Lecture 23

and last!

Differential Privacy

Double Descent

Evaluations

Do well on the final!

Sol 7, 8

OH TA Mon 9:30, 10:30

OH MWF 5:30-6:20 Monday

+ another OH TBA

Differentially Private Machine Learning

Gautam Kamath

NetFlix Prize

- Recommendation engine competition (2006-2009)
- Training data: (anonymized) user ID, movie, rating, date
- Matched with public IMDb data: real name, movie, rating, date
- Class action lawsuit, cancellation of sequel

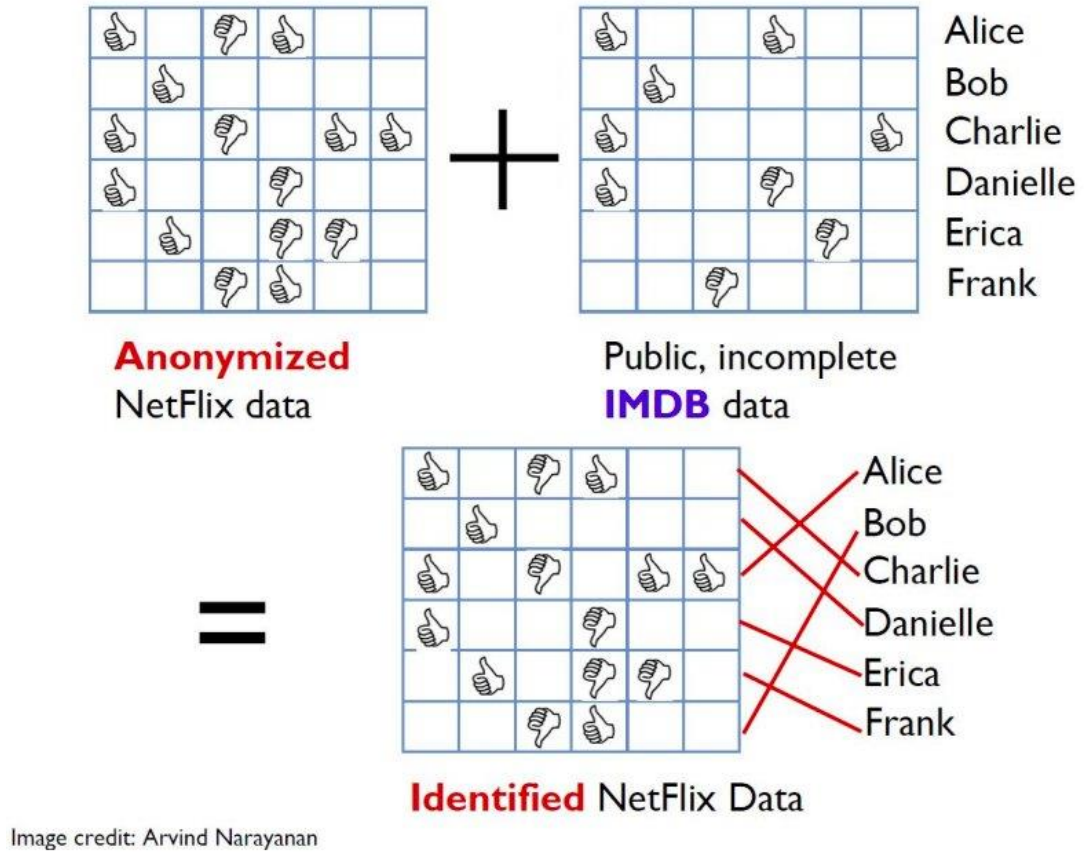


Image credit: Arvind Narayanan

Privacy Concerns

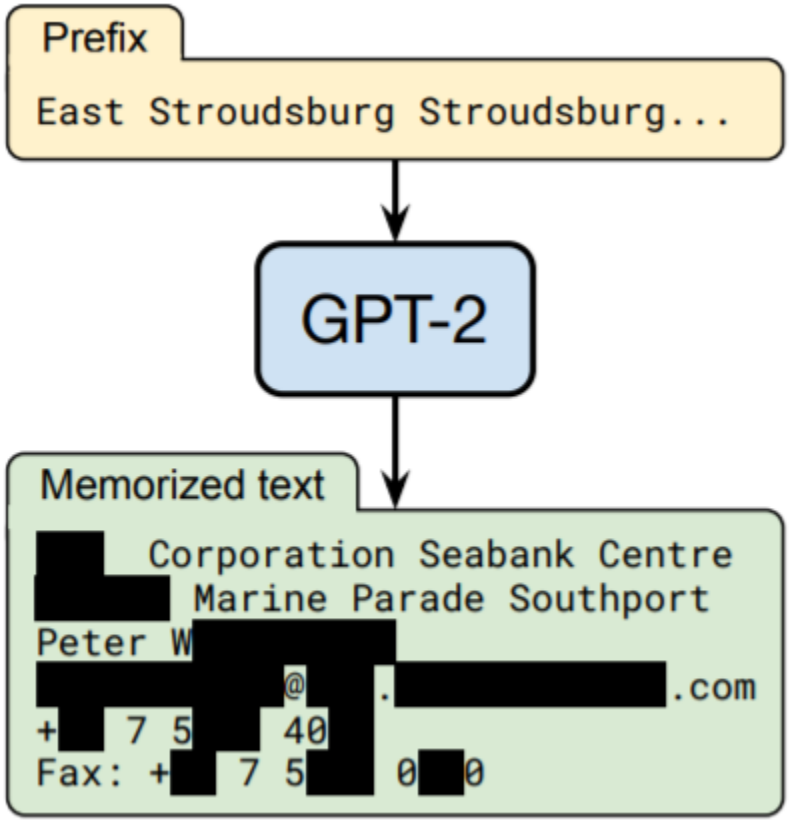
- Machine learning models are trained on very large datasets
- Can be coerced to reproduce training data verbatim!

In our recent paper, we evaluate how large language models *memorize* and *regurgitate* such rare snippets of their training data. **We focus on GPT-2 and find that at least 0.1% of its text generations (a very conservative estimate) contain long verbatim strings that are "copy-pasted" from a document in its training set.**

Blog post: [Wallace, Tramer, Jagielski, Herbert-Voss], 2020

Paper: [Carlini, Tramer, Wallace, Jagielski, Herbert-Voss, Lee, Roberts, Brown, Song, Erlingsson, Oprea, Raffel], 2021

Personal Information



Copyrighted content

Below, we prompt GPT … r and the Philosopher's Stone. **The model corr… ok** (about 240 words) before making its first …

The escape of the Brazilian boa constrictor earned Harry his longest-ever punishment. By the time he was allowed out of his cupboard again, the *summer holidays had started and Dudley had already broken his new video camera, crashed his remote-control aeroplane, and, first time out on his racing bike, knocked down old Mrs Figg as she crossed Privet Drive on her crutches.*

Harry was glad school was over, but there was no escaping Dudley's gang, who visited the house every single day. Piers, Dennis, Malcolm, and Gordon were all big and stupid, but as Dudley was the biggest and stupidest of the lot, he was the leader. The rest of them were all quite happy to join in Dudley's favourite sport: Harry Hunting.

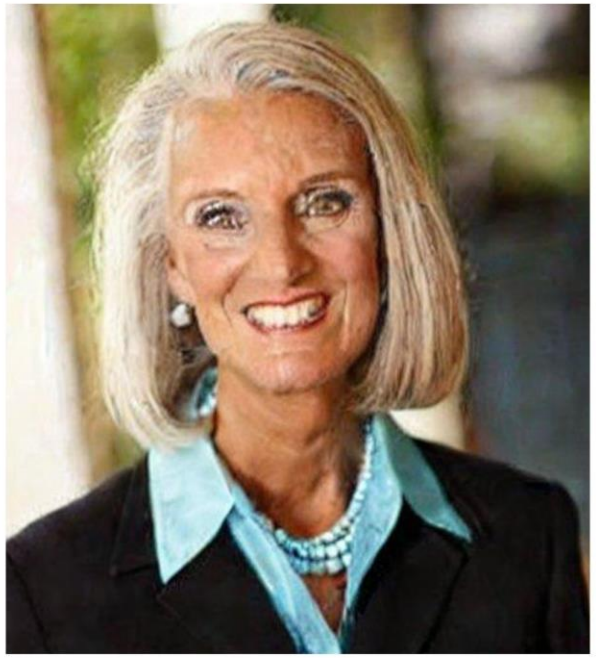
This was why Harry spent as much time as possible out of the house, wandering around and thinking about the end of the holidays, where he could see a tiny ray of hope. When September came he would be going off to secondary school and, for the first time in his life, he wouldn't be with Dudley. Dudley had been accepted at Uncle Vernon's old private school, Smeltings. Piers Polkiss was going there too. Harry, on the other hand, was going to Stonewall High, the local public school. Dudley thought this was very funny.

Also for Diffusion Models

Training Set

Caption: Living in the light with Ann Graham Lotz

Generated Image



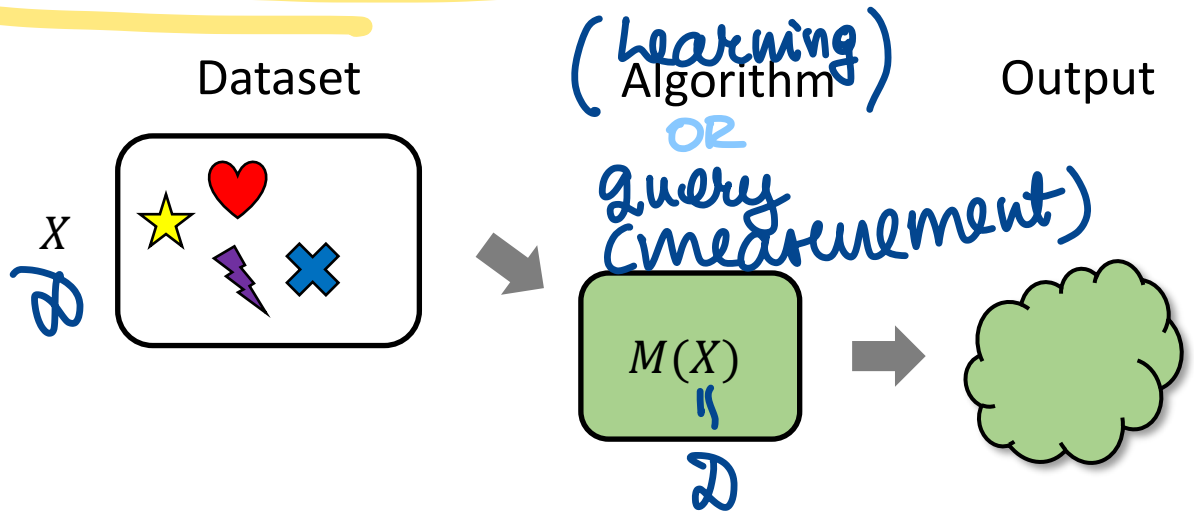
Prompt: Ann Graham Lotz

[Carlini, Hayes, Nasr, Jagielski, Sehwag, Tramer, Balle, Ippolito, Wallace], 2023

Implications for copyright…?

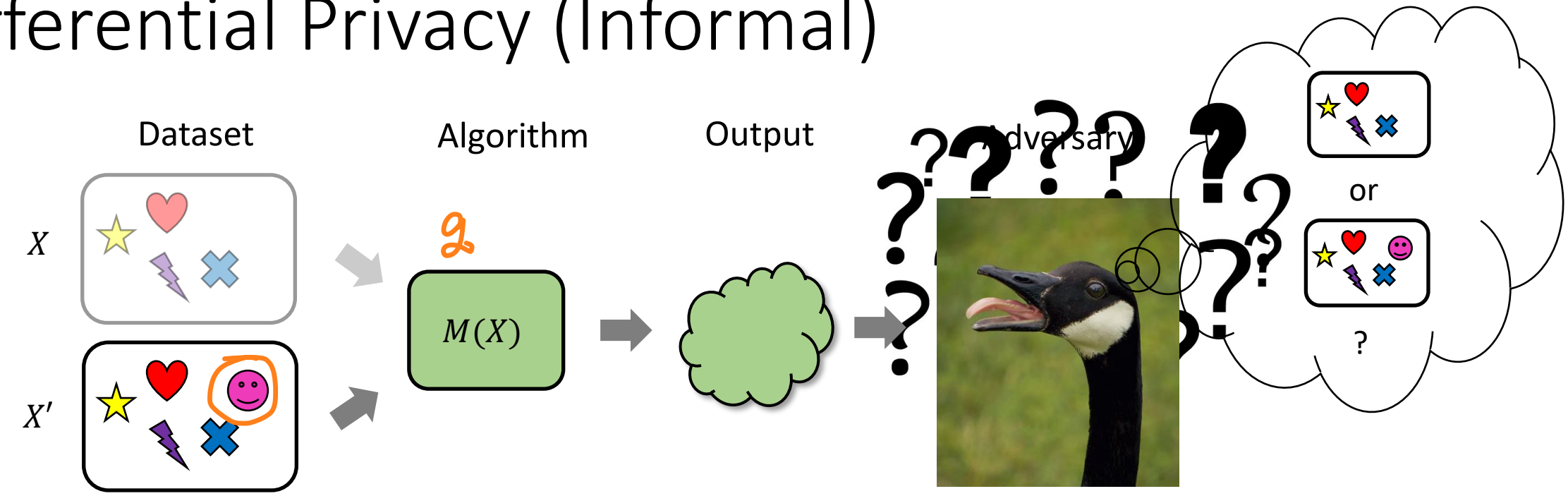
- Getty Images vs. Stability AI?

Differential Privacy (Informal)



[Dwork, McSherry, Nissim, Smith], 2006

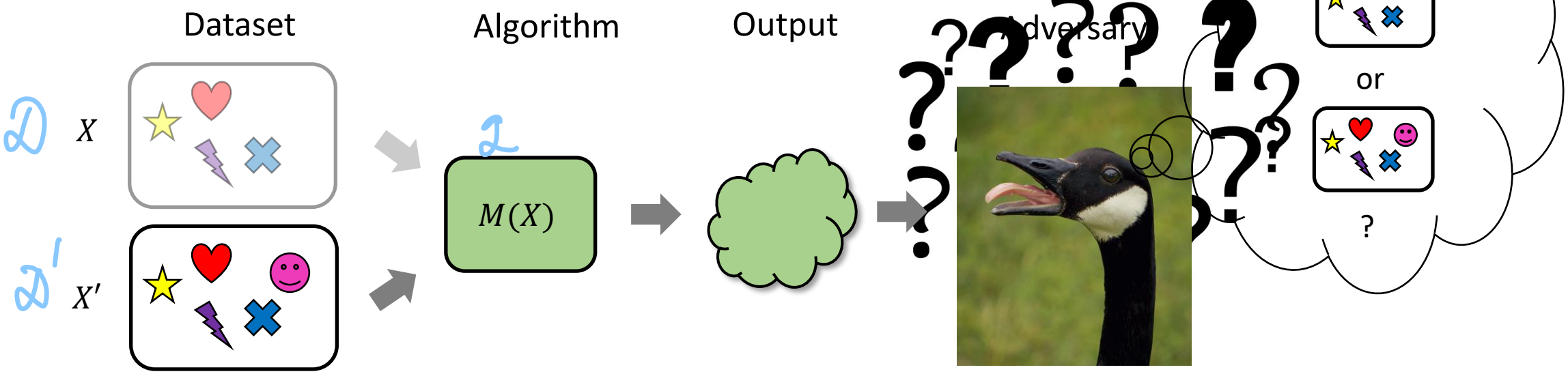
Differential Privacy (Informal)



"An algorithm is differentially private if its distribution over outputs doesn't change much after adding/removing one point."

[Dwork, McSherry, Nissim, Smith], 2006

Differential Privacy



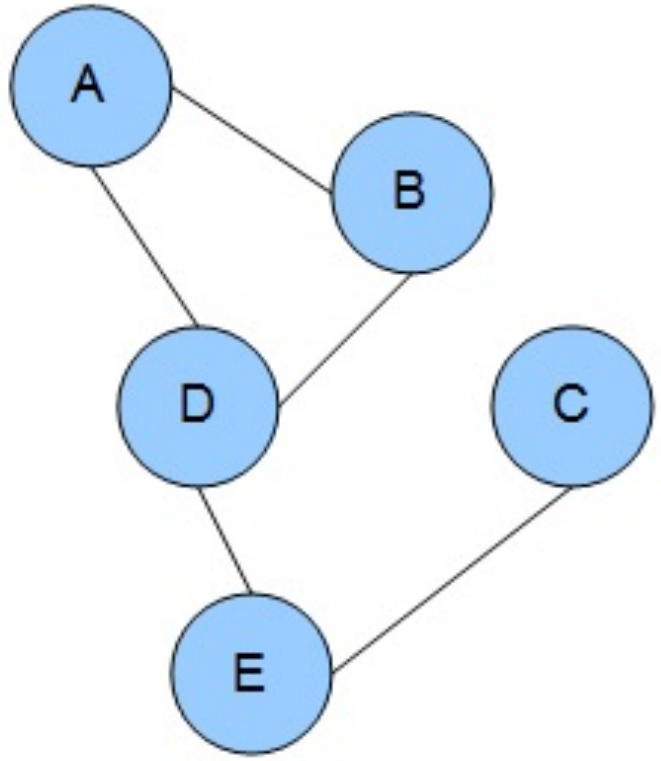
- $M: D^n \rightarrow R$ is (ε, δ)-DP if for all inputs X, X' which differ on one entry:

$$\forall S \subseteq R \qquad \Pr[M(X) \in S] \leq e^{\varepsilon} \Pr[M(X') \in S] + \delta$$

[Dwork, McSherry, Nissim, Smith], 2006

Widely Adopted

Accepted in theory…



And practice!

Google

United States™ Census Bureau

Endorsed by the White House

Administration Priorities The Record

OCTOBER 30, 2023

FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence

BRIEFING ROOM › STATEMENTS AND RELEASES

Today, President Biden is issuing a landmark Executive Order to ensure that America leads the way in seizing the promise and managing the risks of artificial intelligence (AI). The Executive Order establishes new standards for AI safety and security, protects Americans' privacy, advances equity and civil rights, stands up for consumers and workers, promotes innovation and competition, advances American leadership around the world, and more.

(b) Within 365 days of the date of this order, to better enable agencies to use PETs to safeguard Americans' privacy from the potential threats exacerbated by AI, the Secretary of Commerce, acting through the Director of NIST, shall create guidelines for agencies to evaluate the efficacy of differential-privacy-guarantee protections, including for AI. The guidelines shall, at a minimum, describe the significant factors that bear on differential-privacy safeguards and common risks to realizing differential privacy in practice.

Differential Privacy (DMNS06)

$M: D^n \to R$ is (ε, δ)-DP if for all inputs X, X' which differ on one entry:

$$\forall S \subseteq R \qquad \Pr[M(X) \in S] \leq e^{\varepsilon} \Pr[M(X') \in S] + \delta$$

- $\varepsilon \approx 1$ and $\delta < 1/n$
- Worst-case guarantee
- $e^{\varepsilon_1} e^{\varepsilon_2} = e^{\varepsilon_1 + \varepsilon_2}$
- Symmetric definition
- M must be randomized

$$-\varepsilon \leq \ln \frac{\Pr[q(\mathcal{D}) = y]}{\Pr[q(\mathcal{D}') = y]} \leq \varepsilon$$

$\delta > 0$ approximate DP

$\delta = 0$ pure DP

What DP does and does not mean

- Outcome is the same whether or not your data is in the dataset
- Protects against linkage and membership inference attacks
- Does *not* prevent statistics and machine learning
  - "Smoking causes cancer"
- Not suitable when we need to identify a specific individual
- Information-theoretic notion

can learn from D

Properties of Differential Privacy

- Post-processing
  - If $M(X)$ is (ε, δ)-DP, then $f(M(X))$ is (ε, δ)-DP
- Group Privacy
  - If M is (ε, δ)-DP, and X and X' differ in k entries,

$$\forall S \subseteq R \qquad \Pr[M(X) \in S] \leq e^{k\varepsilon} \Pr[M(X') \in S] + \delta$$

- Composition
  - If $M = (M_1, \ldots, M_k)$ is a sequence of k (ε, δ)-DP algorithms
    - M is $(k\varepsilon, k\delta)$-DP (Basic Composition)
    - M is $(O(\sqrt{k}\varepsilon \log(1/\delta')), k\delta + \delta')$-DP (Advanced Composition)

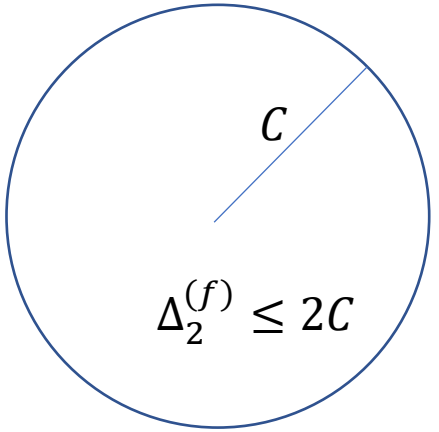
$$e^{\varepsilon_1} \cdot e^{\varepsilon_2} = e^{\varepsilon_1 + \varepsilon_2}$$

Gaussian Mechanism

- ℓ_2-sensitivity of f

$$\Delta_2^{(f)} = \max_{X \sim X'} \|f(X) - f(X')\|_2$$

  - If $\|f(X)\|_2 \leq C$, then $\Delta_2^{(f)} \leq 2C$



- Gaussian Mechanism

$$M(X) = f(X) + (Y_1, \ldots, Y_k)$$

  Where $f(X) \in R^k$, and the Y_i's are $\approx N(0, \Delta^2/\varepsilon^2)$

  - (ε, δ)-DP

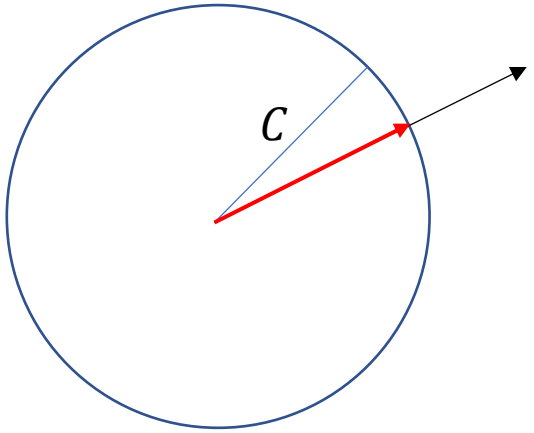
Stochastic Gradient Descent

↳ $B \ll n$

1. Choose a random minibatch B of points from the dataset
2. Compute the average gradient $\frac{1}{|B|}\sum_{(x,y)\in B} \nabla \ell(\theta_t, x, y) = g$ noisy gradient
3. Take a step in the negative direction of the gradient
4. Repeat k times $k = T$ = #iterations

Differentially Private Stochastic Gradient Descent

1. Sample a "lot" of points of (expected) size L by selecting each point to be in the lot with probability L/n
2. For each point in the lot, compute the gradient $\nabla \ell(\theta_t, x, y)$ and "clip" it to have ℓ_2 norm at most C
3. Average the clipped gradients and add Gaussian noise
   - Apply the Gaussian Mechanism
4. Take a step in the negative direction of resulting vector
5. Repeat k times

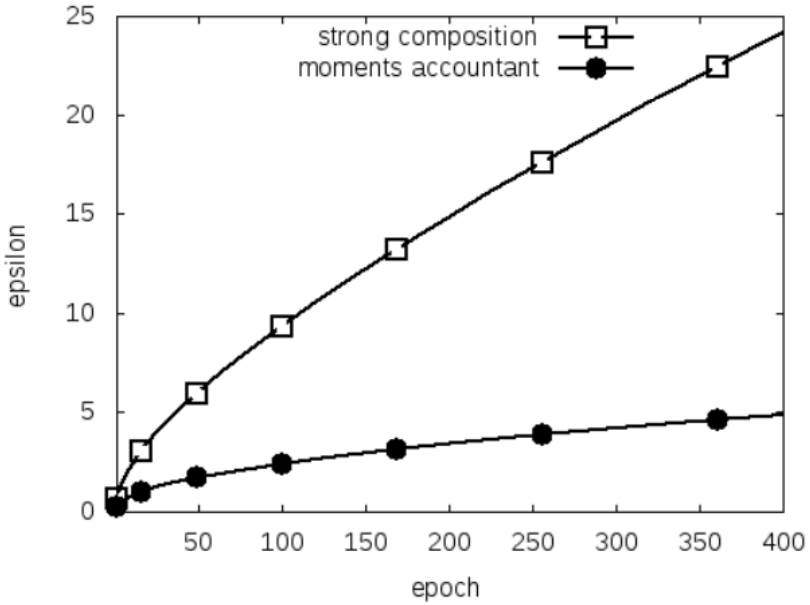


[Song-Chaudhuri-Sarwate '13, Bassily-Smith-Thakurta '14, Abadi-Chu-Goodfellow-McMahan-Mironov-Talwar-Zhang '16]

Privacy of DPSGD (Informal)

$\sigma_{noise} = \frac{c}{\varepsilon}$

- Suppose one step of DPSGD has privacy with parameter ε
- Since we subsample with probability L/n, each step is $\varepsilon L/n$
  - "Privacy amplification by subsampling"
- k steps have privacy with parameter of $\varepsilon\sqrt{k}L/n$ $= \varepsilon_{SGD}(k)$
  - Advanced composition
- Better analysis: "Moments accountant"



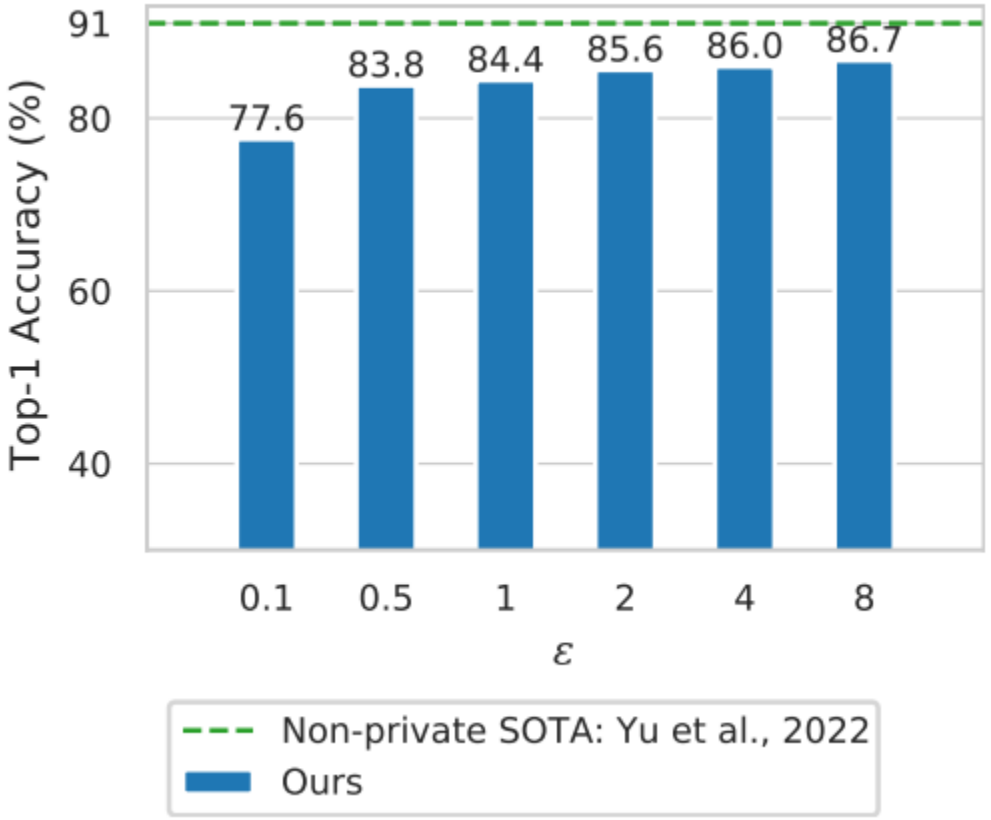
Does it work?

Data	ε-DP	Source	Test Accuracy (%) CNN	ScatterNet+linear	ScatterNet+CNN
MNIST	1.2	Feldman & Zrnic (2020)	96.6	**98.1 ± 0.1**	97.8 ± 0.1
	2.0	Abadi et al. (2016)	95.0	**98.5 ± 0.0**	**98.4 ± 0.1**
	2.32	Bu et al. (2019)	96.6	**98.6 ± 0.0**	98.5 ± 0.0
	2.5	Chen & Lee (2020)	90.0	**98.7 ± 0.0**	98.6 ± 0.0
	2.93	Papernot et al. (2020a)	98.1	**98.7 ± 0.0**	**98.7 ± 0.1**
	3.2	Nasr et al. (2020)	96.1	–	
	6.78	Yu et al. (2019b)	93.2	–	
Fashion-MNIST	2.7	Papernot et al. (2020a)	86.1	**89.5 ± 0.0**	88.7 ± 0.1
	3.0	Chen & Lee (2020)	82.3	**89.7 ± 0.0**	89.0 ± 0.1
CIFAR-10	3.0	Nasr et al. (2020)	55.0	67.0 ± 0.1	**69.3 ± 0.2**
	6.78	Yu et al. (2019b)	44.3	–	–
	7.53	Papernot et al. (2020a)	66.2	–	–
	8.0	Chen & Lee (2020)	53.0	–	–

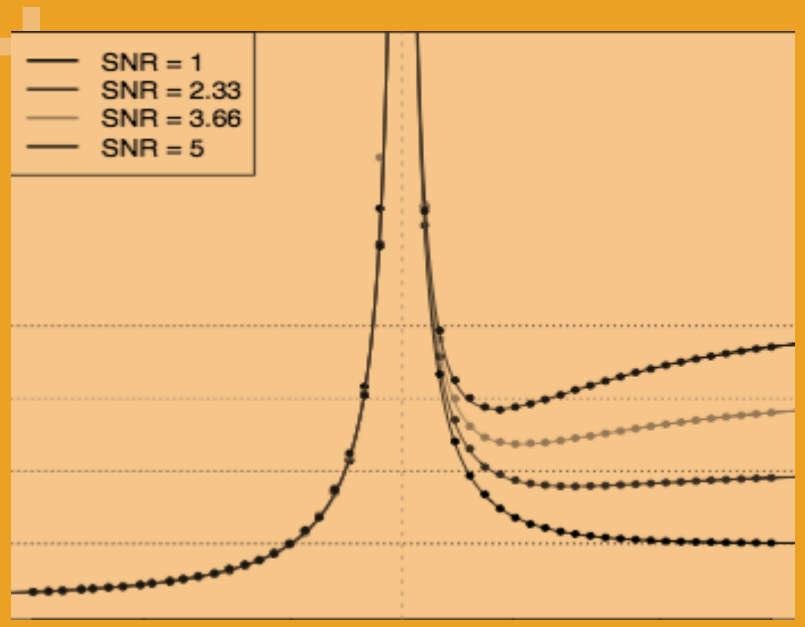
[Tramèr-Boneh ‘21]

Public Data Helps for Private Vision

Dataset	Pre-Training Data	Top-1 Accuracy (%)				δ	Section
		$\varepsilon = 1$	$\varepsilon = 2$	$\varepsilon = 4$	$\varepsilon = 8$		
CIFAR-10	ImageNet	**94.7**	**95.4**	**96.1**	**96.7**	10^{-5}	4.1
CIFAR-100	ImageNet	**70.3**	**74.7**	**79.2**	**81.8**	10^{-5}	4.1
ImageNet	JFT-4B	**84.4**	**85.6**	**86.0**	**86.7**	$8 \cdot 10^{-7}$	4.2
Places-365	JFT-300M	–	–	–	**55.1**	$5 \cdot 10^{-7}$	4.3



[De, Berrada, Hayes, Smith, Balle], 2022



Double Descent

Beyond the Bias-Variance trade-off

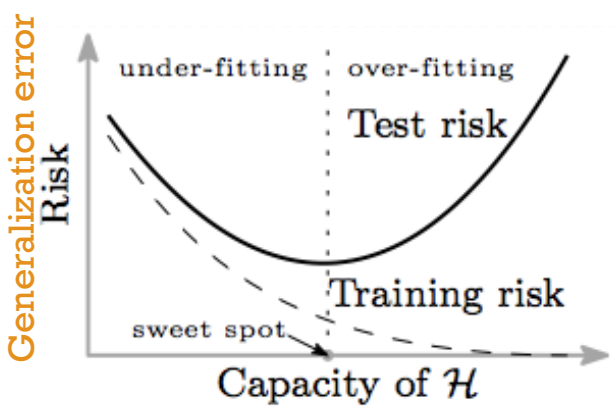
STAT 535+LPL2019

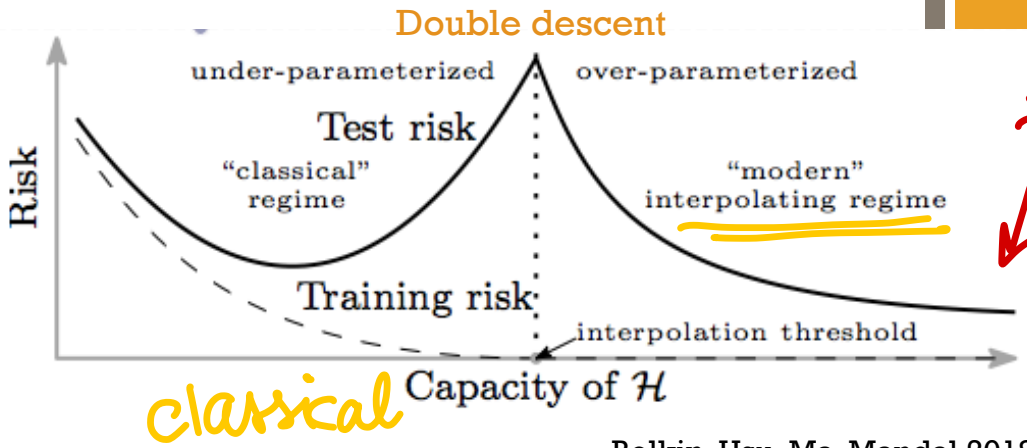
+ CS 480/680 W26

Marina Meila

University of Washington

What is observed

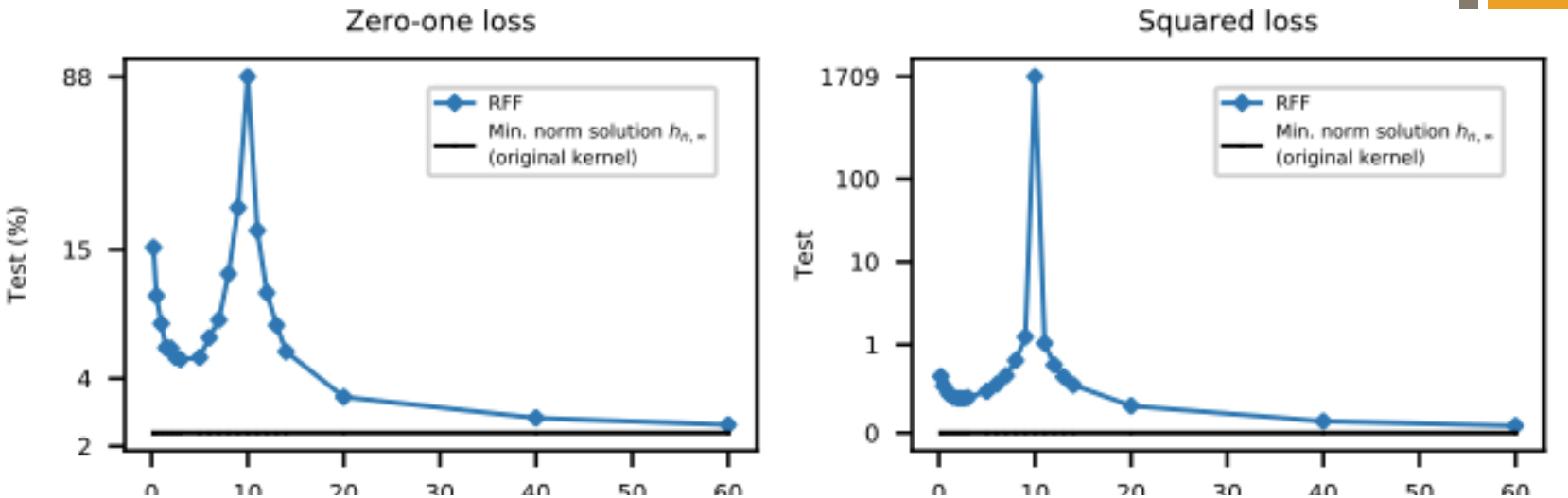




Belkin, Hsu, Ma, Mandal 2018

- Classical regime p < N
- Modern/Deep Learning/High dimensional regime N > n
  - Think N fixed, p increases, gamma=p/N
  - Training error = 0 (interpolation)
  - Test error decreases with p (or gamma)

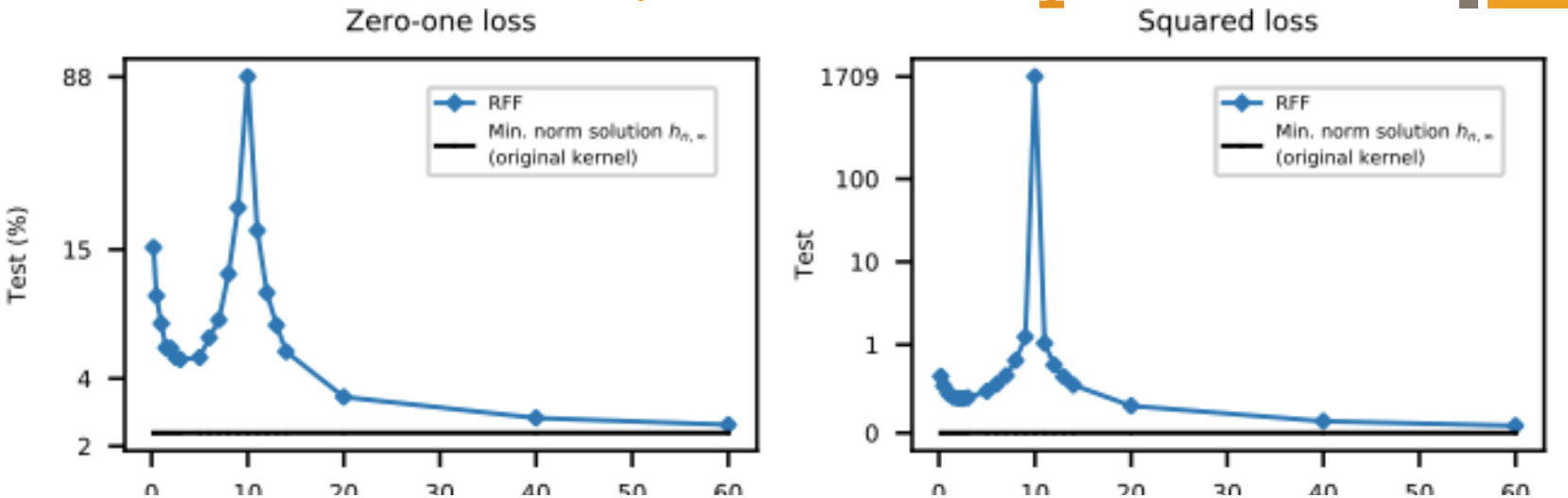
What is observed



Belkin, Hsu, Ma, Mandal 2018

- Double descent curves for the generalization error
  - Random Fourier Features (RFF)
  - ReLU 2 layer networks (with random first layer weights)
  - Random Forests, l2-Adaboost
  - Linear regression
- With and without noise

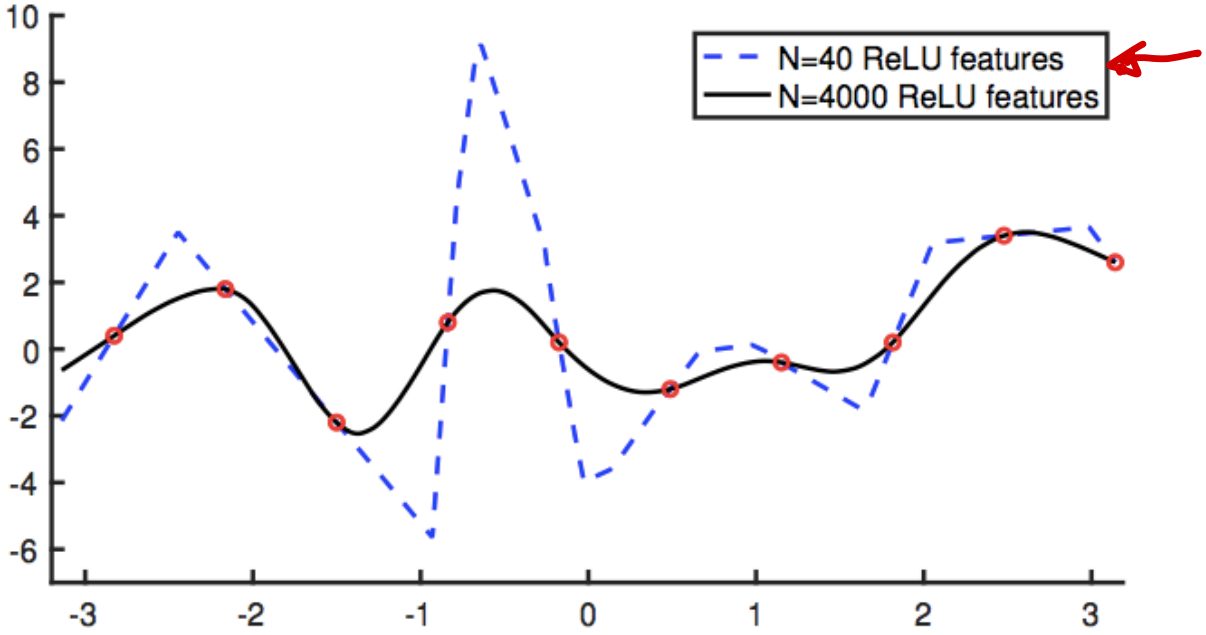
Double descent, the case p > N



Belkin, Hsu, Ma, Mandal 2018

- Model y = <phi(x), beta >
- Large N (cover a compact data domain)
- Features random
- Min-norm solution beta*

Main intuition [Belkin et al.]



ReLu

- The target function h^* is (mostly) smooth
  - i.e. $||h^*||_{RKHS}$ is small
- $p > N$, no noise, hence h_p interpolates data
- Train to minimize $||h_p||$ subject to 0 training error
- Then $||h_p||$ will decrease with p!

Random Fourier Features (RFF)

Random Fourier features. We first consider a popular class of non-linear parametric models called *Random Fourier Features* (*RFF*) [30], which can be viewed as a class of two-layer neural networks with fixed weights in the first layer. The RFF model family $\mathcal{H}_N$ with N (complex-valued) parameters consists of functions $h\colon \mathbb{R}^d \to \mathbb{C}$ of the form

$$h(x) = \sum_{k=1}^{N} a_k \phi(x; v_k) \quad \text{where} \quad \phi(x; v) := e^{\sqrt{-1}\langle v, x \rangle},$$

and the vectors $v_1, \ldots, v_N$ are sampled independently from the standard normal distribution in $\mathbb{R}^d$. (We consider $\mathcal{H}_N$ as a class of real-valued functions with $2N$ real-valued parameters by taking real and imaginary parts separately.) Note that $\mathcal{H}_N$ is a randomized function class, but as $N \to \infty$, the function class becomes a closer and closer approximation to the Reproducing Kernel Hilbert Space (RKHS) corresponding to the Gaussian kernel, denoted by $\mathcal{H}_\infty$.

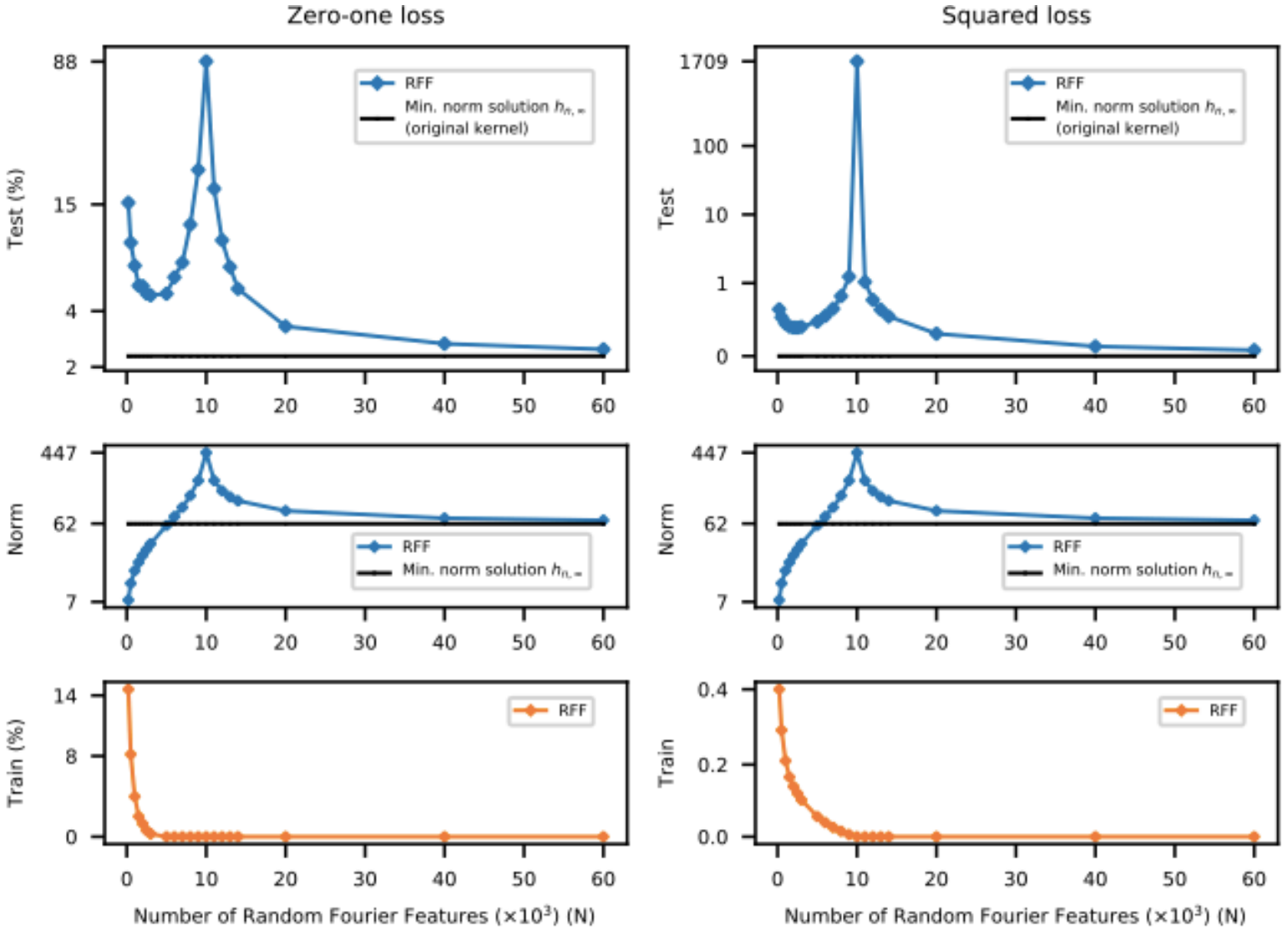
- RFF → $\mathcal{H}_{\text{infinity}}$

Theorem

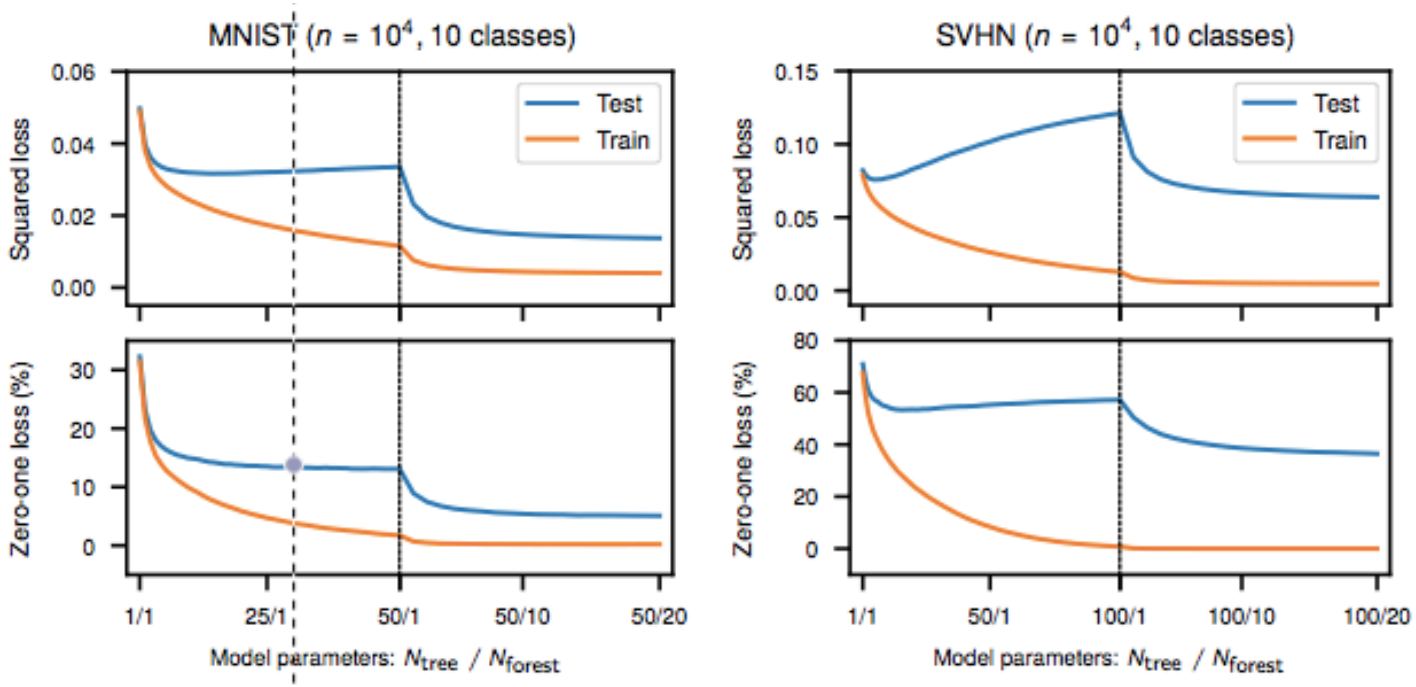
Theorem 1. *Fix any $h^* \in \mathcal{H}_\infty$. Let $(x_1, y_1), \ldots, (x_n, y_n)$ be independent and identically distributed random variables, where x_i is drawn uniformly at random from a compact cube[2] $\Omega \subset \mathbb{R}^d$, and $y_i = h^*(x_i)$ for all i. There exists absolute constants $A, B > 0$ such that, for any interpolating $h \in \mathcal{H}_\infty$ (i.e., $h(x_i) = y_i$ for all i), so that with high probability*

$$\sup_{x \in \Omega} |h(x) - h^*(x)| < Ae^{-B(n/\log n)^{1/d}} \left(\|h^*\|_{\mathcal{H}_\infty} + \|h\|_{\mathcal{H}_\infty} \right).$$

RFF


Zero-one loss
Squared loss
Test (%)
Test
Norm
Train (%)
Train
RFF
Min. norm solution $h_{n,*}$
(original kernel)
Number of Random Fourier Features (×10³) (N)

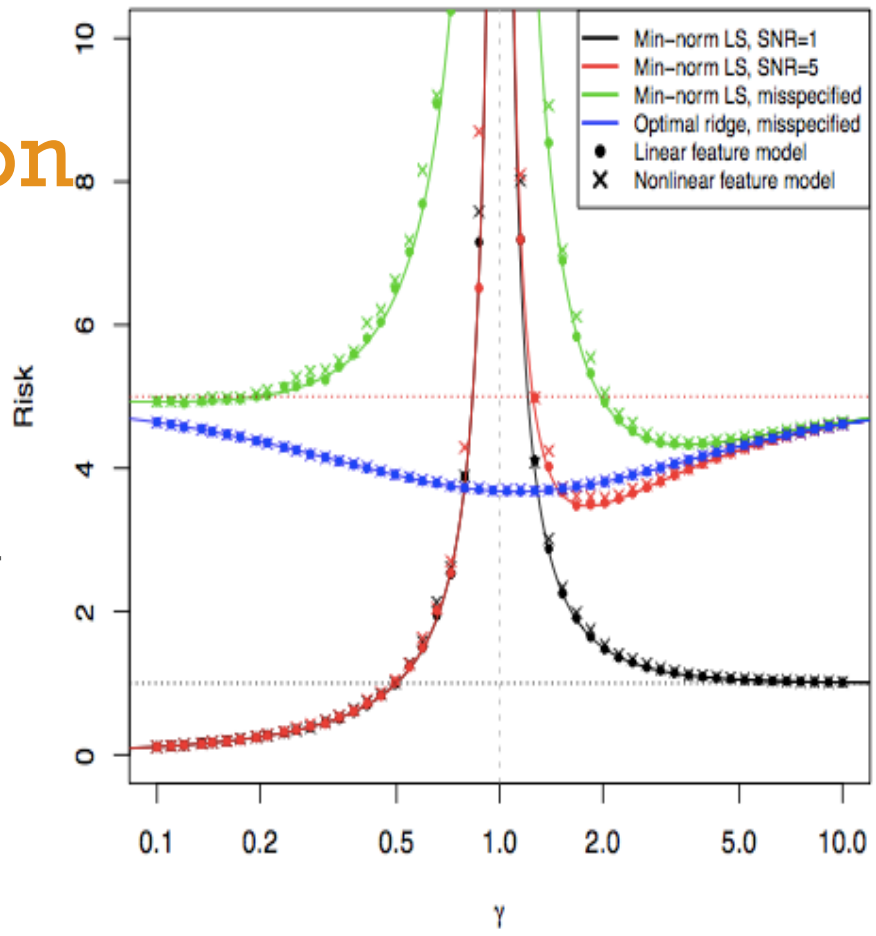
Boosted decision trees



Linear regression

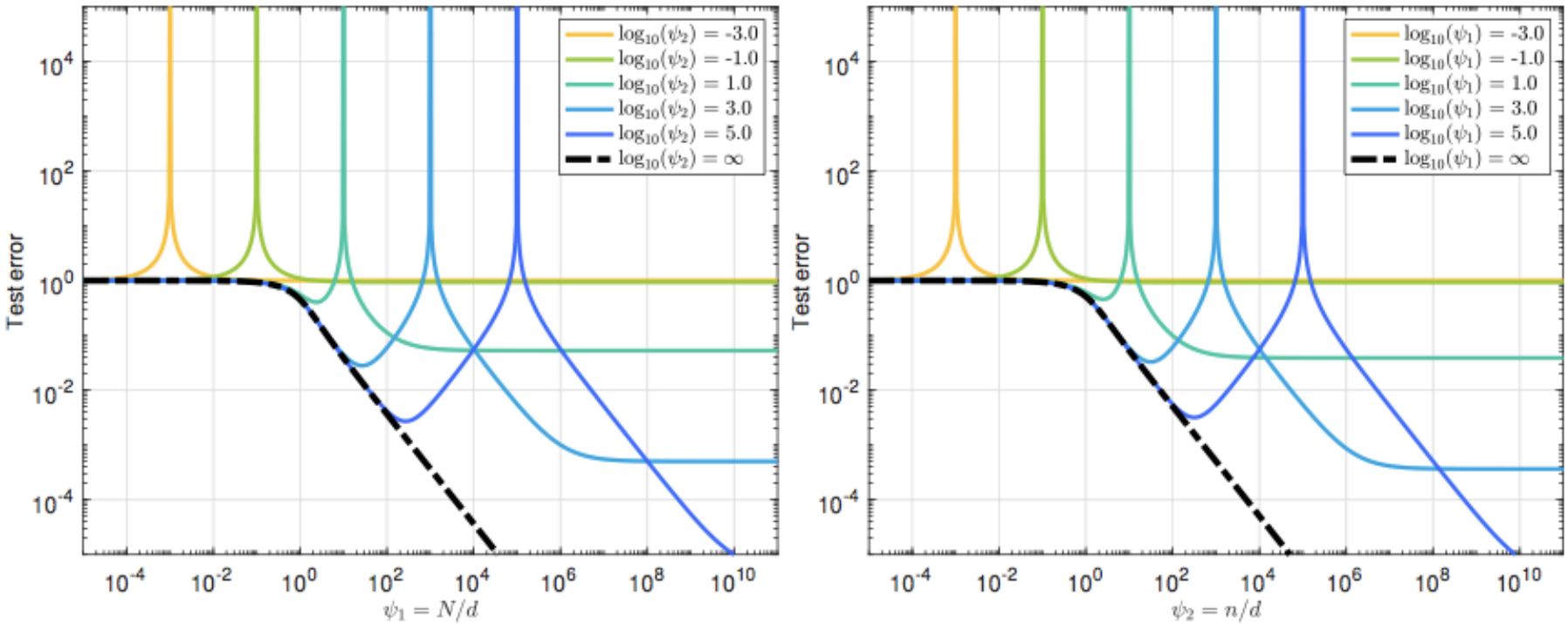
[Hastie, Montanari, Rosset, Tibshirani 2019]

- Linear, nonlinear features behave the same way
- Model correct, misspecified
- Noise level sigma affects asymptotic error
- and optimal N/n



- Double descent is not regularization

Figure 1: Asymptotic risk curves for the linear feature model, as a function of the limiting aspect ratio γ. The risks for min-norm least squares, when SNR $= 1$ and SNR $= 5$, are plotted in black and red, respectively. These two match for $\gamma < 1$ but differ for $\gamma > 1$. The null risks for SNR $= 1$ and SNR $= 5$ are marked by the dotted black and red lines, respectively. The risk for the case of a misspecified model (with significant approximation bias, $a = 1.5$ in (13)), when SNR $= 5$, is plotted in green. Optimally-tuned (equivalently, CV-tuned) ridge regression, in the same misspecified setup, has risk plotted in blue. The points denote finite-sample risks, with $n = 200$, $p = [\gamma n]$, across various values of γ, computed from features X having i.i.d. $N(0, 1)$ entries. Meanwhile, the "x" points mark finite-sample risks for a nonlinear feature model, with $n = 200$, $p = [\gamma n]$, $d = 100$, and $X = \varphi(ZW^T)$, where Z has i.i.d. $N(0, 1)$ entries, W has i.i.d. $N(0, 1/d)$ entries, and $\varphi(t) = a(|t| - b)$ is a "purely nonlinear" activation function, for constants a, b. The theory predicts that this nonlinear risk should converge to the linear risk with p features (regardless of d). The empirical agreement between these two—and the agreement in finite-sample and asymptotic risks—is striking.



- More refined analysis includes noise, non-linearity, data dimension n, ridge regularization lambda [Mei, Montanari 2019]
- When is global minimum in overparametrized regime?
- Enough data N/n > 1
- lambda → 0 (or min-norm LS)
- p >> N
- SNR || beta ||/noise > 1
- Bias, Variance strictly decreasing with p/N to > 0 limit