Generalization Theory for Deep Learning

Rademacher Complexity

Intuition: how well can a classifier class **fit random noise?**

(Empirical) **Rademacher complexity:** For a training set $S = \{x_1, x_2, \ldots, x_n\}$, and a class $\mathcal{F}$, denote:

$$\hat{R}_n(S) = \mathbb{E}_\sigma \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \sigma_i f(x_i) \ .$$

where $\sigma_i \sim$ Unif$\{+1, -1\}$ (Rademacher R.V.).

(Population) **Rademacher complexity:**

$$R_n = \mathbb{E}_S \left[\hat{R}_n(s) \right].$$

Rademacher Complexity Generalization Bound

Theorem: with probability $1 - \delta$ over the choice of a training set, for a bounded loss ℓ, we have

$$\sup_{f \in \mathcal{F}} \left| \underbrace{\frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i)}_{\text{training error}} - \mathbb{E}_{(x,y)\sim D} \left[\ell(f(x), y) \right] \right| = O\left(\frac{\hat{R}_n}{n} + \sqrt{\frac{\log 1/\delta}{n}} \right)$$

$$\hat{R}_n \leq O(VC(\mathcal{F}))$$

and

$$\sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i) - \mathbb{E}_{(x,y)\sim D} \left[\ell(f(x), y) \right] \right| = O\left(\frac{R_n}{n} + \sqrt{\frac{\log 1/\delta}{n}} \right)$$

$S = \{x_1, \ldots, x_n\}$

Norm-based Rademacher complexity bound

σ: ReLU, $\rho = 1$

Theorem: If the activation function is σ is ρ-Lipschitz. Let $\mathcal{F} = \{x \mapsto W_{H+1}\sigma(W_h\sigma(\cdots\sigma(W_1x)\cdots)), \|W_h^T\|_{1,\infty} \le B \forall h \in [H]\}$ then $R_n(\mathcal{S}) \le \underbrace{\|X^\top\|_{2,\infty}}_{O(\sqrt{n})}\underbrace{(2\rho B)^{H+1}}_{\to \text{ Cap of } \mathcal{F}}\sqrt{2\ln d}$ where $X = [x_1, \ldots, x_n] \in \mathbb{R}^{d\times n}$ is the input data matrix.

$x \in \mathbb{R}^d$, $W_1 \in \mathbb{R}^{m\times d}$, $W_2, \ldots, W_H \in \mathbb{R}^{m\times m}$, $W_{H+1} \in \mathbb{R}^m$

$\|W_1^T\|_{1,\infty} = \max_{i=1,\ldots,m} \|W_1^T(:,i)\|_1$

$= \max_{i=1,\ldots,m} \sum_{j=1}^d |W_1^T(j,i)|$

W_1^T $\begin{array}{c|c} & m \\ \hline d & \boxed{} \end{array}$

$\|W_h^T\|_{1,\infty} = \max_{i=1,\ldots,m} \|W_h^T(:,i)\|_1$

X^T $\begin{array}{c|c} & d \\ \hline n & \boxed{} \end{array}$

$\|X^T\|_{2,\infty} = \max_{i=1,\ldots,d} \|X^T(:,i)\|_2$

$= \max_{i=1,\ldots,d} \|(x_1(i), x_2(i), \ldots, x_n(i))\|_2$

Massart Lemma

Lemma: Let V be a set of vectors $\subset \mathbb{R}^d$, we have

$$R_V := \mathbb{E}_{\epsilon \sim \text{Unif}\{1,-1\}^d} \max_{v \in V} \langle \epsilon, v \rangle \leq \sup_{v \in V} \|v\|_2 \sqrt{2 \ln d}.$$

- $\varepsilon \sim \text{Unif}\{1,-1\}^d$
- Fix $u \in V$

Define $X_{u,i} = \varepsilon_i \cdot u_i$

$X_u = \sum_i \varepsilon_i \cdot u_i$

$(X_{u,i})^2 = u_i^2$

$\Rightarrow$ $X_{u,i}$ is u_i^2-subGaussian

$\Rightarrow$ $\|X_u\|_2$ is $\|u\|_2^2$-subGaussian

$\Rightarrow$ $\forall u$, X_u is $\sup_{u \in V} \|u\|_2^2$ subGaussian

$$\mathbb{E}_\varepsilon \max_{u \in V} \langle \varepsilon, u \rangle \leq \underbrace{\sup_{u \in V} \|u\|_2 \cdot \sqrt{2 \ln d}}$$

subGaussian expectation inequality

Properties of Rademacher Complexity

Let V be a set of vectors $\subset \mathbb{R}^d$:

$u_i \in V$

$\sum \lambda_i = 1$

$u \in \mathrm{Conv}(V)$ iff $u = \sum_{i=1}^{n} \lambda_i u_i$, $\lambda_i \geq 0$

1. $R_{\mathrm{Conv}(V)} = R_V$,
2. $R_V = R_{-V}$.

3. Let $V_1, \ldots, V_m$ be m set of vectors such that for anyway $\epsilon \in \{-1, +1\}^d$, $\sup_{v \in V_i} \langle v, \epsilon \rangle \geq 0$ (e.g., $0 \in V_i$), then we have

$$R_{\cup_{i=1}^m V_i} \leq \sum_{i=1}^{m} R_{V_i}.$$

Proof of $(1,\infty)$ norm-based bound

Layer by layer induction

Let $\mathcal{F}_h = \{ x \mapsto \sigma(W_h \sigma(W_{h-1} \cdots \sigma(W_1 x) \cdots) \in \mathbb{R}^m$
$\|W_{h'}^T\|_{1,\infty} \le B,\ h' = 1, \ldots, h \}$

$\mathcal{F}_0 = \{ x \mapsto x \}$

Induction: $\mathrm{Rad}(\mathcal{F}_h|_S) \le \|X^T\|_{2,\infty} (2\rho B)^h \sqrt{2 \ln d}$

take $h = H+1 \Rightarrow$ theorem

(1) $h = 0$ ✓

$\mathcal{F}_0 = \{ x \mapsto (x(1), \ldots, x(d)) : d$ functions $\}$

Massart lemma

$\mathrm{Rad}(\mathcal{F}_0|S) \le \left(\max_{j \in \{1,\ldots,d\}} \|(x_1(j), \ldots, x_n(j))\|_2 \right) \cdot \sqrt{2 \ln d}$

$= \|\cancel{x}^T\|_{2,\infty} \cdot \sqrt{2 \ln d} = \|X^T\|_{2,\infty} \cdot (2\rho B)^0 \sqrt{2 \ln d}$

Proof of $(1,\infty)$ norm-based bound

(2) Induction step, assume hypothesis is true till layer h

$$F_{h+1}|_S = \{x \mapsto \sigma(W_{h+1}\, g(x)),\ g \in F_h,\ \|W_{h+1}\|_{1,\infty} \le \beta\}$$

$$\sigma(W_{h+1}\, g(x)) \in \mathbb{R}^m$$

$$x \mapsto \sigma(W_{h+1}\, g(x))$$

$$[\sigma(W_{h+1}\, g(x))](i), \quad i = 1, \dots, m$$

$$= \sigma(W_{h+1}(i,:)\, g(x))$$

$$\Longleftrightarrow \quad x \mapsto \sigma(\underbrace{\|W_{h+1}(i,:)\|_1}_{\text{magnitude}} \cdot \underbrace{v}_{\text{direction}}{}^T g(x)), \quad v = \frac{W_{h+1}(i,:)}{\|W_{h+1}(i,:)\|_1}$$

$\|v\|_1 = 1$ $\Longleftrightarrow$ $\sum |v(j)| = 1$

$$= \sigma\left(\|W_{h+1}(i,:)\|_1 \cdot \sum_{j=1}^{m} v(j) \cdot [g(x)](j)\right)$$

$$= \sigma\left(\|W_{h+1}(i,:)\|_1 \cdot \underbrace{\sum_{j=1}^{m} |v(j)| \cdot \operatorname{sgn}(v_j)\, [g(x)](j)}\right)$$

$$\Rightarrow F_{h+1}|_S = \{x \mapsto \sigma(\|W_{h+1}(i,:)\|_1 \cdot g(x)),\ g(x) \in \operatorname{conv}(F_h \cup (-F_h))\}$$

Proof of $(1,\infty)$ norm-based bound

Lipshitz & scaling property of Rad.

$$\mathrm{Rad}(\mathcal{F}_{h+1}|_S)$$

$$\leq \rho B \,\mathrm{Rad}\left(\mathrm{conv}\left((\mathcal{F}_h|_S) \cup (-\mathcal{F}_h|_S)\right)\right)$$

$$= \rho B \,\mathrm{Rad}(\mathcal{F}_h|_S \cup (-\mathcal{F}_h|_S)) \quad \text{(property 1)}$$

$$\leq \rho B\left(\mathrm{Rad}(\mathcal{F}_h|_S)\right) + \rho B \cdot \mathrm{Rad}(-\mathcal{F}_h|_S) \quad \text{(property 3)}$$

$$= 2\rho B \,\mathrm{Rad}(\mathcal{F}_h|_S) \quad \text{(property 2)}$$

$$\leq (2\rho B)^{h+1} \|X^T\|_{2,\infty} \sqrt{2\log d}$$

$\square$

Comments on generalization bounds

- When plugged in real values, the bounds are rarely non-trivial (i.e., smaller than 1)
- "*Fantastic Generalization Measures and Where to Find them*" by Jiang et al. '19 : large-scale investigation of the correlation of extant generalization measures with true generalization.

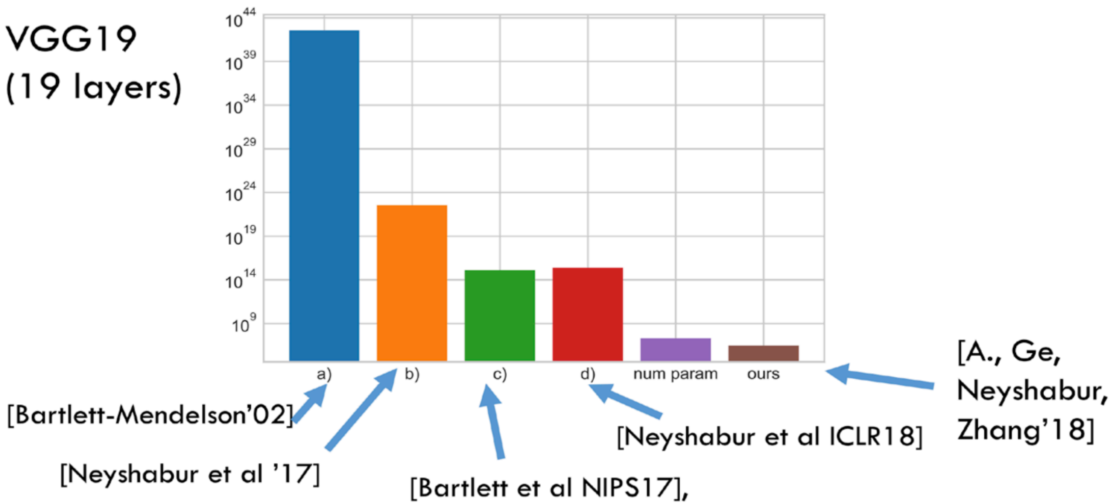


Image credits to Andrej Risteski

Comments on generalization bounds

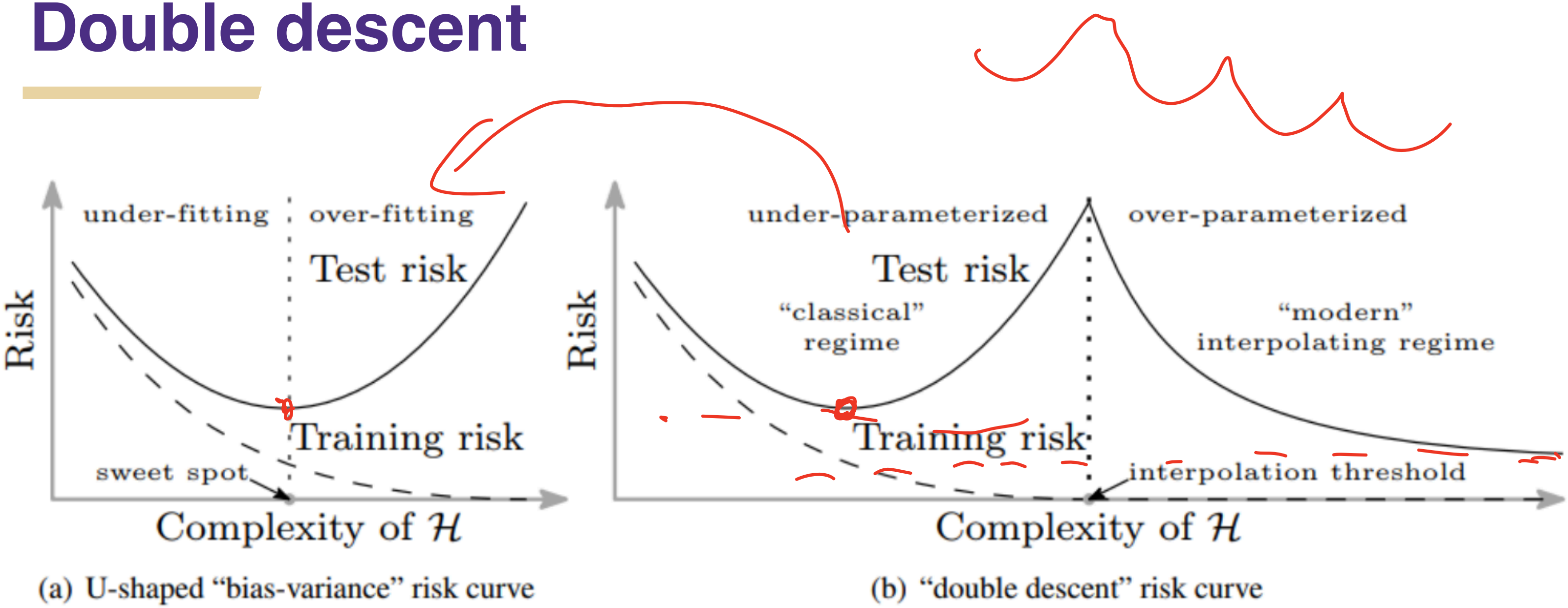
- Uniform convergence may be unable to explain generalization of deep learning [Nagarajan and Kolter, '19]
  - Uniform convergence: a bound for all $f \in \mathcal{F}$
  - Exists example that 1) can generalize, 2) uniform convergence fails.

bias-variance decomposition

K-nearest-neighbor

- Rates:
  - Most bounds: $1/\sqrt{n}$.
  - Local Rademacher complexity: $1/n$.

Double descent

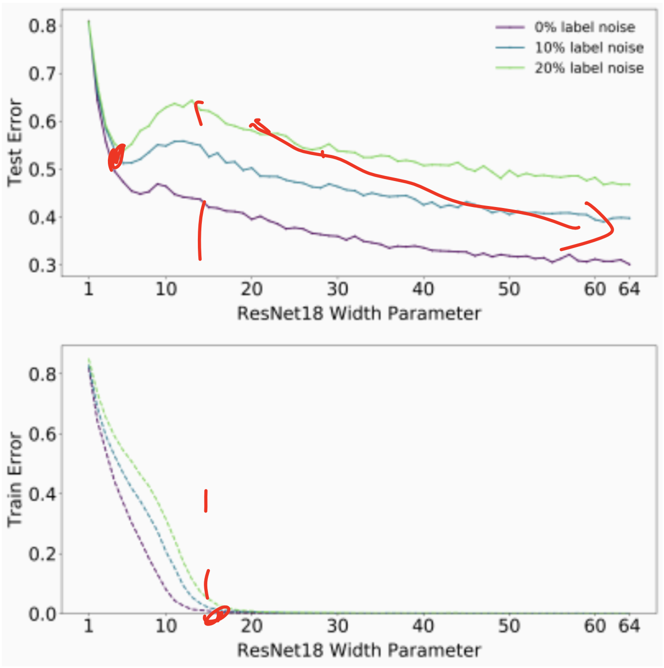


Belkin, Hsu, Ma, Mandal '18

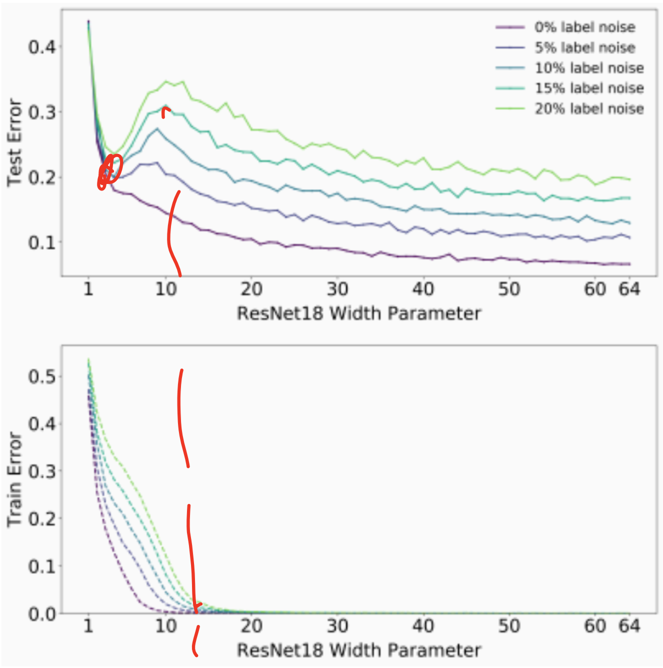
- There are cases where the model gets bigger, yet the (test!) loss goes down, sometimes even lower than in the classical "under-parameterized" regime.
- Complexity: number of parameters.

Double descent

Widespread phenomenon, across architectures (Nakkiran et al. '19):



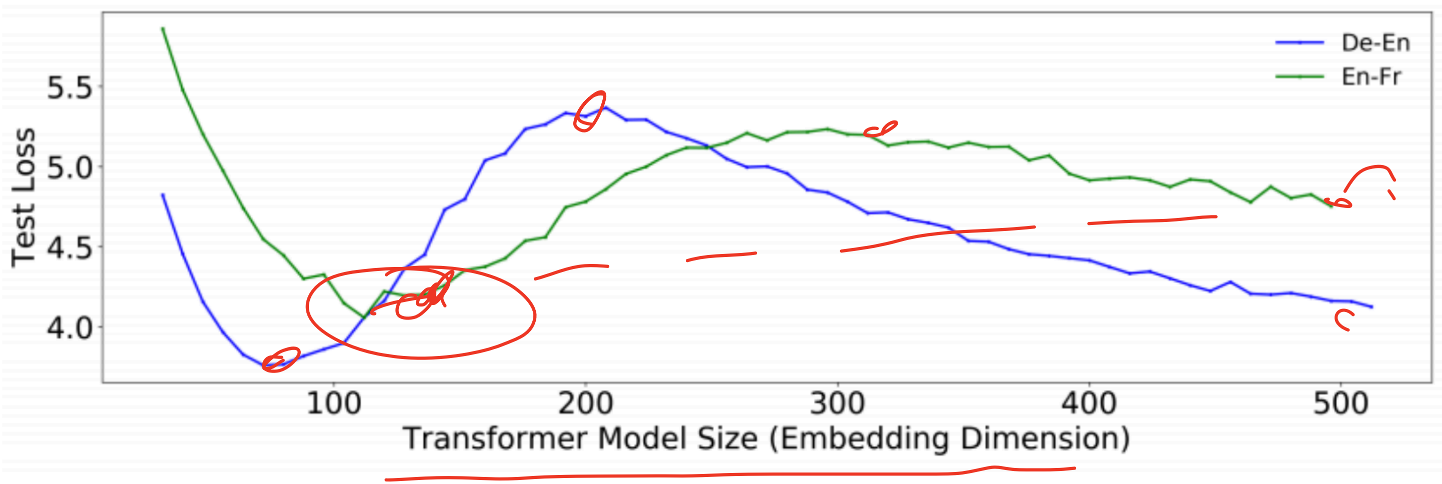
(a) **CIFAR-100.** There is a peak in test error even with no label noise.



(b) **CIFAR-10.** There is a "plateau" in test error around the interpolation point with no label noise, which develops into a peak for added label noise.

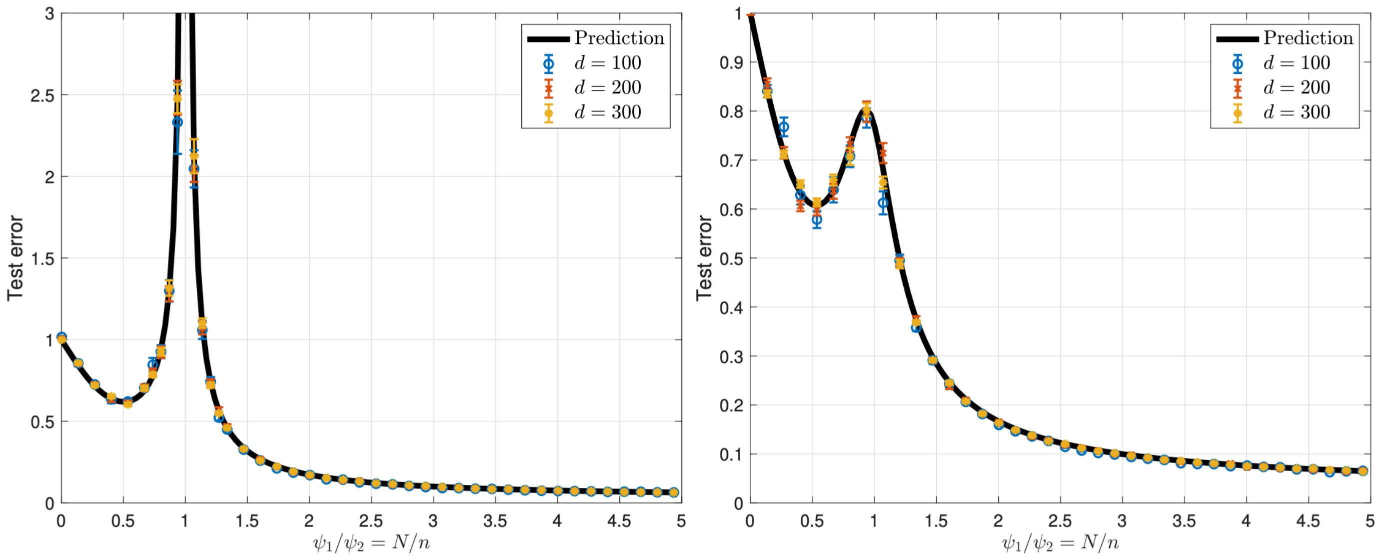
Double descent

Widespread phenomenon, across architectures (Nakkiran et al. ’19):



Double descent

Widespread phenomenon, also in kernels (can be formally proved in some concrete settings [Mei and Montanari ’20]), random forests, etc.



Double descent

Also in other quantities such as train time, dataset, etc (Nakkiran et al. ’19):

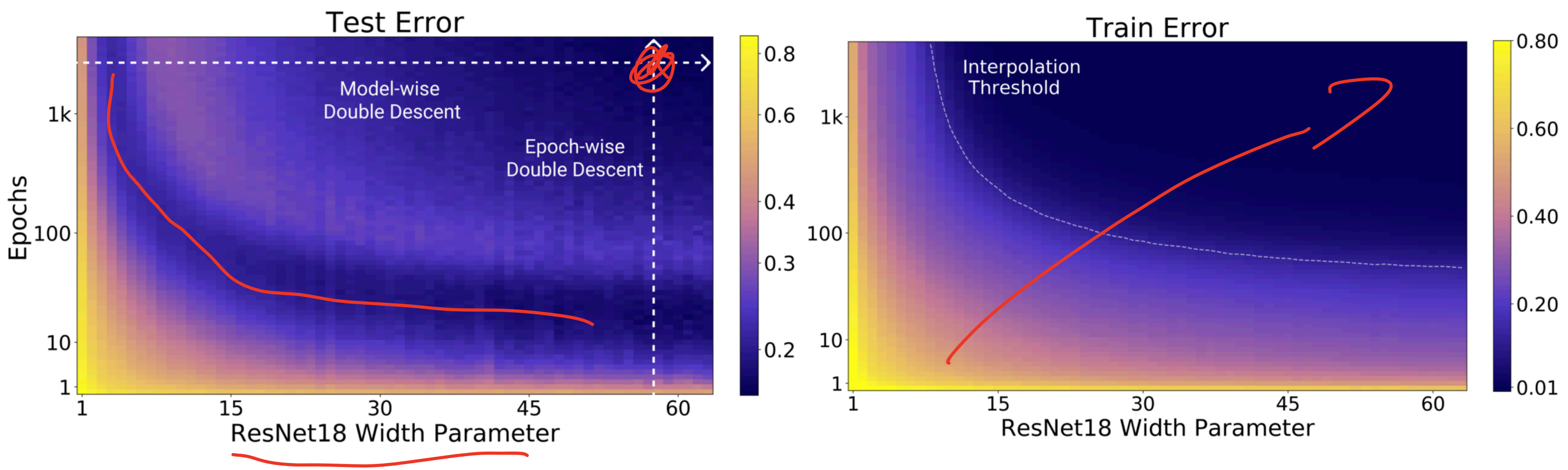
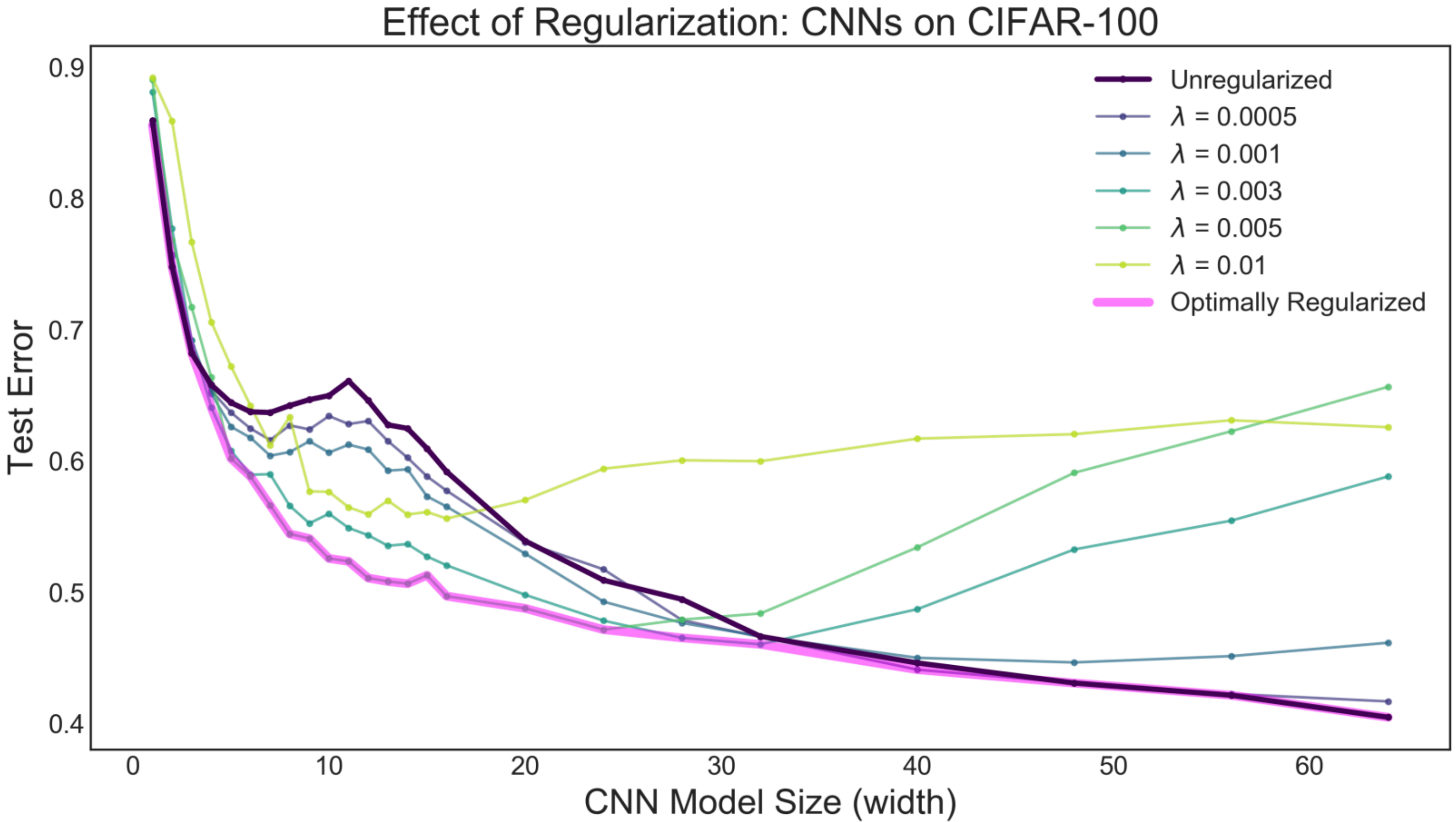


Figure 2: **Left:** Test error as a function of model size and train epochs. The horizontal line corresponds to model-wise double descent–varying model size while training for as long as possible. The vertical line corresponds to epoch-wise double descent, with test error undergoing double-descent as train time increases. **Right** Train error of the corresponding models. All models are Resnet18s trained on CIFAR-10 with 15% label noise, data-augmentation, and Adam for up to 4K epochs.

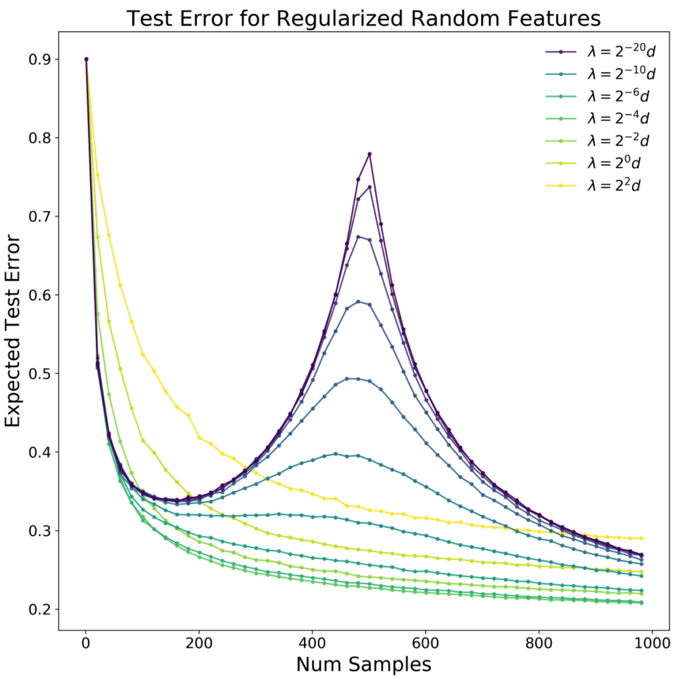
Double descent

Optimal regularization can mitigate double descent [Nakkiran et al. ’21]:

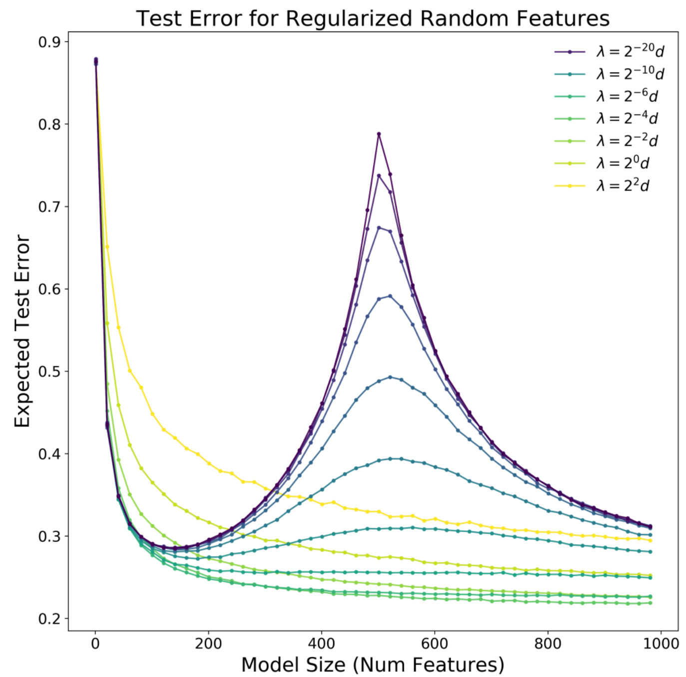


Double descent

Optimal regularization can mitigate double descent [Nakkiran et al. '21]:



a) Test Classification Error vs. Number of Train-
ng Samples.



(b) Test Classification Error vs. Model Size (Number of Random Features).

Implicit Regularization

params >> n

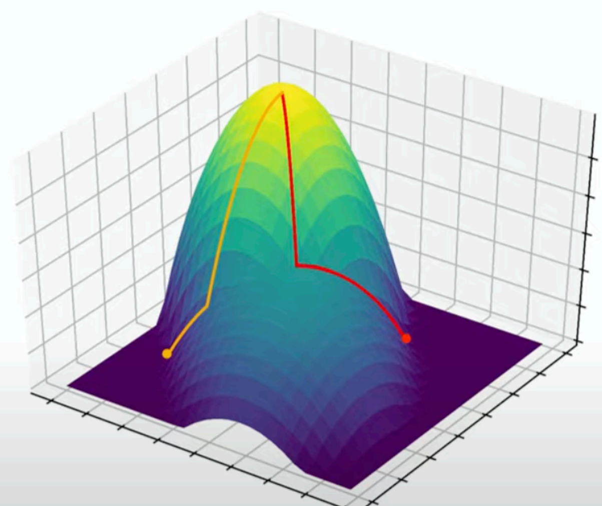
> 1 global min

Different optimization algorithm , initialization, hyper-parameter

➔ Different bias in optimum reached

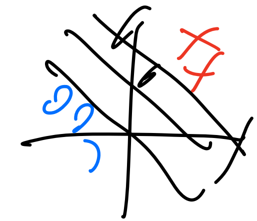
➔ Different Inductive bias

➔ Different generalization properties



Implicit Bias

linear $\bar{u} = \arg\max_{\|w\|_2 \le 1} \min_{i=1,\dots,n} y_i \langle w, x_i \rangle$



Margin: $m_i(w) = y_i \cdot f(w, x_i)$

$(H+1)$-homogeneous $\bar{u} = \dfrac{\min_i m_i}{\|W\|^{H+1}}$

- Linear predictors:
  - Gradient descent, mirror descent, natural gradient descent, steepest descent, etc maximize margins with respect to different norms.

- Non-linear:
  - Gradient descent maximizes margin for homogeneous neural networks.
  - Low-rank matrix sensing: gradient descent finds a low-rank solution.

$M = UU^T$ M: low-rank

U: full-rank

Separation between NN and kernel

- For approximation and optimization, neural network has no advantage over kernel. Why NN gives better performance: generalization.

- [Allen-Zhu and Li '20] Construct a class of functions $\mathscr{F}$ such that $y = f(x)$ for some $f \in \mathscr{F}$:
  - no kernel is sample-efficient;
  - Exists a neural network that is sample-efficient.