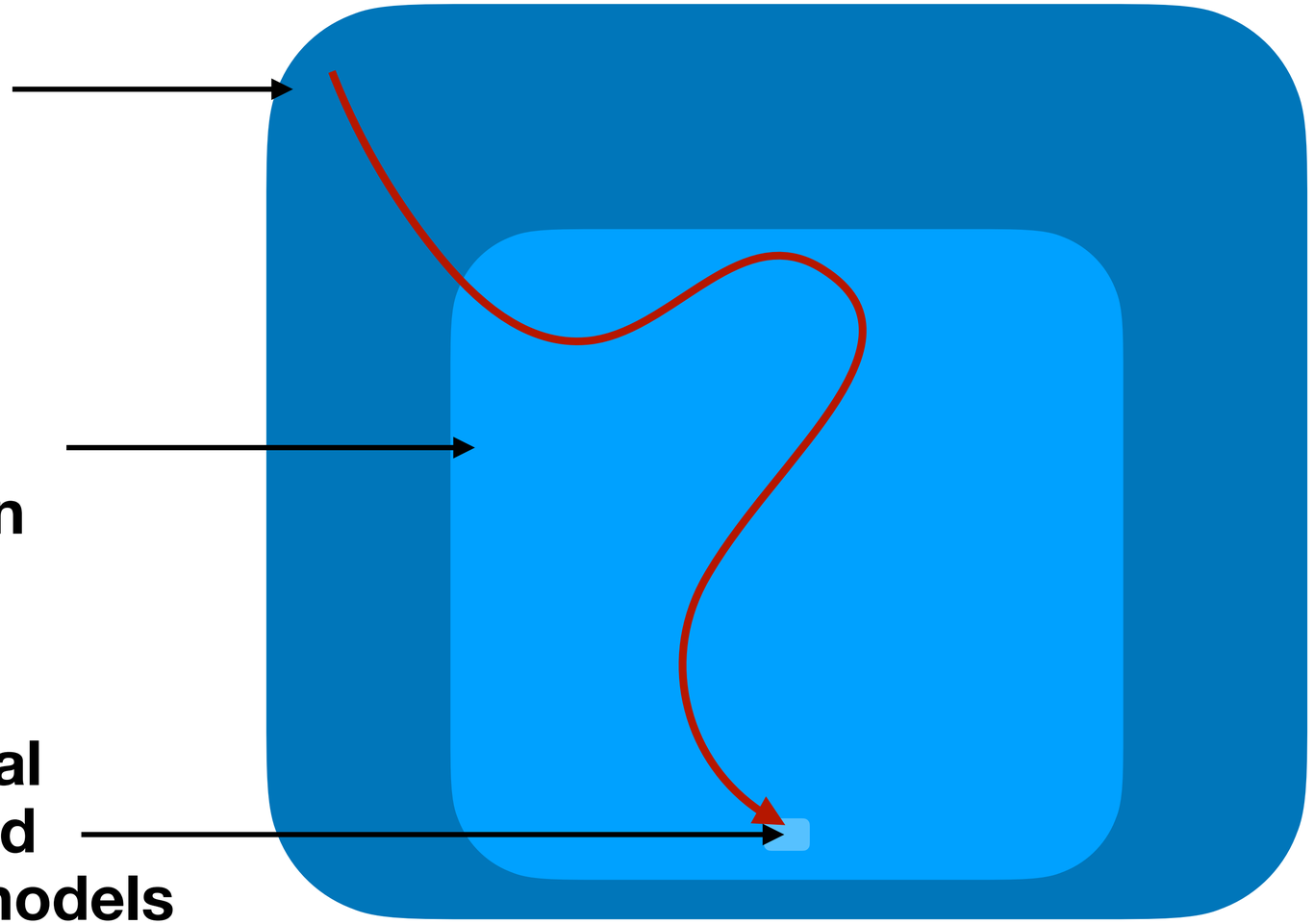
Clayton Sanford ’18

April 1, 2022

Based on work done with Navid Ardeshir and Daniel Hsu.

Talk outline

1. Benign overfitting, double descent, and deep learning theory
2. Benign overfitting for linear regression and classification
3. Behavior of high-dimensional least-squares regression and max-margin classification models



Central issue of deep learning theory

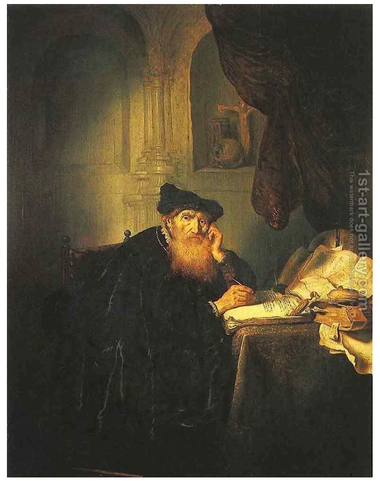
- **Goal of ML theory:**
  - Rigorous mathematical understanding capabilities and limitations of ML algorithms, which translate to practical recommendations for practitioners.
- **Problem:**
  - ML theory is too pessimistic for deep learning; lack of theoretical explanations for neural networks' practical success.
  - Two conflicting narratives for what makes ML models succeed: classical ML theory vs modern deep learning practice.

Supervised learning setting

- Given samples $(x_1, y_1), \ldots, (x_n, y_n) \sim \mathscr{D}$.
- Want to learn $h : \mathbb{R}^d \to \mathscr{Y}$ such $R(h) = \mathbb{E}[\ell(h(x), y)]$ is small for new $(x, y) \sim \mathscr{D}$.
  - $\mathscr{Y} = \{\pm 1\}$ for classification, $\mathscr{Y} = \mathbb{R}$ for regression.
- How? Find $h \in \mathscr{H}$ minimizing training error: $\hat{R}(h) = \frac{1}{n} \sum_{i=1}^{n} \ell(h(x_i), y_i)$.

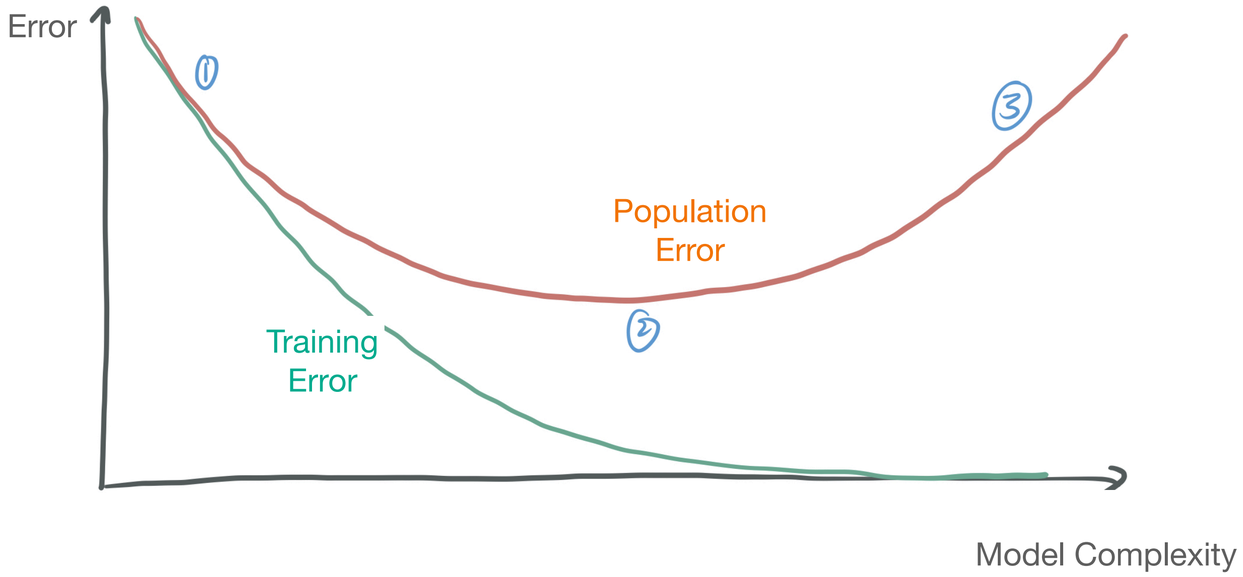
$$\underbrace{R(h)}_{\text{population error}} = \underbrace{\hat{R}(h)}_{\text{training error}} + \underbrace{R(h) - \hat{R}(h)}_{\text{generalization error}}$$

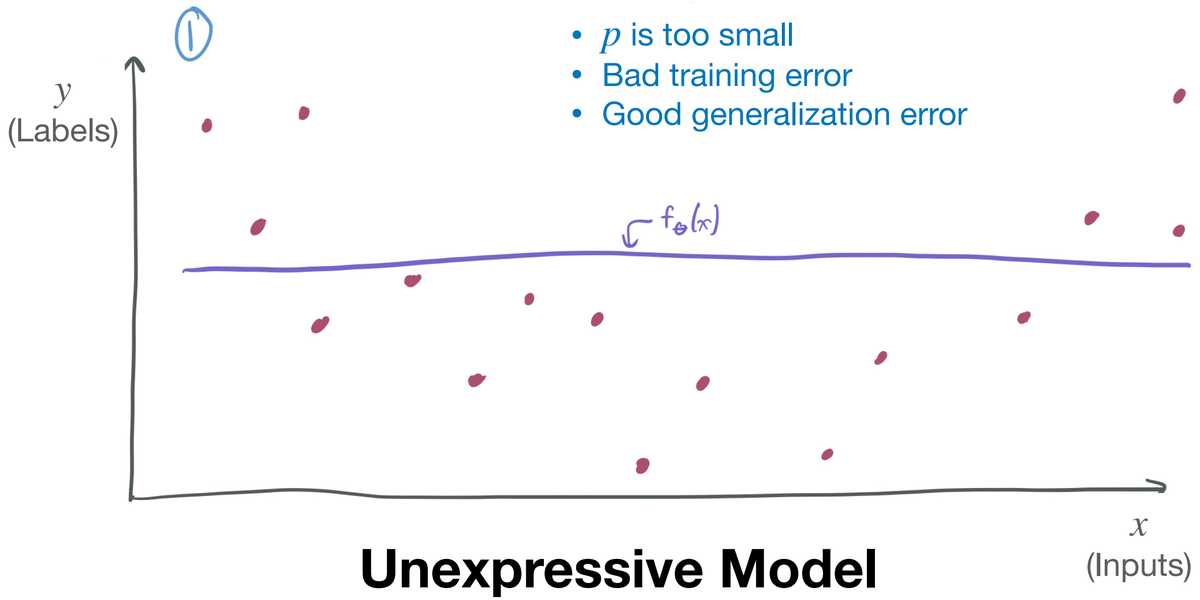
Narrative #1: Classical Theory



- ML textbook trade-offs: greater model complexity requires more samples
- Delicate balance between overfitting (high generalization error $R(h) - \hat{R}(h)$) and over-simplification (high training error $\hat{R}(h)$)

Narrative #1: Classical Theory

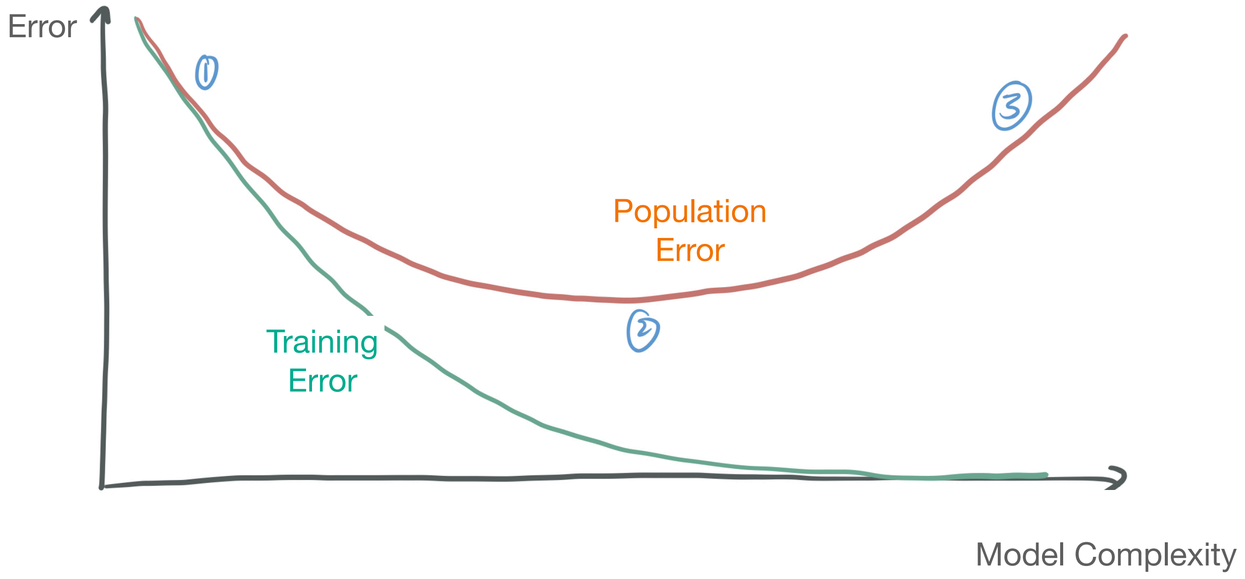


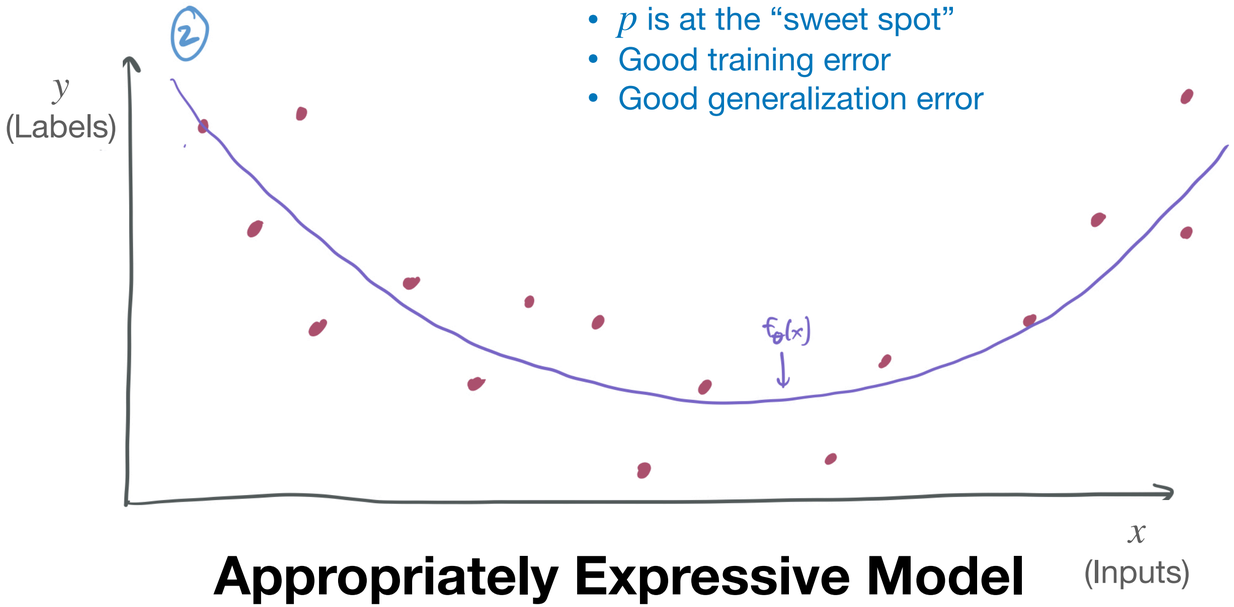


Unexpressive Model

- ML textbook trade-offs: greater model complexity requires more samples
- Delicate balance between overfitting (high generalization error $R(h) - \hat{R}(h)$) and over-simplification (high training error $\hat{R}(h)$)

Narrative #1: Classical Theory



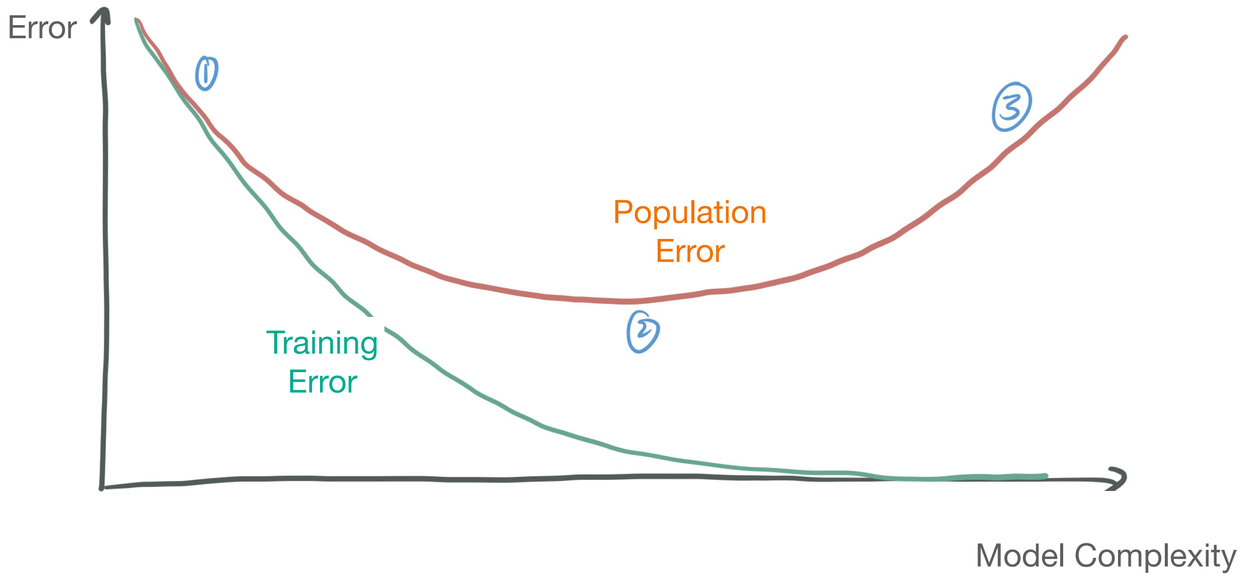


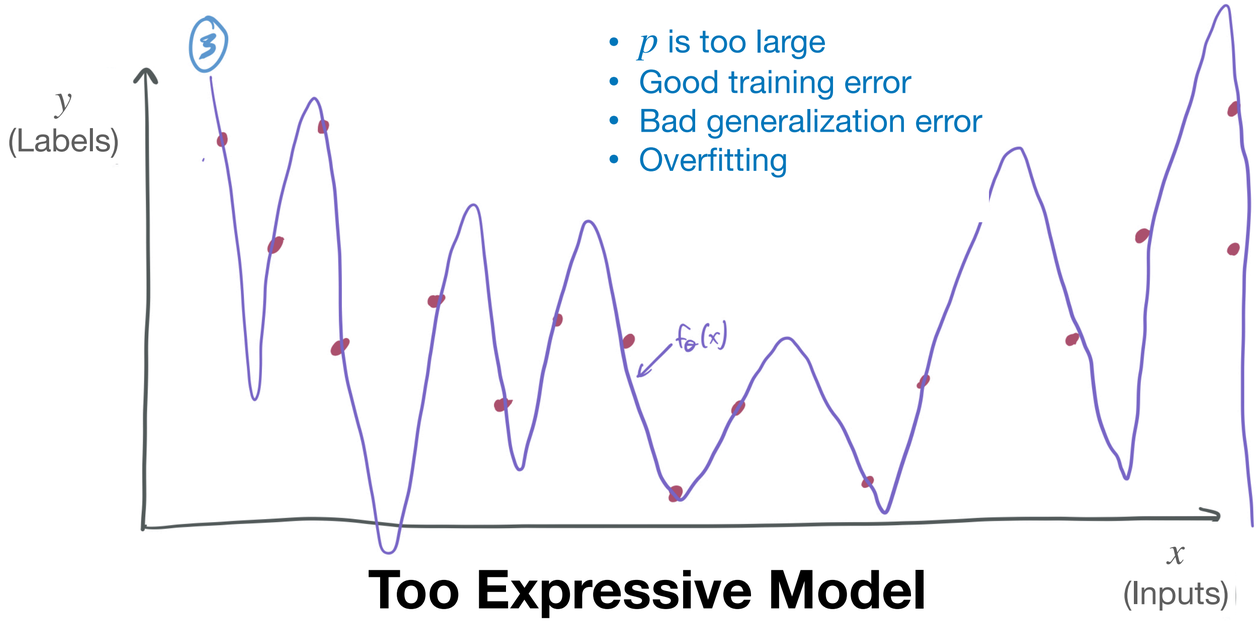
- p is at the "sweet spot"
- Good training error
- Good generalization error

Appropriately Expressive Model

- ML textbook trade-offs: greater model complexity requires more samples
- Delicate balance between overfitting (high generalization error $R(h) - \hat{R}(h)$) and over-simplification (high training error $\hat{R}(h)$)

Narrative #1: Classical Theory

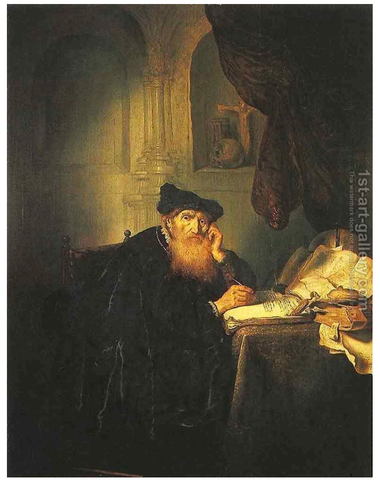




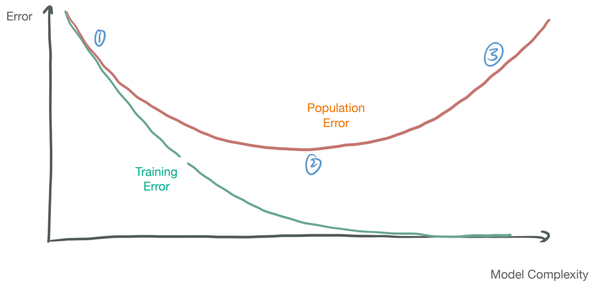
Too Expressive Model

- ML textbook trade-offs: greater model complexity requires more samples
- Delicate balance between overfitting (high generalization error $R(h) - \hat{R}(h)$) and over-simplification (high training error $\hat{R}(h)$)

Narrative #1: Classical Theory



- Made rigorous with measurements of model complexity, like VC-dimension**,** Rademacher complexity, fat-shattering dimension.
- **VC-dimension** measures ability of hypotheses in $\mathscr{H}$ to correctly classify different labelings y_i.
- Generalization bound based on VC-dimension:
  - $R(h) - \hat{R}(h) = O(\sqrt{VC(\mathscr{H})/n})$ for all $h \in \mathscr{H}$.
- Example: Linear classification
  - $VC(\mathscr{H}) = d + 1.$
  - $R(h) - \hat{R}(h) = O(\sqrt{d/n})$ for all $h \in \mathscr{H}$.



Narrative #2: Deep Learning Practice

- **Past decade**: empirical dominance of deep learning over other ML models
- **How to train a neural network:**
  - Initialize a very large model (# params > # samples)
  - Train with gradient descent until convergence to very small training error
  - Necessary tips & tricks: dropout, Adam, batch size, regularization, specialized architectures, choice of loss function, etc.

Clash between narratives

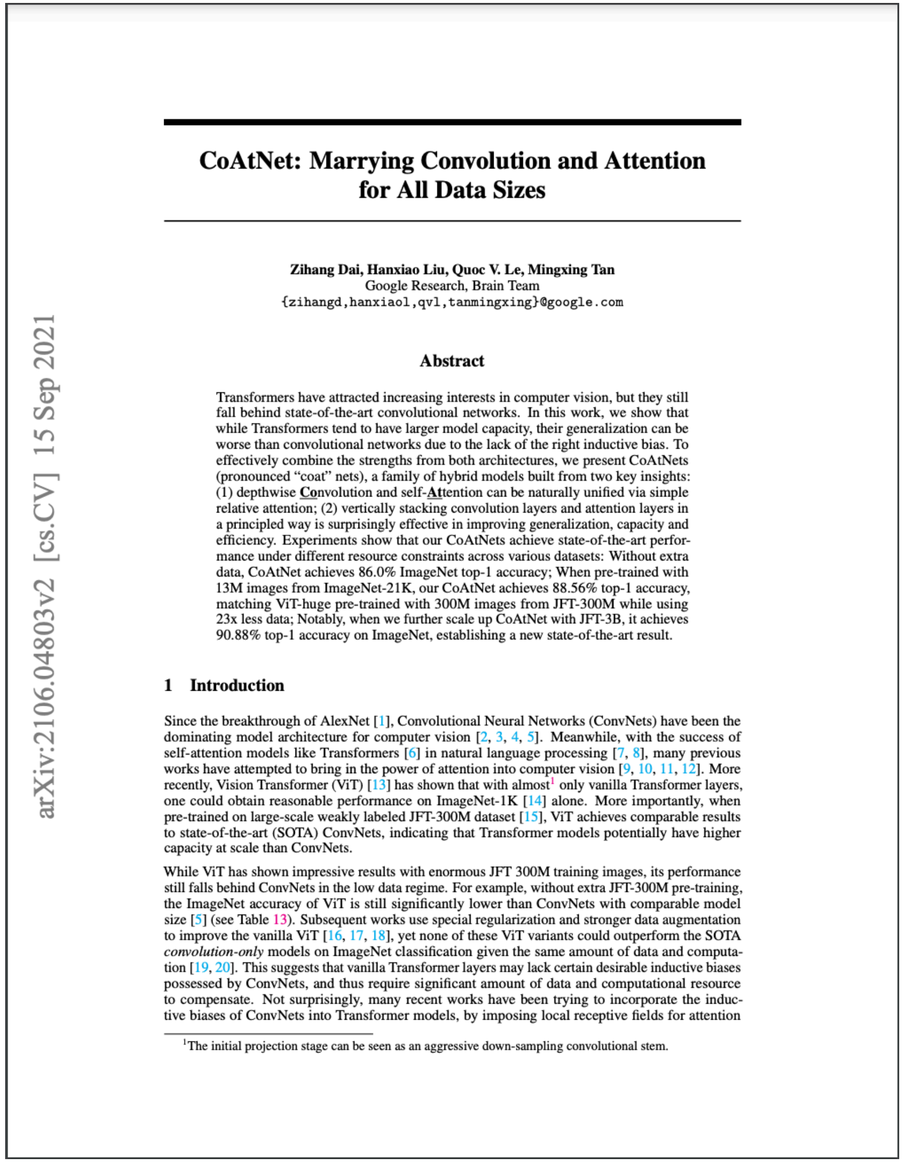
Unprincipled alchemy!

Irrelevant theory!

Clash between narratives

An example

- **CoAtNet:** current holder of SOTA for ImageNet image classification (ignoring ensemble models)
- Achieves **86.09%** accuracy on 1000-class classification by a NN with **168M** parameters and **13M** training samples.
- VC dimension of NNs with fixed depth and w parameters is $\Theta(w \log w)$ [Bartlett, et al '98].
- Generalization bound is vacuous: $R(h) - \hat{R}(h) = \tilde{O}(\sqrt{w/n})$.



CoAtNet: Marrying Convolution and Attention for All Data Sizes

Zihang Dai, Hanxiao Liu, Quoc V. Le, Mingxing Tan
Google Research, Brain Team
{zihangd,hanxiaol,qvl,tanmingxing}@google.com

Abstract

Transformers have attracted increasing interests in computer vision, but they still fall behind state-of-the-art convolutional networks. In this work, we show that while Transformers tend to have larger model capacity, their generalization can be worse than convolutional networks due to the lack of the right inductive bias. To effectively combine the strengths from both architectures, we present CoAtNets (pronounced "coat" nets), a family of hybrid models built from two key insights: (1) depthwise **Co**nvolution and self-**At**tention can be naturally unified via simple relative attention; (2) vertically stacking convolution layers and attention layers in a principled way is surprisingly effective in improving generalization, capacity and efficiency. Experiments show that our CoAtNets achieve state-of-the-art performance under different resource constraints across various datasets: Without extra data, CoAtNet achieves 86.0% ImageNet top-1 accuracy; When pre-trained with 13M images from ImageNet-21K, our CoAtNet achieves 88.56% top-1 accuracy, matching ViT-huge pre-trained with 300M images from JFT-300M while using 23x less data; Notably, when we further scale up CoAtNet with JFT-3B, it achieves 90.88% top-1 accuracy on ImageNet, establishing a new state-of-the-art result.

1 Introduction

Since the breakthrough of AlexNet [1], Convolutional Neural Networks (ConvNets) have been the dominating model architecture for computer vision [2, 3, 4, 5]. Meanwhile, with the success of self-attention models like Transformers [6] in natural language processing [7, 8], many previous works have attempted to bring in the power of attention into computer vision [9, 10, 11, 12]. More recently, Vision Transformer (ViT) [13] has shown that with almost[1] only vanilla Transformer layers, one could obtain reasonable performance on ImageNet-1K [14] alone. More importantly, when pre-trained on large-scale weakly labeled JFT-300M dataset [15], ViT achieves comparable results to state-of-the-art (SOTA) ConvNets, indicating that Transformer models potentially have higher capacity at scale than ConvNets.

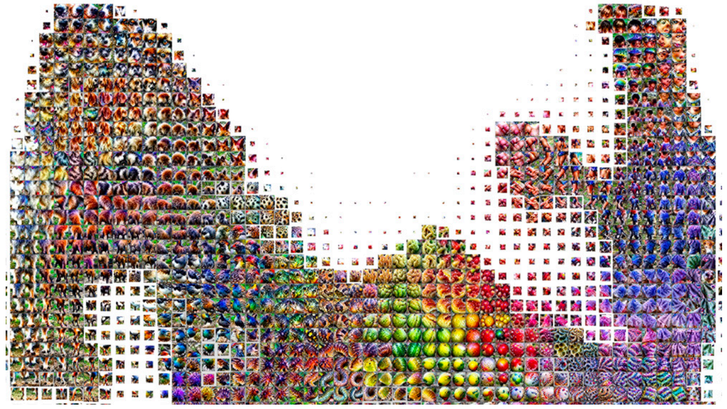
While ViT has shown impressive results with enormous JFT 300M training images, its performance still falls behind ConvNets in the low data regime. For example, without extra JFT-300M pre-training, the ImageNet accuracy of ViT is still significantly lower than ConvNets with comparable model size [5] (see Table 13). Subsequent works use special regularization and stronger data augmentation to improve the vanilla ViT [16, 17, 18], yet none of these ViT variants could outperform the SOTA *convolution-only* models on ImageNet classification given the same amount of data and computation [19, 20]. This suggests that vanilla Transformer layers may lack certain desirable inductive biases possessed by ConvNets, and thus require significant amount of data and computational resource to compensate. Not surprisingly, many recent works have been trying to incorporate the inductive biases of ConvNets into Transformer models, by imposing local receptive fields for attention

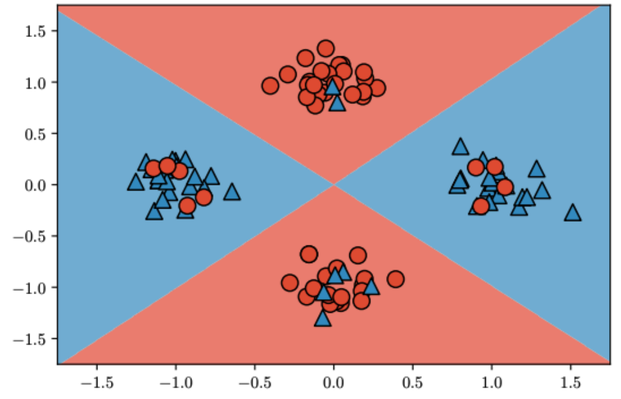
[1]The initial projection stage can be seen as an aggressive down-sampling convolutional stem.

Can they be reconciled?

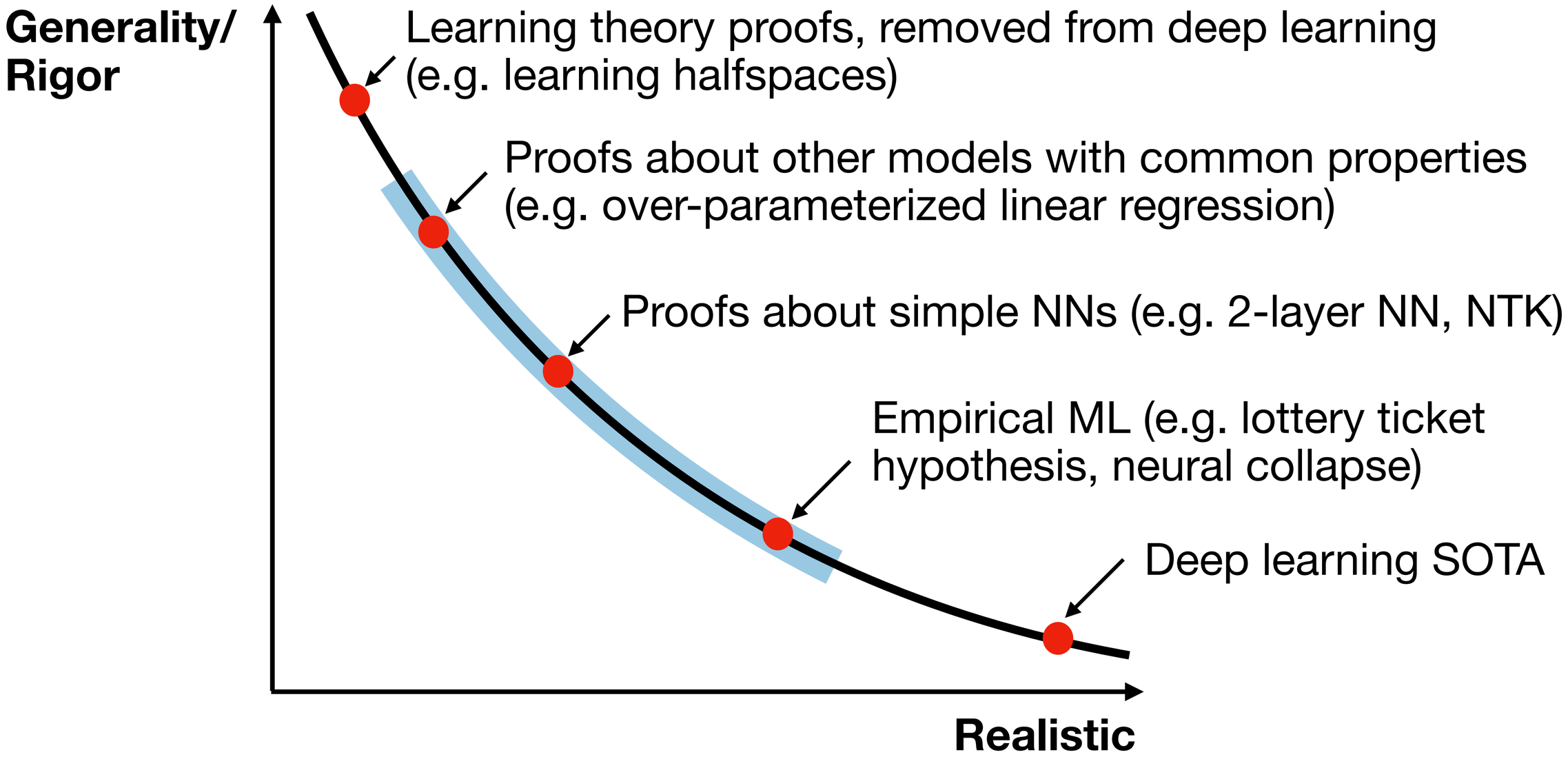
- **Goal:** a theory of **benign overfitting** that mathematically describes why some overfitting models generalize.
- **Challenge:** It's mathematically very difficult to say things about neural networks
- Many other theoretical limitations to reconcile: approximation, optimization, representation learning…

$$
\begin{aligned}
&\sum_{k_0,k_1,k_2} \mathbb{E}\left[\mathcal{O}\left(z^{(\ell+1)}\right) W_{i_0k_0}^{(\ell+1)} W_{i_1k_1}^{(\ell+1)} W_{i_2k_2}^{(\ell+1)} \sigma'^{(\ell)}_{k_0;\delta_0} \sigma'^{(\ell)}_{k_1;\delta_1} \sigma'^{(\ell)}_{k_2;\delta_2} \widehat{\mathrm{d}H}^{(\ell)}_{k_0k_1k_2;\delta_0\delta_1\delta_2}\right] \\
&= \frac{\left(C_W^{(\ell+1)}\right)^2}{n_\ell} \sum_{\delta\in\mathcal{D}} \delta_{i_0i_1} \left\langle\!\!\left\langle \frac{\partial \mathcal{O}}{\partial z^{(\ell+1)}_{i_2;\delta}} \right\rangle\!\!\right\rangle_{G^{(\ell+1)}} \left\{ \frac{1}{n_\ell} \sum_{k,m} \mathbb{E}\left[\sigma'^{(\ell)}_{k;\delta_0} \sigma'^{(\ell)}_{k;\delta_1} \sigma'^{(\ell)}_{m;\delta_2} \sigma^{(\ell)}_{m;\delta} \widehat{\mathrm{d}H}^{(\ell)}_{kkm;\delta_0\delta_1\delta_2}\right] \right\} \\
&\quad + \frac{\left(C_W^{(\ell+1)}\right)^2}{n_\ell} \sum_{\delta\in\mathcal{D}} \delta_{i_0i_2} \left\langle\!\!\left\langle \frac{\partial \mathcal{O}}{\partial z^{(\ell+1)}_{i_1;\delta}} \right\rangle\!\!\right\rangle_{G^{(\ell+1)}} \left\{ \frac{1}{n_\ell} \sum_{k,m} \mathbb{E}\left[\sigma'^{(\ell)}_{k;\delta_0} \sigma'^{(\ell)}_{m;\delta_1} \sigma'^{(\ell)}_{k;\delta_2} \sigma^{(\ell)}_{m;\delta} \widehat{\mathrm{d}H}^{(\ell)}_{kmk;\delta_0\delta_1\delta_2}\right] \right\} \\
&\quad + \frac{\left(C_W^{(\ell+1)}\right)^2}{n_\ell} \sum_{\delta\in\mathcal{D}} \delta_{i_1i_2} \left\langle\!\!\left\langle \frac{\partial \mathcal{O}}{\partial z^{(\ell+1)}_{i_0;\delta}} \right\rangle\!\!\right\rangle_{G^{(\ell+1)}} \left\{ \frac{1}{n_\ell} \sum_{k,m} \mathbb{E}\left[\sigma'^{(\ell)}_{m;\delta_0} \sigma'^{(\ell)}_{k;\delta_1} \sigma'^{(\ell)}_{k;\delta_2} \sigma^{(\ell)}_{m;\delta} \widehat{\mathrm{d}H}^{(\ell)}_{mkk;\delta_0\delta_1\delta_2}\right] \right\} \\
&\quad + O\left(\frac{1}{n^2}\right). \qquad (11.40)
\end{aligned}
$$





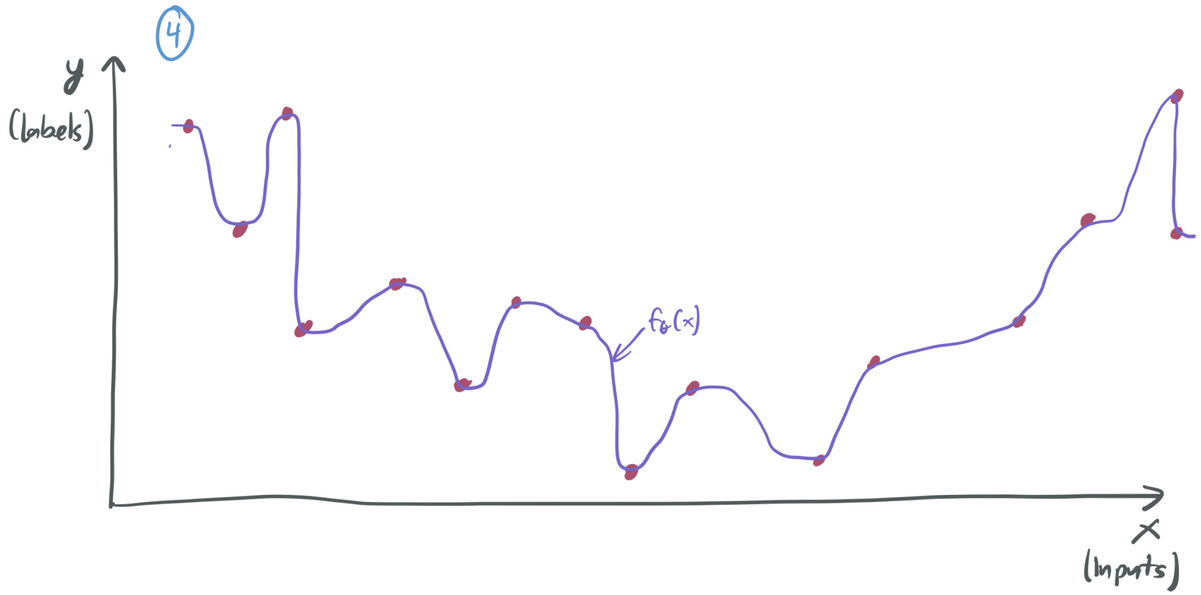
Can they be reconciled? (ctd)



Benign overfitting and double-descent

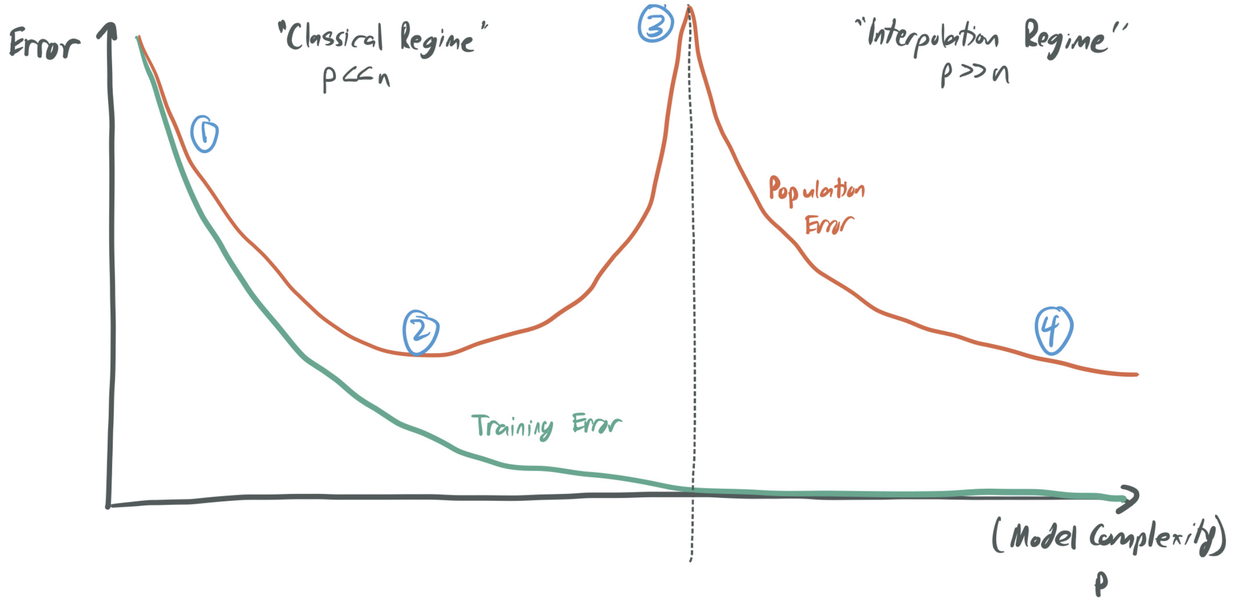
Benign Overfitting

Model generalizes despite over-parameterization and very small training error



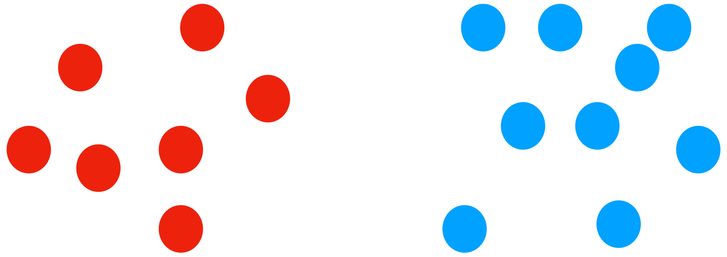
Double Descent

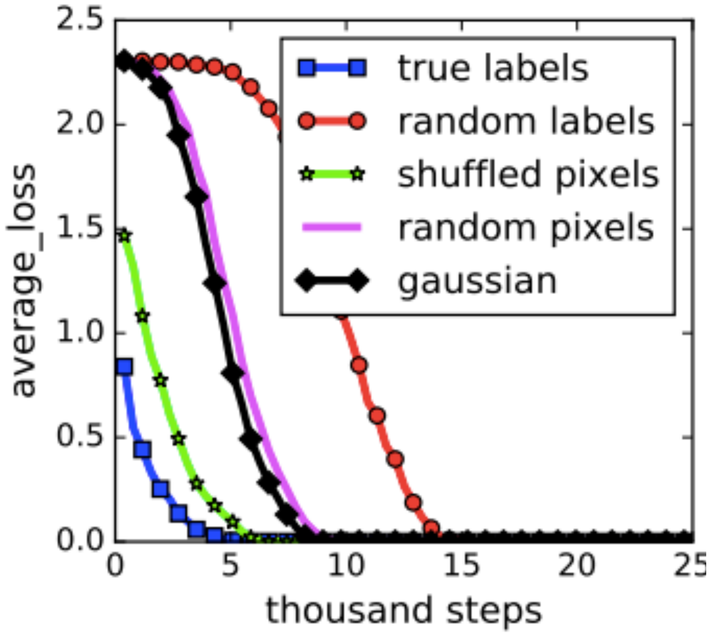
Increasing model complexity beyond initial point of overfitting causes second descent of generalization error



Benign overfitting and double-descent

- **Question:** What if benign overfitting in NNs is caused by simple data patterns that are easy to fit with any model?
- **[ZBHRV17]:** NNs can perfectly classify randomly labeled samples
  - $\Longrightarrow$ benign overfitting can't be a property of dataset alone
- **[BMM18]:** Similar phenomena for kernel regression
  - $\Longrightarrow$ worth studying simpler models than deep neural networks





Where can you find benign overfitting?

- **Least-squares regression** [BHX19, BLLT19, HMRT19, Mitra19, MVSS19]
- Ridge regression [TB20]
- Kernel regression [RZ19, LRZ20]
- **Support vector machines** [MNSBHS20, CL20, **ASH20**]
- Random feature models [MM19]
- Boosting [BFLS98]
- Neural networks (empirical) [NKBYBS19, SGDSBW19]

Linear regression

- Sample $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \mathbb{R}$. $(X, y) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$.
- Learn $x \mapsto \hat{\theta}^T x$.
- **Ordinary least-squares (OLS)** (classical, $n \gg d$):
  - $\hat{\theta} \in \mathbb{R}^d$ minimizes $\sum_{i=1}^{n} (\hat{\theta}^T x_i - y_i)^2$, or $\hat{\theta} = X^\dagger y = (X^T X)^{-1} X^T y$.
- **Minimum-norm interpolation** (interpolation, $d \gg n$):
  - $\hat{\theta} \in \mathbb{R}^d$ minimizes $\|\hat{\theta}\|$ such that $\hat{\theta}^T x_i = y_i$, or $\hat{\theta} = X^\dagger y = X(XX^T)^{-1} Y$.
- Classical generalization bound: $R(h) - \hat{R}(h) \leq O(\sqrt{d/n})$ [Audibert and Catoni, ’10]

Benign overfitting: feature importance

[Bartlett, Long, Lugosi, Tsigler '19]

- Analysis of when benign overfitting occurs for over-parameterized OLS ($d \gg n$).
- Subgaussian x_i with covariance Σ (with eigenvalues $\lambda_1 > \lambda_2 > \ldots > \lambda_d$), optimal weights θ^*, and subgaussian noise σ.
- Depends on **effective ranks** of Σ: $r_k(\Sigma) = \sum_{i>k} \lambda_i / \lambda_{k+1}$ and $R_k(\Sigma) = (\sum_{i>k} \lambda_i)^2 / \sum_{i>k} \lambda_i^2$.
- **Theorem**: With probability 0.99 for $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$:

$$R(h) = O\left(\|\theta^*\|^2 \lambda_1 \left(\sqrt{\frac{r_0(\Sigma)}{n}} + \frac{r_0(\Sigma)}{n} \right) + \sigma^2 \left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \right)$$

- Bound by **bias-variance** decomposition, concentration bounds based on spectrum, analysis of projection operator onto row space of X.

Benign overfitting: feature importance

[Bartlett, Long, Lugosi, Tsigler '19]

- **Theorem**: With probability 0.99 for $k^* = \min\{k \geq 0 : r_k(\Sigma) \geq bn\}$:

$$R(h) = O\left(\|\theta^*\|^2 \lambda_1 \left(\sqrt{\frac{r_0(\Sigma)}{n}} + \frac{r_0(\Sigma)}{n} \right) + \sigma^2 \left(\frac{k^*}{n} + \frac{n}{R_{k^*}(\Sigma)} \right) \right)$$

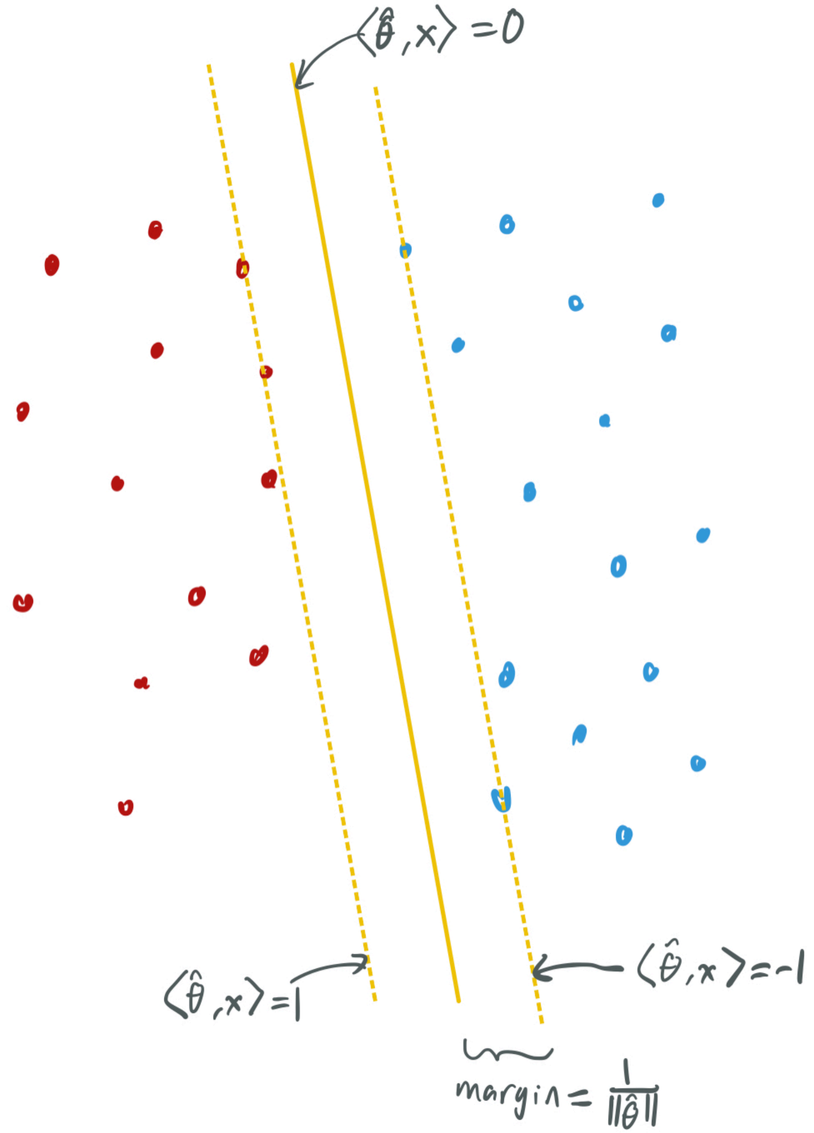
- This bound indicates that benign overfitting occurs when…
  - The variances $\lambda_1, \ldots, \lambda_d$ decay not too fast but not too slow. (e.g. occurs for $\lambda_i = 1/(i \log^2(i+1))$)
  - Output y depends mostly on high-variance directions of x
  - $\Longrightarrow$ Benign overfitting occurs when there are many low-importance low-variance features that cancel one another out

Limitations of benign overfitting in linear regression

- Beaten by properly regularized ridge regression
- Bounds are distribution-dependent [Bartlett and Long ’20]

Hard SVM or maximum-margin classification

- Linearly separable
  $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^d \times \{-1, 1\}$.
- Learn $x \mapsto \text{sign}(\hat{\theta}^T x)$.
- $\hat{\theta} \in \mathbb{R}^d$ minimizes $\|\hat{\theta}\|_2$ such that
  $y_i \hat{\theta}^T x_i \geq 1$.
- x_i is a **support vector** if $\hat{\theta}^T x_i = y_i$.
- Classical generalization bounds rely on bounding number of support vectors.



Motivation

Or: It's 2022, who cares about SVMs??

- Gradient descent sometimes biased in favor of max-margin classifier
  - Classification tasks with logistic loss function converge to the max margin solution [Soudry et al, '17], [Ji & Telgarsky '19]
  - Wide two-layer neural nets with logistic loss function also converge to max margin solutions [Chizat et al, '20]

SVM benign overfitting by connection to OLS [MNSBHS20]

Exhibits benign overfitting for SVM linear classification of Gaussian data with bi-level variances:

1. For certain distributions, over-parameterized OLS regression for $y_i \in \mathbb{R}$ has benign overfitting. [BLLT19]

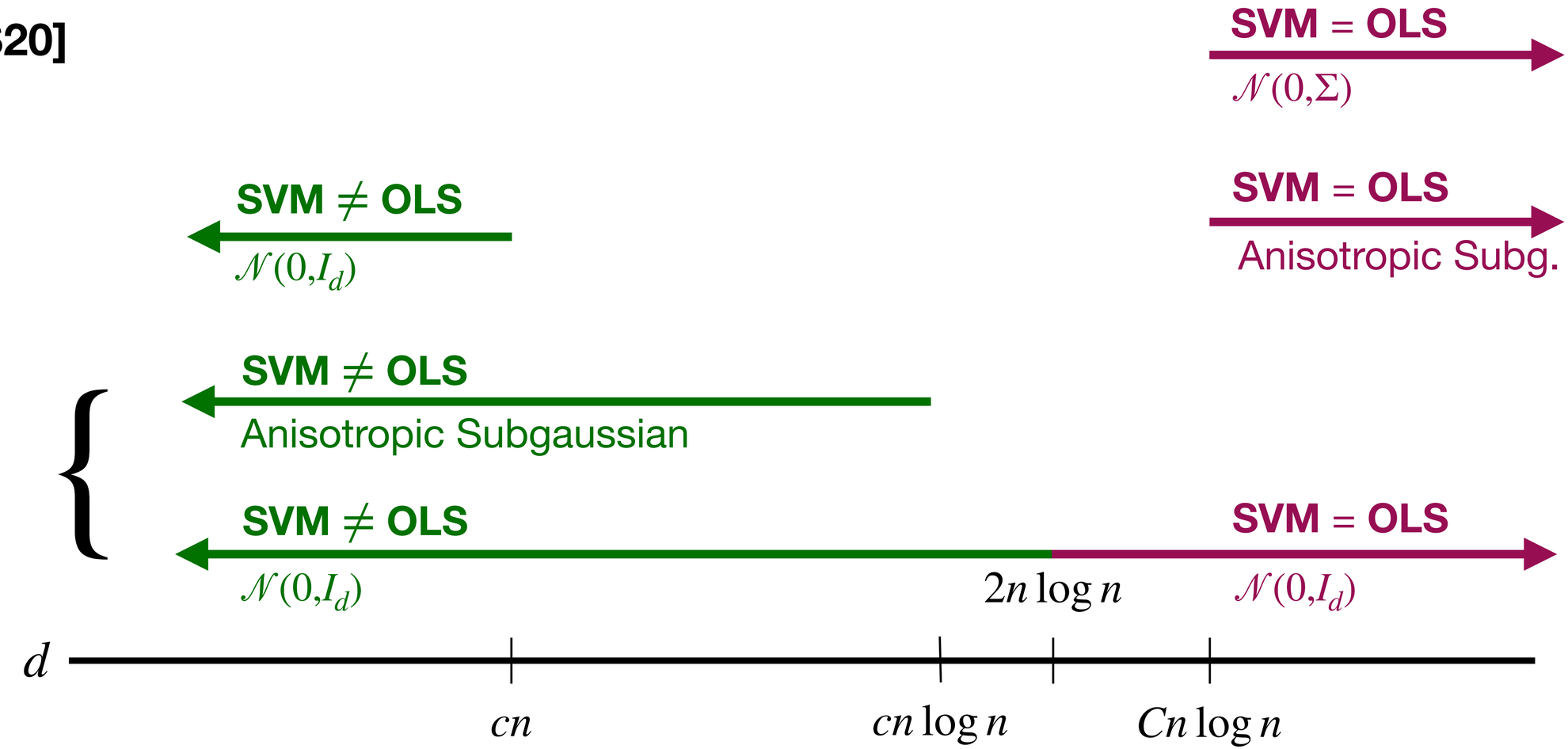
2. Then, "OLS classification" with $y_i \in \{\pm 1\}$ and prediction rule $x \mapsto \operatorname{sign}(x^T \theta_{OLS})$ also has benign overfitting.

3. Given dimension $d = \Omega(n^{3/2} \log n)$, same weights returned by OLS classification and SVM: $\theta_{OLS} = \theta_{SVM}$.

Tightening [MNSBHS20]…

Question: For what $d = d(n)$ do we have **SVM** = **OLS** with high probability?

- **[MNSBHS20]**
- **[HMX20]**
- **[ASH21]**



Our setting

(More general in the paper!)

- **Data model**: Labels are fixed and features are standard Gaussian

$$\boldsymbol{x}_i \sim \mathcal{N}(0, I_d),\ y_i \in \{\pm 1\},\ 1 \le i \le n$$

- **Question**: For what $d = d(n)$ do we have **SVM** = **OLS** with high probability?

$$\begin{aligned} &\min \ \|\theta\|_2 \\ &\text{s.t. } \ y_i \boldsymbol{x}_i^\intercal \theta \ge 1 \end{aligned} \qquad\qquad \begin{aligned} &\min \ \|\theta\|_2 \\ &\text{s.t. } \ \boldsymbol{x}_i^\intercal \theta = y_i \end{aligned}$$

- **SVM** = **OLS** if and only if every sample $\boldsymbol{x}_i$ is a support vector (i.e. $\boldsymbol{x}_i^T \theta = y_i$).

Features for sample i	$\boldsymbol{x}_i$
Label for sample i	y_i
Collection of Features except sample i	$\boldsymbol{X}_{\setminus i}$
Collection of labels except sample i	$y_{\setminus i}$

Proof ideas

Key lemma [HMX20]

$$\max_{i \le n} \left\{ \left\langle \, y_i \boldsymbol{x}_i \, , \, \boldsymbol{X}_{\setminus i}^\intercal \left(\boldsymbol{X}_{\setminus i} \boldsymbol{X}_{\setminus i}^\intercal \right)^{-1} y_{\setminus i} \, \right\rangle \right\} < 1 \iff \text{All samples are support vectors (SVM = OLS)}$$

- **Proof**: Analysis of $(\boldsymbol{X}_{\setminus i} \boldsymbol{X}_{\setminus i}^\intercal)^{-1}$ with Cramer's rule.
- Intuition: $w_{OLS}^{(i)} = \boldsymbol{X}_{\setminus i}^\intercal \left(\boldsymbol{X}_{\setminus i} \boldsymbol{X}_{\setminus i}^\intercal \right)^{-1} y_{\setminus i} \implies \langle \, y_i \boldsymbol{x}_i \, , \, w_{OLS}^{(i)} \, \rangle < 1$ then $\mathbf{x}_i$ is "necessary" in **SVM**

Features for sample i	$\boldsymbol{x}_i$
Label for sample i	y_i
Collection of Features except sample i	$\boldsymbol{X}_{\setminus i}$
Collection of labels except sample i	$y_{\setminus i}$

Proof ideas

- **Question:** For what values $d = d(n)$ do we have the following with high probability?

$$\max_{i \le n} \underbrace{\{\langle y_i \boldsymbol{x}_i \, , \, \boldsymbol{X}_{\setminus i}^{\intercal} (\boldsymbol{X}_{\setminus i} \boldsymbol{X}_{\setminus i}^{\intercal})^{-1} y_{\setminus i} \rangle\}}_{z_i} < 1$$

- $z_i \mid \boldsymbol{X}_{\setminus i} \sim \mathcal{N}(0, \|\boldsymbol{X}_{\setminus i}(\boldsymbol{X}_{\setminus i}\boldsymbol{X}_{\setminus i}^T)^{-1} y_{\setminus i}\|_2^2)$.
- $\|\boldsymbol{X}_{\setminus i}(\boldsymbol{X}_{\setminus i}\boldsymbol{X}_{\setminus i}^T)^{-1} y_{\setminus i}\|_2^2 = y_{\setminus i}^T (\boldsymbol{X}_{\setminus i}\boldsymbol{X}_{\setminus i}^T)^{-1} y_{\setminus i}$
- Gaussian concentration: $\boldsymbol{X}_{\setminus i}\boldsymbol{X}_{\setminus i}^T \approx d I_{n-1}$ and $(\boldsymbol{X}_{\setminus i}\boldsymbol{X}_{\setminus i}^T)^{-1} \approx 1/d \cdot I_{n-1}$.
- $\|\boldsymbol{X}_{\setminus i}(\boldsymbol{X}_{\setminus i}\boldsymbol{X}_{\setminus i}^T)^{-1} y_{\setminus i}\|_2^2 \approx \|y_{\setminus i}\|_2^2 / d = (n-1)/d$.
- $z_i \mid \boldsymbol{X}_{\setminus i}$ behaves roughly as $\mathcal{N}\left(0, \, (n-1)/d\right)$

Proof ideas

Features for sample i	$\boldsymbol{x}_i$
Label for sample i	y_i
Collection of Features except sample i	$\boldsymbol{X}_{\setminus i}$
Collection of labels except sample i	$y_{\setminus i}$
$\langle y_i \boldsymbol{x}_i \, , \, \boldsymbol{X}_{\setminus i}^{\mathsf{T}} (\boldsymbol{X}_{\setminus i} \boldsymbol{X}_{\setminus i}^{\mathsf{T}})^{-1} y_{\setminus i} \rangle$	z_i

- **Question:** For what values $d = d(n)$ do we have $\max_{i \le n} z_i < 1$ with high probability?
- $z_i \mid X_{\setminus i}$ behaves roughly as $\mathcal{N}\left(0, \, (n-1)/d\right)$.
- **If** z_i's were independent: $\max_{i \le n} z_i = \Theta_p\left(\sqrt{2n \log(n)/d}\right)$.
  - $\Longrightarrow d = \Theta(n \log(n))$ is a critical threshold
- Remainder of proof: showing that same threshold occurs because **dependence** among z_i's is weak.

Proof ideas

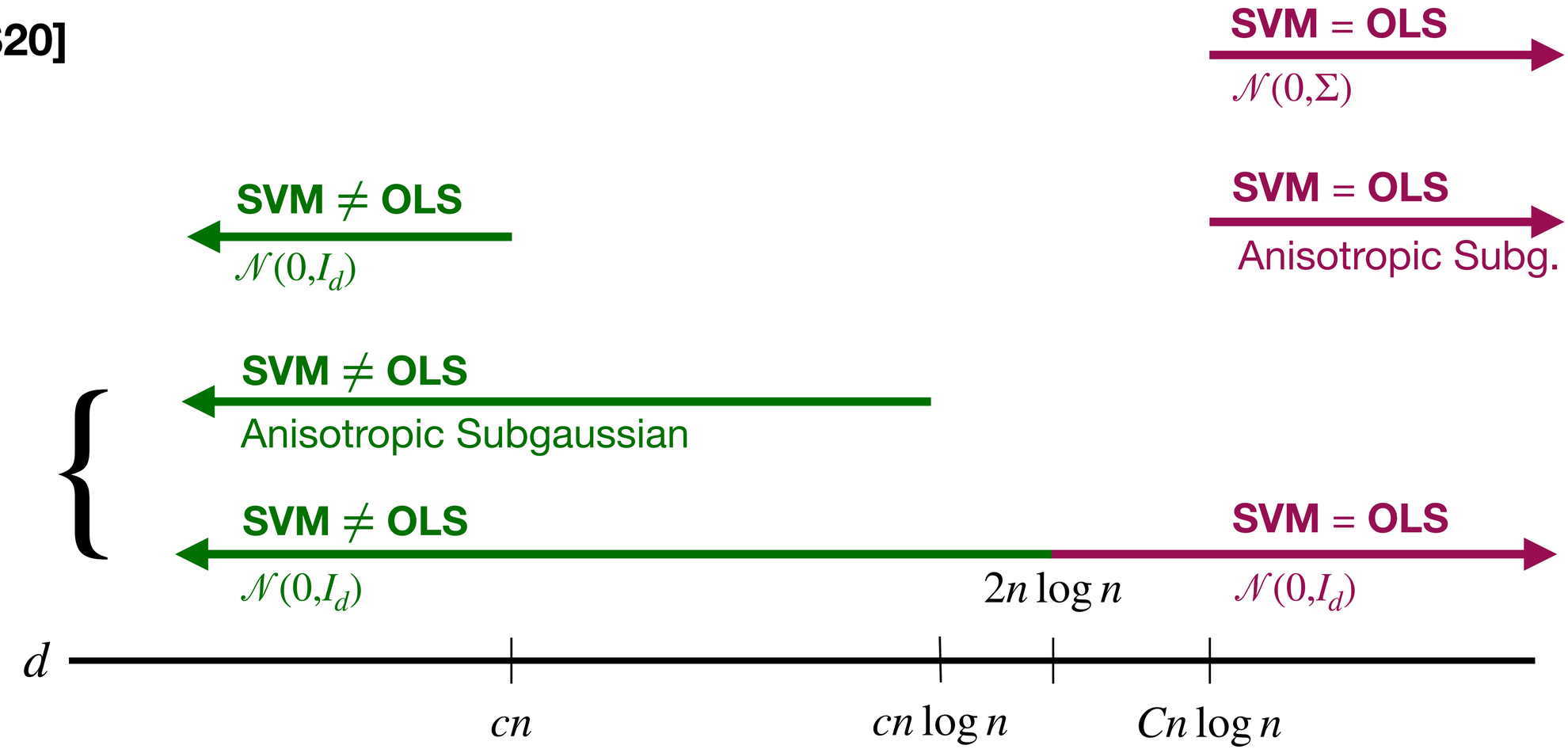
Features for sample i	$\boldsymbol{x}_i$
Label for sample i	y_i
Collection of Features except sample i	$\boldsymbol{X}_{\setminus i}$
Collection of labels except sample i	$y_{\setminus i}$
$\langle y_i \boldsymbol{x}_i \, , \, \boldsymbol{X}_{\setminus i}^\intercal (\boldsymbol{X}_{\setminus i} \boldsymbol{X}_{\setminus i}^\intercal)^{-1} y_{\setminus i} \rangle$	z_i

- **Question:** For what values $d = d(n)$ do we have $\max_{i \le n} z_i < 1$ with high probability?
- Idea: Split z_i into three terms, which can be more easily controlled by considering a subsample of $m \ll n$ samples. $z_i = z_i^{(1)} + z_i^{(2)} + z_i^{(3)}$:
  - $z_i^{(1)} := \langle y_i \boldsymbol{x}_i \, , \, \boldsymbol{X}_{\setminus i}^\intercal [(\boldsymbol{X}_{\setminus i} \boldsymbol{X}_{\setminus i}^\intercal)^{-1} - 1/d \cdot I_{n-1}] y_{\setminus i} \rangle$
  - $z_i^{(2)} := 1/d \cdot \langle y_i \boldsymbol{x}_i \, , \, \boldsymbol{X}_{[m]\setminus i}^\intercal y_{[m]\setminus i} \rangle$
  - $z_i^{(3)} := 1/d \cdot \langle y_i \boldsymbol{x}_i \, , \, \boldsymbol{X}_{\setminus [m]}^\intercal y_{\setminus [m]} \rangle$
- $\max_{i \le n} z_i \ge 1 \Longleftarrow \max_{i \le m} (z_i^{(1)} + z_i^{(2)} + z_i^{(3)}) \ge 1 \Longleftarrow$
  $\max_{i \le m} |z_i^{(1)}| \le 1 \Longleftarrow$ subgaussian concentration
  $\max_{i \le m} |z_i^{(2)}| \le 1 \Longleftarrow$ subgaussian concentration
  $\max_{i \le m} z_i^{(3)} \ge 3 \Longleftarrow$ anti-concentration: Berry-Esseen

Our results, recap

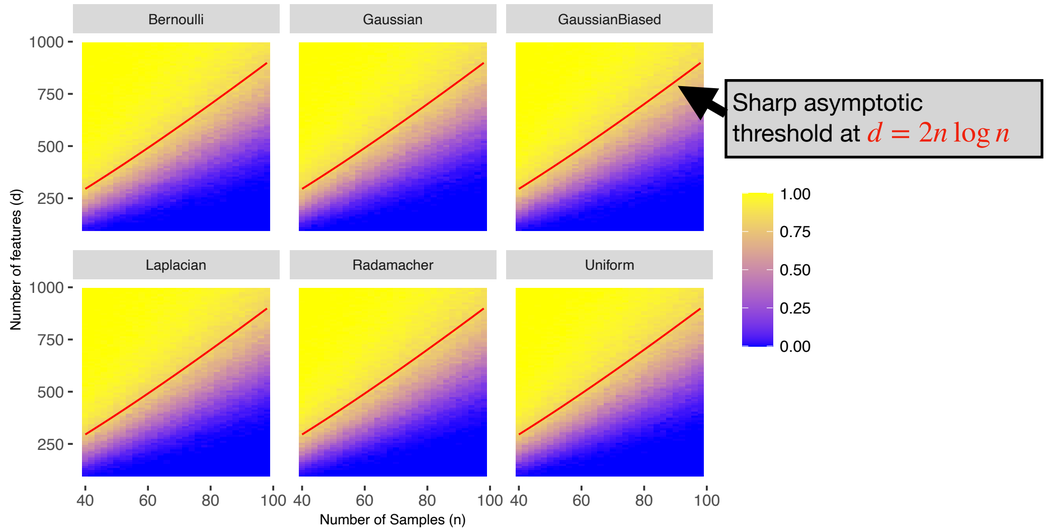
Question: For what $d = d(n)$ do we have **SVM = OLS** with high probability?

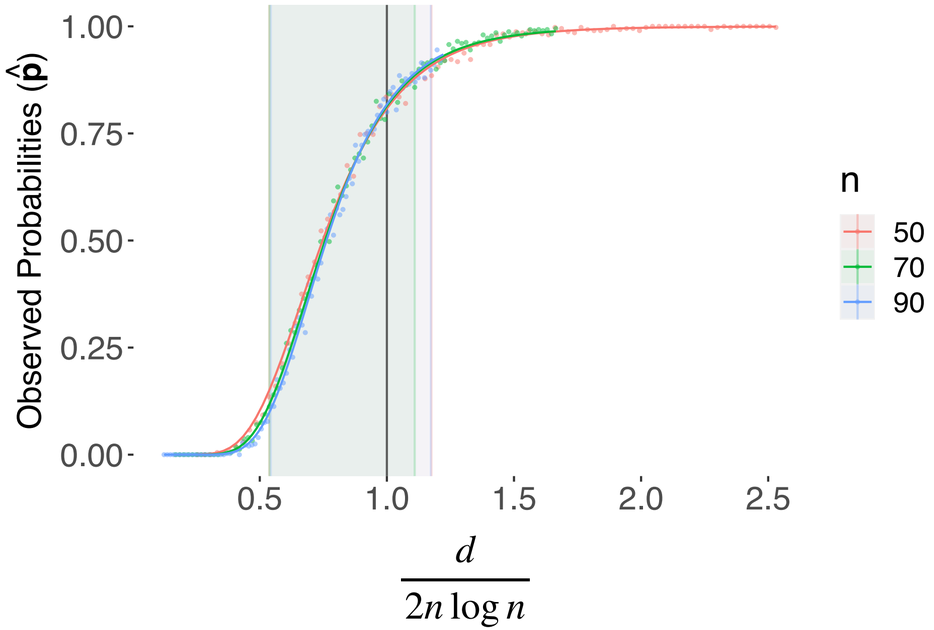
- **[MNSBHS20]**
- **[HMX20]**
- **[ASH21]**



Other contributions

- Extension to **anisotropic subgaussian** inputs and **linearly-separable** outputs
- Universality of **SVP** phenomenon (**SVM** = **OLS**) under different feature distributions
- Asymptotic rate of convergence of **SVM** = **OLS** threshold to $d = 2n \log n$
- A geometric interpretation of when **SVM** = **OLS**
- Statistically rigorous study of empirical universality (à la [Donoho & Tanner, '09])

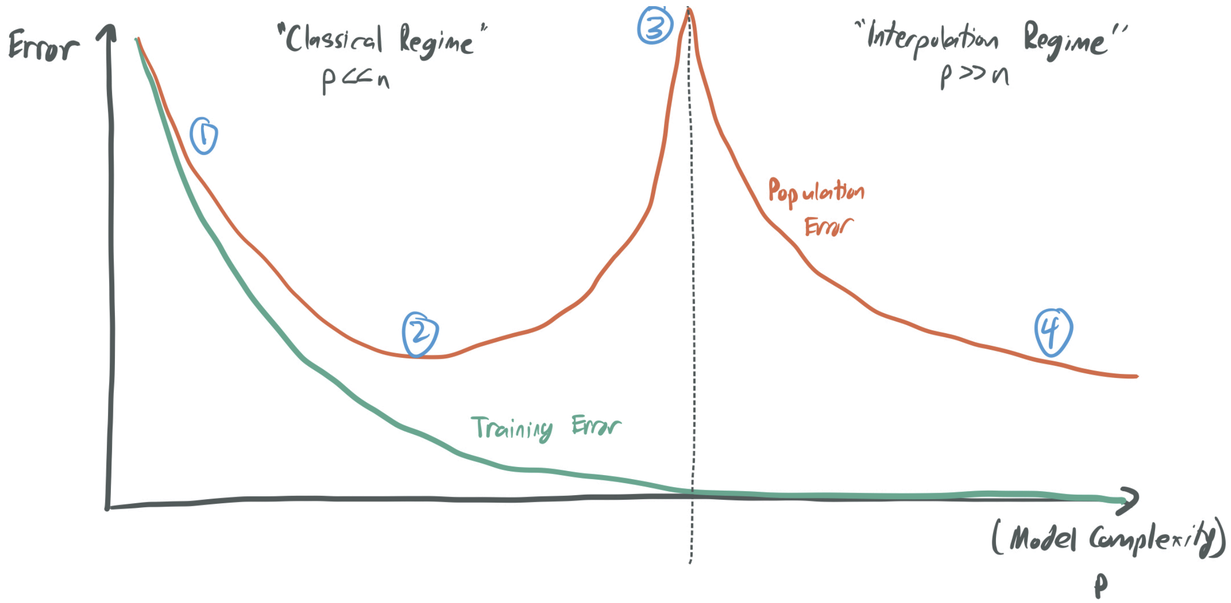




Zooming back out…

Questions to ask about over-parameterization/benign overfitting/neural networks

- When does benign overfitting occur for more complex neural networks?
    - Especially those behaving very differently from linear/kernel models.
- Which data assumptions are realistic to make for ML applications?
- What is the role of voting/averaging in benign overfitting?
- What is the implicit bias of gradient descent, and does it connect to simpler models?



Thank you
(Really, thank you)