Double Descent / Implicit Regularization + PAC-Bayes

CPSC 532S: Modern Statistical Learning Theory

30 March 2022

cs.ubc.ca/~dsuth/532S/22/

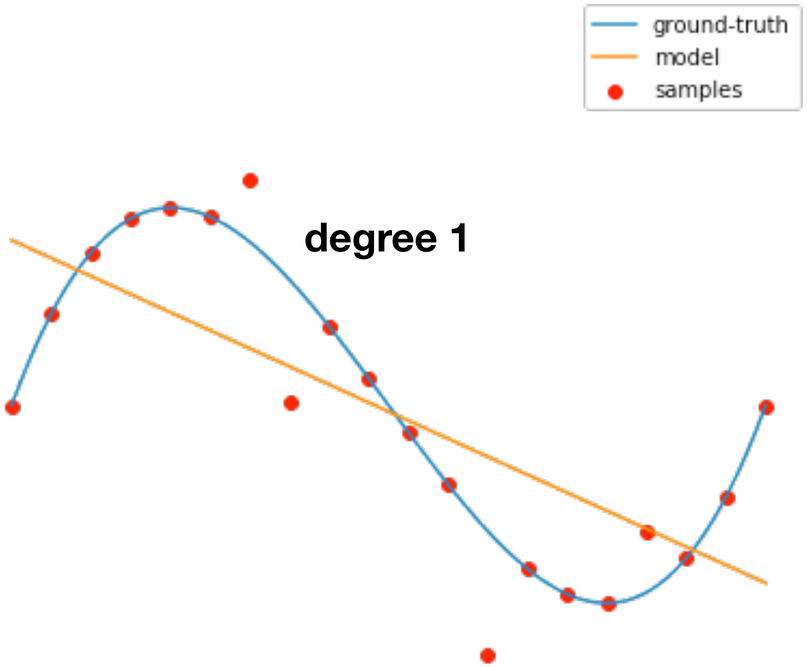
Admin

- A3 late-late deadline is tonight
- A4 out, due next Friday – late deadline Sunday, late-late deadline Tuesday

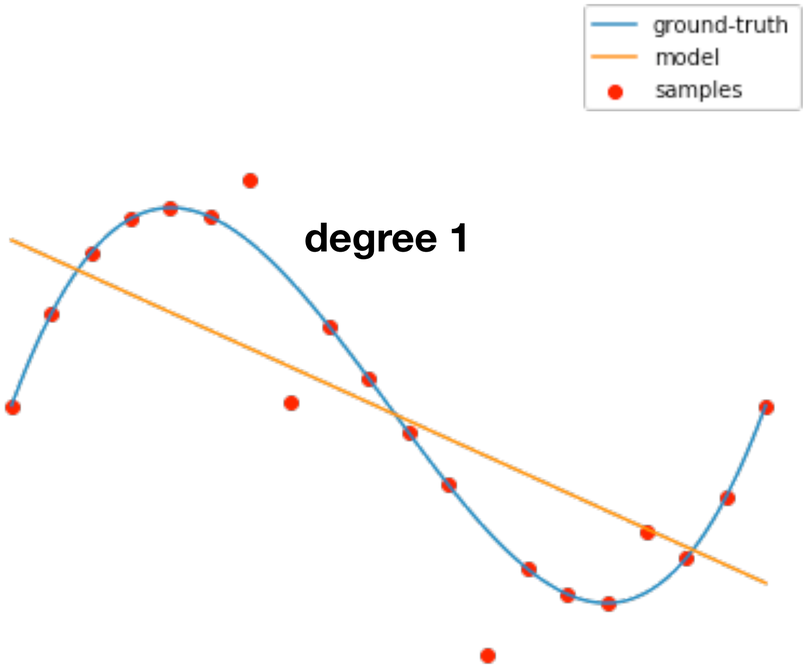
- Last regular class is Monday
  - Overview of online learning, privacy (~equivalent), connection to stability

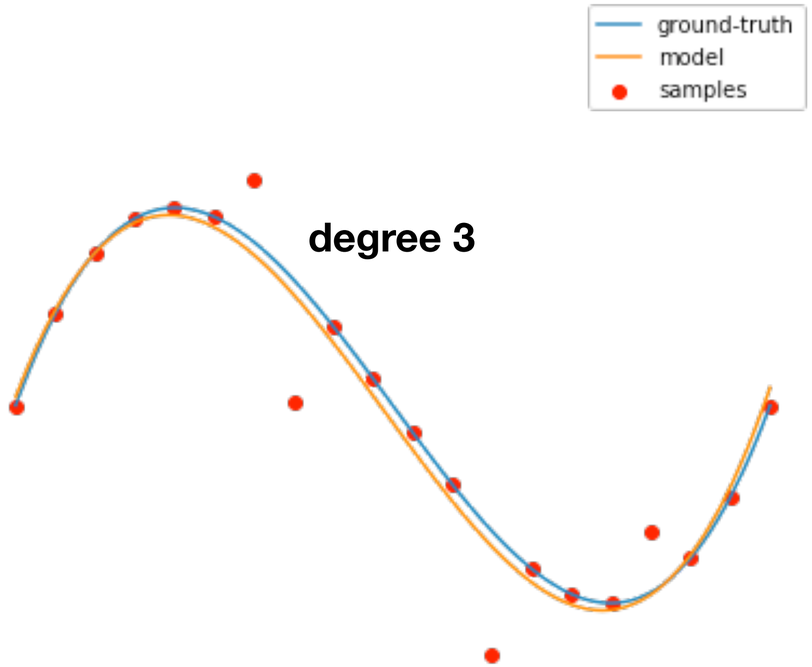
- Project presentations in class next Wednesday
  - Details were posted on Piazza a bit ago
  - Will post schedule tonight
- Project reports due next Friday – usual late policy (-5 Saturday, -10 Sunday)

- Final will be take-home, available over most of the finals period
  - Exact dates TBA – contact me if this really matters to you for whatever reason

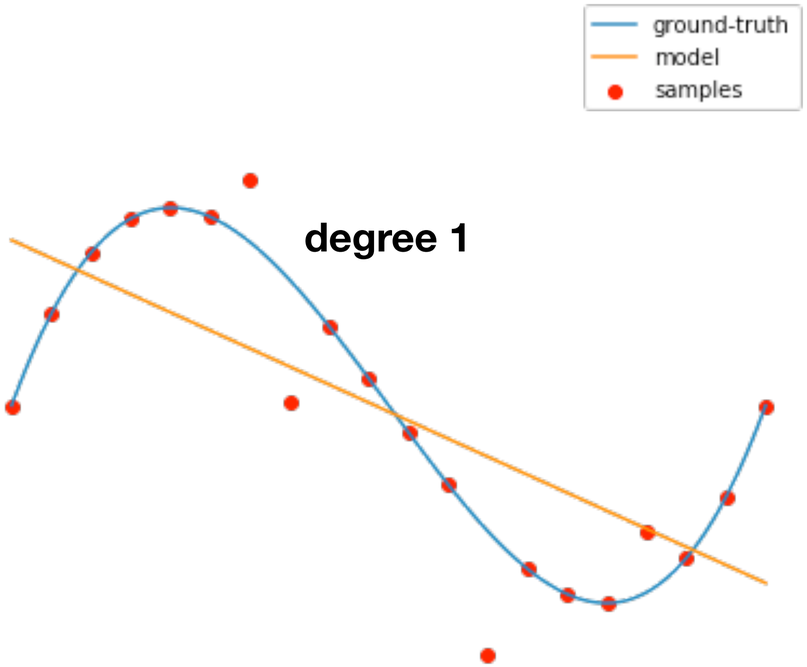


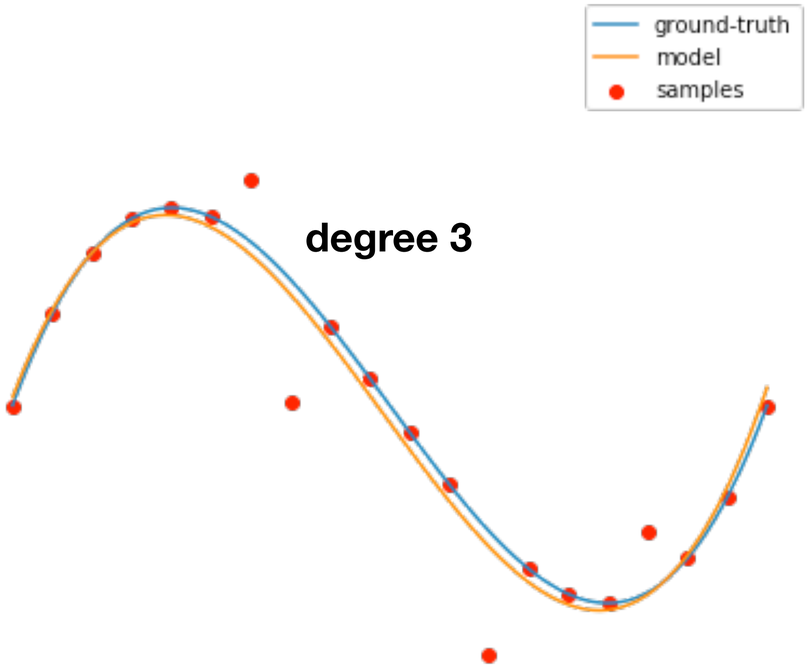
Nakkiran et al. blog post's companion notebook

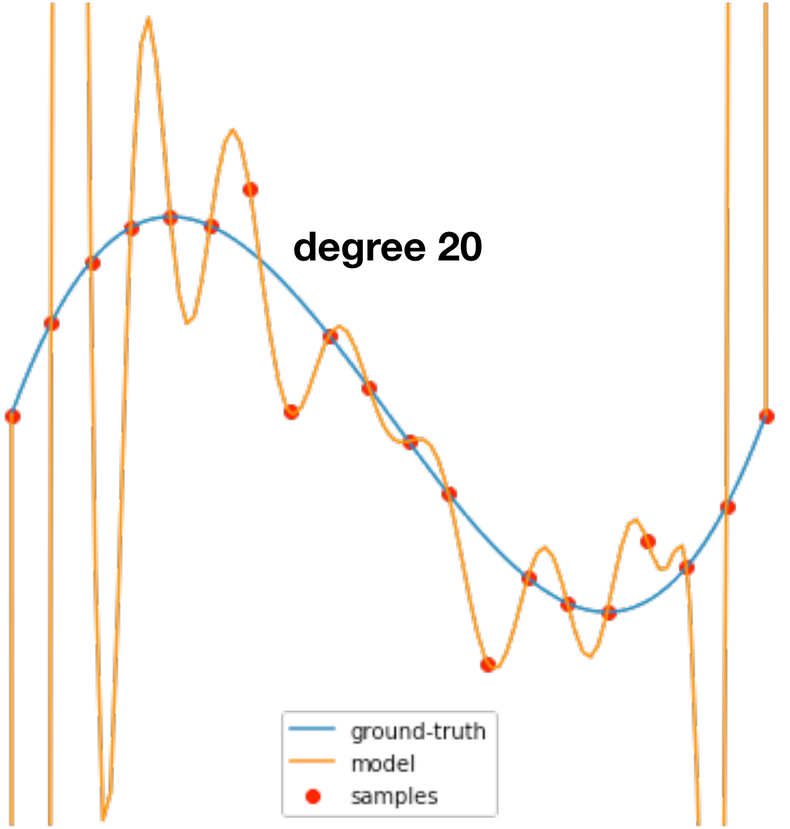




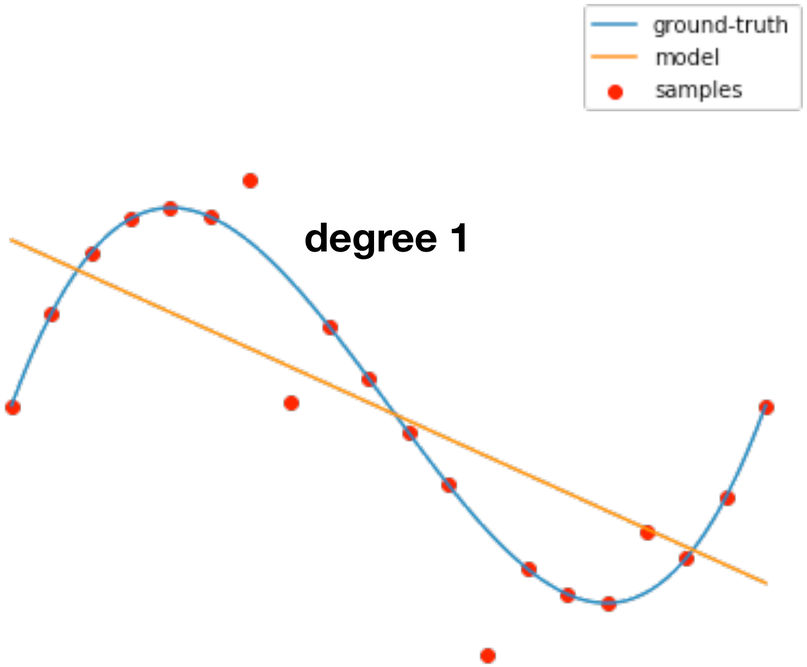
Nakkiran et al. blog post's companion notebook

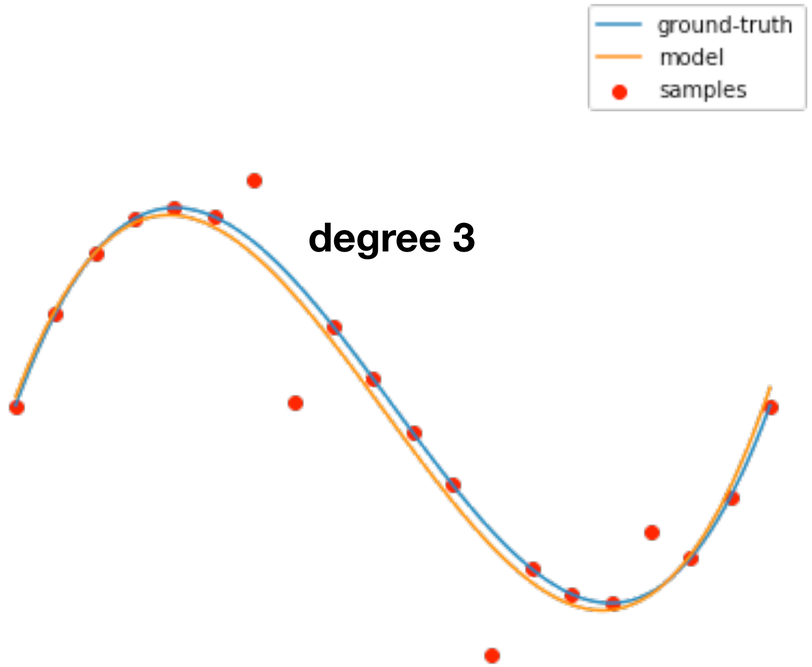


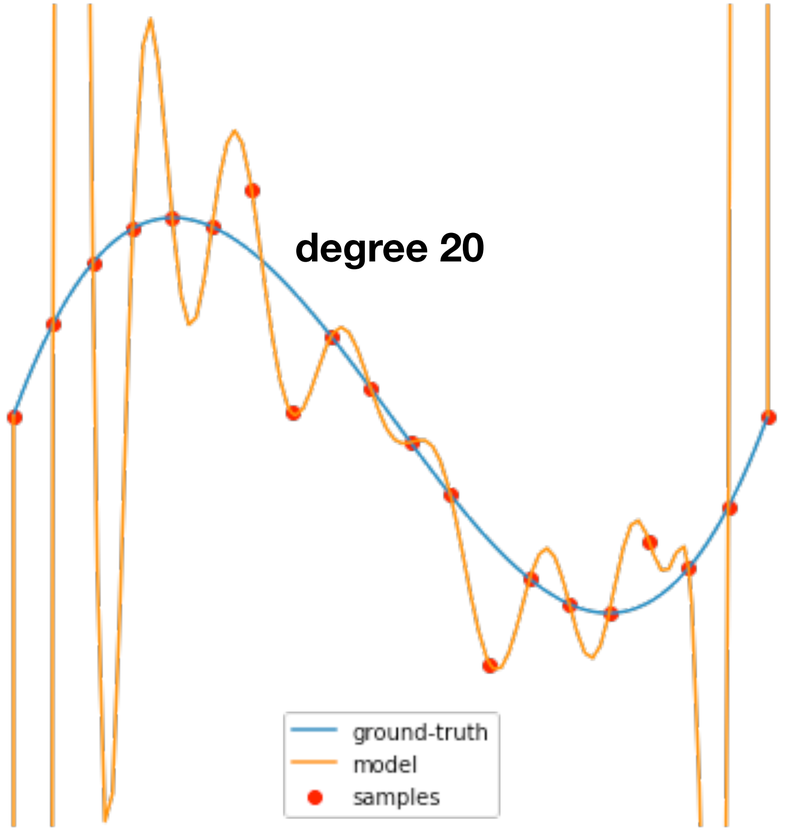


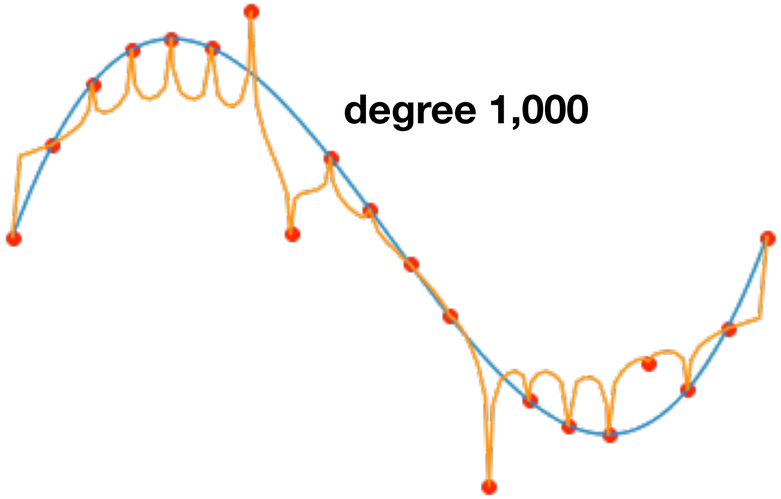


Nakkiran et al. blog post's companion notebook

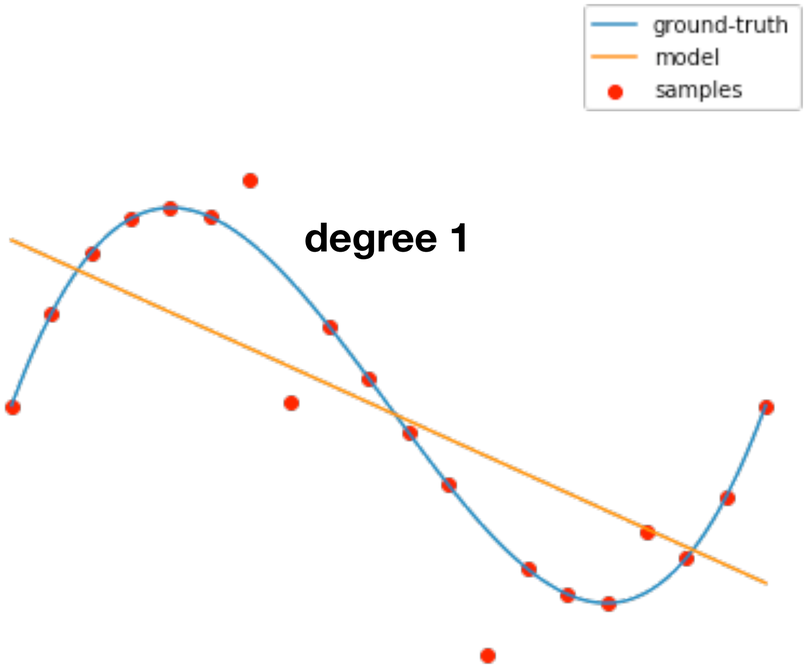


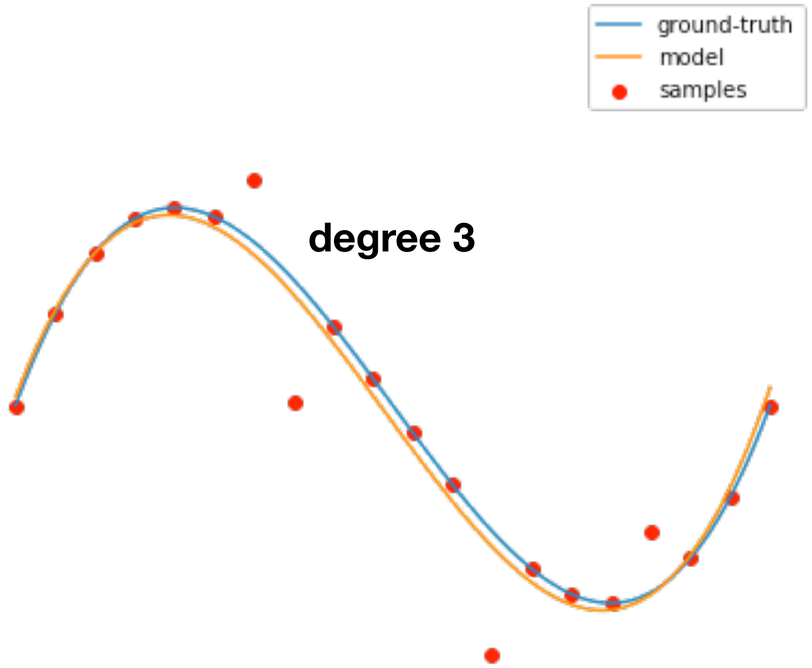


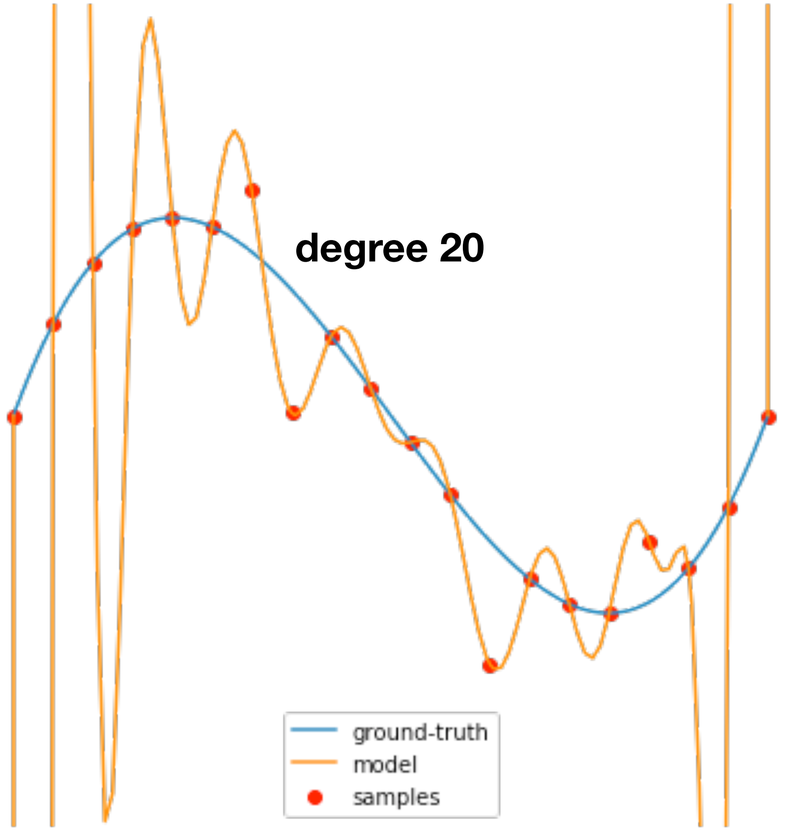


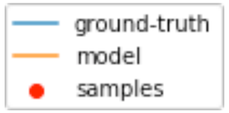


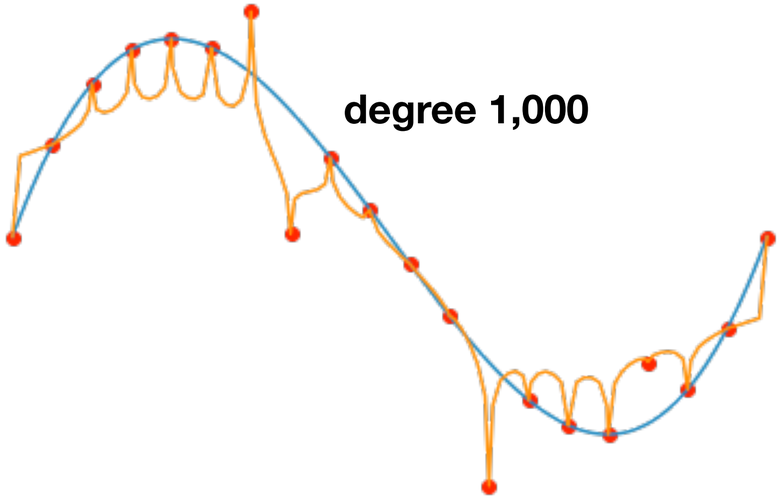
Nakkiran et al. blog post's companion notebook











Important: this is the *minimum norm* solution!
(There are also terrible degree-1,000 ERMs)

Nakkiran et al. blog post's companion notebook

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2$$

$X \in \mathbb{R}^{n \times d}$ with $n < d$,
of rank n

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2$$

$X \in \mathbb{R}^{n \times d}$ with $n < d$, of rank n

$$\nabla L_S(w) = \frac{1}{n} X^\top (Xw - y)$$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2$$

$X \in \mathbb{R}^{n \times d}$ with $n < d$, of rank n

$$\nabla L_S(w) = \frac{1}{n} X^\top (Xw - y) \qquad w_0 = 0$$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2$$

$X \in \mathbb{R}^{n \times d}$ with $n < d$, of rank n

$$\nabla L_S(w) = \frac{1}{n} X^\top (Xw - y) \qquad w_0 = 0$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1})$$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2$$

$X \in \mathbb{R}^{n \times d}$ with $n < d$, of rank n

$$\nabla L_S(w) = \frac{1}{n} X^\top (Xw - y) \qquad w_0 = 0$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n} X^\top X\right) w_{k-1} + \frac{\eta}{n} X^\top y$$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2$$

$X \in \mathbb{R}^{n\times d}$ with $n < d$, of rank n

$$\nabla L_S(w) = \frac{1}{n}X^\top(Xw - y) \qquad w_0 = 0$$

$$\tilde{\eta} = \frac{\eta}{n}$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y$$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2$$

$X \in \mathbb{R}^{n\times d}$ with $n < d$, of rank n

$$\nabla L_S(w) = \frac{1}{n}X^\top(Xw - y) \qquad w_0 = 0$$

$$\tilde{\eta} = \frac{\eta}{n}$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y$$

$X = U\Sigma V^\top$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2$$

$X \in \mathbb{R}^{n\times d}$ with $n < d$, of rank n

$$\nabla L_S(w) = \frac{1}{n}X^\top(Xw - y) \qquad w_0 = 0$$

$$\tilde{\eta} = \frac{\eta}{n}$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y$$

$$X = U\Sigma V^\top = U\tilde{\Sigma}\tilde{V}^\top$$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2$$ $X \in \mathbb{R}^{n\times d}$ with $n < d$, of rank n $$\nabla L_S(w) = \frac{1}{n}X^\top(Xw - y) \quad w_0 = 0$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y \qquad \tilde{\eta} = \frac{\eta}{n}$$

$X = U\Sigma V^\top = U\tilde{\Sigma}\tilde{V}^\top$ $\tilde{\Sigma} = [\Sigma \quad 0] \in \mathbb{R}^{n\times d}$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2$$ $X \in \mathbb{R}^{n\times d}$ with $n < d$, of rank n $$\nabla L_S(w) = \frac{1}{n}X^\top(Xw - y) \quad w_0 = 0$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y \qquad \tilde{\eta} = \frac{\eta}{n}$$

$X = U\Sigma V^\top = U\tilde{\Sigma}\tilde{V}^\top$ $\tilde{\Sigma} = [\Sigma \quad 0] \in \mathbb{R}^{n\times d}$ $\tilde{V} = \begin{bmatrix} V & V_2 \end{bmatrix} \in \mathbb{R}^{d\times d}$; $\tilde{V}^\top\tilde{V} = I = \tilde{V}\tilde{V}^\top$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2$$ $X \in \mathbb{R}^{n\times d}$ with $n < d$, of rank n $$\nabla L_S(w) = \frac{1}{n}X^\top(Xw - y) \quad w_0 = 0$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y \qquad \tilde{\eta} = \frac{\eta}{n}$$

$$X = U\Sigma V^\top = U\tilde{\Sigma}\tilde{V}^\top \quad \tilde{\Sigma} = [\Sigma \quad 0] \in \mathbb{R}^{n\times d} \quad \tilde{V} = \begin{bmatrix} V & V_2 \end{bmatrix} \in \mathbb{R}^{d\times d};\ \tilde{V}^\top\tilde{V} = I = \tilde{V}\tilde{V}^\top$$

$$w_k = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\tilde{V}\tilde{\Sigma}^\top\tilde{\Sigma}\tilde{V}^\top)^\ell\ \tilde{V}\tilde{\Sigma}^\top U^\top y$$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2$$ $X \in \mathbb{R}^{n\times d}$ with $n < d$, of rank n $$\nabla L_S(w) = \frac{1}{n}X^\top(Xw - y) \quad w_0 = 0$$

$$\tilde{\eta} = \frac{\eta}{n}$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y$$

$$X = U\Sigma V^\top = U\tilde{\Sigma}\tilde{V}^\top \quad \tilde{\Sigma} = [\Sigma \quad 0] \in \mathbb{R}^{n\times d} \quad \tilde{V} = \begin{bmatrix} V & V_2 \end{bmatrix} \in \mathbb{R}^{d\times d};\ \tilde{V}^\top\tilde{V} = I = \tilde{V}\tilde{V}^\top$$

$$w_k = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\tilde{V}\tilde{\Sigma}^\top\tilde{\Sigma}\tilde{V}^\top)^\ell\ \tilde{V}\tilde{\Sigma}^\top U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}\tilde{V}\left(I - \tilde{\eta}\begin{bmatrix}\Sigma^2 & 0\\ 0 & 0\end{bmatrix}\right)^\ell \tilde{V}^\top\ \tilde{V}\tilde{\Sigma}^\top U^\top y$$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2 \quad X \in \mathbb{R}^{n\times d} \text{ with } n < d, \text{ of rank } n \quad \nabla L_S(w) = \frac{1}{n}X^\top(Xw - y) \quad w_0 = 0$$

$$\tilde{\eta} = \frac{\eta}{n}$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y$$

$$X = U\Sigma V^\top = U\tilde{\Sigma}\tilde{V}^\top \quad \tilde{\Sigma} = [\Sigma \quad 0] \in \mathbb{R}^{n\times d} \quad \tilde{V} = \begin{bmatrix} V & V_2 \end{bmatrix} \in \mathbb{R}^{d\times d};\ \tilde{V}^\top\tilde{V} = I = \tilde{V}\tilde{V}^\top$$

$$w_k = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\tilde{V}\tilde{\Sigma}^\top\tilde{\Sigma}\tilde{V}^\top)^\ell\ \tilde{V}\tilde{\Sigma}^\top U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}\tilde{V}\left(I - \tilde{\eta}\begin{bmatrix}\Sigma^2 & 0\\ 0 & 0\end{bmatrix}\right)^\ell \tilde{V}^\top\ \tilde{V}\tilde{\Sigma}^\top U^\top y$$

$$= \tilde{\eta}\sum_{\ell=0}^{k-1}\begin{bmatrix} V & V_2 \end{bmatrix}\begin{bmatrix}(I - \tilde{\eta}\Sigma^2)^\ell & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}\Sigma\\ 0\end{bmatrix} U^\top y$$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2 \quad X \in \mathbb{R}^{n\times d} \text{ with } n < d, \text{ of rank } n \quad \nabla L_S(w) = \frac{1}{n}X^\top(Xw - y) \quad w_0 = 0$$

$$\tilde{\eta} = \frac{\eta}{n}$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y$$

$$X = U\Sigma V^\top = U\tilde{\Sigma}\tilde{V}^\top \quad \tilde{\Sigma} = [\Sigma \quad 0] \in \mathbb{R}^{n\times d} \quad \tilde{V} = \begin{bmatrix} V & V_2 \end{bmatrix} \in \mathbb{R}^{d\times d};\ \tilde{V}^\top\tilde{V} = I = \tilde{V}\tilde{V}^\top$$

$$w_k = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\tilde{V}\tilde{\Sigma}^\top\tilde{\Sigma}\tilde{V}^\top)^\ell\ \tilde{V}\tilde{\Sigma}^\top U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}\tilde{V}\left(I - \tilde{\eta}\begin{bmatrix}\Sigma^2 & 0\\ 0 & 0\end{bmatrix}\right)^\ell \tilde{V}^\top\ \tilde{V}\tilde{\Sigma}^\top U^\top y$$

$$= \tilde{\eta}\sum_{\ell=0}^{k-1}\begin{bmatrix} V & V_2\end{bmatrix}\begin{bmatrix}(I - \tilde{\eta}\Sigma^2)^\ell & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}\Sigma\\ 0\end{bmatrix} U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1} V(I - \tilde{\eta}\Sigma^2)^\ell\Sigma U^\top y$$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2 \quad X \in \mathbb{R}^{n\times d} \text{ with } n < d, \text{ of rank } n \quad \nabla L_S(w) = \frac{1}{n}X^\top(Xw - y) \quad w_0 = 0$$

$$\tilde{\eta} = \frac{\eta}{n}$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y$$

$$X = U\Sigma V^\top = U\tilde{\Sigma}\tilde{V}^\top \quad \tilde{\Sigma} = [\Sigma \quad 0] \in \mathbb{R}^{n\times d} \quad \tilde{V} = \begin{bmatrix} V & V_2 \end{bmatrix} \in \mathbb{R}^{d\times d};\ \tilde{V}^\top\tilde{V} = I = \tilde{V}\tilde{V}^\top$$

$$w_k = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\tilde{V}\tilde{\Sigma}^\top\tilde{\Sigma}\tilde{V}^\top)^\ell\ \tilde{V}\tilde{\Sigma}^\top U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}\tilde{V}\left(I - \tilde{\eta}\begin{bmatrix}\Sigma^2 & 0\\ 0 & 0\end{bmatrix}\right)^\ell \tilde{V}^\top\ \tilde{V}\tilde{\Sigma}^\top U^\top y$$

$$= \tilde{\eta}\sum_{\ell=0}^{k-1}\begin{bmatrix} V & V_2\end{bmatrix}\begin{bmatrix}(I - \tilde{\eta}\Sigma^2)^\ell & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}\Sigma\\ 0\end{bmatrix} U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1} V(I - \tilde{\eta}\Sigma^2)^\ell \Sigma U^\top y$$

$$= \tilde{\eta}V\left[\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\Sigma^2)^\ell\right]\Sigma U^\top y$$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2 \quad X \in \mathbb{R}^{n\times d} \text{ with } n < d, \text{ of rank } n \quad \nabla L_S(w) = \frac{1}{n}X^\top(Xw - y) \quad w_0 = 0$$

$$\tilde{\eta} = \frac{\eta}{n}$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y$$

$$X = U\Sigma V^\top = U\tilde{\Sigma}\tilde{V}^\top \quad \tilde{\Sigma} = [\Sigma \quad 0] \in \mathbb{R}^{n\times d} \quad \tilde{V} = \begin{bmatrix} V & V_2 \end{bmatrix} \in \mathbb{R}^{d\times d};\ \tilde{V}^\top\tilde{V} = I = \tilde{V}\tilde{V}^\top$$

$$w_k = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\tilde{V}\tilde{\Sigma}^\top\tilde{\Sigma}\tilde{V}^\top)^\ell\ \tilde{V}\tilde{\Sigma}^\top U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}\tilde{V}\left(I - \tilde{\eta}\begin{bmatrix}\Sigma^2 & 0\\ 0 & 0\end{bmatrix}\right)^\ell \tilde{V}^\top\ \tilde{V}\tilde{\Sigma}^\top U^\top y$$

$$= \tilde{\eta}\sum_{\ell=0}^{k-1}\begin{bmatrix} V & V_2\end{bmatrix}\begin{bmatrix}(I - \tilde{\eta}\Sigma^2)^\ell & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}\Sigma\\ 0\end{bmatrix} U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1} V(I - \tilde{\eta}\Sigma^2)^\ell \Sigma U^\top y$$

$$= \tilde{\eta}V\left[\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\Sigma^2)^\ell\right]\Sigma U^\top y$$

Neumann series: $\sum_{\ell=0}^{\infty} A^\ell = (I - A)^{-1}$ when $\|A\|_{\text{op}} < 1$

$L_S(w) = \frac{1}{2n}\|Xw - y\|^2$ $X \in \mathbb{R}^{n\times d}$ with $n < d$, of rank n $\nabla L_S(w) = \frac{1}{n}X^\top(Xw - y)$ $w_0 = 0$

$\tilde{\eta} = \frac{\eta}{n}$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y$$

$X = U\Sigma V^\top = U\tilde{\Sigma}\tilde{V}^\top$ $\tilde{\Sigma} = [\Sigma \quad 0] \in \mathbb{R}^{n\times d}$ $\tilde{V} = \begin{bmatrix} V & V_2 \end{bmatrix} \in \mathbb{R}^{d\times d}$; $\tilde{V}^\top\tilde{V} = I = \tilde{V}\tilde{V}^\top$

$$w_k = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\tilde{V}\tilde{\Sigma}^\top\tilde{\Sigma}\tilde{V}^\top)^\ell\, \tilde{V}\tilde{\Sigma}^\top U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}\tilde{V}\left(I - \tilde{\eta}\begin{bmatrix}\Sigma^2 & 0\\ 0 & 0\end{bmatrix}\right)^\ell \tilde{V}^\top\, \tilde{V}\tilde{\Sigma}^\top U^\top y$$

$$= \tilde{\eta}\sum_{\ell=0}^{k-1}\begin{bmatrix} V & V_2\end{bmatrix}\begin{bmatrix}(I - \tilde{\eta}\Sigma^2)^\ell & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}\Sigma\\ 0\end{bmatrix} U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1} V(I - \tilde{\eta}\Sigma^2)^\ell \Sigma U^\top y$$

$$= \tilde{\eta} V\left[\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\Sigma^2)^\ell\right]\Sigma U^\top y$$

Neumann series: $\sum_{\ell=0}^{\infty} A^\ell = (I - A)^{-1}$ when $\|A\|_{\text{op}} < 1$ $\tilde{\eta} < 2/\lambda_{\max}(\Sigma^2)$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2 \quad X \in \mathbb{R}^{n\times d} \text{ with } n < d, \text{ of rank } n \quad \nabla L_S(w) = \frac{1}{n}X^\top(Xw - y) \quad w_0 = 0$$

$$\tilde{\eta} = \frac{\eta}{n}$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y$$

$$X = U\Sigma V^\top = U\tilde{\Sigma}\tilde{V}^\top \quad \tilde{\Sigma} = [\Sigma \quad 0] \in \mathbb{R}^{n\times d} \quad \tilde{V} = \begin{bmatrix} V & V_2 \end{bmatrix} \in \mathbb{R}^{d\times d};\ \tilde{V}^\top\tilde{V} = I = \tilde{V}\tilde{V}^\top$$

$$w_k = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\tilde{V}\tilde{\Sigma}^\top\tilde{\Sigma}\tilde{V}^\top)^\ell\, \tilde{V}\tilde{\Sigma}^\top U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}\tilde{V}\left(I - \tilde{\eta}\begin{bmatrix}\Sigma^2 & 0\\ 0 & 0\end{bmatrix}\right)^\ell \tilde{V}^\top\, \tilde{V}\tilde{\Sigma}^\top U^\top y$$

$$= \tilde{\eta}\sum_{\ell=0}^{k-1}\begin{bmatrix} V & V_2\end{bmatrix}\begin{bmatrix}(I - \tilde{\eta}\Sigma^2)^\ell & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}\Sigma\\ 0\end{bmatrix} U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1} V(I - \tilde{\eta}\Sigma^2)^\ell \Sigma U^\top y$$

$$= \tilde{\eta}V\left[\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\Sigma^2)^\ell\right]\Sigma U^\top y \to \tilde{\eta}V\left[I - (I - \tilde{\eta}\Sigma^2)\right]^{-1}\Sigma U^\top y$$

Neumann series: $\sum_{\ell=0}^{\infty} A^\ell = (I - A)^{-1}$ when $\|A\|_{\text{op}} < 1$ $\quad \tilde{\eta} < 2/\lambda_{\max}(\Sigma^2)$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2 \quad X \in \mathbb{R}^{n\times d} \text{ with } n < d, \text{ of rank } n \quad \nabla L_S(w) = \frac{1}{n}X^\top(Xw - y) \quad w_0 = 0$$

$$\tilde{\eta} = \frac{\eta}{n}$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y$$

$$X = U\Sigma V^\top = U\tilde{\Sigma}\tilde{V}^\top \quad \tilde{\Sigma} = [\Sigma \quad 0] \in \mathbb{R}^{n\times d} \quad \tilde{V} = \begin{bmatrix} V & V_2 \end{bmatrix} \in \mathbb{R}^{d\times d};\ \tilde{V}^\top\tilde{V} = I = \tilde{V}\tilde{V}^\top$$

$$w_k = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\tilde{V}\tilde{\Sigma}^\top\tilde{\Sigma}\tilde{V}^\top)^\ell\, \tilde{V}\tilde{\Sigma}^\top U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}\tilde{V}\left(I - \tilde{\eta}\begin{bmatrix}\Sigma^2 & 0\\ 0 & 0\end{bmatrix}\right)^\ell \tilde{V}^\top\, \tilde{V}\tilde{\Sigma}^\top U^\top y$$

$$= \tilde{\eta}\sum_{\ell=0}^{k-1}\begin{bmatrix} V & V_2\end{bmatrix}\begin{bmatrix}(I - \tilde{\eta}\Sigma^2)^\ell & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}\Sigma\\ 0\end{bmatrix} U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1} V(I - \tilde{\eta}\Sigma^2)^\ell \Sigma U^\top y$$

$$= \tilde{\eta}V\left[\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\Sigma^2)^\ell\right]\Sigma U^\top y \to \tilde{\eta}V\left[I - (I - \tilde{\eta}\Sigma^2)\right]^{-1}\Sigma U^\top y = V\Sigma^{-1}U^\top y$$

Neumann series: $\sum_{\ell=0}^{\infty} A^\ell = (I - A)^{-1}$ when $\|A\|_{\text{op}} < 1$ $\quad \tilde{\eta} < 2/\lambda_{\max}(\Sigma^2)$

$$L_S(w) = \frac{1}{2n}\|Xw - y\|^2 \quad X \in \mathbb{R}^{n\times d} \text{ with } n < d, \text{ of rank } n \quad \nabla L_S(w) = \frac{1}{n}X^\top(Xw - y) \quad w_0 = 0$$

$$\tilde{\eta} = \frac{\eta}{n}$$

$$w_k = w_{k-1} - \eta \nabla L_S(w_{k-1}) = \left(I - \frac{\eta}{n}X^\top X\right) w_{k-1} + \frac{\eta}{n}X^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}X^\top X)^\ell X^\top y$$

$$X = U\Sigma V^\top = U\tilde{\Sigma}\tilde{V}^\top \quad \tilde{\Sigma} = [\Sigma \quad 0] \in \mathbb{R}^{n\times d} \quad \tilde{V} = \begin{bmatrix} V & V_2 \end{bmatrix} \in \mathbb{R}^{d\times d};\ \tilde{V}^\top\tilde{V} = I = \tilde{V}\tilde{V}^\top$$

$$w_k = \tilde{\eta}\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\tilde{V}\tilde{\Sigma}^\top\tilde{\Sigma}\tilde{V}^\top)^\ell\ \tilde{V}\tilde{\Sigma}^\top U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1}\tilde{V}\left(I - \tilde{\eta}\begin{bmatrix}\Sigma^2 & 0\\ 0 & 0\end{bmatrix}\right)^\ell \tilde{V}^\top\ \tilde{V}\tilde{\Sigma}^\top U^\top y$$

$$= \tilde{\eta}\sum_{\ell=0}^{k-1}\begin{bmatrix} V & V_2\end{bmatrix}\begin{bmatrix}(I - \tilde{\eta}\Sigma^2)^\ell & 0\\ 0 & I\end{bmatrix}\begin{bmatrix}\Sigma\\ 0\end{bmatrix} U^\top y = \tilde{\eta}\sum_{\ell=0}^{k-1} V(I - \tilde{\eta}\Sigma^2)^\ell \Sigma U^\top y$$

$$= \tilde{\eta}V\left[\sum_{\ell=0}^{k-1}(I - \tilde{\eta}\Sigma^2)^\ell\right]\Sigma U^\top y \to \tilde{\eta}V\left[I - (I - \tilde{\eta}\Sigma^2)\right]^{-1}\Sigma U^\top y = V\Sigma^{-1}U^\top y = X^\dagger y$$

Neumann series: $\sum_{\ell=0}^{\infty} A^\ell = (I - A)^{-1}$ when $\|A\|_{\text{op}} < 1$ $\quad \tilde{\eta} < 2/\lambda_{\max}(\Sigma^2)$

Implicit regularization of gradient descent

- We just showed that gradient descent for OLS with X of rank n,
  starting from zero with $\eta < 2n / \sigma_{\max}(X)^2$,
  converges to the minimum-norm interpolator $X^\dagger y$

Implicit regularization of gradient descent

- We just showed that gradient descent for OLS with X of rank n,
  starting from zero with $\eta < 2n / \sigma_{\max}(X)^2$,
  converges to the minimum-norm interpolator $X^\dagger y$
  - If we track $w_0 \neq 0$ in same analysis, get $w_\infty = V_2 V_2^\top w_0 + X^\dagger y$ (proof)

Implicit regularization of gradient descent

- We just showed that gradient descent for OLS with X of rank n,
  starting from zero with $\eta < 2n / \sigma_{\max}(X)^2$,
  converges to the minimum-norm interpolator $X^\dagger y$
  - If we track $w_0 \neq 0$ in same analysis, get $w_\infty = V_2 V_2^\top w_0 + X^\dagger y$ (proof)

- So, the 1,000-degree polynomial picture is what GD would give

Implicit regularization of gradient descent

- We just showed that gradient descent for OLS with X of rank n,
  starting from zero with $\eta < 2n / \sigma_{\max}(X)^2$,
  converges to the minimum-norm interpolator $X^\dagger y$
  - If we track $w_0 \neq 0$ in same analysis, get $w_\infty = V_2 V_2^\top w_0 + X^\dagger y$ (proof)

- So, the 1,000-degree polynomial picture is what GD would give
- Does this same idea hold for other losses / models? **Not necessarily.**

Implicit regularization of gradient descent

- We just showed that gradient descent for OLS with X of rank n, starting from zero with $\eta < 2n / \sigma_{\max}(X)^2$, converges to the minimum-norm interpolator $X^\dagger y$
  - If we track $w_0 \neq 0$ in same analysis, get $w_\infty = V_2 V_2^\top w_0 + X^\dagger y$ (proof)

- So, the 1,000-degree polynomial picture is what GD would give
- Does this same idea hold for other losses / models? **Not necessarily.**
  - Logistic regression:

Implicit regularization of gradient descent

- We just showed that gradient descent for OLS with X of rank n,
  starting from zero with $\eta < 2n / \sigma_{\max}(X)^2$,
  converges to the minimum-norm interpolator $X^\dagger y$
  - If we track $w_0 \neq 0$ in same analysis, get $w_\infty = V_2 V_2^\top w_0 + X^\dagger y$ (proof)

- So, the 1,000-degree polynomial picture is what GD would give
- Does this same idea hold for other losses / models? **Not necessarily.**
  - Logistic regression:
    - Separable: norm diverges in direction of max-margin separator (Soudry et al.)

Implicit regularization of gradient descent

- We just showed that gradient descent for OLS with X of rank n,
  starting from zero with $\eta < 2n / \sigma_{\max}(X)^2$,
  converges to the minimum-norm interpolator $X^\dagger y$
  - If we track $w_0 \neq 0$ in same analysis, get $w_\infty = V_2 V_2^\top w_0 + X^\dagger y$ (proof)

- So, the 1,000-degree polynomial picture is what GD would give
- Does this same idea hold for other losses / models? **Not necessarily.**
  - Logistic regression:
    - Separable: norm diverges in direction of max-margin separator (Soudry et al.)
    - Non-separable: biased towards max-margin, but complicated (Ji/Telgarsky)

Implicit regularization of gradient descent

- We just showed that gradient descent for OLS with X of rank n,
  starting from zero with $\eta < 2n / \sigma_{\max}(X)^2$,
  converges to the minimum-norm interpolator $X^\dagger y$
  - If we track $w_0 \neq 0$ in same analysis, get $w_\infty = V_2 V_2^\top w_0 + X^\dagger y$ (proof)

- So, the 1,000-degree polynomial picture is what GD would give
- Does this same idea hold for other losses / models? **Not necessarily.**
  - Logistic regression:
    - Separable: norm diverges in direction of max-margin separator (Soudry et al.)
    - Non-separable: biased towards max-margin, but complicated (Ji/Telgarsky)
    - Also see Telgarsky notes section 10

Implicit regularization of gradient descent

- We just showed that gradient descent for OLS with X of rank n,
  starting from zero with $\eta < 2n / \sigma_{\max}(X)^2$,
  converges to the minimum-norm interpolator $X^\dagger y$
  - If we track $w_0 \neq 0$ in same analysis, get $w_\infty = V_2 V_2^\top w_0 + X^\dagger y$ (proof)

- So, the 1,000-degree polynomial picture is what GD would give
- Does this same idea hold for other losses / models? **Not necessarily.**
  - Logistic regression:
    - Separable: norm diverges in direction of max-margin separator (Soudry et al.)
    - Non-separable: biased towards max-margin, but complicated (Ji/Telgarsky)
    - Also see Telgarsky notes section 10
  - Matrix factorization models: conjectured min nuclear norm, slightly controversial

Implicit regularization of gradient descent

- We just showed that gradient descent for OLS with X of rank n,
  starting from zero with $\eta < 2n / \sigma_{\max}(X)^2$,
  converges to the minimum-norm interpolator $X^\dagger y$
  - If we track $w_0 \neq 0$ in same analysis, get $w_\infty = V_2 V_2^\top w_0 + X^\dagger y$ (proof)

- So, the 1,000-degree polynomial picture is what GD would give
- Does this same idea hold for other losses / models? **Not necessarily.**
  - Logistic regression:
    - Separable: norm diverges in direction of max-margin separator (Soudry et al.)
    - Non-separable: biased towards max-margin, but complicated (Ji/Telgarsky)
    - Also see Telgarsky notes section 10
  - Matrix factorization models: conjectured min nuclear norm, slightly controversial
  - Deep learning: ???

Double descent

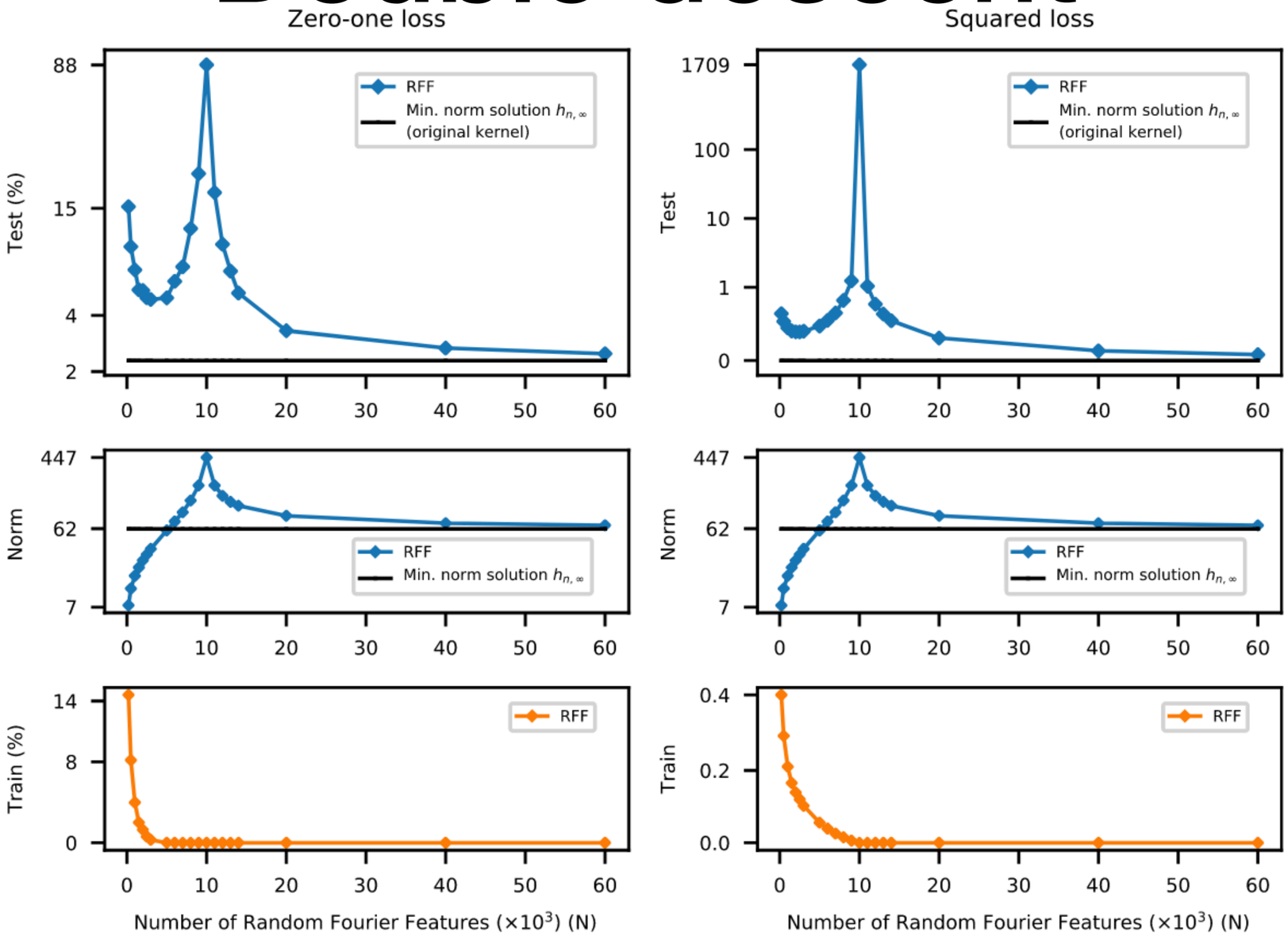


Fig. 2. Double-descent risk curve for the RFF model on MNIST. Shown are test risks (log scale), coefficient ℓ_2 norms (log scale), and training risks of the RFF model predictors $h_{n,N}$ learned on a subset of MNIST ($n = 10^4$, 10 classes). The interpolation threshold is achieved at $N = 10^4$.

Belkin/Hsu/Ma/Mandal, PNAS 2019

Double descent

Classical regime (left of peak): unique ERM

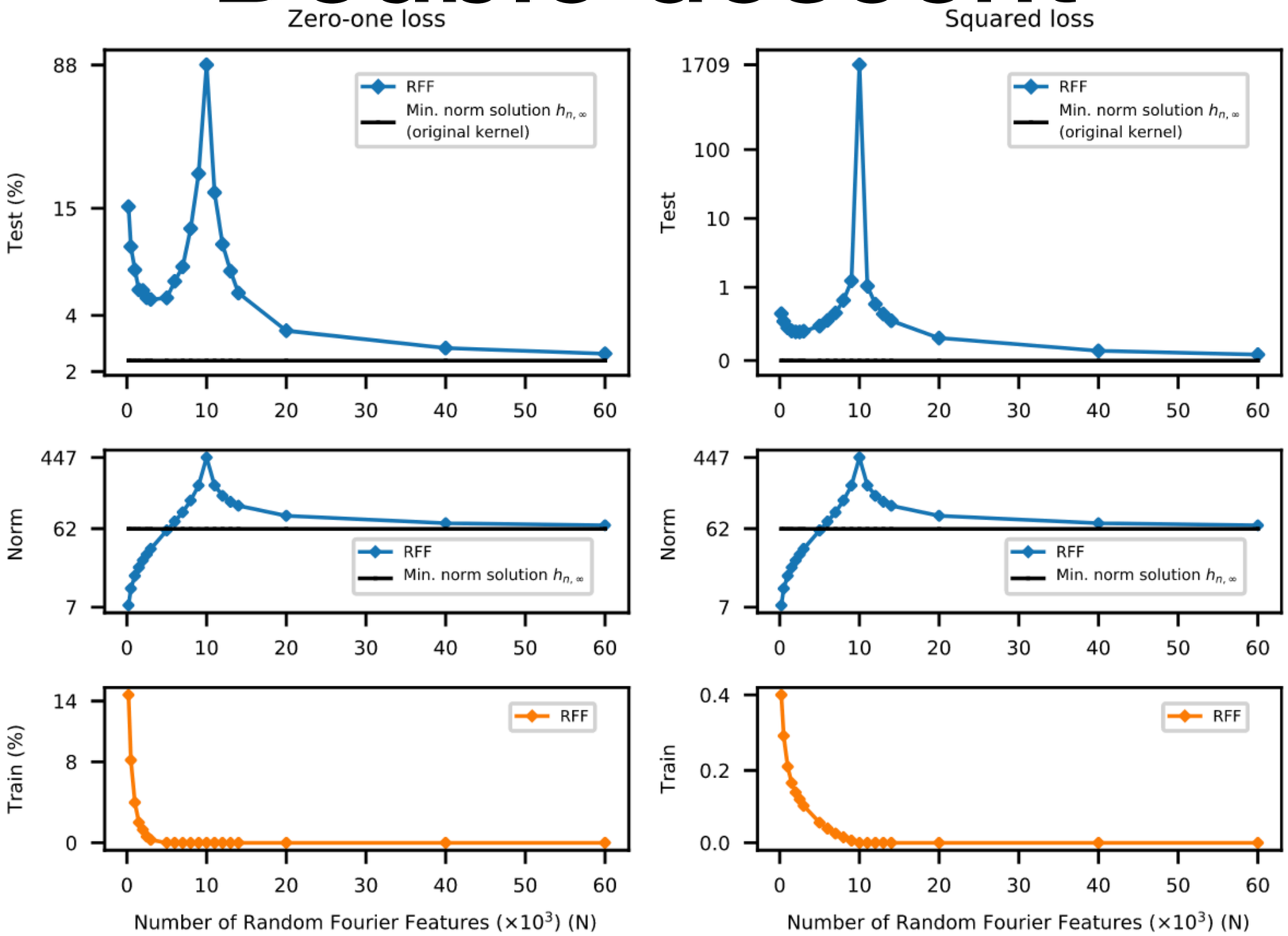
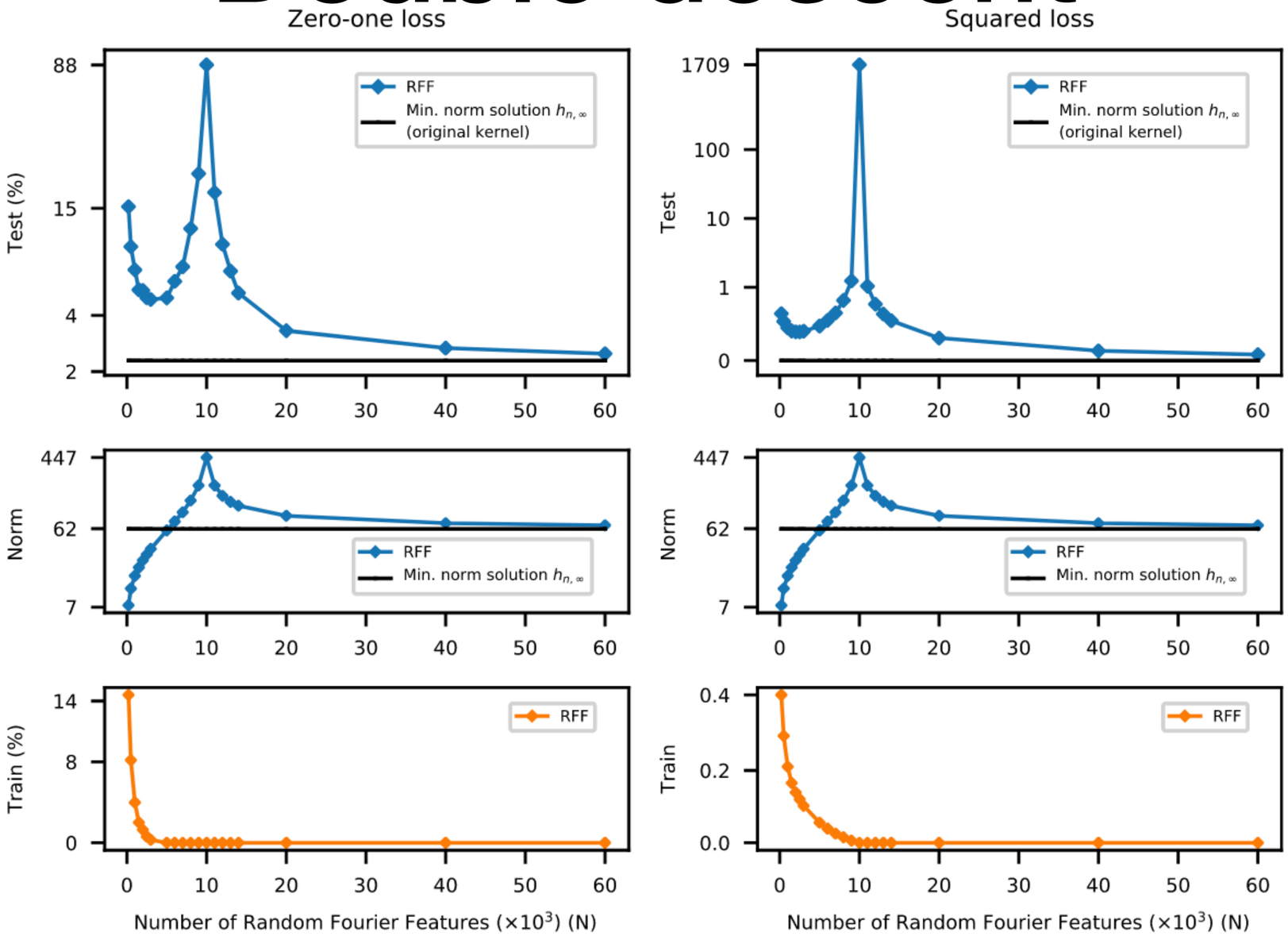


Fig. 2. Double-descent risk curve for the RFF model on MNIST. Shown are test risks (log scale), coefficient ℓ_2 norms (log scale), and training risks of the RFF model predictors $h_{n,N}$ learned on a subset of MNIST ($n = 10^4$, 10 classes). The interpolation threshold is achieved at $N = 10^4$.

Belkin/Hsu/Ma/Mandal, PNAS 2019

Double descent

Classical regime (left of peak): unique ERM



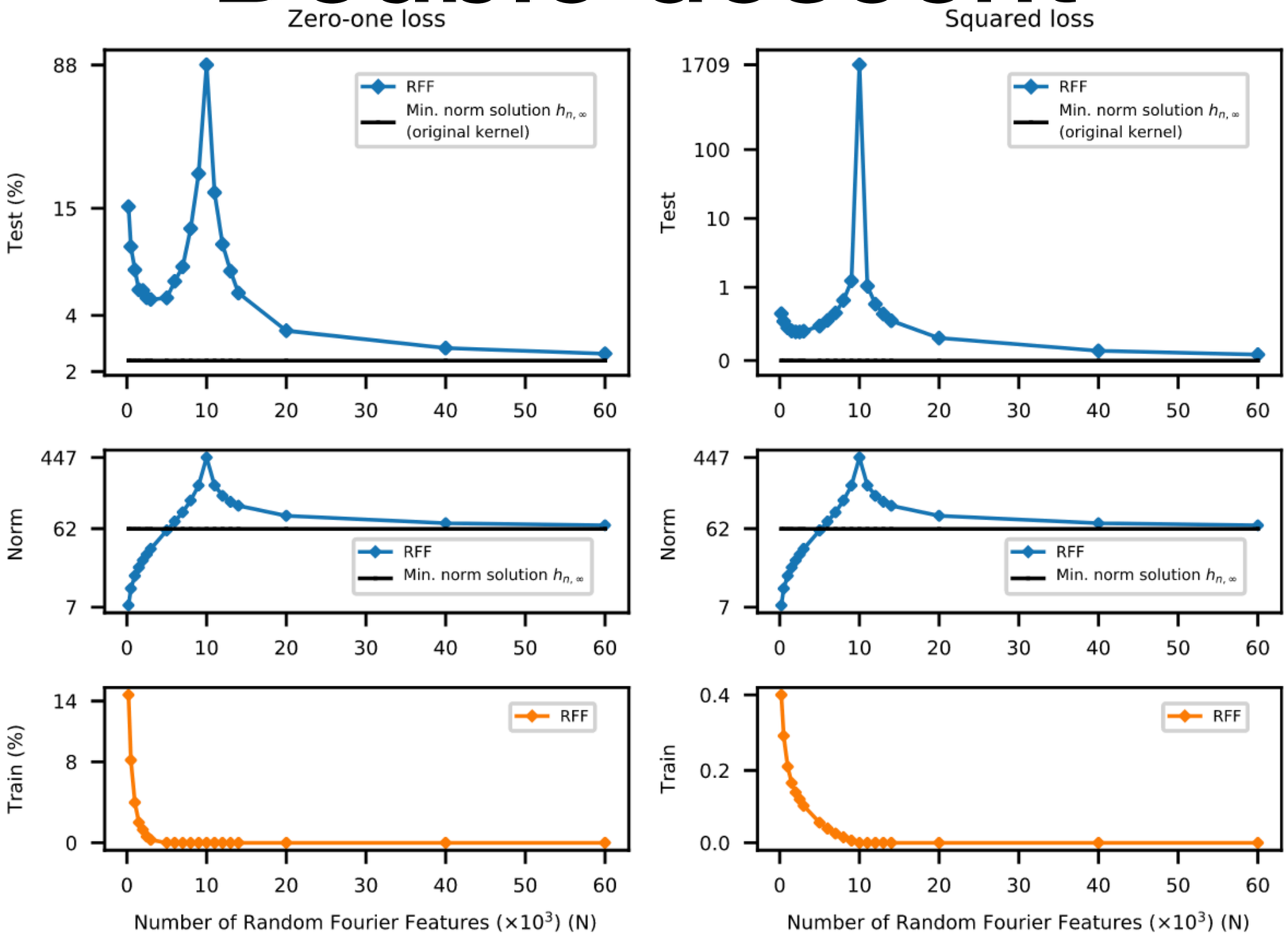
Interpolating regime (right of peak): many possible interpolators

Fig. 2. Double-descent risk curve for the RFF model on MNIST. Shown are test risks (log scale), coefficient ℓ_2 norms (log scale), and training risks of the RFF model predictors $h_{n,N}$ learned on a subset of MNIST ($n = 10^4$, 10 classes). The interpolation threshold is achieved at $N = 10^4$.

Belkin/Hsu/Ma/Mandal, PNAS 2019

Double descent

Classical regime (left of peak): unique ERM



Interpolating regime (right of peak): many possible interpolators

which one we get depends on alg.'s implicit bias

Fig. 2. Double-descent risk curve for the RFF model on MNIST. Shown are test risks (log scale), coefficient ℓ_2 norms (log scale), and training risks of the RFF model predictors $h_{n,N}$ learned on a subset of MNIST ($n = 10^4$, 10 classes). The interpolation threshold is achieved at $N = 10^4$.

Belkin/Hsu/Ma/Mandal, PNAS 2019

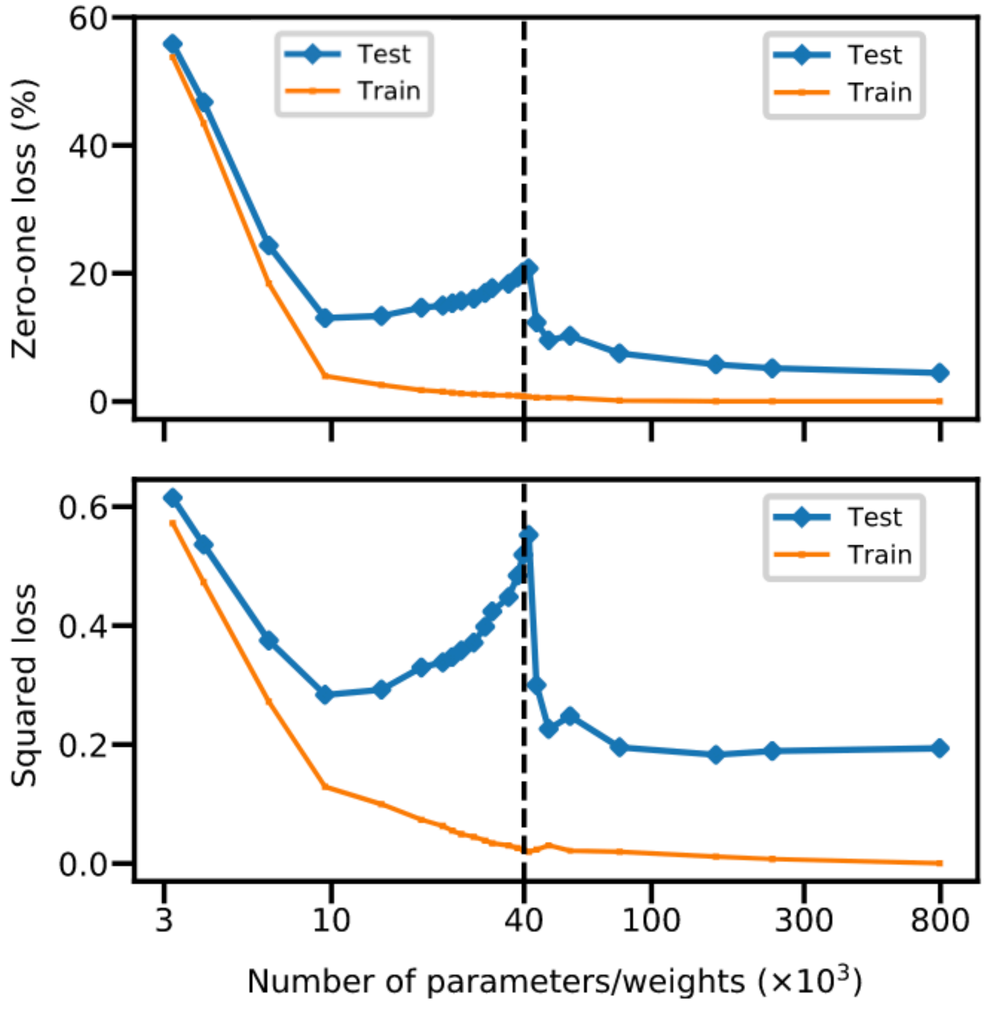


Fig. 3. Double-descent risk curve for a fully connected neural network on MNIST. Shown are training and test risks of a network with a single layer of H hidden units, learned on a subset of MNIST ($n = 4 \cdot 10^3$, $d = 784$, $K = 10$ classes). The number of parameters is $(d + 1) \cdot H + (H + 1) \cdot K$. The interpolation threshold (black dashed line) is observed at $n \cdot K$.

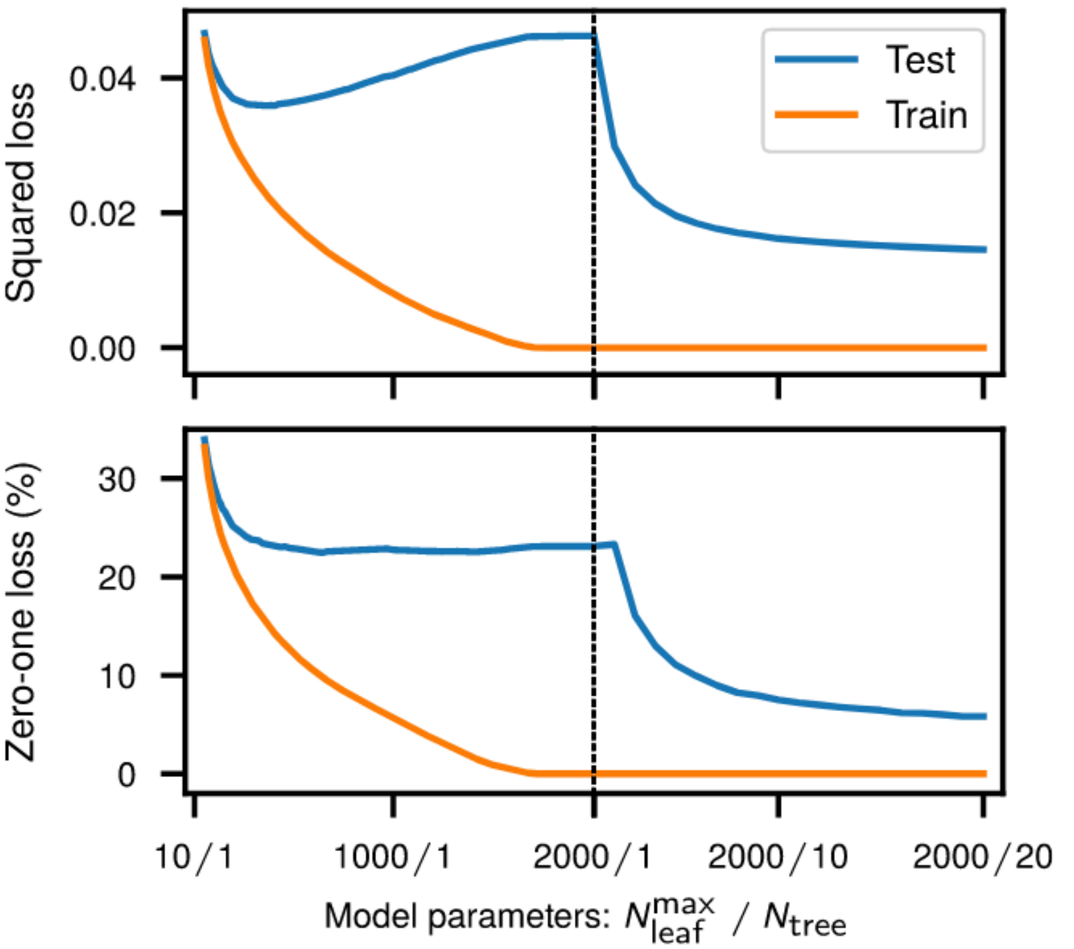
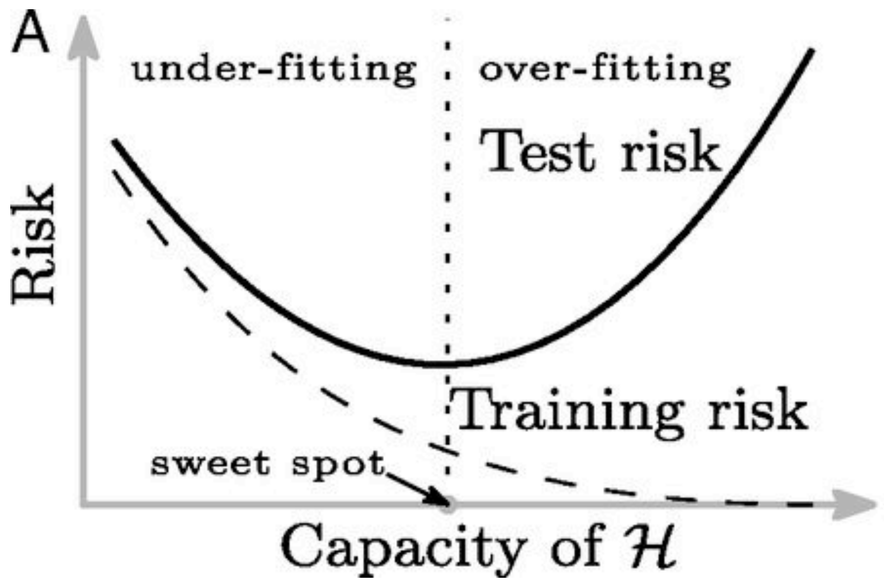
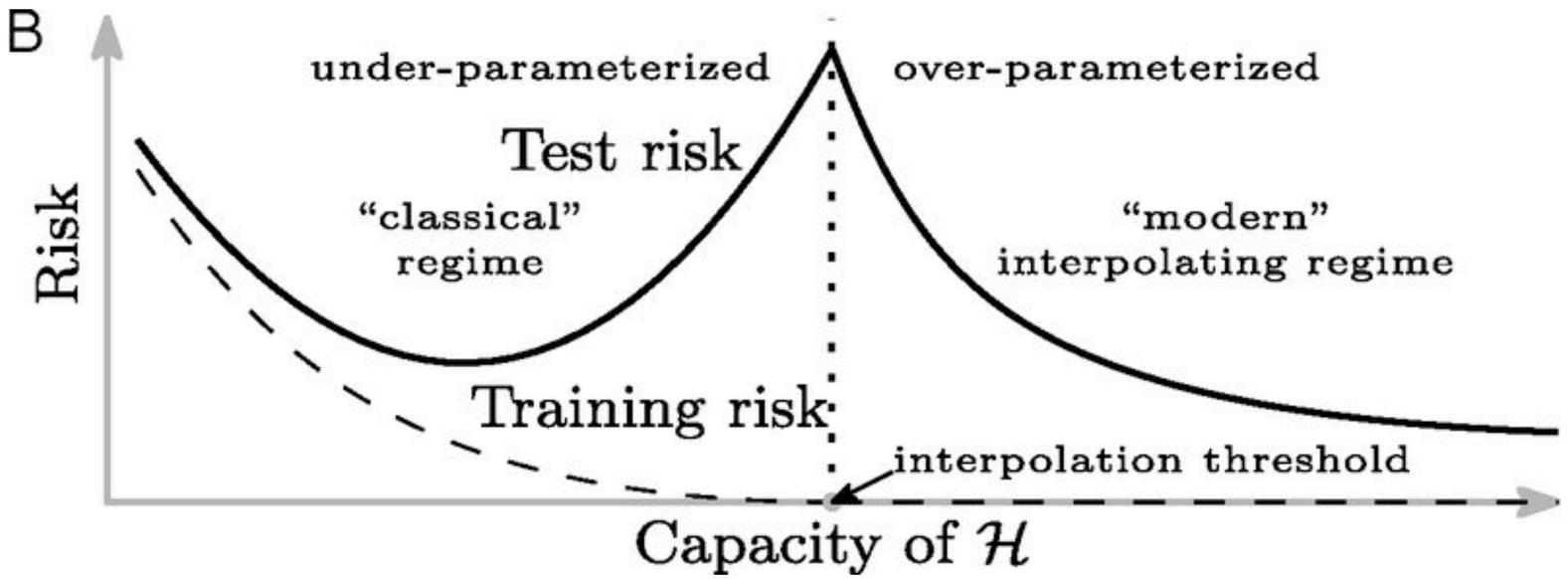


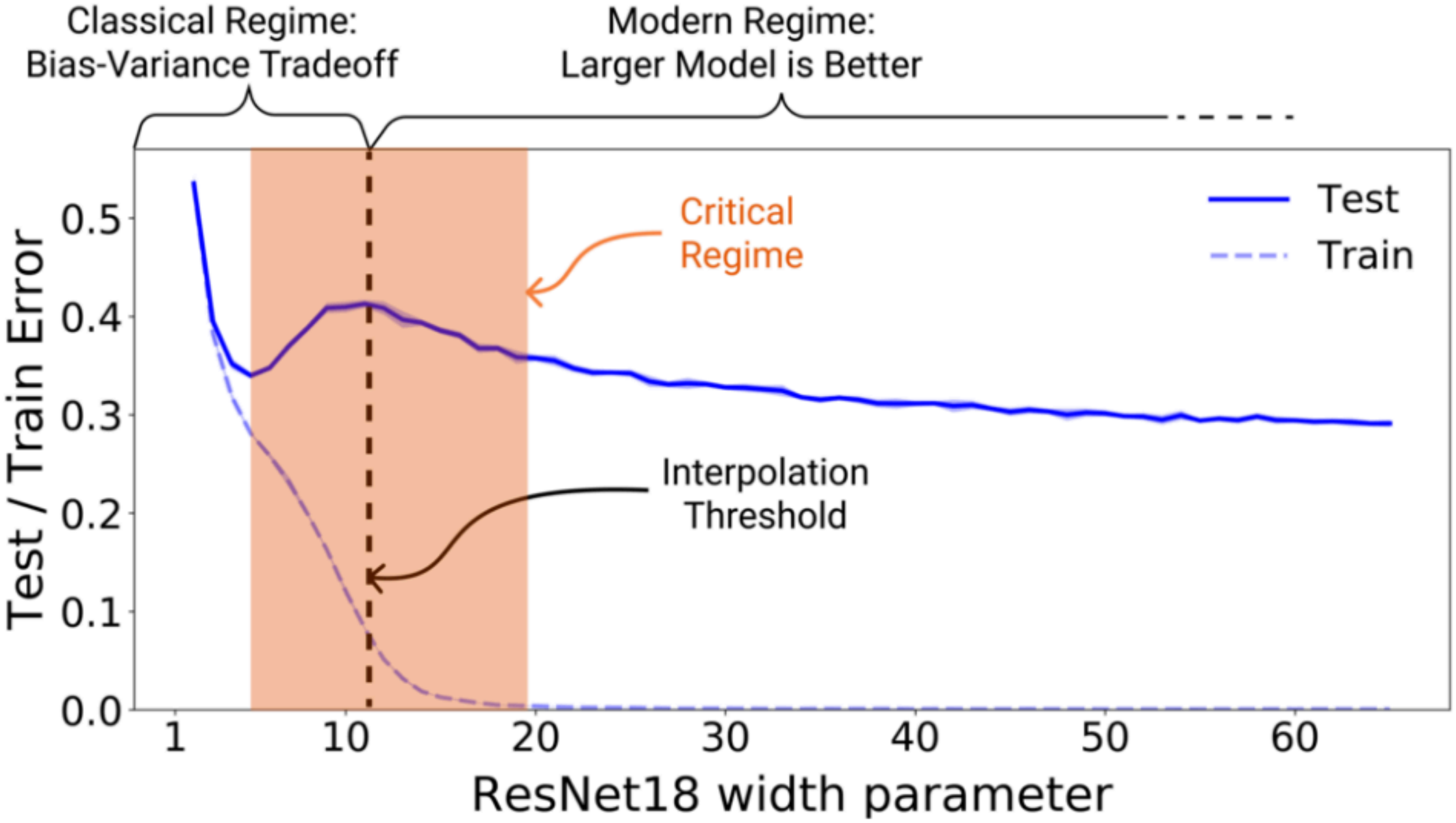
Fig. 4. Double-descent risk curve for random forests on MNIST. The double-descent risk curve is observed for random forests with increasing model complexity trained on a subset of MNIST ($n = 10^4$, 10 classes). Its complexity is controlled by the number of trees N_{tree} and the maximum number of leaves allowed for each tree $N_{\text{leaf}}^{\max}$.

Belkin/Hsu/Ma/Mandal, PNAS 2019

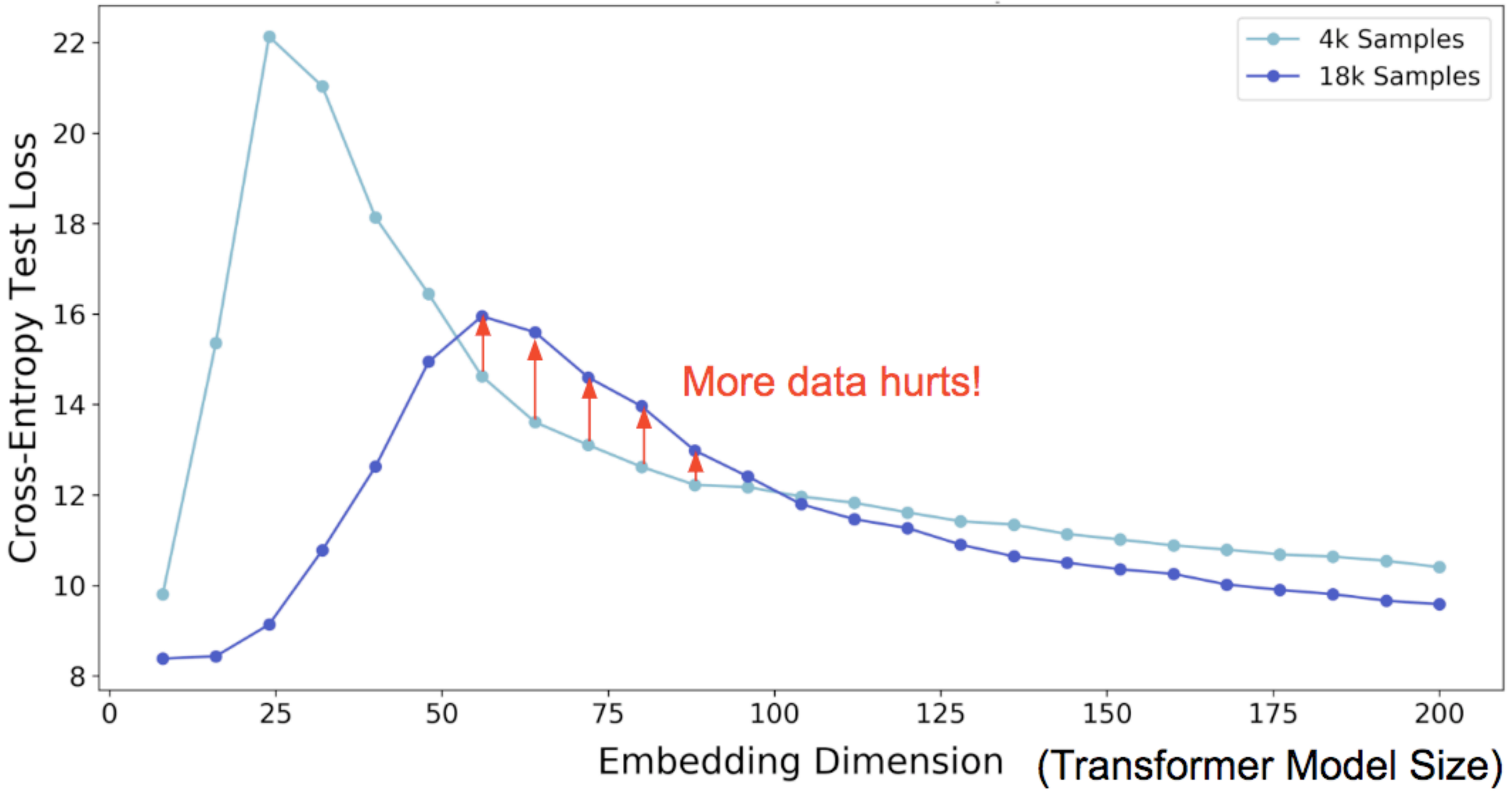




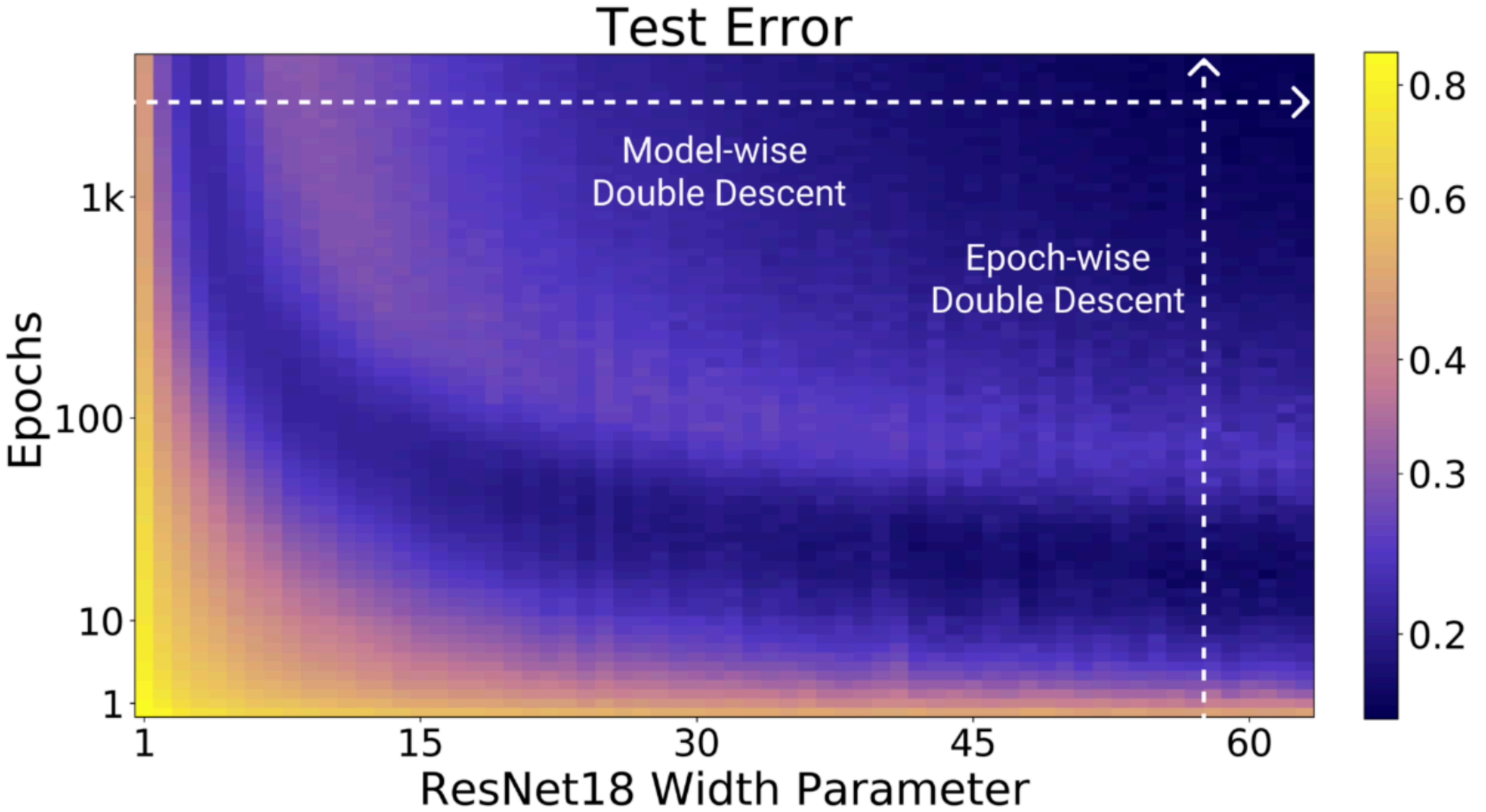
Belkin/Hsu/Ma/Mandal, PNAS 2019



Nakkiran et al. ICLR-20



Nakkiran et al. ICLR-20



Nakkiran et al. ICLR-20

Definition 1 (Effective Model Complexity) *The* Effective Model Complexity *(EMC) of a training procedure $\mathcal{T}$, with respect to distribution $\mathcal{D}$ and parameter $\epsilon > 0$, is defined as:*

$$\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T}) := \max\{n \mid \mathbb{E}_{S\sim\mathcal{D}^n}[\mathrm{Error}_S(\mathcal{T}(S))] \leq \epsilon\}$$

where $\mathrm{Error}_S(M)$ is the mean error of model M on train samples S.

Our main hypothesis can be informally stated as follows:

Hypothesis 1 (Generalized Double Descent hypothesis, informal) *For any natural data distribution $\mathcal{D}$, neural-network-based training procedure $\mathcal{T}$, and small $\epsilon > 0$, if we consider the task of predicting labels based on n samples from $\mathcal{D}$ then:*

Under-paremeterized regime. *If $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T})$ is sufficiently smaller than n, any perturbation of $\mathcal{T}$ that increases its effective complexity will decrease the test error.*

Over-parameterized regime. *If $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T})$ is sufficiently larger than n, any perturbation of $\mathcal{T}$ that increases its effective complexity will decrease the test error.*

Critically parameterized regime. *If $\mathrm{EMC}_{\mathcal{D},\epsilon}(\mathcal{T}) \approx n$, then a perturbation of $\mathcal{T}$ that increases its effective complexity might decrease* **or increase** *the test error.*

Nakkiran et al. ICLR-20

(pause)

A road to PAC-Bayes

- Bayesians say:
  - Start with a prior distribution $\pi(h)$ on choice of hypothesis

A road to PAC-Bayes

- Bayesians say:
  - Start with a prior distribution $\pi(h)$ on choice of hypothesis
  - Observe data S with likelihood $\mathscr{L}(S \mid h)$

A road to PAC-Bayes

- Bayesians say:
  - Start with a prior distribution $\pi(h)$ on choice of hypothesis
  - Observe data S with likelihood $\mathscr{L}(S \mid h)$
  - End up with posterior distribution $\rho(h \mid S) \propto \mathscr{L}(S \mid h)\,\pi(h)$

A road to PAC-Bayes

- Bayesians say:
  - Start with a prior distribution $\pi(h)$ on choice of hypothesis
  - Observe data S with likelihood $\mathscr{L}(S \mid h)$
  - End up with posterior distribution $\rho(h \mid S) \propto \mathscr{L}(S \mid h)\,\pi(h)$
  - Make predictions/decision based on posterior mean/median, MAP, single draw, …

A road to PAC-Bayes

- Bayesians say:
  - Start with a prior distribution $\pi(h)$ on choice of hypothesis
  - Observe data S with likelihood $\mathscr{L}(S \mid h)$
  - End up with posterior distribution $\rho(h \mid S) \propto \mathscr{L}(S \mid h)\,\pi(h)$
  - Make predictions/decision based on posterior mean/median, MAP, single draw, …
- This is optimal if you believe in your prior + likelihood! 😊

A road to PAC-Bayes

- Bayesians say:
  - Start with a prior distribution $\pi(h)$ on choice of hypothesis
  - Observe data S with likelihood $\mathscr{L}(S \mid h)$
  - End up with posterior distribution $\rho(h \mid S) \propto \mathscr{L}(S \mid h)\,\pi(h)$
  - Make predictions/decision based on posterior mean/median, MAP, single draw, …
- This is optimal if you believe in your prior + likelihood! 😊
  - Frequentists say: “but how good is it actually???”

A road to PAC-Bayes

- Bayesians say:
  - Start with a prior distribution $\pi(h)$ on choice of hypothesis
  - Observe data S with likelihood $\mathscr{L}(S \mid h)$
  - End up with posterior distribution $\rho(h \mid S) \propto \mathscr{L}(S \mid h)\,\pi(h)$
  - Make predictions/decision based on posterior mean/median, MAP, single draw, …
- This is optimal if you believe in your prior + likelihood! 😊
  - Frequentists say: “but how good is it actually???”
  - What if your model class / prior / … are wrong?

A road to PAC-Bayes

- Bayesians say:
  - Start with a prior distribution $\pi(h)$ on choice of hypothesis
  - Observe data S with likelihood $\mathscr{L}(S \mid h)$
  - End up with posterior distribution $\rho(h \mid S) \propto \mathscr{L}(S \mid h)\,\pi(h)$
  - Make predictions/decision based on posterior mean/median, MAP, single draw, …
- This is optimal if you believe in your prior + likelihood! 😊
  - Frequentists say: “but how good is it actually???”
  - What if your model class / prior / … are wrong?
- Tempered likelihood (Zhang 06) / SafeBayes (Grünwald 12):

A road to PAC-Bayes

- Bayesians say:
  - Start with a prior distribution $\pi(h)$ on choice of hypothesis
  - Observe data S with likelihood $\mathscr{L}(S \mid h)$
  - End up with posterior distribution $\rho(h \mid S) \propto \mathscr{L}(S \mid h)\,\pi(h)$
  - Make predictions/decision based on posterior mean/median, MAP, single draw, …
- This is optimal if you believe in your prior + likelihood! 😊
  - Frequentists say: “but how good is it actually???”
  - What if your model class / prior / … are wrong?
- Tempered likelihood (Zhang 06) / SafeBayes (Grünwald 12):
  - If your model is misspecified, can be provably better to use $\mathscr{L}^\lambda$ for $\lambda < 1$

A road to PAC-Bayes

- Bayesians say:
  - Start with a prior distribution $\pi(h)$ on choice of hypothesis
  - Observe data S with likelihood $\mathscr{L}(S \mid h)$
  - End up with posterior distribution $\rho(h \mid S) \propto \mathscr{L}(S \mid h)\,\pi(h)$
  - Make predictions/decision based on posterior mean/median, MAP, single draw, …
- This is optimal if you believe in your prior + likelihood! 😊
  - Frequentists say: “but how good is it actually???”
  - What if your model class / prior / … are wrong?
- Tempered likelihood (Zhang 06) / SafeBayes (Grünwald 12):
  - If your model is misspecified, can be provably better to use $\mathscr{L}^\lambda$ for $\lambda < 1$
  - No longer quite Bayesian inference, but turns a prior into a posterior

A road to PAC-Bayes

- Bayesians say:
  - Start with a prior distribution $\pi(h)$ on choice of hypothesis
  - Observe data S with likelihood $\mathscr{L}(S \mid h)$
  - End up with posterior distribution $\rho(h \mid S) \propto \mathscr{L}(S \mid h)\,\pi(h)$
  - Make predictions/decision based on posterior mean/median, MAP, single draw, …
- This is optimal if you believe in your prior + likelihood! 😊
  - Frequentists say: “but how good is it actually???”
  - What if your model class / prior / … are wrong?
- Tempered likelihood (Zhang 06) / SafeBayes (Grünwald 12):
  - If your model is misspecified, can be provably better to use $\mathscr{L}^\lambda$ for $\lambda < 1$
  - No longer quite Bayesian inference, but turns a prior into a posterior
- PAC-Bayes: analyzes *any* prior-posterior pair (potentially even totally unrelated)

PAC-Bayes: McAllester bound

- We start with some prior π (independent of the data S) on hypotheses

PAC-Bayes: McAllester bound

- We start with some prior π (independent of the data S) on hypotheses
- Our learning algorithm sees S and gives us a posterior ρ

PAC-Bayes: McAllester bound

- We start with some prior π (independent of the data S) on hypotheses
- Our learning algorithm sees S and gives us a posterior ρ
- We'll analyze $L_{\mathcal{D}}(\rho) = \mathbb{E}_{h\sim\rho}[L_{\mathcal{D}}(h)]$ based on $L_S(\rho) = \mathbb{E}_{h\sim\rho}[L_S(h)]$

PAC-Bayes: McAllester bound

- We start with some prior π (independent of the data S) on hypotheses
- Our learning algorithm sees S and gives us a posterior ρ
- We'll analyze $L_{\mathcal{D}}(\rho) = \mathbb{E}_{h\sim\rho}[L_{\mathcal{D}}(h)]$ based on $L_S(\rho) = \mathbb{E}_{h\sim\rho}[L_S(h)]$
- McAllester-style bound (SSBD theorem 31.1):

PAC-Bayes: McAllester bound

- We start with some prior π (independent of the data S) on hypotheses
- Our learning algorithm sees S and gives us a posterior ρ
- We'll analyze $L_{\mathcal{D}}(\rho) = \mathbb{E}_{h\sim\rho}[L_{\mathcal{D}}(h)]$ based on $L_S(\rho) = \mathbb{E}_{h\sim\rho}[L_S(h)]$
- McAllester-style bound (SSBD theorem 31.1):
  - If $\ell(h, z) \in [0,1]$, with probability at least $1-\delta$ over $S \sim \mathcal{D}^n$,

$$L_{\mathcal{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\frac{n}{\delta}}{2(n-1)}}$$

where $\mathrm{KL}(\rho\|\pi) = \mathbb{E}_{h\sim\rho} \log\frac{\rho(h)}{\pi(h)}$ (the usual KL divergence)

PAC-Bayes: McAllester bound

- We start with some prior π (independent of the data S) on hypotheses
- Our learning algorithm sees S and gives us a posterior ρ
- We'll analyze $L_{\mathcal{D}}(\rho) = \mathbb{E}_{h\sim\rho}[L_{\mathcal{D}}(h)]$ based on $L_S(\rho) = \mathbb{E}_{h\sim\rho}[L_S(h)]$
- McAllester-style bound (SSBD theorem 31.1):
  - If $\ell(h, z) \in [0,1]$, with probability at least $1-\delta$ over $S \sim \mathcal{D}^n$,

$$L_{\mathcal{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\frac{n}{\delta}}{2(n-1)}}$$

    where $\mathrm{KL}(\rho\|\pi) = \mathbb{E}_{h\sim\rho} \log\frac{\rho(h)}{\pi(h)}$ (the usual KL divergence)

  - Proved in SSBD chapter 31 (not bad at all)

What learning algorithm?

$$L_{\mathscr{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log \frac{n}{\delta}}{2(n-1)}}$$

- What's the best learning algorithm, according to this bound?

What learning algorithm?

$$L_{\mathscr{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log \frac{n}{\delta}}{2(n-1)}}$$

- What's the best learning algorithm, according to this bound?
  - Turns out to be the **Gibbs posterior**: $\rho(h) \propto \exp(-\lambda L_S(h))\ \pi(h)$

What learning algorithm?

$$L_{\mathscr{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log \frac{n}{\delta}}{2(n-1)}}$$

- What's the best learning algorithm, according to this bound?
  - Turns out to be the **Gibbs posterior**: $\rho(h) \propto \exp(-\lambda L_S(h))\ \pi(h)$
  - Same as tempered likelihood / SafeBayes if $\mathscr{L}(S \mid h) = -\log L_S(h) + \text{const}$

What learning algorithm?

$$L_{\mathscr{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log \frac{n}{\delta}}{2(n-1)}}$$

- What's the best learning algorithm, according to this bound?
  - Turns out to be the **Gibbs posterior**: $\rho(h) \propto \exp(-\lambda L_S(h))\ \pi(h)$
  - Same as tempered likelihood / SafeBayes if $\mathscr{L}(S \mid h) = -\log L_S(h) + \text{const}$
  - Typical choice (see 340): e.g. squared loss ↔ Gaussian likelihood

What learning algorithm?

$$L_{\mathscr{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log \frac{n}{\delta}}{2(n-1)}}$$

- What's the best learning algorithm, according to this bound?
  - Turns out to be the **Gibbs posterior**: $\rho(h) \propto \exp(-\lambda L_S(h))\ \pi(h)$
  - Same as tempered likelihood / SafeBayes if $\mathscr{L}(S \mid h) = -\log L_S(h) + \text{const}$
  - Typical choice (see 340): e.g. squared loss $\leftrightarrow$ Gaussian likelihood

- But the bound applies to **any** prior-posterior pair (with π independent of S)

What learning algorithm?

$$L_{\mathscr{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log \frac{n}{\delta}}{2(n-1)}}$$

- What's the best learning algorithm, according to this bound?
  - Turns out to be the **Gibbs posterior**: $\rho(h) \propto \exp(-\lambda L_S(h))\ \pi(h)$
  - Same as tempered likelihood / SafeBayes if $\mathscr{L}(S \mid h) = -\log L_S(h) + \text{const}$
  - Typical choice (see 340): e.g. squared loss ↔ Gaussian likelihood

- But the bound applies to **any** prior-posterior pair (with π independent of S)
  - For instance: could learn a $\hat{h}$ with (S)GD and then add noise to it

What learning algorithm?

$$L_{\mathscr{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log \frac{n}{\delta}}{2(n-1)}}$$

- What's the best learning algorithm, according to this bound?
  - Turns out to be the **Gibbs posterior**: $\rho(h) \propto \exp(-\lambda L_S(h))\ \pi(h)$
  - Same as tempered likelihood / SafeBayes if $\mathscr{L}(S \mid h) = -\log L_S(h) + \text{const}$
  - Typical choice (see 340): e.g. squared loss $\leftrightarrow$ Gaussian likelihood

- But the bound applies to **any** prior-posterior pair (with π independent of S)
  - For instance: could learn a $\hat{h}$ with (S)GD and then add noise to it
  - If $\hat{h}$ is in a **flat minimum**, then $\hat{h}$ + noise will still be good

What learning algorithm?

$$L_{\mathscr{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \log\frac{n}{\delta}}{2(n-1)}}$$

- What's the best learning algorithm, according to this bound?
  - Turns out to be the **Gibbs posterior**: $\rho(h) \propto \exp(-\lambda L_S(h))\ \pi(h)$
  - Same as tempered likelihood / SafeBayes if $\mathscr{L}(S \mid h) = -\log L_S(h) + \text{const}$
  - Typical choice (see 340): e.g. squared loss $\leftrightarrow$ Gaussian likelihood

- But the bound applies to **any** prior-posterior pair (with π independent of S)
  - For instance: could learn a $\hat{h}$ with (S)GD and then add noise to it
  - If $\hat{h}$ is in a **flat minimum**, then $\hat{h}$ + noise will still be good
  - But note that if $\rho \to$ point mass and π continuous, $\mathrm{KL}(\rho\|\pi) \to \infty$

What prior?

$$L_{\mathscr{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log \frac{n}{\delta}}{2(n-1)}}$$

- What's the best prior?
  - Bound on generalization gap is better if ρ is "closer" to π

What prior?

$$L_{\mathcal{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log \frac{n}{\delta}}{2(n-1)}}$$

- What's the best prior?
  - Bound on generalization gap is better if ρ is "closer" to π
    - S didn't make us "change our mind" too much – similar to MDL

What prior?

$$L_{\mathcal{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log \frac{n}{\delta}}{2(n-1)}}$$

- What's the best prior?
  - Bound on generalization gap is better if ρ is "closer" to π
    - S didn't make us "change our mind" too much – similar to MDL
  - But we also want a good ρ, i.e. average training loss $L_S(\rho)$ should be small

What prior?

$$L_{\mathscr{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log \frac{n}{\delta}}{2(n-1)}}$$

- What's the best prior?
  - Bound on generalization gap is better if ρ is "closer" to π
    - S didn't make us "change our mind" too much – similar to MDL
  - But we also want a good ρ, i.e. average training loss $L_S(\rho)$ should be small
- Notice π only shows up in the bound – nothing to do with the learning algorithm

What prior?

$$L_{\mathscr{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log \frac{n}{\delta}}{2(n-1)}}$$

- What's the best prior?
  - Bound on generalization gap is better if ρ is "closer" to π
    - S didn't make us "change our mind" too much – similar to MDL
  - But we also want a good ρ, i.e. average training loss $L_S(\rho)$ should be small
- Notice π only shows up in the bound – nothing to do with the learning algorithm
  - We could potentially pick a prior that actually **depends on** $\mathscr{D}$

What prior?

$$L_{\mathscr{D}}(\rho) - L_S(\rho) \leq \sqrt{\frac{\mathrm{KL}(\rho \| \pi) + \log \frac{n}{\delta}}{2(n-1)}}$$

- What's the best prior?
  - Bound on generalization gap is better if ρ is "closer" to π
    - S didn't make us "change our mind" too much – similar to MDL
  - But we also want a good ρ, i.e. average training loss $L_S(\rho)$ should be small
- Notice π only shows up in the bound – nothing to do with the learning algorithm
  - We could potentially pick a prior that actually **depends on** $\mathscr{D}$
  - …as long as we can still bound $\mathrm{KL}(\rho \| \pi)$

Other forms of PAC-Bayes bounds

- Linear bound: $L_{\mathscr{D}}(\rho) \leq \frac{1}{\beta} L_S(\rho) + \frac{\mathrm{KL}(\rho\|\pi) + \log\frac{1}{\delta}}{2\beta(1-\beta)n}$ for any $\beta \in (0,1)$

- Catoni bound: for $\alpha > 1$, $\Phi_\gamma^{-1}(x) = (1-\exp(-\gamma x))/(1-\exp(-\gamma))$,

  $$L_{\mathscr{D}}(\rho) \leq \inf_{\lambda>1} \Phi_{\lambda/n}^{-1}\left(L_S(\rho) + \frac{\alpha}{\lambda}\left[\mathrm{KL}(\rho\|\pi) - \log\varepsilon + 2\log\frac{\log(\alpha^2\lambda)}{\log\alpha}\right]\right)$$

  - Can be much tighter (unfortunately) if $\mathrm{KL}(\rho\|\pi)/n$ is big

- Also variants based on general f-divergences, Wasserstein, …

NON-VACUOUS GENERALIZATION BOUNDS AT THE IMAGENET SCALE: A PAC-BAYESIAN COMPRESSION APPROACH

Wenda Zhou
Columbia University
New York, NY
wz2335@columbia.edu

Victor Veitch
Columbia University
New York, NY
victorveitch@gmail.com

Morgane Austern
Columbia University
New York, NY
ma3293@columbia.edu

Ryan P. Adams
Princeton University
Princeton, NJ
rpa@princeton.edu

Peter Orbanz
Columbia University
New York, NY
porbanz@stat.columbia.edu

- Pre-pick a coding scheme to represent networks (e.g. compress the weights)
- Train a network with SGD, sparsify it/etc to $\hat{h}$, then add a little noise to weights

NON-VACUOUS GENERALIZATION BOUNDS AT THE IMAGENET SCALE: A PAC-BAYESIAN COMPRESSION APPROACH

Wenda Zhou
Columbia University
New York, NY
wz2335@columbia.edu

Victor Veitch
Columbia University
New York, NY
victorveitch@gmail.com

Morgane Austern
Columbia University
New York, NY
ma3293@columbia.edu

Ryan P. Adams
Princeton University
Princeton, NJ
rpa@princeton.edu

Peter Orbanz
Columbia University
New York, NY
porbanz@stat.columbia.edu

- Pre-pick a coding scheme to represent networks (e.g. compress the weights)
- Train a network with SGD, sparsify it/etc to $\hat{h}$, then add a little noise to weights

Table 1: Summary of bounds obtained from compression

Dataset	Orig. size	Comp. size	Robust. Adj.	Eff. Size	Error Bound Top 1	Error Bound Top 5
MNIST	168.4 KiB	8.1 KiB	1.88 KiB	6.23 KiB	$< 46\,\%$	NA
ImageNet	5.93 MiB	452 KiB	102 KiB	350 KiB	$< 96.5\,\%$	$< 89\,\%$

Derandomizing PAC-Bayes

- In practice, we don’t actually use randomized predictors (usually)

Derandomizing PAC-Bayes

- In practice, we don't actually use randomized predictors (usually)
- Possible to "derandomize" to a high-probability bound on $L_{\mathcal{D}}(h) - L_S(h)$:

Derandomizing PAC-Bayes

- In practice, we don't actually use randomized predictors (usually)
- Possible to "derandomize" to a high-probability bound on $L_{\mathcal{D}}(h) - L_S(h)$:
  - Show convergence of $L_{\mathcal{D}}(h)$ to $\mathbb{E}_{h\sim\rho} L_{\mathcal{D}}(h)$, $L_S(h)$ to $\mathbb{E}_{h\sim\rho} L_S(h)$, under ρ

Derandomizing PAC-Bayes

- In practice, we don't actually use randomized predictors (usually)
- Possible to "derandomize" to a high-probability bound on $L_{\mathcal{D}}(h) - L_S(h)$:
  - Show convergence of $L_{\mathcal{D}}(h)$ to $\mathbb{E}_{h\sim\rho}L_{\mathcal{D}}(h)$, $L_S(h)$ to $\mathbb{E}_{h\sim\rho}L_S(h)$, under ρ
  - Or, use a margin-type loss to show 0-1 error doesn't change under ρ

Derandomizing PAC-Bayes

- In practice, we don't actually use randomized predictors (usually)
- Possible to "derandomize" to a high-probability bound on $L_{\mathcal{D}}(h) - L_S(h)$:
  - Show convergence of $L_{\mathcal{D}}(h)$ to $\mathbb{E}_{h\sim\rho}L_{\mathcal{D}}(h)$, $L_S(h)$ to $\mathbb{E}_{h\sim\rho}L_S(h)$, under ρ
  - Or, use a margin-type loss to show 0-1 error doesn't change under ρ
- But…these then become "two-sided" bounds

Derandomizing PAC-Bayes

- In practice, we don't actually use randomized predictors (usually)
- Possible to "derandomize" to a high-probability bound on $L_{\mathcal{D}}(h) - L_S(h)$:
  - Show convergence of $L_{\mathcal{D}}(h)$ to $\mathbb{E}_{h\sim\rho}L_{\mathcal{D}}(h)$, $L_S(h)$ to $\mathbb{E}_{h\sim\rho}L_S(h)$, under ρ
  - Or, use a margin-type loss to show 0-1 error doesn't change under ρ
- But…these then become "two-sided" bounds
  - Subject to the Nagarajan/Kolter failure mode (their Appendix J)

Uniform convergence may be unable to explain generalization in deep learning

---

Vaishnavh Nagarajan
Department of Computer Science
Carnegie Mellon University
Pittsburgh, PA
vaishnavh@cs.cmu.edu

J. Zico Kolter
Department of Computer Science
Carnegie Mellon University &
Bosch Center for Artificial Intelligence
Pittsburgh, PA
zkolter@cs.cmu.edu

Recap

- Double descent
  - Classical behaviour, then descend again after interpolation peak
  - Highly dependent on learning algorithm's implicit regularization

- PAC-Bayes
  - A Primer on PAC-Bayesian Learning (Guedj 2019)
  - Tightest bounds we know for deep learning (afaik)…but still not *that* tight
  - Still, significant practical issues, not always super explanatory
  - Lots of ongoing fancy variations