TRUSTWORTHY MACHINE LEARNING THROUGH THE LENS OF HIGH-DIMENSIONAL PROBABILITY

Marco Mondelli, Institute of Science and Technology Austria (ISTA)

DARMSTADT SPRING SCHOOL IN PROBABILITY
"ARTIFICIAL INTELLIGENCE: PROBABILISTIC CONCEPTS"

OVERVIEW

→ Role of overparameterization

*) How many parameters to interpolate the labels?

*) How many parameters to interpolate robustly?

*) How many parameters to reconstruct training data?

→ Dynamics of differentially private algorithms

*) Privacy for free (when enough samples)

*) Optimal rate via aggressive clipping

BASIC SETTING AND TERMINOLOGY

Supervised learning

response / label $\in \mathbb{R}$

Training data : $\{(x_i, y_i)\}_{i \leq n} \overset{iid}{\sim} \mathbb{P}(\mathbb{R}^d \times \mathbb{R})$

vector of covariates in $\mathbb{R}^d$ — unknown

Goal : Find $f : \mathbb{R}^d \longrightarrow \mathbb{R}$ to predict y_{test} given x_{test}

with $(x_{test}, y_{test}) \sim \mathbb{P}$

test distribution = training distribution
(no distribution shift)

Measure performance via test error / population error / population risk :

$$R(f) = \mathbb{E}\left[\ell\left(y_{test}, f(x_{test})\right)\right]$$

loss function $\ell : \mathbb{R} \times \mathbb{R} \longrightarrow \mathbb{R}$

expectation over the distribution of (x_{test}, y_{test})

Running example : Square loss

$$R(f) = \mathbb{E}\left[| y_{test} - f(x_{test}) |^2 \right]$$

Parametric models (neural networks) : $f(x) = f(x; \theta) \quad \theta \in \mathbb{R}^p$

vector of parameters
(weights of the neural network)

Empirical risk minimization (ERM), standard approach to learn $f(\cdot\,;\theta)$

empirical

$$\hat{R}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \ell(y_i, f(x_i;\theta))$$

samples

$$\text{Solve} \quad \min_{\theta} \hat{R}_n(\theta)$$

We will solve the optimization problem via (several variants of) gradient descent:

*) Gradient descent (GD) (Full batch)

$$\theta^{k+1} = \theta^k - \eta_k \nabla_\theta \hat{R}_n(\theta^k) = \theta^k - \eta_k \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(y_i, f(x_i;\theta^k))$$

step size, can choose different step sizes for different groups of parameters

*) Gradient flow (GF)

$$\dot{\theta}_t = - \nabla_\theta \hat{R}_n(\theta_t)$$

When η_k is small, GD $\approx$ GF

*) Stochastic gradient descent (SGD) (Online, 1-pass)

$$\theta^{k+1} = \theta^k - \eta_k \nabla_\theta \ell(y_i, f(x_i;\theta^k)) \qquad i \sim \mathrm{Unif}(\{1,\dots,n\})$$

Later in the mini-course also:

*) Differentially-private gradient descent / stochastic gradient descent

HOW MANY PARAMETERS TO INTERPOLATE THE LABELS? (MEMORIZATION #1)

$$\exists \theta \text{ s.t. } \quad y_i = f(x_i, \theta) \quad \forall i \in \{1, \dots, n\}$$

Classical problem dating back to [Cover, 1965]: for a neural network with a single neuron and sign activation, label interpolation for data in generic position is possible if and only if $p/n > 1/2$.

*) Single neuron and sign activation: $f(x, \theta) = \text{sign}(\langle x, w \rangle)$
(spherical perceptron)

*) Data in generic position: Every subset of $x_1, \dots, x_n$ containing p or fewer vectors is linearly independent (satisfied by i.i.d. data with high probability).

More complex architectures? θ obtained by training via gradient descent?

<u>IDEA</u>: Under certain conditions, $f(x, \theta)$ is close to its linearization

$$f_{lin}(x, \theta) = f(x, \underbrace{\theta_0}_{\text{initialization}}) + \langle \theta - \theta_0, \nabla_\theta f(x, \theta_0) \rangle$$

throughout the training (GD/GF) dynamics.

$f_{lin}(x, \theta)$ is <u>linear</u> in θ so its GD training dynamics θ^k_{lin} can be solved explicitly.

Some more notation:

*) $y = \begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n$

$$*) \quad f_{lin}(\theta) = \begin{bmatrix} f_{lin}(x_1, \theta) \\ \vdots \\ f_{lin}(x_n, \theta) \end{bmatrix} \in \mathbb{R}^n$$

We now compute :

$f_{lin}(x_i, \theta)$ is linear in θ

$$\nabla_\theta f_{lin}(\theta) = \begin{bmatrix} \nabla_\theta f_{lin}(x_1, \theta) \\ \vdots \\ \nabla_\theta f_{lin}(x_1, \theta) \end{bmatrix} = \begin{bmatrix} \nabla_\theta f(x_1, \theta_0) \\ \vdots \\ \nabla_\theta f(x_n, \theta_0) \end{bmatrix} = \Phi \in \mathbb{R}^{n \times p}$$

does not depend on θ !

Then, we have

$$\theta_{lin}^{k+1} = \theta_{lin}^k - \eta_k \nabla_\theta \hat{R}_n(\theta_{lin}^k)$$

square loss, constant step size $\eta_k \equiv \eta$

$$= \theta_{lin}^k - \eta \nabla_\theta \frac{1}{2} \| y - f_{lin}(\theta_{lin}^k) \|^2$$

$$= \theta^k - \eta \Phi^T (f_{lin}(\theta_{lin}^k) - y)$$

$$= \theta^k - \eta \Phi^T (\Phi \theta_{lin}^k - y)$$

pick $\theta_0 = 0$ and f such that $f(x_i, \theta_0) = 0$ for all x_i.

Define $r^k = \Phi \theta_{lin}^k - y$. Then,

$$r^{k+1} = (I - \eta \Phi \Phi^T) r^k \implies \| r^{k+1} \|_2 \leq \| I - \eta \Phi \Phi^T \| \, \| r^k \|_2$$

$$\leq \left(1 - \frac{\lambda_{min}(\Phi \Phi^T)}{\| \Phi \|^2} \right) \| r^k \|_2$$

pick $\eta \leq \frac{1}{\| \Phi \|^2}$

Loss converges geometrically to 0 as long as $\lambda_{min}(\underbrace{\Phi \Phi^T}) > 0$.

Neural tangent kernel (NTK) $\in \mathbb{R}^{n \times n}$

NTK introduced by [Jacot, Gabriel, Hongler, 2018] and then lots of follow-up work [Du, Zhai, Poczos, Singh, 2019; Chizat, Oyallon, Bach, 2019; ...]

ONE FORMAL RESULT

Let $f(\vartheta) = \begin{bmatrix} f(x_1, \vartheta) \\ \vdots \\ f(x_n, \vartheta) \end{bmatrix} \in \mathbb{R}^n$, $J(\vartheta) \in \mathbb{R}^{n \times p}$ the corresponding Jacobian,

generic parametric model

$$(J(\vartheta))_{ij} = \frac{\partial f(x_i, \vartheta)}{\partial \vartheta_j}$$

and $\ell(y_i, f(x_i; \vartheta)) = \frac{1}{2}(y_i - f(x_i, \vartheta))^2$ the square loss.

Consider a set $\mathcal{D} \subset \mathbb{R}^p$ containing the initialization ϑ_0 such that

(A1) $\quad \alpha \leq \sigma_{min}(J(\vartheta)) \leq \| J(\vartheta) \| \leq \beta \qquad \forall \vartheta \in \mathcal{D}$

Bounded spectrum of the Jacobian

(A2) $\quad \| J(\vartheta_1) - J(\vartheta_2) \| \leq \dfrac{\alpha^2}{2\beta} \qquad \forall \vartheta_1, \vartheta_2 \in \mathcal{D}$

Bounded deviations of the Jacobian

<u>THEOREM</u> [Oymak, Soltanolkotabi, 2019] Assume that (A1) and (A2) hold when $\mathcal{D}$ is a ball centered at ϑ_0 with radius $R = 4 \| f(\vartheta_0) - y \|_2 / \alpha$. Run GD with constant step size $\eta \leq \frac{1}{2\beta^2}$ and initialization ϑ_0. Then,

$$\| f(\vartheta^k) - y \|_2^2 \leq \left(1 - \frac{\eta \alpha^2}{2}\right)^k \| f(\vartheta_0) - y \|_2^2 .$$

<u>INTERPRETATION</u> Loss converges geometrically to 0 as long as

$$\lambda_{min}(\Phi \Phi^T) = \alpha^2 > 0 .$$

Much more is true :

*) $\| \theta^k - \theta^k_{lin} \|_2 \ll \| \theta^k - \theta_0 \|_2$ for all k until convergence

GD dynamics close to GD dynamics of the linearization

*) $\| f(\cdot, \theta^k) - f_{lin}(\cdot, \theta^k_{lin}) \|_{L^2} \ll \| f(\cdot, \theta^k) \|_{L^2}$

with $\| g \|_{L^2} = (\mathbb{E}\, g^2(x))^{1/2}$ where $\mathbb{E}$ is taken on a test point.

Model learned by GD close to model learned by the linearized flow on test points

*) Total gradient path is bounded :

$$\sum_{k=0}^{\infty} \| \theta^{k+1} - \theta^k \|_2 \leq \frac{4 \| f(\theta_0) - y \|_2}{\alpha}$$

Gradient descent iterates remain close to initialization (they never leave a neighborhood of radius $\frac{4}{\alpha} \| f(\theta_0) - y \|_2$ around initialization)

*) Gradient descent follows a short path :

$$\| \theta^k - \theta_0 \| \leq 4 \frac{\beta}{\alpha} \| \theta^* - \theta_0 \|$$

$$\sum_{k=0}^{\infty} \| \theta^{k-1} - \theta^k \| \leq 4 \frac{\beta}{\alpha} \| \theta^* - \theta_0 \| ,$$

with θ^* a global optimum of the loss CLOSEST (in ℓ_2) to initialization.

GD follows an almost direct route from initialization to a global optimum (the length of the GD path is within a factor of the distance between initialization and CLOSEST global optimum)

⟶ Also known as "lazy" training [Chizat, Oyallon, Bach, 2019]

⟶ This "linear regime" is discussed in detail in the review [Bartlett, Montanari, Rakhlin, 2021], see Section 5 therein.

BOUNDING THE SMALLEST EIGENVALUE OF THE NTK

Key quantity in this analysis is $\alpha^2 = \lambda_{min}(\Phi \Phi^T)$

Here, Φ is the Jacobian at initialization which is a random quantity (weights at initialization are random, data is also random).

If $p < n$, then $\lambda_{min}(\Phi \Phi^T) = 0$ ..

How large does p have to be so that $\lambda_{min}(\Phi\Phi^T) > 0$?

→ Two-layer networks

$$f(x,\theta) = \underbrace{a^T}_{\in \mathbb{R}^N} \underbrace{\phi}_{\text{activation function applied component-wise}}(\underbrace{W}_{\in \mathbb{R}^{N\times d}} \overbrace{x}^{\in \mathbb{R}^d}) = \sum_{i=1}^N a_i \phi(\langle w_i, x\rangle)$$

d = input dimension, N = # neurons, $p = Nd$

$p \gg n \implies \lambda_{min}(\Phi\Phi^T) > 0$ [Montanari, Zhong, 2022]

(improvement upon earlier work by [Soltanolkotabi, Javanmard, Lee, 2018])

This is optimal (up to poly-log factors)!

→ Deep networks

$$f(x,\theta) := f_L(x,\theta) \quad \text{with} \quad f_\ell(x,\theta) = \begin{cases} x & \ell = 0 \\ \phi(W_\ell^T f_{\ell-1}) & \ell \in \{1,\dots,L-1\} \\ W_L^T f_{L-1} & \ell = L \end{cases}$$

$W_\ell \in \mathbb{R}^{N_{\ell-1}\times N_\ell}$, $N_0 = d$, $N_L = 1$, N_ℓ = # neurons at layer ℓ ($\ell \in \{1,\dots,L-1\}$)

A direct calculation gives

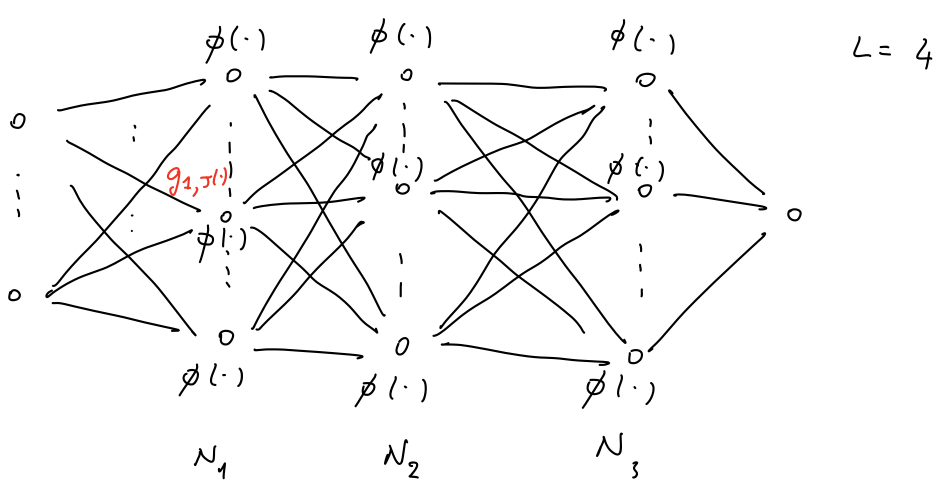
$$\Phi \Phi^T = \sum_{\ell=0}^{L-1} F_\ell F_\ell^T \underbrace{\circ}_{\text{Hadamard (component-wise) product}} B_{\ell+1} B_{\ell+1}^T$$

*) $F_\ell = \begin{bmatrix} f_\ell(x_1) \\ \vdots \\ f_\ell(x_n) \end{bmatrix} \in \mathbb{R}^{n \times N_\ell}$ feature matrix of layer ℓ

*) $B_\ell \in \mathbb{R}^{n \times N_\ell}$ backpropagation matrix

$$(B_\ell)_{i:} = \begin{cases} \Sigma_\ell(x_i) \left(\prod_{\ell'=\ell+1}^{L-1} W_{\ell'} \Sigma_{\ell'}(x_i) \right) W_L , & \ell \in \{1, .., L-2\} \\ \Sigma_{L-1}(x_i) W_L , & \ell = L-1 \\ 1 , & \ell = L \end{cases}$$

$$\Sigma_\ell(x) = \mathrm{diag}\left(\left[\phi'(\underbrace{g_{\ell,j}(x)}_{\text{pre-activation neuron}}) \right]_{j=1}^{N_\ell} \right)$$



Weight assumption : $(W_\ell)_{IJ} \overset{iid}{\sim} \mathcal{N}(0, 1/n_{\ell-1})$, $(W_L)_i \overset{iid}{\sim} \mathcal{N}(0,1)$

standard in practice (He, LeCun initialization)

Data assumptions :

$$\left.\begin{aligned} &(a)\ \int \|x\|_2 \, dP_x(x) = \Theta(\sqrt{d}) \\ &(b)\ \int \|x\|_2^2 \, dP_x(x) = \Theta(d) \\ &(c)\ \int \Big\| x - \int x' \, dP_x(x') \Big\|_2^2 \, dP_x(x) = \Omega(d) \end{aligned}\right\}\ \text{scaling}$$

$$(d)\ \mathbb{P}\left(\left| \varphi(x) - \int \varphi(x') \, dP_x(x') \right| > t \right) \le 2e^{-t^2/2c \cdot \mathrm{Lip}(\varphi)^2}, \quad \forall \varphi \text{ Lipschitz}$$

dimension - independent constant

(C-)Lipschitz concentration holds in a variety of settings :

*) Standard Gaussian

*) Uniform distribution on the sphere and the hypercube

*) Data produced by a fully connected neural network with well-conditioned weights and Lipschitz activations

*) Distributions satisfying log- Sobolev (with dimension - independent constant)

EXTRA :

A probability measure μ satisfies the log- Sobolev inequality with constant $C > 0$ if for any smooth function f

$$\mathrm{Ent}_\mu(f^2) \le C \int |\nabla f(x)|^2 \, d\mu(x) ,$$

with $\mathrm{Ent}_\mu(f) = \int f \ln f \, d\mu - \int f \ln\left(\int f \, d\mu\right) d\mu$

We now give a lower bound on $\lambda_{min}(\Phi \Phi^T)$ in two settings:

1. Single wide layer $\quad N_K \gg n \quad$ for some $K \in \{1, .., L-1\}$

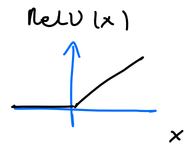
2. Minimum overparameterization (smallest overall number of neurons)

   $N_{L-2} N_{L-1} \gg n$

(1) Single wide layer $N_k \gg n$ for some $k \in \{1, \dots, L-1\}$

\+ no exponential bottleneck before the wide layer

$$\left(\prod_{l=1}^{k-2} \log N_l \ll \min_{l \in \{0, \dots, k-1\}} N_l \right)$$



THEOREM (Simplification of [Nguyen, M., Montufar, 2021]). Let $\phi(\cdot) = \mathrm{ReLU}(\cdot)$.

Assume $N_k = \tilde{\Omega}(n)$ and $N_l = \tilde{\Theta}(1)$ for $l \neq k$, where $\sim$ omits polylog factors. Then, with high probability,

$$\lambda_{min}(\Phi \Phi^T) = \Theta(N_k).$$

INTERPRETATION One layer whose width is linear in the number of samples suffices to guarantee well-behavedness of the NTK.

PROOF SKETCH.

STEP 1: From NTK to feature matrices.

For PSD matrices P, Q, it holds

$$\lambda_{min}(P \circ Q) \geq \lambda_{min}(P) \min_i Q_{ii}$$

$$\Downarrow$$

$$\lambda_{min}(\Phi \Phi^T) \geq \sum_{l=0}^{L-1} \lambda_{min}(F_l F_l^T) \min_{i \in \{1, \dots, n\}} \|(B_{l+1})_{i:}\|_2^2$$

$$\geq \lambda_{min}(F_k F_k^T) \min_{i \in \{1, \dots, n\}} \|(B_{k+1})_{i:}\|_2^2$$

$$\|(B_{k+1})_{i:}\|_2^2 = \| \underbrace{\Sigma_{k+1}(x_i) \left(\prod_{l=k+2}^{L-1} W_l \Sigma_l(x_i) \right)}_{B} W_L \|_2^2$$

$$\approx \mathbb{E}_{W_L} \|B W_L\|_2^2 = \|B\|_F^2 = \Theta(1)$$

Hanson-Wright $(A = B^2)$ induction + Bernstein

STEP 2: Concentration of smallest eigenvalue

truncation + matrix Chernoff

$$\lambda_{min}(F_k F_k^T) \geq \frac{1}{4}\lambda_{min}\left(\mathbb{E}_{W_k} F_k F_k^T\right)$$

$$= \frac{N_k}{4}\lambda_{min}\left(\mathbb{E}_{w\sim N(0,\frac{1}{N_{k-1}}I)} \phi(F_{k-1}w)\,\phi(F_{k-1}w)^T\right)$$

$\phi(\cdot)$ is homogeneous

$$= \frac{N_k}{4N_{k-1}}\lambda_{min}\left(D\, \mathbb{E}_{\hat w\sim N(0,I)}\left[\phi(\hat F_{k-1}w)\,\phi(\hat F_{k-1}w)^T\right] D\right)$$

$D = diag(\|(F_k)_{1:}\|_2, \dots, \|(F_k)_{n:}\|_2)$

ϕ_r = r-th Hermite coefficient of ϕ

$$= \frac{N_k}{4N_{k-1}}\lambda_{min}\left(D\left(\phi_0^2\, 1_n 1_n^T + \sum_{s=1}^{\infty} \phi_s^2 (\hat F_{k-1}\hat F_{k-1}^T)^{\circ s}\right)D\right)$$

$$\geq \frac{N_k}{4N_{k-1}}\phi_r^2\, \lambda_{min}\left(D(\hat F_{k-1}\hat F_{k-1}^T)^{\circ r}D\right)$$

$$= \frac{N_k}{4N_{k-1}}\phi_r^2\,\lambda_{min}\left(D^{-(r-1)}(F_{k-1}F_{k-1}^T)^{\circ r}D^{-(r-1)}\right)$$

$$\geq \underbrace{\frac{N_k}{4N_{k-1}}}_{\text{polylog}(n)}\phi_r^2\, \frac{\lambda_{min}\left((F_{k-1}F_{k-1}^T)^{\circ r}\right)}{\underbrace{\max_{i\in\{1,\dots,n\}}\|(F_{k-1})_{i:}\|_2^{2(r-1)}}_{\text{polylog}(n)}}$$

$$= \tilde\Theta\left(N_k\, \lambda_{min}\left((F_{k-1}F_{k-1}^T)^{\circ r}\right)\right)$$

STEP 3: Concentration of Gram matrix of features

Let $\tilde{F}_{k-1} = F_{k-1} - \mathbb{E}_{(x_1,\dots,x_n)} F_{k-1}$, $\mu = \mathbb{E}_{x} f_{k-1}(x) \in \mathbb{R}^{N_{k-1}}$ and $\Lambda = \mathrm{diag}\left(F_{k-1}\mu - \|\mu\|^2 1_n\right)$.

Then,

$$F_{k-1}F_{k-1}^T = \tilde{F}_{k-1}\tilde{F}_{k-1}^T + \|\mu\|^2 1_n 1_n + \Lambda 1_n 1_n^T + 1_n 1_n^T \Lambda$$

$$= \tilde{F}_{k-1}\tilde{F}_{k-1}^T + \underbrace{\left(\|\mu\| 1_n + \frac{\Lambda 1_n}{\|\mu\|}\right)\left(\|\mu\| 1_n + \frac{\Lambda 1_n}{\|\mu\|}\right)^T}_{\succeq 0} - \frac{\Lambda 1_n 1_n^T \Lambda}{\|\mu\|^2}$$

$$\succeq \tilde{F}_{k-1}\tilde{F}_{k-1}^T - \frac{\Lambda 1_n 1_n^T \Lambda}{\|\mu\|^2}$$

$$\lambda_{min}\left(\left(F_{k-1}F_{k-1}^T\right)^{\circ r}\right) \geq \lambda_{min}\left(\left(\tilde{F}_{k-1}\tilde{F}_{k-1}^T - \frac{\Lambda 1_n 1_n^T \Lambda}{\|\mu\|^2}\right)^{\circ r}\right)$$

$$= \lambda_{min}\left(\left(\overbrace{\tilde{F}_{k-1}\tilde{F}_{k-1}^T}^{\approx I}\right)^{\circ r}\right)\left(1 + o(1)\right) = \Theta(1).$$

$$\left(\left(\tilde{F}_{k-1}\tilde{F}_{k-1}^T\right)^{\circ r}\right)_{ij} = \langle \underbrace{\tilde{f}_{k-1}(x_i)}_{=\, f_{k-1}(x_i) - \mathbb{E}_{x_i} f_{k-1}(x_i)}, \tilde{f}_{k-1}(x_j)\rangle^r \approx \delta_{ij}$$

by picking large enough r

centered feature vector at layer $k-1$ computed with input x_i

$f_{k-1}(\cdot)$ is Lipschitz, so by Assumption (d) on the data,

$$\langle \tilde{f}_{k-1}(x_i), \tilde{f}_{k-1}(x_j)\rangle \ll \|\tilde{f}_{k-1}(x_i)\| \, \|\tilde{f}_{k-1}(x_j)\| \quad \forall i \neq j$$

This gives the lower bound $\lambda_{min}(\Phi\Phi^T) = \Omega(N_k)$.

The upper bound is easy :

$$\lambda_{min}(\Phi\Phi^T) \leq (\Phi\Phi^T)_{1,1} = \sum_{\ell=0}^{L-1} \|(F_\ell)_{1:}\|^2 \overbrace{\|(B_{\ell+1})_{1:}\|^2}^{\Theta(1)}$$

$$= \Theta(N_K) .$$

$\boxtimes$

[Extra :

Hanson - Wright inequality.

THEOREM Let $x \in \mathbb{R}^n$ be a random vector with independent, mean zero, sub-Gaussian coordinates (with sub-Gaussian norm of constant order). Let $A \in \mathbb{R}^{n\times n}$. Then,

$$\mathbb{P}(|x^T A x - \mathbb{E}\, x^T A x| > t) \leq 2 e^{-c \min\left(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|}\right)}$$

Matrix Chernoff.

THEOREM. Let $A \in \mathbb{R}^{n\times m}$ and assume $\|(A_{:j})(A_{:j})^T\| \leq t^2$. Then,

$$\mathbb{P}(\lambda_{min}(AA^T) \leq (1-\epsilon)\lambda_{min}(\mathbb{E} AA^T)) \leq n\left[\frac{e^{-\epsilon}}{(1-\epsilon)^{1-\epsilon}}\right]^{\frac{\lambda_{min}(\mathbb{E} AA^T)}{t^2}}$$

Hermite expansion.

THEOREM Let $\phi : \mathbb{R} \to \mathbb{R}$ with Hermite coefficients $\{\phi_r\}_{r\in\mathbb{N}}$, let $u, v \in \mathbb{R}^d$ such that $\|u\|_2 = \|v\|_2 = 1$. Then,

$$\mathbb{E}_{w\sim N(0,I_d)} \phi(\langle w, u\rangle)\phi(\langle w, v\rangle) = \sum_{r\in\mathbb{N}} \phi_r^2 \langle u, v\rangle^r .$$

]

② Minimum overparameterization (smallest number of neurons) $N_{L-2} N_{L-1} \gg n$

THEOREM [Bombari, Amani, M., 2022] Let ϕ be non-linear, Lipschitz, with Lipschitz derivative. Assume $N_\ell = O(N_{\ell-1})$ for $\ell \in \{1, \cdots, L-1\}$ and $n \operatorname{polylog}(n) = o(N_{L-2} N_{L-1})$. Then, with high probability,

$$\lambda_{min}(\Phi \Phi^T) = \Omega(N_{L-2} N_{L-1}) \quad \text{and} \quad \lambda_{min}(\Phi \Phi^T) = O(d N_{L-1})$$

INTERPRETATION $\tilde{\Omega}(\sqrt{n})$ neurons and thus $\tilde{\Omega}(n)$ parameters are enough to interpolate n data points. This is in agreement with back-of-the-envelope calculations on CIFAR-10 and ImageNet :

*) CIFAR-10 : $n = 50000$ images and 10^6 parameters suffice to fit random labels

*) ImageNet : $n = 1.2 \cdot 10^6$ images and $2.4 \, 10^7$ parameters suffice to fit random labels

PROOF SKETCH. Upper bound as before. Let's focus on the lower bound.

If $N_\ell < n$, then $\lambda_{min}(F_\ell F_\ell^T) = 0$ and step 1 of the previous approach fails.

$$\Phi \Phi^T = \sum_{\ell=0}^{L-1} F_\ell F_\ell^T \circ B_{\ell+1} B_{\ell+1}^T \succcurlyeq \underbrace{F_{L-2} F_{L-2}^T \circ B_{L-1} B_{L-1}^T}_{\text{need to consider the Hadamard product together}} := \Phi_{L-2} \Phi_{L-2}^T$$

Let $A, B \in \mathbb{R}^{n \times \sqrt{n}}$. Then, $AA^T, BB^T \in \mathbb{R}^{n\times n}$ have rank $\sqrt{n}$,

but $AA^T \circ BB^T \in \mathbb{R}^{n \times n}$ can have rank n.

$$(\Phi_{L-2})_{i:} = \underbrace{f_{L-2}(x_i) \otimes \text{diag}(W_L)\,\sigma'(W_{L-1}^T f_{L-2}(x_i))}_{}$$

Kronecker (tensor) product

rows are iid w.r.t. the randomness in $(x_1, \dots, x_n)$.

THEOREM (Simplification of [Adamczak, Litvak, Pajor, Tomczak-Jaegermann, 2011]).

Let Φ be a matrix with i.i.d. rows having ℓ_2 norm ℓ and sub-exponential norm ψ. Then, with high probability,

$$\lambda_{min}(\Phi \Phi^T) \geq \ell^2 - \tilde{O}(\psi^2 \ell \sqrt{n})$$

STEP 1 : Centering

*) Necessary otherwise the sub-exponential norm is too large (dominated by the mean).

*) Need to center F_{L-2} and B_{L-1} together, otherwise cross-terms become too large.

$$\lambda_{min}(F_{L-2} F_{L-2}^T \circ B_{L-1} B_{L-1}^T) \geq \lambda_{min}(\overbrace{\tilde{F}_{L-2} \tilde{F}_{L-2}^T \circ \tilde{B}_{L-1} \tilde{B}_{L-1}^T}^{\tilde{\Phi}_{L-2}^{(1)} \tilde{\Phi}_{L-2}^{(1)T}}) - o(N_{L-2} N_{L-1})$$

$$\tilde{F}_{L-2} = F_{L-2} - \mathbb{E}_{x_1, \dots, x_n} F_{L-2}, \quad \tilde{B}_{L-1} = B_{L-1} - \mathbb{E}_{x_1, \dots, x_n} B_{L-1}$$

$$\left(\tilde{\Phi}^{(1)}_{L-2}\right)_{i:} = \underbrace{\tilde{f}_{L-2}(x_i)}_{= \, f_{L-2}(x_i) - \mathbb{E}_{x_i} f_{L-2}(x_i)} \otimes \operatorname{diag}(W_L)\, \underbrace{\tilde{\sigma}'\left(W_{L-1}^T f_{L-2}(x_i)\right)}_{= \, \sigma'(W_{L-1}^T f_{L-2}(x_i)) - \mathbb{E}_{x_i} \sigma'(W_{L-1}^T f_{L-2}(x_i))}$$

this is still not 0-mean w.r.t. x_i because of the quadratic terms

$$x \otimes x = [\, x_1^2 \;\; x_1 x_2 \;\; x_1 x_3 \;\cdots\; x_1 x_n \;\; x_2 x_1 \;\; x_2^2 \;\cdots\; x_n^2 \,]$$

not 0-mean

$$\left(\tilde{\Phi}_{L-2}\right)_{i:} = \left(\tilde{\Phi}^{(1)}_{L-2}\right)_{i:} - \mathbb{E}_{x_i}\left(\tilde{\Phi}^{(1)}_{L-2}\right)_{i:}$$

$$\tilde{F}_{L-2}\tilde{F}_{L-2}^T \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^T = \tilde{\Phi}_{L-2}\tilde{\Phi}_{L-2}^T + \underbrace{\text{rank-}1}_{} + \text{PSD}$$

operator norm bounded via (generalized) Hanson-Wright

$$\Downarrow$$

$$\lambda_{\min}\left(\tilde{F}_{L-2}\tilde{F}_{L-2}^T \circ \tilde{B}_{L-1}\tilde{B}_{L-1}^T\right) \geq \lambda_{\min}\left(\tilde{\Phi}_{L-2}\tilde{\Phi}_{L-2}^T\right) + o(N_{L-1}N_{L-2})$$

STEP 2 : Control norms

$$\ell = \left\| \left(\tilde{\Phi}_{L-2}\right)_{i:} \right\| = \Theta\left(\sqrt{N_{L-1}N_{L-2}}\right)$$

$$\psi = \left\| \left(\tilde{\Phi}_{L-2}\right)_{i:} \right\|_{\psi_1} = \tilde{O}(1)$$

sub-exponential norm

Applying the result by Adamczak et al. gives

$$\lambda_{\min}\left(\tilde{\Phi}_{L-2}\tilde{\Phi}_{L-2}^T\right) \geq N_{L-1}N_{L-2} - \tilde{O}\left(\sqrt{n N_{L-1} N_{L-2}}\right) = \Omega(N_{L-1}N_{L-2})$$

$\boxtimes$

HOW MANY PARAMETERS TO INTERPOLATE ROBUSTLY ?

Machine learning models are vulnerable to adversarial perturbations, and this has been known for over 10 years now [Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, Fergus, 2014].

*) Prototypical examples: given the image of e.g. a cat, one can construct a perturbation which is imperceptible to the human eye but makes the neural network classify the image as e.g. a gibbon.

Immense literature on the topic, including "adversarial training" methods aimed at reducing this effect [Madry, Makelov, Schmidt, Tsipras, Vladu, 2018].

Our focus is on the ROLE OF OVERPARAMETERIZATION:

*) How many parameters NECESSARY to have robustness ?

*) How many parameters SUFFICIENT to have robustness ?

A Necessary Condition: Universal Law of Robustness

Theorem ([Bubeck, Sellke, 2021]). Let $\mathcal{F}$ be a class of functions from $\mathbb{R}^d \to \mathbb{R}$ let $\{(x_i, y_i)\}_{i=1}^n$ be input-output pairs in $\mathbb{R}^d \times [-1,1]$ and fix $\epsilon, \delta \in (0,1)$.
Assume that

1. $\mathcal{F} = \{ f_w, w \in \mathcal{W}, \mathrm{Lip}(f_w) \leq L \}$ with $\mathcal{W} \subseteq \mathbb{R}^p$, $\mathrm{diam}(\mathcal{W}) \leq W$ and for any $w_1, w_2 \in \mathcal{W}$
$$\| f_{w_1} - f_{w_2} \| \leq J \| w_1 - w_2 \|$$

2. The distribution μ of the covariates x_i can be written as $\mu = \sum_{\ell=1}^{k} \alpha_\ell \mu_\ell$, where $\alpha_\ell \geq 0$, $\sum_{\ell=1}^{k} \alpha_\ell = 1$, $k \log(8k/\delta) \leq c \cdot n \epsilon^2$ and each μ_ℓ satisfies Lipschitz concentration:
$$\mathbb{P}_{x \sim \mu_\ell} \left[| f(x) - \mathbb{E}[f] | \geq t \right] \leq 2 e^{-\frac{c d t^2}{\mathrm{Lip}(f)^2}}.$$

3. The expected conditional variance of the output is strictly positive, i.e.
$$\sigma^2 \equiv \mathbb{E}_\mu \mathrm{Var}[y_i | x_i] > 0$$

4. The dimension d is large compared to ϵ, i.e., $d \geq C_1 \left(\frac{c L^2 \sigma^2}{\epsilon^2} \right)$.

Then, with probability at least $1-\delta$ with respect to the sampling of the data, one has simultaneously for all $f \in \mathcal{F}$:
$$\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - y_i)^2 \leq \sigma^2 - \epsilon$$
$$\Downarrow$$
$$\mathrm{Lip}(f) \geq \frac{\epsilon}{\sigma} \sqrt{c} \sqrt{\frac{nd}{p \log(1 + W J \epsilon^{-1} c^{-1}) + \log(4/\delta)}}.$$

INTERPRETATION.

adversarial perturbation which is "imperceptible"

$$\underbrace{x_{adv}}_{\text{adversarial example close to } x} = x + \Delta_{adv}, \qquad \|\Delta_{adv}\| \leq \delta \cdot \|x\|$$

$$\begin{aligned}
|f(x_{adv},\vartheta) - f(x,\vartheta)| &\simeq |\nabla_x f(x,\vartheta)^T \Delta_{adv}| \\
&\leq \|\Delta_{adv}\| \; \|\nabla_x f(x,\vartheta)\| \\
&\leq \delta \cdot \|x\| \cdot \|\nabla_x f(x,\vartheta)\| \\
&\leq c\,\delta\, \mathrm{Lip}(f)
\end{aligned}$$

*) under Assumption 2, $\|x\| = \Theta(1)$

*) $\sup_x \|\nabla_x f(x,\vartheta)\| = \mathrm{Lip}(f)$

Let us take Δ_{adv} aligned with the direction of the gradient (w.r.t. x) having the largest norm. Then, if $\mathrm{Lip}(f) \gg 1$, $f(x_{adv})$ is far away from $f(x)$.

RESTATING THE RESULT. Consider a family of functions $\mathcal{F}$ that admits a Lipschitz parametrization by p parameters each of size poly(n,d) (Assumption 1 with $J = \text{poly}(n,d)$). Assume that the distribution of the data features is a mixture of (not too many) Lipschitz-concentrated distributions (Assumption 2). Then, if the network fits below the noise level $\sigma^2 > 0$ (e.g., it interpolates the training data), then, with high probability, $p = \Omega(nd)$ is NECESSARY to have a robust predictor, namely, such that $f(x_{adv})$ has to be close to $f(x)$ for any choice of Δ_{adv}.

IDEA OF THE PROOF. Consider an f that interpolates random ± 1 labels

*) Lipschitz concentration implies that either 0-level set or 1-level set of f have probability $\leq e^{-\frac{d}{Lip(f)^2}}$

*) Probability of fitting all n points is $\leq e^{-\frac{nd}{Lip(f)^2}}$

*) Union bound over a function class of size N gives probability of

$$N e^{-\frac{nd}{Lip(f)^2}} = e^{\log N - \frac{nd}{Lip(f)^2}}$$

*) Discretization argument gives that $\log N = O(p)$ for smoothly parameterized family of functions with p parameters

Need $p - \frac{nd}{Lip(f)^2} = \Omega(1)$ (otherwise probability of finding a fitting function is very small)

$$\Downarrow$$

$$Lip(f) = \Omega\left(\sqrt{\frac{nd}{p}}\right).$$

<u>PROOF</u> For simplicity pick $k=1$ and assume σ, ϵ are constants >0 (independent of n, d, p) so that we don't need to track their dependency in the calculations. We use c to denote a constant >0 independent of n, d, p whose value may change from passage to passage.

STEP 1 : If f fits below noise level, then the predictions are correlated with noise.

Let $\underbrace{g(x)}_{\text{target function}} = \mathbb{E}[y|x]$ and $\underbrace{z_i}_{\text{noise part of the observed label } y_i} = y_i - g(x_i)$. Then,

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n}(y_i - f(x_i))^2 \leq \sigma^2 - \epsilon\right) \tag{1}$$

$$\leq 2e^{-cn\epsilon^2} + \mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^{n} f(x_i) z_i \geqslant \epsilon/4\right)$$

We now prove (1).

*) (z_i^2) are iid. with mean σ^2 $(\sigma^2 = \mathrm{Var}[y|x])$ and $|z_i|^2 \leq 4$ $(y_i \in [-1,1])$. Thus, by Hoeffding's inequality,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} z_i^2 \leq \sigma^2 - \frac{\epsilon}{6}\right) \leq e^{-cn\epsilon^2}$$

*) $(z_i\, g(x_i))$ are iid with mean 0 $(\mathbb{E}[z_i | x_i] = 0)$ and $|z_i\, g(x_i)| \leq 2$ $(y_i \in [-1,1])$. Thus, by Hoeffding's inequality,

$$\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} z_i\, g(x_i) \leq -\frac{\epsilon}{6}\right) \leq e^{-cn\epsilon^2}$$

*) Define $z = \frac{1}{\sqrt{n}}(z_1, \dots, z_n)$, $G = \frac{1}{\sqrt{n}}(g(x_1), \dots, g(x_n))$,

$F = \frac{1}{\sqrt{n}}(f(x_1), \dots, f(x_n))$.

We can rewrite (1) as

$$\mathbb{P}(\exists f \in \mathcal{F} : \|G+z-F\|^2 \le \sigma^2 - \epsilon) \le 2e^{-cn\epsilon^2} + \mathbb{P}(\exists f\in\mathcal{F} : \langle F, z\rangle \ge \frac{\epsilon}{4})$$

Note that we have just proved that, with probability at least $1-2e^{-cn\epsilon^2}$,

$$\|z\|^2 \ge \sigma^2 - \frac{\epsilon}{6} \quad \text{and} \quad \langle z, G\rangle \ge -\frac{\epsilon}{6}.$$

Thus,

$$\sigma^2-\epsilon \ge \|G+z-F\|^2 = \|z\|^2 + 2\langle z, G-F\rangle + \|G-F\|^2$$

$$= \underbrace{\|z\|^2}_{\ge \sigma^2-\epsilon/6} + 2\underbrace{\langle z,G\rangle}_{\ge -\epsilon/6} - 2\langle z,F\rangle + \underbrace{\|G-F\|^2}_{\ge 0} \ge \sigma^2 - \frac{\epsilon}{2} - 2\langle z,F\rangle$$

$$\Downarrow$$

$\langle F, z\rangle \ge \epsilon/4$, which gives (1).

STEP 2: Chance that f fits below noise level is $e^{\log|\mathcal{F}| - c\frac{nd}{L^2}}$

All functions in $\mathcal{F}$ are L-Lipschitz

$$\mathbb{P}\left(\exists f\in\mathcal{F}: \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \le \sigma^2-\epsilon\right) \le 4e^{-cn\epsilon^2} + 2e^{\log|\mathcal{F}| - c\frac{\epsilon^2 nd}{L^2}} \quad (2)$$

We now prove (2).

*) f L-Lipschitz + distribution of x_i satisfying $\mathbb{P}(|f(x_i) - \mathbb{E}[f]| > t) \le 2e^{-\frac{cdt^2}{L^2}}$

$$\Downarrow$$

$\frac{f(x_i)-\mathbb{E}[f]}{L}\sqrt{d}$ is c-subGaussian

$$\Downarrow$$

$\frac{f(x_i)-\mathbb{E}[f]}{L}\sqrt{d}\, z_i$ is c-subGaussian $(|z_i|\le 2)$

+ it is 0 mean $\left(\mathbb{E}\left[(f(x_i) - \mathbb{E}[f]) z_i \right] = 0 \quad \text{as} \quad \mathbb{E}[z_i \mid x_i] = 0 \right)$

$$\Downarrow$$

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{f(x_i) - \mathbb{E}[f]}{L} \sqrt{d}\, z_i \quad \text{is} \quad c\text{- subGaussian}$$

$$\Downarrow$$

$$\mathbb{P}\left(\sqrt{\frac{d}{nL^2}} \sum_{i=1}^{n} (f(x_i) - \mathbb{E}[f]) z_i \geq t \right) \leq 2 e^{-ct^2}$$

$$\Downarrow$$

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - \mathbb{E}[f]) z_i \geq \frac{\epsilon}{8} \right) \leq 2 e^{-c \frac{\epsilon^2 n d}{L^2}}$$

*) $\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}[f] z_i \geq \frac{\epsilon}{8} \right) \leq \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} z_i \right| \geq \frac{\epsilon}{8} \right)$ $(\mathbb{E}[f] \in [-1,1])$

$\leq 2 e^{-cn\epsilon^2}$ (Hoeffding since $|z_i| \leq 2$)

Putting everything together, we have

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n} \sum_{i=1}^{n} (y_i - f(x_i))^2 \leq \sigma^2 - \epsilon \right)$$

$$\underbrace{\leq}_{(1)} 2 e^{-cn\epsilon^2} + \mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n} \sum_{i=1}^{n} f(x_i) z_i \geq \epsilon/4 \right)$$

$$\leq 2 e^{-cn\epsilon^2} + \underbrace{|\mathcal{F}|}_{\text{union bound}} \mathbb{P}\left(\frac{1}{n} \sum_{i=1}^{n} (f(x_i) - \mathbb{E}[f]) z_i \geq \frac{\epsilon}{8} \right) + \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} z_i \right| \geq \frac{\epsilon}{8} \right)$$

$\underbrace{\leq}_{\text{calculation above}} 2|\mathcal{F}|\, e^{-c \frac{\epsilon^2 n d}{L^2}} + 4 e^{-cn\epsilon^2}$, which gives (2).

STEP 3 : Covering argument (ϵ - net)

Let $\mathcal{W}_\epsilon$ be an $\frac{\epsilon}{16J}$ - net of $\mathcal{W}$. Then,

$$|\mathcal{W}_\epsilon| \le \left(1 + \frac{JW}{c\epsilon}\right)^p .$$

Apply (2) to $\mathcal{F}_\epsilon = \{ f_w, \ w \in \mathcal{W}_\epsilon \}$:

$$\mathbb{P}\left(\exists f \in \mathcal{F}_\epsilon : \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \le \sigma^2 - \epsilon/2\right) \le 4 e^{-cn\epsilon^2} + 2 e^{p\log(1+JW\epsilon^{-1}c^{-1}) - c\frac{\epsilon^2 nd}{L^2}}$$

If $\|f-h\|_\infty \le \frac{\epsilon}{16} \le 1$, $\|f\|_\infty \le 1$, $\|h\|_\infty \le 1$ and $|y_i| \le 1$ for all i, then

$$\frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \le \frac{1}{n}\sum (y_i - h(x_i))^2 + \frac{\epsilon}{2}$$

$$(y_i - f(x_i))^2 = (y_i - h(x_i) + (h(x_i) - f(x_i)))^2$$

$$= (y_i - h(x_i))^2 + (h(x_i) - f(x_i))^2 + 2(y_i - h(x_i))(h(x_i) - f(x_i))$$

$$\le (y_i - h(x_i))^2 + \left(\frac{\epsilon}{16}\right)^2 + 2\cdot 2\,\frac{\epsilon}{16} \le (y_i - h(x_i))^2 + \frac{\epsilon}{2}$$

For any $w \in \mathcal{W}$, $\exists\, w' \in \mathcal{W}_\epsilon$ s.t. $\|w - w'\| \le \frac{\epsilon}{16J}$

$$\Downarrow$$

For any $f_w \in \mathcal{F}$, $\exists\, f_{w'} \in \mathcal{F}_\epsilon$ s.t. $\|f_w - f_{w'}\|_\infty \le \frac{\epsilon}{16}$

$$\Downarrow$$

$$\frac{1}{n}\sum_{i=1}^n (y_i - f_w(x_i))^2 \le \frac{1}{n}\sum (y_i - f_{w'}(x_i))^2 + \frac{\epsilon}{2}$$

$$\Downarrow$$

$$\mathbb{P}\left(\exists f \in \mathcal{F} : \frac{1}{n}\sum_{i=1}^n (y_i - f(x_i))^2 \le \sigma^2 - \epsilon\right) \le 4 e^{-cn\epsilon^2} + 2 e^{p\log(1+JW\epsilon^{-1}c^{-1}) - c\frac{\epsilon^2 nd}{L^2}}$$

$$4 e^{-cn\epsilon^2} \leq \frac{\delta}{2} \qquad \left(\log(8/\delta) \leq c \cdot n \epsilon^2 \right)$$

$$2 e^{p \log(1 + JW\epsilon^{-1} c^{-1}) - c \frac{\epsilon^2 nd}{L^2}} = \frac{\delta}{2} \implies L \geq \epsilon \sqrt{c} \sqrt{\frac{nd}{p \log(1 + JW \epsilon^{-1} c^{-1}) + \log(4/\delta)}}$$

(In the statement we had $\frac{\epsilon}{\sigma}$ in place of ϵ, but we have assumed σ bounded away from 0 for simplicity. The paper [Bubeck, Sellke, 2021] has a slightly more refined analysis that captures the dependence on $\frac{\epsilon}{\sigma}$).

$\boxtimes$

EXTRA :

Hoeffding's inequality for bounded random variables.

THEOREM Let $X_1, \dots, X_N$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for all i. Then, for every $t > 0$, we have

$$\mathbb{P}\left(\sum_{i=1}^{N} (X_i - \mathbb{E} X_i) \geq t \right) \leq e^{-2t^2 / \sum_{i=1}^{N} (M_i - m_i)^2}$$

ϵ-net : $\mathcal{N} \subseteq K$ is an ϵ-net of K if every point in K is within distance ϵ of some point of $\mathcal{N}$

Formally, $\forall x \in K \quad \exists x_0 \in \mathcal{N} : \underbrace{\| x - x_0 \|} \leq \epsilon$

valid for general metric spaces

$(d(x, x_0) \leq \epsilon)$

THEOREM Let $\mathcal{N}$ be an ϵ-net of the Euclidean ball in n dimensions with unit radius. Then,

$$\left(\frac{1}{\epsilon} \right)^n \leq |\mathcal{N}| \leq \left(\frac{2}{\epsilon} + 1 \right)^n$$

A SUFFICIENT CONDITION FOR NTK FEATURES

Let's go back to :

$$x_{adv} = x + \Delta_{adv} , \quad \| \Delta_{adv} \| \leq \delta \cdot \|x\|$$

$$| f(x_{adv}, \vartheta) - f(x,\vartheta) | \simeq | \nabla_x f(x,\vartheta)^T \Delta_{adv} | \leq \delta \cdot \underbrace{\|x\| \cdot \| \nabla_x f(x,\vartheta) \|}_{= S_f(x)}$$

sensitivity of the model evaluated in x

The model f is ROBUST if $S_f(x) = \overbrace{O(1)}^{\text{size of the labels}}$ for most test samples x. This means that the output change $|f(x_{adv},\vartheta) - f(x,\vartheta)|$ is of the same order as $|f(x,\vartheta)|$. In contrast, if $S_f(x) \gg 1$, we expect the model to be vulnerable to adversarial perturbations.

NTK FEATURES. Consider the two-layer network

$$f_{NN}(x, w) = \sum_{i=1}^{k} \phi(\langle w_i^{(1)}, x \rangle) - \sum_{i=1}^{k} \phi(\langle w_i^{(2)}, x \rangle)$$

*) $2k$ neurons

*) Activation function ϕ

*) # parameters $p = 2kd$, $W^{(1)} = [w_1^{(1)}, \dots, w_k^{(1)}]$, $W^{(2)} = [w_1^{(2)}, \dots, w_k^{(2)}]$
$w = [W^{(1)}, W^{(2)}]$

Initialization $w_0 = [W_0^{(1)}, W_0^{(2)}]$ with $[W_0^{(1)}]_{ij} \overset{iid}{\sim} N(0, 1/d)$ and $W_0^{(2)} = W_0^{(1)}$.

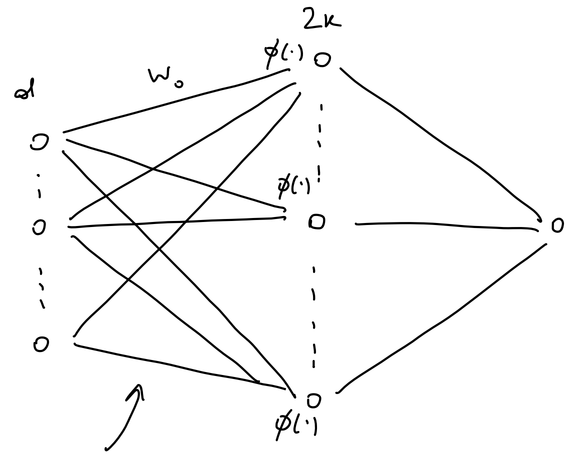
Under certain conditions, the GD dynamics of $f_{NN}(x,w)$ is close to the GD dynamics of its linearization around w_0

$$\underbrace{f_{NN}(x, w_0)}_{= 0 \text{ by symmetry}} + \langle \nabla_w f_{NN}(x, w_0), w - w_0 \rangle$$

Define $\quad f_{NTK}(x,\theta) = \Phi_{NTK}(x)^T \theta, \quad \Phi_{NTK}(x) = \nabla_w f_{NN}(x, w_0)$

*) $\theta \in \mathbb{R}^p$, $p = 2kd$, vector of trainable parameters

$$\underset{\text{obtained by running GD from initialization } \theta_0}{\theta^*} = \theta_0 + \Phi_{NTK}^T \left(\Phi_{NTK} \Phi_{NTK}^T \right)^{-1} y, \quad \text{with} \quad \Phi_{NTK} = \begin{bmatrix} \Phi_{NTK}(x_1) \\ \vdots \\ \Phi_{NTK}(x_n) \end{bmatrix} \in \mathbb{R}^{n \times p}$$



"lazy" training happens here

$$S_{f_{NTK}}(x) = \|x\| \cdot \| \nabla_x \Phi_{NTK}^T(x) \Phi_{NTK} \left(\Phi_{NTK} \Phi_{NTK}^T \right)^{-1} y + \underbrace{\nabla_x \Phi_{NTK}^T(x) \theta_0}_{= 0 \text{ by symmetry}} \|$$

Data assumptions:

*) $\|x_i\| = \sqrt{d}$ for all i $\qquad$ data on the sphere of radius $\sqrt{d}$

*) $\mathbb{E}\, x_i = 0$ $\qquad$ data centred

*) The data distribution P_x satisfies $\underset{\text{dimension-independent constant}}{c}$- Lipschitz concentration

$$\mathbb{P}\left(\left| \varphi(x) - \int \varphi(x') \, dP_x(x') \right| > t \right) \leq 2e^{-t^2/2c \cdot Lip(\varphi)^2}$$

<u>THEOREM</u> [Bombari, Kiyani, M., 2023]. Let ϕ be non-linear, even, with Lipschitz derivative. Assume $n \,\text{polylog}(n) = o(kd)$, $k = O(d)$, $n = O(k)$. Then, with high probability,

$$S_{f_{NTK}}(x) = O\left(\log k \sqrt{\frac{nd}{p}} \right).$$

INTERPRETATION

*) Same condition on overparameterization as in the lower bound by [Bubeck, Sellke, 2021]

*) Lower bound is for Lipschitz constant $\left(\sup_x \|\nabla_x f(x)\| \right)$, while the upper bound above holds with high probability over a test point.

PROOF SKETCH

$$S_{f_{NTK}}(x) \leq \|x\|_2 \; \|\nabla_x \Phi_{NTK}^T(x) \, \Phi_{NTK}^T\|_{op} \; \|(\Phi_{NTK} \Phi_{NTK}^T)^{-1}\|_{op} \; \|y\|$$

*) $\|x\|_2 = \sqrt{d}$ data normalization

*) $\|y\| = \sqrt{n}$ iid data with $O(1)$ y_i

*) $$\|(\Phi_{NTK} \Phi_{NTK}^T)^{-1}\|_{op} = \frac{1}{\lambda_{min}(\Phi_{NTK} \Phi_{NTK}^T)} = O\left(\frac{1}{dk}\right)$$

previous analysis

*) $$\|\nabla_x \Phi_{NTK}^T(x) \, \Phi_{NTK}^T\|_{op} = O\left(\log k \, (\sqrt{k} + \sqrt{n})\sqrt{d}\right)$$

direct calculation via chain rule

Thus, $$S_{f_{NTK}}(x) = O\left(\sqrt{nd} \; \frac{1}{dk} \; \log k \, \overbrace{(\sqrt{k} + \sqrt{n})}^{O(\sqrt{k})} \sqrt{d}\right)$$

$$= O\left(\log k \sqrt{\frac{nd}{\underbrace{dk}_{\text{"}p\text{"}}}}\right).$$

$\boxtimes$

Both model (NTK features) and activation function (even) matter for robustness...

RANDOM FEATURES (RF) [Rahimi, Recht, 2007]

$$f_{RF}(x, \theta) = \Phi_{RF}^T(x)\,\theta \;, \qquad \Phi_{RF}(x) = \phi(Vx)$$

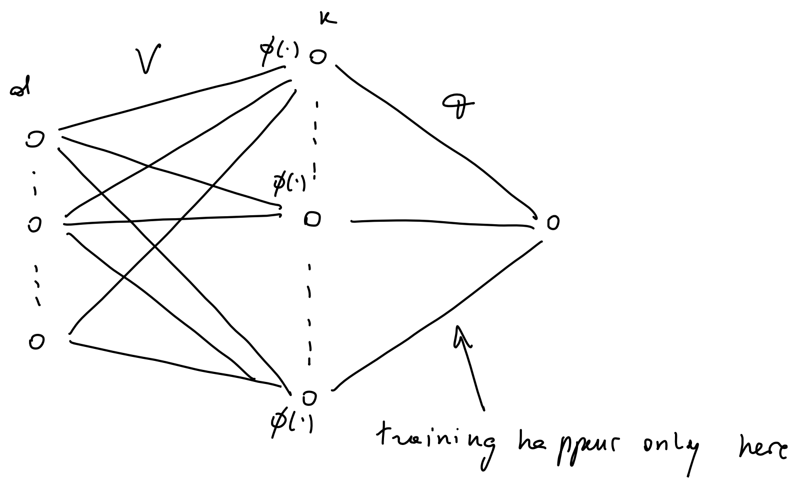
*) $V \in \mathbb{R}^{k \times d}$ $\qquad V_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1/d)$

*) $\theta \in \mathbb{R}^p$, $p = k$, vector of trainable parameters

$$\theta^* = \Phi_{RF}^T \left(\Phi_{RF} \Phi_{RF}^T \right)^{-1} y, \quad \text{with} \quad \Phi_{RF} = \begin{bmatrix} \Phi_{RF}(x_1) \\ \vdots \\ \Phi_{RF}(x_n) \end{bmatrix} = \phi(X V^T) \in \mathbb{R}^{n \times k}$$

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n \times d}$$

obtained by running GD from initialization $\theta_0 = 0$



*) RF model can be regarded as a two-layer neural network with first-layer weights frozen at initialization.

*) Simplest model allowing a degree of freedom in the number of parameters (in linear regression, # parameters = input dimension)

*) Lots of theoretical work on RF model, and we'll see more later in this course...

$$S_{f_{RF}}(x) = \|x\| \cdot \| \nabla_x \Phi_{RF}^T(x)\, \Phi_{RF}^T \left(\Phi_{RF} \Phi_{RF}^T \right)^{-1} y \|$$

THEOREM [Bombari, Kiyani, M., 2023]. Assume the data assumptions hold, and let $y_i = g(x_i) + \epsilon_i$ with ϵ_i zero-mean and sub-Gaussian. Let ϕ be non-linear, Lipschitz, with two Lipschitz derivatives and such that $\mathbb{E}_{G\sim N(0,1)} \phi'(G) \neq 0$. Assume $n \operatorname{polylog}(n) = o(k)$, $d \operatorname{polylog}(d) = O(k)$, and $n \operatorname{polylog}(n) = o(d^{3/2})$. Then, with high probability,

$$S_{f_{RF}}(x) = \Omega(n^{1/6}) \gg 1 .$$

INTERPRETATION

*) RF model NOT ROBUST for any degree of overparameterization

*) $\mathbb{E}_{G\sim N(0,1)} \phi'(G) \neq 0$ may be fundamental (confirmed by simulations)

PROOF SKETCH

STEP 1: Fitting the noise lower bounds the sensitivity

Let $\mathbb{E}[\epsilon^2] = \varepsilon^2$ and $A(x) = \nabla_x \Phi_{RF}^T(x)\, \Phi_{RF}^T (\Phi_{RF} \Phi_{RF}^T)^{-1} \in \mathbb{R}^{d\times n}$.

Then, with high probability,

$$S_{f_{RF}}(x) \geq \frac{\varepsilon}{2} \|x\| \, \|A(x)\|_F$$

Idea: Decompose y_i into $g(x_i)$ and ϵ_i + use Hanson-Wright

$$S_{f_{RF}}(x) = \|x\| \cdot \|A(x)\, y\| = \|x\| \cdot \| A(x) \underbrace{(g(X) + \epsilon)}_{\begin{bmatrix} g(x_1)+\epsilon_1 \\ \vdots \\ g(x_n)+\epsilon_n \end{bmatrix}} \|$$

remove dependence on x from $A(x)$ for simplicity of notation

$$\|Ay\|^2 = \epsilon^T A^T A \epsilon + g(X)^T A^T A\, g(X) + 2 g(X)^T A^T A \epsilon$$
$$\geq \epsilon^T A^T A \epsilon + g(X)^T A^T A\, g(X) - \varepsilon \|A g(X)\| \, \|A\|_F$$

holds with high probability upon concentrating $2 g(X)^T A^T A \epsilon$

$$= \epsilon^T A^T A \epsilon + \mathcal{M}^2 - \varepsilon \mathcal{M} \|A\|_F$$

$\mathcal{M} := \|A g(x)\|$

$$\geq \epsilon^T A^T A \epsilon - \frac{\varepsilon^2 \|A\|_F^2}{4} \geq \frac{\varepsilon^2}{2} \|A\|_F^2 - \frac{\varepsilon^2}{4} \|A\|_F^2 = \frac{\varepsilon^2}{4} \|A\|_F^2$$

minimize over $\mathcal{M}$

holds with high probability upon concentrating $\epsilon^T A^T(x) A(x) \epsilon$ via Hanson-Wright

STEP 2: Split interaction and kernel components

Let $\underbrace{\mathcal{I}_{RF}(x)}_{\text{interaction matrix}} = \nabla_x \Phi_{RF}^T(x) \tilde{\Phi}_{RF}^T \in \mathbb{R}^{n \times n}$ with $\tilde{\Phi}_{RF} = \Phi_{RF} - \mathbb{E}_{x_1, \dots, x_n} \Phi_{RF}$.

Then, with high probability,

$$\|A(x)\|_F \geq \frac{\|\mathcal{I}_{RF}(x)\|_F}{\lambda_{max}(\tilde{\Phi}_{RF} \tilde{\Phi}_{RF}^T)} - O\left(\frac{\sqrt{n+d}}{d}\right) = \Omega\left(\frac{1}{k} \frac{d}{n+d} \|\mathcal{I}_{RF}(x)\|_F - \frac{\sqrt{n+d}}{d}\right)$$

$$\|A(x)\|_F \leq \frac{\|\mathcal{I}_{RF}(x)\|_F}{\lambda_{min}(\tilde{\Phi}_{RF} \tilde{\Phi}_{RF}^T)} + O\left(\frac{\sqrt{n+d}}{d}\right) = O\left(\frac{1}{k} \|\mathcal{I}_{RF}(x)\|_F + \frac{\sqrt{n+d}}{d}\right)$$

Ideas:

*) Center $A(x)$

$\nabla_x \Phi_{RF}^T(x) \Phi_{RF}^T (\Phi_{RF} \Phi_{RF}^T)^{-1}$

$$\left| \|A(x)\|_F - \|\nabla_x \Phi_{RF}^T(x) \tilde{\Phi}_{RF}^T (\tilde{\Phi}_{RF} \tilde{\Phi}_{RF}^T)^{-1}\|_F \right| = O\left(\frac{\sqrt{n+d}}{d}\right)$$

(extract two rank-1 terms via Sherman-Morrison + various estimates)

*) Bounds on the spectrum of $\tilde{\Phi}_{RF} \tilde{\Phi}_{RF}^T$

$\nabla_x \Phi_{RF}^T(x) \tilde{\Phi}_{RF}^T$

$$\underbrace{\frac{\|\mathcal{I}_{RF}(x)\|_F}{\lambda_{max}(\tilde{\Phi}_{RF} \tilde{\Phi}_{RF}^T)}}_{O(k(n+d)/d)} \leq \|\nabla_x \Phi_{RF}^T(x) \tilde{\Phi}_{RF}^T (\tilde{\Phi}_{RF} \tilde{\Phi}_{RF}^T)^{-1}\|_F \leq \frac{\|\mathcal{I}_{RF}(x)\|_F}{\underbrace{\lambda_{min}(\tilde{\Phi}_{RF} \tilde{\Phi}_{RF}^T)}_{\Omega(k)}}$$

STEP 3 : Estimating $\| \mathcal{I}_{RF}(x) \|_F$

$$\left| \, \| \mathcal{I}_{RF}(x) \|_F - \Big(\mathbb{E}_{G\sim N(0,1)} \phi'(G) \Big)^2 \frac{k\sqrt{n}}{\sqrt{d}} \right| = o\left(\frac{k\sqrt{n}}{\sqrt{d}} \right)$$

If $\mathbb{E}_{G\sim N(0,1)} \phi'(G) = 0$, then $\| \mathcal{I}_{RF}(x) \|_F$ is of lower order and the sensitivity can also be much smaller.

Idea :

$$\| \mathcal{I}_{RF}(x) \|_F^2 \overset{\text{direct calculation}}{=} \sum_{i=1}^{n} \| V^T \operatorname{diag}(\phi'(Vx)) \tilde{\phi}(Vx_i) \|^2$$

$$\Big(\tilde{\phi}(Vx_i) = \phi(Vx_i) - \mathbb{E}_{x_i} \phi(Vx_i) \Big)$$

*) $\operatorname{diag}(\phi'(Vx))$ and $\tilde{\phi}(Vx_i)$ depend on V only via a single projection

*) We split $\operatorname{diag}(\phi'(Vx))$ and $\tilde{\phi}(Vx_i)$ via Taylor expansion into a component correlated with V and an independent one

*) Correlated components are computed exactly and independent ones are negligible

$$\| \mathcal{I}_{RF}(x) \|_F^2 \approx n \overbrace{\| V^T \operatorname{diag}(\phi'(Vx)) \tilde{\phi}(Vx_1) \|^2}^{\sum_{j=1}^{d} \left(V^T \operatorname{diag}(\phi'(Vx)) \tilde{\phi}(Vx_1) \right)_j^2}$$

$$= n \sum_{j=1}^{d} \left[\left| (x_1)_j \Big(\mathbb{E}_{G\sim N(0,1)} \phi'(G) \Big)^2 \frac{k}{d} \right| + o\left(\frac{k}{d} \right) \right]^2$$

$$= n \cancel{d} \Big(\mathbb{E}_{G\sim N(0,1)} \phi'(G) \Big)^4 \frac{k^2}{d^{\cancel{2}}} (1 + o(1))$$

Putting everything together:

$$S_{f_{NF}}(x) = \|x\| \cdot \| A(x)\, y \|$$

Step 1

$$= \Omega\left(\sqrt{d}\, \| A(x) \|_F \right)$$

Step 2

$$= \Omega\left(\sqrt{d} \left[\frac{1}{\kappa} \frac{d}{n+d} \| \mathcal{I}_{NF}(x) \|_F - \frac{\sqrt{n+d}}{d} \right] \right)$$

Step 3

$$= \Omega\left(\sqrt{d} \left[\frac{1}{\kappa} \frac{d}{n+d} \frac{\kappa \sqrt{n}}{\sqrt{d}} - \frac{\sqrt{n+d}}{d} \right] \right)$$

$$= \Omega\left(n^{1/6} \right)$$

$$\frac{1}{\kappa} \frac{d}{n+d} \frac{\kappa \sqrt{n}}{\sqrt{d}} = \frac{\sqrt{nd}}{n+d} \gg \frac{\sqrt{n+d}}{d} \iff d^{3/2} n^{1/2} \gg (n+d)^{3/2} = \Theta(\max(n^{3/2}, d^{1/2}))$$

Distinguish two cases:

① $n \leq d$

$$(n+d)^{3/2} = \Theta(d^{3/2}) \ll d^{3/2} n^{1/2}$$

$$\sqrt{d} \left[\frac{1}{\kappa} \frac{d}{n+d} \frac{\kappa \sqrt{n}}{\sqrt{d}} - \frac{\sqrt{n+d}}{d} \right] = \Omega\left(\sqrt{d} \frac{\sqrt{nd}}{d} \right) = \Omega(\sqrt{n})$$

② $n > d$

$$(n+d)^{3/2} = \Theta(n^{3/2}) \ll n^{1/2} d^{3/2} \iff n \ll d^{3/2} \quad (\text{by hypothesis})$$

$$\sqrt{d} \left[\frac{1}{\kappa} \frac{d}{n+d} \frac{\kappa \sqrt{n}}{\sqrt{d}} - \frac{\sqrt{n+d}}{d} \right] = \Omega\left(\sqrt{d} \frac{\sqrt{nd}}{n} \right) = \Omega\left(\frac{d}{\sqrt{n}} \right) = \Omega\left(n^{1/6} \right)$$

$n \ll d^{3/2}$

HOW MANY PARAMETERS TO RECONSTRUCT TRAINING DATA? (MEMORIZATION #2)

[start with slides]

RANDOM FEATURES (RF)

$$f_{RF}(x, \theta) = \Phi_{RF}^T(x)\,\theta \quad , \qquad \Phi_{RF}(x) = \phi(Vx)$$

$*)\quad V \in \mathbb{R}^{p \times d} \qquad V_{ij} \overset{iid}{\sim} N(0, 1/d)$

ASSUMPTIONS :

(B1) Let the training samples $(x_i)_{i=1}^n$ be iid, sub-Gaussian (with sub-Gaussian norm having constant order) and ℓ_2 norm equal to $\sqrt{d}$.

(data distribution)

(B2) The activation function $\phi : \mathbb{R} \to \mathbb{R}$ is non-linear, Lipschitz, with Lipschitz derivative. Letting ϕ_ℓ denote its ℓ-th Hermite coefficient, we assume $\phi_0 = \phi_2 = 0$, $\phi_1 \neq 0$ and that there exist two non-zero Hermite coefficients of order ≥ 3 with different parity.

(activation function)

resolves sign ambiguity in the reconstruction

(more on this later)

(B3) $n = O(d)$ and $p = \tilde{\Omega}(nd)$

(overparameterization)

THEOREM [Iurada, Bombari, Tommasi, M., 2026] Let (B1)-(B2)-(B3) hold. Let $\hat{X} = \begin{bmatrix} \hat{x}_1 \\ \vdots \\ \hat{x}_n \end{bmatrix} \in \mathbb{R}^{n \times d}$ be such that $\|\hat{x}_i\| = \sqrt{d}$ and, for all $i \in \{1, .., n\}$, $\phi(Vx_i) \in \text{span}\{\phi(V\hat{x}_1), .., \phi(V\hat{x}_n)\}$. Then, with high probability, for any $\hat{\imath} \in \{1, .., n\}$, there exists $i \in \{1, .., n\}$ such that

$$\| \hat{x}_{\hat{\imath}} - x_i \| = o(\sqrt{d})$$

Each row of $\hat{X}$ is close to a training sample

<u>INTERPRETATION</u> If random features of training samples are spanned by random features of $\hat{X}$, the rows of $\hat{X}$ must be close to original training samples, as long as $p \gg dn$.

*) Existence of two non-zero Hermite coefficients with different parity is a necessary condition to recover the sign of the training samples.

⟶ For example, if ϕ is either even or odd, the problem is under-determined in terms of the signs of the $\hat{x}_i$'s, as span$\{\phi(V\hat{x}_1), \dots, \phi(V\hat{x}_n)\}$ does not depend on them.

⟶ Evident in numerical simulations (for $\phi = \mathrm{ReLU}$, where $\phi_{2\ell+1} = 0$ for $\ell > 1$, negatives of training samples may be reconstructed).

⟶ Can drop the condition if only interested in overlap:

$$\frac{|\langle \hat{x}_{\hat{i}}, x_i \rangle|}{\|\hat{x}_{\hat{i}}\| \, \|x_i\|} = 1 - o(1)$$

PROOF SKETCH.

STEP 1 : Decompose features of $\hat{X}$ into features of training samples

For any $\hat{x}$ row of $\hat{X}$, we have

$$\phi(V\hat{x}) = \sum_{i=1}^{n} a_i \, \phi(Vx_i) \quad \text{for some } (a_i)_{i=1}^{n} \qquad (3)$$

*) Using ideas similar to those discussed when bounding the smallest eigenvalue of the NTK, we have that $\lambda_{min}(\Phi\Phi^T) > 0$, with $\Phi = \begin{bmatrix} \phi(Vx_1) \\ \vdots \\ \phi(Vx_n) \end{bmatrix}$.

*) By hypothesis, $\phi(Vx_i) \in \text{span}\{\phi(V\hat{x}_1), \dots, \phi(V\hat{x}_n)\}$ for all i, which gives

$$\text{span}\{\phi(Vx_1), \dots, \phi(Vx_n)\} \subseteq \text{span}\{\phi(V\hat{x}_1), \dots, \phi(V\hat{x}_n)\}$$

As $\dim(\text{span}\{\phi(Vx_1), \dots, \phi(Vx_n)\}) = n$ $\left(\lambda_{min}(\Phi\Phi^T) > 0\right)$, the two spans are in fact equal, which gives

$$\phi(V\hat{x}) \in \text{span}\{\phi(Vx_1), \dots, \phi(Vx_n)\} \quad \text{for any } \hat{x} \text{ row of } \hat{X}$$

this is equivalent to (3)

STEP 2 : Looking at the decomposition in one direction

Let $\tilde{\phi}(t) = \phi(t) - \phi_1 t$ be obtained by removing the linear part in $\phi(\cdot)$.

Linear part is problematic and, in fact, if ϕ was linear, then there is a simple counterexample to the statement:

Pick $\hat{x}_i = \sqrt{d} \dfrac{x_i + x_{i+1}}{\| x_i + x_{i+1} \|}$ $i \in \{1, \dots, n-1\}$, $\hat{x}_n = \sqrt{d} \dfrac{x_n - x_1}{\| x_n - x_1 \|}$.

Then $Vx_i \in \text{span}\{V\hat{x}_1, \dots, V\hat{x}_n\}$ for all $i \in \{1, \dots, n\}$, but $(\hat{x}_i)_{i=1}^{n}$ are not close to $(x_i)_{i=1}^{n}$.

$$V = \begin{bmatrix} v_1 \\ \vdots \\ v_p \end{bmatrix} \in \mathbb{R}^{p \times d}$$

$$\tilde{\phi}(V\hat{x})^T \phi(V\hat{x}) \overset{(3)}{=} \sum_{i=1}^{n} a_i \, \tilde{\phi}(V\hat{x})^T \phi(V x_i) = \sum_{i=1}^{n} a_i \sum_{j=1}^{p} \tilde{\phi}(\langle v_j, \hat{x} \rangle)\, \phi(\langle v_j, x_i \rangle)$$

$$= \mathbb{E}_V \left[\sum_{i=1}^{n} a_i \sum_{j=1}^{p} \tilde{\phi}(\langle v_j, \hat{x} \rangle)\, \phi(\langle v_j, x_i \rangle) \right] \left(1 + \tilde{O}\left(\|a\| \sqrt{\frac{dn}{p}} \right) \right)$$

Bernstein inequality as $\tilde{\phi}(\langle v_j, \hat{x} \rangle)\, \phi(\langle v_j, x_i \rangle)$ is sub-exponential

$$= p \sum_{i=1}^{n} a_i \sum_{\ell=3}^{\infty} \phi_\ell^2 \frac{\langle \hat{x}, x_i \rangle^\ell}{d^\ell} \left(1 + \tilde{O}\left(\|a\| \sqrt{\frac{dn}{p}} \right) \right)$$

Hermite expansion :

$$\mathbb{E}_{v_j} \tilde{\phi}(\langle v_j, \hat{x} \rangle)\, \phi(\langle v_j, x_i \rangle) = \mathbb{E}_{\bar{v} \sim N(0, I_d)} \tilde{\phi}\left(\left\langle \bar{v}, \frac{\hat{x}}{\sqrt{d}} \right\rangle\right) \phi\left(\left\langle \bar{v}, \frac{x_i}{\sqrt{d}} \right\rangle\right)$$

$$= \sum_{\ell \in \mathbb{N}} \phi_\ell \tilde{\phi}_\ell \frac{\langle x_i, \hat{x} \rangle^\ell}{d^\ell} = \sum_{\ell \geq 3} \phi_\ell^2 \frac{\langle \hat{x}, x_i \rangle^\ell}{d^\ell}$$

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \qquad \text{Hadamard product}$$

$$\leq p \|a\| \sum_{\ell=3}^{\infty} \phi_\ell^2 \left\| \frac{(X\hat{x})^{\circ \ell}}{d^\ell} \right\| \left(1 + \tilde{O}\left(\|a\| \sqrt{\frac{dn}{p}} \right) \right)$$

$$\leq p \sum_{\ell=3}^{\infty} \phi_\ell^2 \left\| \frac{(X\hat{x})^{\circ \ell}}{d^\ell} \right\| \left(1 + \tilde{O}\left(\sqrt{\frac{dn}{p}} \right) \right)$$

$$\left| \|a\|^2 - 1 \right| = \tilde{O}\left(\sqrt{\frac{dn}{p}} \right)$$

$$\leq p \sum_{\ell=3}^{\infty} \phi_\ell^2 \Big(\underbrace{\max_i \left| \frac{x_i^T \hat{x}}{d} \right|}_{:= C \, \in [0,1]} \Big)^{\ell - 1} \left(1 + \tilde{O}\left(\sqrt{\frac{dn}{p}} \right) \right)$$

$\hat{x}$ can't be too correlated with too many x_i's

$$\leq p \underbrace{\sum_{\ell=3}^{\infty} \phi_\ell^2}_{\bar{\phi}^2} C^2 \left(1 + \tilde{O}\left(\sqrt{\frac{dn}{p}} \right) \right)$$

However, we also have

(similar argument with Bernstein + Hermite expansion)

$$\tilde{\phi}(V\hat{x})^T \phi(V\hat{x}) = \bar{\phi}^2 p(1+o(1))$$

This implies $C = 1 + o(1)$, so $\left|\frac{x_i^T \hat{x}}{d}\right|$ is close to 1 for some i.

*) Bounds need to hold uniformly over $\hat{x}$, so we need (a few) ϵ-net arguments.

*) To pass from overlap $\left(\left|\frac{x_i^T \hat{x}}{d}\right| = 1 + o(1) \right)$ to ℓ_2 distance $\left(\|\hat{x} - x_i\| = o(\sqrt{d}) \right)$, we use the existence of Hermite coefficients with different parities to resolve the ambiguity on the sign.

With a similar argument, we have that $\left| \bar{\phi}^2 - a_i \sum_{\ell=3}^{\infty} \phi_\ell^2 \frac{(x_i^T \hat{x})^\ell}{d^\ell} \right| = o(1)$.

This is possible only if $(x_i^T \hat{x})^\ell$ has the same sign for all $\ell \geq 3$ s.t. $\phi_\ell \neq 0$.

$\square$

EXTRA:

Bernstein inequality

THEOREM Let $X_1, \dots, X_N$ be independent, mean zero, sub-exponential random variables. Then, for every $t \geq 0$, we have

$$\mathbb{P}\left(\left| \sum_{i=1}^N X_i \right| > t \right) \leq 2 \exp\left[-c \min\left(\frac{t^2}{\sum_{i=1}^N \|X_i\|_{\psi_1}^2}, \frac{t}{\max_i \|X_i\|_{\psi_1}} \right) \right],$$

where $\|\cdot\|_{\psi_1}$ denotes the sub-exponential norm.

*) Result above allows for duplicates (multiple rows of $\hat{X}$ corresponding to the same training sample). We can further show that ALL training samples are reconstructed (i.e., no duplicates) for $n = 2$.

proof quite ad-hoc for $n = 2$. For $n = 3$, we would need to consider separately the case in which $\hat{x}_1$, $\hat{x}_2$ and $\hat{x}_3$ are close and the case in which just a pair of reconstructions is close. Thus, # cases increases combinatorially with n.

[OPEN PROBLEM] Show all training samples are reconstructed for general n.

*) Assumption that $\phi(Vx_i) \in \operatorname{span}\{\phi(V\hat{x}_1), \dots, \phi(V\hat{x}_n)\}$ is quite strong ...

$$\phi(Vx_i) \in \overbrace{\operatorname{span}\{\phi(V\hat{x}_1), \dots, \phi(V\hat{x}_n)\}}^{S}$$

$$\Downarrow$$

$$\| P_{\hat{X}}^{\perp} \phi(Vx_i) \|_2 = 0 \quad \text{with } P = \text{projector on } S$$

$$\theta^* = \Phi_{RF}^T (\Phi_{RF} \Phi_{RF}^T)^{-1} y, \quad \text{with } \Phi_{RF} = \begin{bmatrix} \Phi_{RF}(x_1) \\ \vdots \\ \Phi_{RF}(x_n) \end{bmatrix} = \phi(XV^T) \in \mathbb{R}^{n \times p}, \quad X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n \times d}$$

obtained by running GD from initialization $\theta_0 = 0$

θ^* is a linear combination of $\phi(Vx_i)$ so instead of optimizing for $\| P_{\hat{X}}^{\perp} \phi(Vx_i) \|_2 = 0$ for all i (which we cannot do in practice since we don't know $\phi(Vx_i)$), we just optimize for $\| P_{\hat{X}}^{\perp} \theta^* \|_2 = 0$.

$$\hat{X}^* = \underset{\hat{X} : \|\hat{x}_i\| = \sqrt{d}}{\operatorname{argmin}} \| P_{\hat{X}}^{\perp} \theta^* \|^2$$

*) Numerical evidence (shown before) that $\hat{X}^*$ reconstructs X when $p \gg dn$.

*) When $p \gg nd$, $\phi(Vx_i)$ approximately lies in span $\{\phi(V\hat{x}_1), \dots, \phi(V\hat{x}_n)\}$.

[OPEN PROBLEM] Show that optimizing $\| P_{\hat{X}}^{\perp} \theta^* \|$ leads to the reconstruction of the training samples.

[OPEN PROBLEM] Show similar results beyond the RF model (e.g. NTK model).

RECAP

* $p \gg n$ gives interpolation

* $p \gg dn$ gives smooth interpolation (no adversarial examples) but it allows an adversary to reconstruct the training dataset from the parameters of the trained network.

How to defend from this adversary?

One option is DIFFERENTIAL PRIVACY which will be the subject of the rest of the course.

DEEP LEARNING WITH DIFFERENTIAL PRIVACY

What is differential privacy?

Textbook: [Dwork, Roth, 2014] "The Algorithmic Foundations of Differential Privacy"

DEFINITION A dataset D' is adjacent to a dataset D if they differ by only one sample.

DEFINITION ((ε, δ)-DP) [Dwork, McSherry, Nissim, Smith, 2006]. A randomized algorithm A satisfies (ε, δ)-differential privacy if, for any pair of adjacent datasets D, D' and for any subset of the parameter space $S \subseteq \mathbb{R}^p$, we have

$$\mathbb{P}\left(A(D) \in S\right) \leq e^{\varepsilon}\, \mathbb{P}\left(A(D') \in S\right) + \delta$$

probability over the randomness induced by the algorithm

holds uniformly on all adjacent datasets D, D'

A is private if one cannot distinguish $A(D)$ from $A(D')$. This means that given the output of the algorithm, it is (information-theoretically) impossible to recover a single training sample.

*) $(\varepsilon, \delta) = (0, 0) \Rightarrow$ perfect privacy $\left(\mathbb{P}(A(D) \in S) = \mathbb{P}(A(D') \in S)\right)$

*) $\varepsilon \gg 1$ or $\delta = 1 \Rightarrow$ no privacy at all

*) In practice $\varepsilon \sim 1$, $\delta \ll 1/n$

number of training samples

Many other notions in the literature. We will see later Renyi-DP and zero-concentrated DP.

How to enforce it ?

Essentially add noise ...

*) Different types of noise : Laplace mechanism / noise, Gaussian mechanism / noise

*) Different ways of adding it : inside the empirical risk minimization objective (objective perturbation : perturb the problem and then solve it), after calculating the ERM solution (output perturbation : perturb the solution) during the iterative algorithm (DP-GD / DP-SGD : perturb the iterations)

Here, we focus uniquely on adding noise to gradient descent-type algorithms.

Differentially-private gradient descent (DP-GD)

$$\theta^{t+1} = \theta^t - \eta \frac{1}{n}\sum_{i=1}^{n} \nabla_\theta \ell(y_i, f(x_i;\theta^t)) \Big/ \max\left(1, \frac{\|\nabla_\theta \ell(y_i, f(x_i;\theta^t))\|_2}{C_{clip}}\right)$$

$$+ \sqrt{\eta}\, \frac{2 C_{clip}}{n}\, \sigma\, \mathcal{N}(0, I_p)$$

→ Gradient is clipped (its norm is at most C_{clip})

→ Additive Gaussian noise (Gaussian mechanism for privacy)

Informally, this avoids that the algorithm memorizes each individual data point.

More formally, we have the following :

Proposition For any $\delta \in (0,1)$, $\varepsilon \in (0, 8\log(1/\delta))$, if we set $\sigma \geq \sqrt{\eta T}\, \frac{\sqrt{8\log(1/\delta)}}{\varepsilon}$, then the output θ^T of DP-GD is (ε, δ)-differentially private.

*) Adaptation of the analysis from [Abadi, Chu, Goodfellow, 2016]

What's the cost of differential privacy?

Excess population risk:

$$R_E = \mathbb{E}\left[\ell\left(y_{test}, f(x_{test}; \theta^T)\right)\right] - \mathbb{E}\left[\ell\left(y_{test}, f(x_{test}; \theta^*)\right)\right]$$

solution of DP-GD with privacy guarantees   solution of GD with no privacy guarantees

Excess empirical risk:

$$\hat{R}_E = \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i; \theta^T)) - \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f(x_i; \theta^*))$$

More privacy $\Rightarrow$ $\varepsilon \downarrow$, $\delta \downarrow$

$\Rightarrow$ $\sigma \uparrow$ (or equivalently $C_{clip} \downarrow$)

proposition above

$\Rightarrow$ $R_E, \hat{R}_E \uparrow$ (worse performance)

Existing bounds on excess empirical/population risk tend to degrade as p grows:

*) $\hat{R}_E = \tilde{O}\left(\frac{\sqrt{p}}{n\varepsilon}\right)$ (objective perturbation, constrained, strongly convex (ε,δ)-DP optimization) [Kifer, Smith, Thakurta, 2012]

*) $\hat{R}_E = \beta L \,\tilde{O}\left(\frac{\sqrt{p}}{n\varepsilon}\right)$ (DP-GD, β = diameter of optimization domain, L = Lipschitz constant of the loss), $R_E = \tilde{O}\left(\frac{p^{1/4}}{\sqrt{n\varepsilon}}\right)$

[Bassily, Smith, Thakurta, 2014]

*) $\mathbb{E}[R_\varepsilon] = \rho L \; \tilde{O}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{p}}{n\varepsilon}\right)$ [Bassily, Feldman, Talwar, Thakurta, 2019]

expectation over randomness of algorithm and training data

*) Lots of work on unconstrained optimization too, but still dependence on dimension p

Noise introduced by DP-GD increases with dimension p (each entry of $\theta \in \mathbb{R}^p$ perturbed with standard Gaussian noise $\Rightarrow$ ℓ_2 norm of perturbation scales as $\sqrt{p}$).

*) DP algorithms acting over lower-dimensional subspaces

We will show that this bad dependence does not have to be there!

DP Tools and a Proof of the Proposition

Definition Let $\mu : \mathcal{D} \longrightarrow \mathbb{R}^p$ be an arbitrary, deterministic, p-dimensional function, where $\mathcal{D}$ represents the space of datasets. Its ℓ_2 sensitivity is defined as

$$\Delta_2 \mu = \sup_{D \text{ adjacent } D'} \| \mu(D) - \mu(D') \|_2$$

Definition The Gaussian mechanism $\mathcal{M} : \mathcal{D} \longrightarrow \mathbb{R}^p$ with parameter ρ is the randomized mechanism that adds noise $N(0, \rho^2)$ to each of the p components of the output of μ, i.e.,

$$\mathcal{M}(\cdot) := \mu(\cdot) + \rho \, N(0, I_p)$$

The privacy guarantees for the Gaussian mechanism are given by the following result.

Theorem [Thm A.1, Dwork, Roth, 2014] For every $\varepsilon \in (0,1)$ and $\delta > 0$, the Gaussian mechanism $\mathcal{M}$ with parameter

$$\rho \geqslant \frac{\sqrt{2 \log(1.25/\delta)} \, \Delta_2 \mu}{\varepsilon}$$

is (ε, δ)-differentially private.

In our setting,

$$\mu_{\theta^{t-1}, D} = \theta^{t-1} - \frac{\eta}{n} \sum_{(x_i, y_i \in D)} \nabla_\theta \, \ell(y_i, f(x_i; \theta^{t-1})) \Big/ \max\left(1, \frac{\| \nabla_\theta \ell(y_i, f(x_i; \theta^{t-1}))\|_2}{C_{clip}}\right)$$

Let us now compute the sensitivity.

$$\Delta_2 \mu = \sup_{\theta^{t-1} \in \mathbb{R}^p,\ D \text{ adjacent with } D'} \| \mu_{\theta^{t-1}, D} - \mu_{\theta^{t-1}, D'} \|_2$$

$$= \sup_{\substack{\theta^{t-1} \in \mathbb{R}^p \\ D \text{ adjacent } D'}} \frac{\eta}{n} \Big\| \sum_{(x_i, y_i \in D)} \nabla_\theta \ell(y_i, f(x_i; \theta^{t-1})) / \max\Big(1, \frac{\|\nabla_\theta \ell(y_i, f(x_i; \theta^{t-1}))\|_2}{C_{clip}}\Big)$$

$$- \sum_{(x_i, y_i \in D')} \nabla_\theta \ell(y_i, f(x_i; \theta^{t-1})) / \max\Big(1, \frac{\|\nabla_\theta \ell(y_i, f(x_i; \theta^{t-1}))\|_2}{C_{clip}}\Big) \Big\|_2$$

$$= \frac{\eta}{n} \sup_{\theta^{t-1} \in \mathbb{R}^p, (x,y), (x',y')} \Big\| \nabla_\theta \ell(y, f(x; \theta^{t-1})) / \max\Big(1, \frac{\|\nabla_\theta \ell(y, f(x; \theta^{t-1}))\|_2}{C_{clip}}\Big)$$

$$- \nabla_\theta \ell(y', f(x'; \theta^{t-1})) / \max\Big(1, \frac{\|\nabla_\theta \ell(y', f(x'; \theta^{t-1}))\|_2}{C_{clip}}\Big) \Big\|_2$$

$$\leq \frac{2\eta}{n} \sup_{\theta^{t-1} \in \mathbb{R}^p, (x,y)} \Big\| \nabla_\theta \ell(y, f(x; \theta^{t-1})) / \max\Big(1, \frac{\|\nabla_\theta \ell(y, f(x; \theta^{t-1}))\|_2}{C_{clip}}\Big) \Big\|_2$$

$$\leq \frac{2\eta}{n} C_{clip}.$$

For 1 step, $\quad \rho \gtrsim \frac{\sqrt{\log 1/\delta}}{\varepsilon} \quad \frac{C_{clip}\, \eta}{n}$

DP-GD enforced $\quad \sqrt{\eta} \, \frac{2 C_{clip}}{n} \sigma \gtrsim \sqrt{\eta} \, \frac{C_{clip}}{n} \, \frac{\sqrt{\log 1/\delta} \, \sqrt{\eta T}}{\varepsilon}$

This matches for one step !

Now, we need to compose T independent Gaussian mechanisms.

Attempt 1 : Naive composition

Theorem [Thm 3.16, Dwork, Roth, 2014] For $t \in \{1, .., T\}$, let $\mathcal{M}_t$ be (ε', δ')-differentially private. Then, their composition is $(T\varepsilon', T\delta')$-differentially private.

This would give $\sigma \gtrsim \sqrt{\eta} \frac{\sqrt{\log(T/\delta)}}{\varepsilon} T$

☹ Too much ! We want $\sigma \gtrsim \sqrt{\eta T} \frac{\sqrt{8\log(1/\delta)}}{\varepsilon}$

Attempt 2: Advanced composition

Theorem [Thm 3.20, Dwork, Roth, 2014] For $t \in \{1, .., T\}$, let $\mathcal{M}_t$ be (ε', δ')-differentially private. Then, their composition is $(\varepsilon, T\delta' + \delta)$-differentially private, with $\varepsilon = \sqrt{2T\log(1/\delta)}\,\varepsilon' + T\varepsilon'(e^{\varepsilon'} - 1)$

This requires $\delta' < 1/T$, which would give at least

$$\sigma \gtrsim \sqrt{\eta T} \sqrt{\log T}$$

☹ also too much since we will look at the limit $T \to +\infty$, $\eta \to 0$ and $\eta T \to t$

Attempt 3 : Moment accountant

<u>Definition</u> For adjacent datasets $D, D' \in \mathcal{D}$, a randomized mechanism $M_t : \mathbb{R}^p \times \mathcal{D} \longrightarrow \mathbb{R}^p$, auxiliary input $\theta^{t-1} \in \mathbb{R}^p$, the <u>privacy loss</u> at the output θ is defined as

probability density function

$$\gamma(\theta; M_t, \theta^{t-1}, D, D') = \log \frac{p(M_t(\theta^{t-1}, D) = \theta)}{p(M_t(\theta^{t-1}, D') = \theta)}$$

We also define :

$$\alpha_{M_t}(\lambda; \theta^{t-1}, D, D') = \log \mathbb{E}_{\theta \sim M_t(\theta^{t-1}, D)} [\exp(\lambda \gamma(\theta; M_t, \theta^{t-1}, D, D'))]$$

log of moment-generating function of the privacy loss evaluated at λ

$$\alpha_{M_t}(\lambda) = \sup_{\theta^{t-1} \in \mathbb{R}^p,\ D \text{ adjacent to } D'} \alpha_{M_t}(\lambda; \theta^{t-1}, D, D')$$

supremum over all possible θ^{t-1} and adjacent datasets D, D'

<u>Theorem</u> [Abadi, Chu, Goodfellow, 2016] Let A consist of a sequence of independent mechanisms $M_1, \dots, M_T$. Then,

1. <u>Composability</u>. For any λ,

$$\alpha_A(\lambda) \leq \sum_{t=1}^{T} \alpha_{M_t}(\lambda).$$

2. <u>Tail bound</u>. For any $\varepsilon > 0$, A is (ε, δ)-differentially private for

$$\delta = \inf_{\lambda} \exp(\alpha_A(\lambda) - \lambda\varepsilon)$$

*) Moment accountant exploits the independence of the mechanisms while (advanced) composition allows for arbitrary correlations between them.

In our setting,

$$\mathcal{M}_t \sim \mathcal{N}\left(\mu_{\theta^{t-1}, D}, \rho^2\right) \quad \text{with} \quad \rho = \sqrt{\eta}\, \frac{2\, C_{\text{clip}}}{n}\, \sigma .$$

Thus,

$$\gamma(\theta; \mathcal{M}_t, \theta^{t-1}, D, D') = \log \frac{p(\mathcal{M}_t(\theta^{t-1}, D) = \theta)}{p(\mathcal{M}_t(\theta^{t-1}, D') = \theta)}$$

$$= -\frac{1}{2\rho^2}\left(\|\theta - \mu_{\theta^{t-1}, D}\|_2^2 - \|\theta - \mu_{\theta^{t-1}, D'}\|_2^2 \right)$$

$$= -\frac{1}{2\rho^2}\left(2\theta^T(\mu_{\theta^{t-1}, D'} - \mu_{\theta^{t-1}, D}) + \|\mu_{\theta^{t-1}, D}\|_2^2 - \|\mu_{\theta^{t-1}, D'}\|_2^2 \right)$$

$$= -\frac{1}{2\rho^2}\Big(2(\theta - \mu_{\theta^{t-1}, D})^T(\mu_{\theta^{t-1}, D'} - \mu_{\theta^{t-1}, D}) - \|\mu_{\theta^{t-1}, D}\|_2^2$$

$$+ 2\mu_{\theta^{t-1}, D}^T \mu_{\theta^{t-1}, D'} - \|\mu_{\theta^{t-1}, D'}\|_2^2 \Big)$$

$$= -\frac{1}{2\rho^2}\left(2(\theta - \mu_{\theta^{t-1}, D})^T \Delta_{\theta^{t-1}, D, D'} - \|\Delta_{\theta^{t-1}, D, D'}\|_2^2 \right)$$

$$\Delta_{\theta^{t-1}, D, D'} = \mu_{\theta^{t-1}, D'} - \mu_{\theta^{t-1}, D}$$

$$\alpha_{\mathcal{M}_t}(\lambda; \theta^{t-1}, D, D') = \log \mathbb{E}_{\theta \sim \mathcal{M}_t(\theta^{t-1}, D)}\left[\exp\left(\lambda \gamma(\theta; \mathcal{M}_t, \theta^{t-1}, D, D')\right) \right]$$

$$= \frac{\|\Delta_{\theta^{t-1}, D, D'}\|_2^2}{2\rho^2}(\lambda + \lambda^2)$$

some calculations...

$$\alpha_{M_t}(\lambda) = \sup_{\theta^{t-1} \in \mathbb{R}^p,\ \mathcal{D} \text{ adjacent to } \mathcal{D}'} \alpha_{M_t}(\lambda ; \theta^{t-1}, \mathcal{D}, \mathcal{D}')$$

$$= \frac{(\Delta_2 \mu)^2}{2\rho^2}(\lambda + \lambda^2) \leq \frac{\eta}{2\sigma^2}(\lambda + \lambda^2)$$

$$\Delta_2 \mu \leq \frac{2\eta\, C_{clip}}{n}$$

$$\rho = \sqrt{\eta}\, \frac{2C_{clip}}{n}\, \sigma$$

so $\dfrac{\Delta_2 \mu}{\rho} \leq \sqrt{\eta}/\sigma$

Let A be the output of DP-GD after T steps. Then, by composability,

$$\alpha_A(\lambda) \leq \frac{\eta T}{2\sigma^2}(\lambda + \lambda^2).$$

Thus,

$$\exp(\alpha_A(\lambda) - \lambda \varepsilon) \leq \exp\left(\frac{\eta T}{2\sigma^2}(\lambda + \lambda^2) - \lambda\varepsilon\right)$$

$$\leq \exp\left(\frac{\varepsilon^2}{16 \log(1/\delta)}(\lambda + \lambda^2) - \lambda \varepsilon\right)$$

$$\sigma \geq \sqrt{\eta T}\, \frac{\sqrt{8 \log(1/\delta)}}{\varepsilon}$$

$$\Downarrow$$

$$\frac{\eta T}{2\sigma^2} \leq \frac{\varepsilon^2}{16 \log(1/\delta)}$$

$$= \exp\left(\frac{\varepsilon^2}{16 \log(1/\delta)}\lambda^2 - \left(1 - \frac{\varepsilon}{16 \log(1/\delta)}\right)\lambda\varepsilon\right)$$

$$\leq \exp\left(\frac{\varepsilon^2}{16 \log(1/\delta)}\lambda^2 - \frac{\lambda\varepsilon}{2}\right)$$

$$\varepsilon \in (0, 8 \log(1/\delta))$$

$$\inf_{\lambda} \exp\left(\alpha_A(\lambda) - \lambda \varepsilon\right) \leq \exp\left(\frac{\varepsilon^2}{16\log(1/\delta)}\lambda_*^2 - \frac{\lambda_* \varepsilon}{2}\right)$$

$$= \exp\left(\log 1/\delta - 2\log(1/\delta)\right) = \delta$$

pick $\lambda_* = 4\log(1/\delta)/\varepsilon$

By the tail bound, DP-GD is (ε, δ)-differentially private and the proof of the proposition is complete. $\square$

PRIVACY FOR FREE IN THE OVERPARAMETERIZED REGIME

RANDOM FEATURES (RF)

$$f_{RF}(x, \vartheta) = \Phi_{RF}^T(x)\,\vartheta \quad , \qquad \Phi_{RF}(x) = \phi(Vx)$$

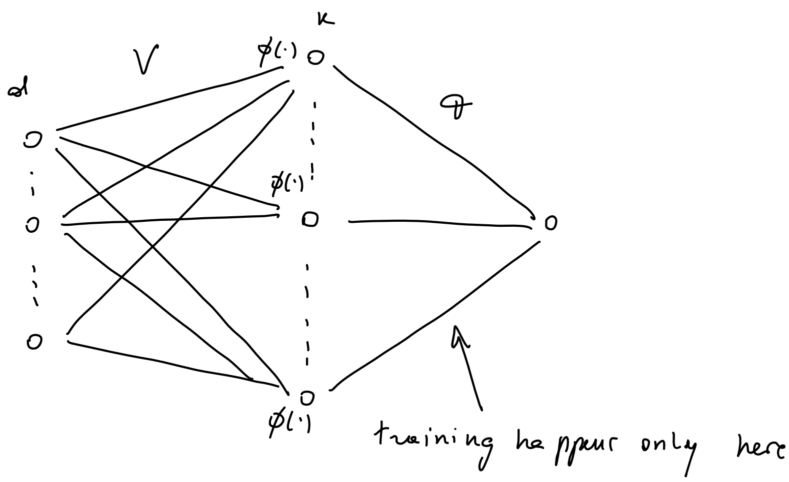
*) $V \in \mathbb{R}^{k \times d}$ $\qquad V_{ij} \overset{iid}{\sim} \mathcal{N}(0, 1/d)$

*) $\vartheta \in \mathbb{R}^p$, $p = k$, vector of trainable parameters

$$\vartheta^* = \Phi_{RF,n}^T \left(\Phi_{RF,n} \Phi_{RF,n}^T \right)^{-1} y \, , \quad \text{with} \quad \Phi_{RF,n} = \begin{bmatrix} \Phi_{RF}(x_1) \\ \vdots \\ \Phi_{RF}(x_n) \end{bmatrix} = \phi(X V^T) \in \mathbb{R}^{n \times k}$$

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \in \mathbb{R}^{n \times d}$$

obtained by running GD from initialization $\vartheta_0 = 0$



Data assumptions :

(a) $\int \|x\|_2 \, dP_x(x) = \sqrt{d}$ $\leftarrow$ scaling

(b) $\|x\|_{\psi_2} = O(1)$ $\leftarrow$ sub Gaussian data (with subGaussian norm bdd by universal constant)

(c) $\lambda_{min}\left(\mathbb{E}_{x \sim P_x}[x x^T] \right) = \Omega(1)$ $\leftarrow$ well conditioned covariance

(d) $|y| \leq C$ $\leftarrow$ bounded labels

Square loss : $\qquad \ell(y, \hat{y}) = (y - \hat{y})^2$

THEOREM [Bombari, M., 2025] Let $\phi(\cdot)$ be nonlinear, Lipschitz such that $\phi_0 = \phi_2 = 0$, $\phi_1 \neq 0$. Let $n = O(\sqrt{p})$, $n = \tilde{\Omega}(d)$, $n = \tilde{o}(d^{3/2})$ and consider a privacy budget $\delta \in (0,1)$, $\varepsilon \in (0, 8\log(1/\delta))$, $\frac{\varepsilon}{\sqrt{\log 1/\delta}} \gg \frac{d}{n}$.

Then, by setting properly C_{clip}, σ, T, we have that θ^T is (ε,δ)-differentially private and

$$R_E = \tilde{O}\left(\frac{d}{n\varepsilon} + \sqrt{\frac{d}{n}} + \sqrt{\frac{n}{d^{3/2}}} \right).$$

INTERPRETATION If $d \ll n \ll d^{3/2}$, then privacy comes for free (in the sense that the excess population risk R_E is $o(1)$) as long as $\varepsilon \gg \frac{d}{n}$, for any degree of overparameterization.

→ $\varepsilon = \Theta(1)$ usual in practice. Here we can even guarantee the strong privacy requirement $\varepsilon = o(1)$

→ dependence on δ only logarithmic (and hidden in $\tilde{O}(\cdot)$)

→ no dependence on p as long as $p \gtrsim n^2$. We expect this can be relaxed to $p \gtrsim n$ (which is the number of parameters to interpolate as seen before)

→ $d \ll n \ll d^{3/2}$ corresponds to standard datasets, such as CIFAR-10 ($n = 5 \cdot 10^4$, $d \approx 3 \cdot 10^3$) or ImageNet ($n \approx 1.3 \cdot 10^6$, $d \approx 9 \cdot 10^4$). We also expect that it can be relaxed to $d \ll n \ll d^2$ and, in fact, to $d^{\ell} \ll n \ll d^{\ell+1}$, as explained later.

→ $\phi_0 = \phi_2 = 0$ can again be potentially relaxed and role of $\phi_1 \neq 0$ clarified later.

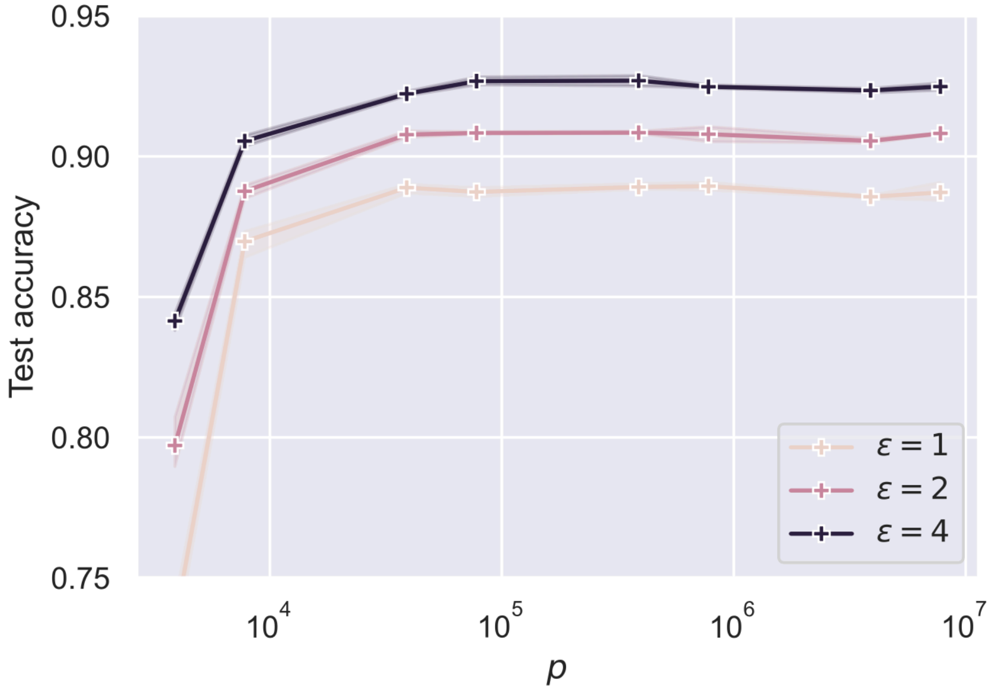
→ Proof describes correct scaling of hyperparameters C_{clip}, σ, T

IDEA OF THE ARGUMENT IN TWO CARTOONS

CARTOON 1 : Test loss of DP-GD on MNIST for two-layer networks

(a) n fixed

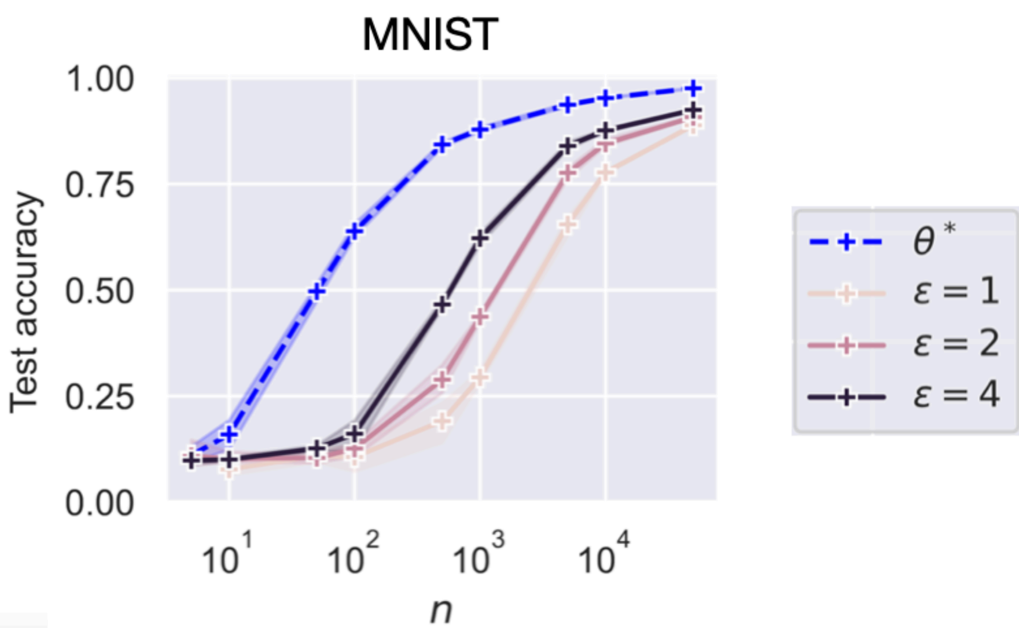
Test accuracy as a function of p



Test accuracy increases until the network is wide enough and then plateaus.

(b) p fixed

Test accuracy as a function of n



As # samples n grows, gap between DP-GD and GD shrinks

CARTOON 2 : Test loss of (non-private) GD on RF regression.

Before the actual cartoon, let us start with a result from the review by [Misiakiewicz, Montanari, 2024]

THEOREM Let $x_i \sim \mathrm{Unif}(\mathbb{S}^{d-1}(\sqrt{d}))$, $y_i = f_\star(x_i) + \varepsilon_i$, $f_\star \in L^2$, $\varepsilon_i \perp x_i$, $\mathbb{E}[\varepsilon_i] = 0$, $\mathbb{E}[\varepsilon_i^2] = \tau^2$. Assume $d^{\ell_1+\delta} \leq n \leq d^{\ell_1+1-\delta}$, $d^{\ell_2+\delta} \leq p \leq d^{\ell_2+1-\delta}$, $\max(p/n, n/p) \gtrsim d^\delta$ for some integers ℓ_1, ℓ_2 and constant $\delta > 0$. Denote $\ell = \min(\ell_1, \ell_2)$. Furthermore assume that $\phi(\cdot)$ has all Hermite coefficients bounded away from 0. Then,

$$\mathbb{E}_x \left| f_\star(x) - f_{RF}(x, \theta^*) \right|^2 = \underbrace{\| P_{>\ell} f_\star \|_{L^2}^2}_{} + o(1) \left(\| f_\star \|_{L^{2+}}^2 + \tau^2 \right)$$

$P_{>\ell} = I - P_{\leq \ell}$, with $P_{\leq \ell} : L^2 \to L^2$ the orthogonal projector onto the subspace of polynomials of degree at most ℓ

(project $f_\star$ into the space of degree-ℓ polynomials and take the L^2 norm of the residual)

we expect the result on DP-GD to hold under these milder conditions

$p \gtrsim n$ and $d \ll n \ll d^2$ correspond to $\ell_2 \geq \ell_1 = 1$ which gives $\ell = \min(\ell_1, \ell_2) = 1$. Then, the result above says that the test error of the RF model is flat as a function of n in the whole regime $d \ll n \ll d^2$ (when $p \geq n$).

Loss plateaus for $d \ll n \ll d^2$ $\Longrightarrow$ $\Theta(d)$ samples used to achieve utility and the surplus to achieve privacy.

More generally :

If $p \gg n$, then the test error of the RF model is well approximated by the test error of the corresponding $p = \infty$ kernel method
[Mei, Misiakiewicz, Montanari, 2022].

Formally, recall that $\Phi_{RF,n} = \begin{bmatrix} \Phi_{RF}(x_1) \\ \vdots \\ \Phi_{RF}(x_n) \end{bmatrix} = \begin{bmatrix} \phi^T(Vx_1) \\ \vdots \\ \phi^T(Vx_n) \end{bmatrix} \in \mathbb{R}^{n \times n}$, so that

$$f_{RF}(x, \theta^*) = (\phi(Vx))^T \, \Phi_{RF,n}^T \left(\Phi_{RF,n} \Phi_{RF,n}^T \right)^{-1} y$$

$$= \left[\langle \phi(Vx), \phi(Vx_1) \rangle, \dots, \langle \phi(Vx), \phi(Vx_n) \rangle \right] K_{RF,n}^{-1} \, y$$

with $\quad (K_{RF,n})_{ij} = \langle \phi(Vx_i), \phi(Vx_j) \rangle$.

The corresponding $p = \infty$ kernel estimator is :

$$f_{RF,p=\infty}(x) = \left[\mathbb{E}_v \left[\phi(v^T x) \cdot \phi(v^T x_1) \right], \dots, \mathbb{E}_v \left[\phi(v^T x) \cdot \phi(v^T x_n) \right] \right] K_n^{-1} \, y$$

with $\quad (K_n)_{ij} = \mathbb{E}_v \left[\phi(v^T x_i) \cdot \phi(v^T x_j) \right]$

[Mei, Misiakiewicz, Montanari, 2022] show that, when $p \gtrsim n$,

$$R_{RF} := \mathbb{E}_x \left| f_*(x) - f_{RF}(x, \theta^*) \right|^2 \approx \mathbb{E}_x \left| f_*(x) - f_{RF,p=\infty}(x) \right|^2 =: R_{RF,\, p=\infty}$$

this is the test loss of a kernel ridge(less) regression estimator

Test loss of Kernel Ridge Regression R_{KRR} :

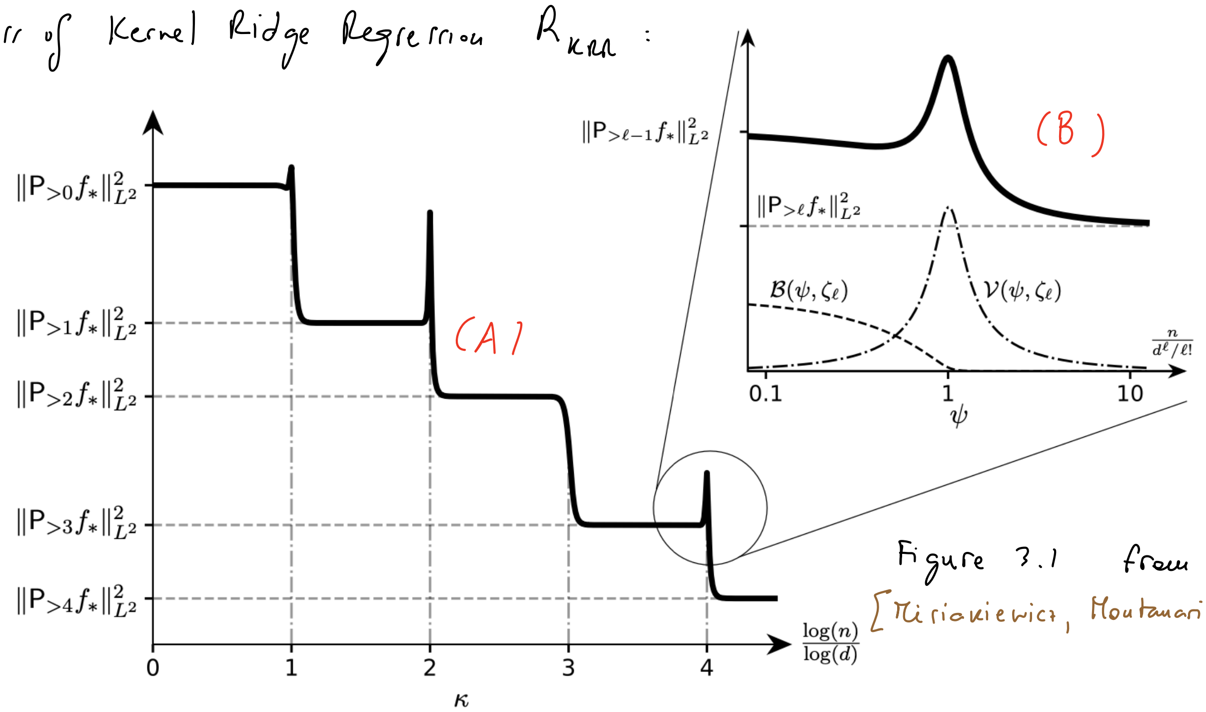


Figure 3.1 from [Misiakiewicz, Montanari, 2024]

(A) If $d^{\ell+\delta} \leq n \leq d^{\ell+1-\delta}$ for some integer ℓ and constant $\delta > 0$, then

$$R_{KRR} = \| P_{>\ell} f_* \|^2_{L^2} + o(1) \left(\| f_* \|^2_{L^2} + \tau^2 \right)$$

we expect the result on DP-GD to hold under this milder condition

If $\underbrace{d^\ell \ll n \ll d^{\ell+1}}$, then the model fits the best degree-ℓ approximation of the target function.

$\longrightarrow$ Proved first by [Ghorbani, Mei, Misiakiewicz, Montanari, 2021] and generalized to any RKHS under a spectral gap assumption in [Mei, Misiakiewicz, Montanari, 2022].

(b) If $\frac{n}{d^\ell \ell!} \longrightarrow \psi$ for some integer ℓ and constant $\psi > 0$, then

$$R_{KRR} = \| \underbrace{P_\ell} f_* \|^2_{L^2} \mathcal{B}(\psi) + \left(\| P_{>\ell} f_* \|^2_{L^2} + \tau^2 \right) \mathcal{V}(\psi) + \| P_{>\ell} f_* \|^2_{L^2}$$
$$+ o(1) \left(\| f_* \|^2_{L^2} + \tau^2 \right)$$

$P_\ell = P_{\leq \ell} P_{>\ell-1}$

⟶ Proved by [Xiao, Hu, Misiakiewicz, Lu, Pennington, 2022]

Proof of this result based on diagonalization of inner-product kernels on the sphere and expansion in spherical harmonics.

To analyze DP-GD, need non-asymptotic control (in n, d, p) of the whole trajectory of the algorithm. We will next discuss the outline of the proof.

PROOF OUTLINE

DP-GD :

$$\vartheta^{t+1} = \vartheta^t - \eta \frac{1}{n} \sum_{i=1}^{n} \nabla_\vartheta \ell(y_i, f(x_i;\vartheta^t)) \Big/ \max\left(1, \frac{\|\nabla_\vartheta \ell(y_i, f(x_i;\vartheta^t))\|_2}{C_{clip}}\right)$$

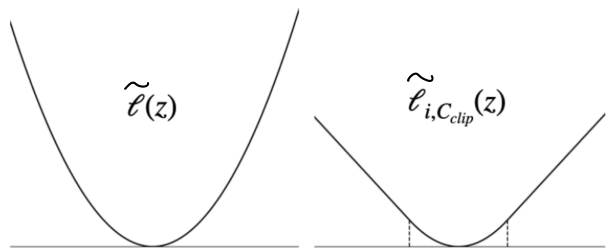
$$+ \sqrt{\eta} \, \frac{2 C_{clip}}{n} \sigma N(0, I_p)$$

$$= \vartheta^t - \eta \nabla_\vartheta \hat{R}_n^{(clip)}(\vartheta^t) + \sqrt{\eta} \, \frac{2 C_{clip}}{n} \sigma N(0, I_p)$$

$$\hat{R}_n^{(clip)}(\vartheta) = \frac{1}{n} \sum_{i=1}^{n} \ell_{i, C_{clip}}(y_i, f(x_i;\vartheta))$$ clipped / Huber loss

$$\ell(y_i, f(x_i, \vartheta)) = \tilde{\ell}(y_i - f(x_i, \vartheta))$$

$$\ell_{i, C_{clip}}(y_i, f(x_i, \vartheta)) = \tilde{\ell}_{i, C_{clip}}(y_i - f(x_i, \vartheta))$$



This is the Euler–Maruyama discretization scheme of the SDE :

$$d\Theta(t) = -\nabla \hat{R}_n^{(clip)}(\Theta(t))\, dt + \overbrace{\frac{2 C_{clip}}{n} \sigma}^{:= \Sigma} \underbrace{dB(t)}_{\text{p-dimensional Wiener process}}$$

By the earlier proposition, if $\Sigma \geq \frac{2 C_{clip}}{n} \sqrt{T} \frac{\sqrt{8 \log(1/\delta)}}{\varepsilon}$, then

$\Theta(T)$ is (ε, δ)-differentially private. It remains to compute the test error of $\Theta(T)$.

Consider now the SDE with the usual quadratic loss $\hat{R}_n(\cdot)$:

$$d\hat{\Theta}(t) = -\nabla \hat{R}_n(\Theta(t))\,dt + \Sigma\, dB(t)$$

$$= -\frac{2\,\Phi_{RF}^{T}}{n}\left(\Phi_{RF}\,\hat{\Theta}(t) - y\right)dt + \Sigma\, dB(t)$$

This is easier to control, as it is an Ornstein-Uhlenbeck (OU) process.

Let $\mathcal{C} := \{ \theta : \quad \| \nabla_\theta \ell(y_i, f(x_i; \theta)) \| < C_{clip} \quad \forall i \}$

(subset of parameter space where clipping does not happen, i.e., $\hat{R}_n(\theta) = \hat{R}_n^{(clip)}(\theta)$

If $\hat{\Theta}(t)$ stays in $\mathcal{C}$, then $\Theta(t) = \hat{\Theta}(t)$.

STEP 1 : $\hat{\Theta}(t) \in \mathcal{C} \ \forall t \in [0,T]$ with $C_{clip} = \sqrt{p} \log^2 n$ and $T = \frac{d}{p} \log^2 n$

*) Computing $\nabla_\theta \ell(\cdot,\cdot)$, we can express $\mathcal{C}$ as

$$\mathcal{C} = \left\{ \theta : \quad | \phi^T(Vx_i)\,\theta - y_i | < \frac{C_{clip}}{2\,\|\phi(Vx_i)\|} \quad \forall i \right\}$$

*) $$\hat{\Theta}(t) = \underbrace{\mathbb{E}_B\left[\hat{\Theta}(t) \right]}_{:= \hat{\theta}(t) \text{ expectation over the OU process giving the usual gradient flow}} + \tilde{\Theta}(t)$$

We show that, with high probability (at least $1 - \exp(-\Omega(\log^2 n)) - \exp(-\Omega(p))$)

(1.A) $$\sup_{t\in[0,T]} | \phi^T(Vx_i)\,\hat{\theta}(t) - y_i | = O(\log n)$$

(1.B) $$\sup_{t\in[0,T]} | \phi^T(Vx_i)\,\tilde{\Theta}(t) | = O(\log n)$$

(1.A) + (1.B) give the desired claim as $$\frac{C_{clip}}{2\|\phi(Vx_i)\|} = \Theta(\log^2 n)$$

PROOF OF (1.A) VIA LEAVE - ONE - OUT

*) $$\hat{\theta}(t) = \left(1 - e^{-2\Phi_{RF}^T \Phi_{RF} \frac{1}{n} t}\right) \underbrace{\Phi_{RF}^{+}}_{\text{Moore - Penrose inverse}} y \qquad (\hat{\Theta}(0) = \hat{\theta}(0) = 0)$$

$$\Downarrow$$

$$y_i - \phi^T(Vx_i)\,\hat{\theta}(t) = \phi^T(Vx_i)\, e^{-2\Phi_{RF}^T \Phi_{RF} \frac{1}{n} t}\, \Phi_{RF}^{+} y$$

Consider the leave-one-out quantities $\Phi_{RF,-i}$ and y_{-i} obtained respectively from Φ_{RF} and y after removing the i-th sample.

*) $$\| \Phi_{RF}^{+} y - \Phi_{RF,-i}^{+} y_{-i} \| = \tilde{O}\left(\frac{1}{\sqrt{p}}\right)$$

GD sensitivity + lower bound on $\lambda_{min}(\Phi_{RF} \Phi_{RF}^T)$

*) Leave-one-out at the exponent :

$$\sup_{t \in [0,T]} \left| \phi^T(Vx_i)\, e^{-2\Phi_{RF}^T \Phi_{RF} \frac{1}{n} t}\, \Phi_{RF,-i}^{+} y_{-i} \right|$$

$$\leq 2 \sup_{t \in [0,T]} \left| \phi^T(Vx_i)\, e^{-2\Phi_{RF,-i}^T \Phi_{RF,-i} \frac{1}{n} t}\, \Phi_{RF,-i}^{+} y_{-i} \right|$$

By Lie's product formula

$$\left| \phi^T(Vx_i)\, e^{-2\Phi_{RF}^T \Phi_{RF} \frac{1}{n} t}\, \Phi_{RF,-i}^{+} y_{-i} \right|$$

$$= \lim_{s \to +\infty} \left| \phi^T(Vx_i) \underbrace{\left(e^{-\frac{2\phi(Vx_i)\phi^T(Vx_i)}{ns} t}\, e^{-2\Phi_{RF,-i}^T \Phi_{RF,-i} \frac{1}{ns} t} \right)^s}_{(:= \Pi(s)} \Phi_{RF,-i}^{+} y_{-i} \right|$$

Expand the r-th power in $\Pi(r) = \left(\left(I + \alpha(r) \frac{\phi(Vx_i)\, \phi^T(Vx_i)}{\|\phi(Vx_i)\|^2} \right) A(s) \right)^s$

$$A(r) = e^{-2\, \Phi_{RF,-i}^T \Phi_{RF,-i} \frac{1}{ns} t}, \qquad \alpha(s) = -1 + e^{-\frac{2\|\phi(Vx_i)\|^2}{ns} t}$$

*) Bound $\sup_{t \in [0,T]} \Big| \underbrace{\phi^T(Vx_i)} \underbrace{e^{-2\, \Phi_{RF,-i}^T \Phi_{RF,-i} \frac{1}{n} t}\, \Phi_{RF,-i}^{+}\, y_{-i}} \Big|$

these two pieces are independent now!

by Dudley's chaining tail inequality

Upper bound the ϵ-covering number $N(\epsilon, \tilde{T})$ of the set $\tilde{T} \subseteq \mathbb{R}^d$ described by the curve $\gamma(t) = V^T e^{-2\, \Phi_{RF,-i}^T \Phi_{RF,-i} \frac{1}{n} t}\, \Phi_{RF,-i}^{+}\, y_{-i}$:

$$\int_0^\infty \sqrt{\log N(\tilde{T}, \epsilon)}\, d\epsilon = O(\log n), \qquad \mathrm{diam}(\tilde{T}) = O(1)$$

Proof of (1.B) by Comparison Inequalities

Consider the auxiliary process $dz_i(t) = \phi^T(Vx_i)\,\Sigma\, dB(t)$.

This is obtained by removing the attractive drift $-2\,\Phi_{RF}^T(\Phi_{RF}\hat{\Theta}(t)-y)/n$ to the SDE, and it is easier to analyze (a Wiener process).

$$\sup_{t\in[0,T]} |\phi^T(Vx_i)\,\tilde{\Theta}(t)| \leq \mathbb{E}_B\left[\sup_{t\in[0,T]} |\phi^T(Vx_i)\,\tilde{\Theta}(t)|\right] + \log n$$

Borell-Tis + variance $= O(1)$

$$\leq \mathbb{E}_{z_i}\left[\sup_{t\in[0,T]} |z_i(t)|\right] + \log n$$

Sudakov-Fernique

(removing the drift increases variance, so process less concentrated around the mean)

$$\leq O(1) + \log n$$

$\mathbb{E}_{z_i}\left[\sup_{t\in[0,T]} |z_i(t)|\right] = O(1)$ since the variance of the Wiener process is $\Sigma^2\, T\, \|\phi(Vx_i)\|^2 = O(1)$.

STEP 2 : Control noise and early stopping in $\hat{\Theta}(t)$

$$\hat{\Theta}(t) = \hat{\vartheta}(t) + \tilde{\Theta}(t) = \overset{\text{(non-private) GD solution}}{\vartheta^*} + \overbrace{(\hat{\vartheta}(t) - \vartheta^*)}^{\text{early stopping}} + \overbrace{\tilde{\Theta}(t)}^{\text{noise}}$$

(2.A) $$\mathbb{E}_x\left[\left(\underbrace{\phi^T(Vx)\,\tilde{\Theta}(T)}_{}\right)^2\right] = \tilde{O}\left(\frac{d^2}{\varepsilon^2 n^2}\right)$$

Gaussian with variance proportional to $\|\phi(Vx)\|^2$, T, Σ^2

so claim follows from choice of hyperparameters

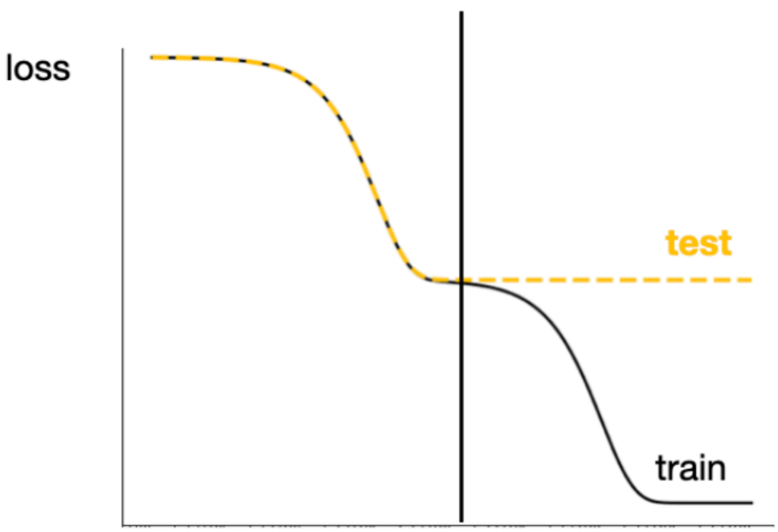
(2.B) $$\mathbb{E}_x\left[\left(\phi^T(Vx)\,(\hat{\vartheta}(T) - \vartheta^*)\right)^2\right] = \tilde{O}\left(\frac{d}{n} + \frac{n}{d^{3/2}}\right)$$

Gap between d-th and $(d+1)$-th eigenvalue of $\Phi_{RF}\Phi_{RF}^T$

$\left(\lambda_d(\Phi_{RF}\Phi_{RF}^T) = \Omega\left(p\frac{n}{d}\right),\ \lambda_{d+1}(\Phi_{RF}\Phi_{RF}^T) = O(p)\right)$

$\Downarrow$

$T = \frac{d}{p}\log^2 n$ is enough time to reach the GD plateau and more time would not help anyway resulting only in overfitting (reduction in training error without improvement in test error)



Formally, we decompose:

$$\phi^T(Vx)\,(\hat{\theta}(T) - \theta^*) = \phi^T(Vx)\,(P_\Lambda + P_\Lambda^\perp)\,(\hat{\theta}(T) - \theta^*)$$

where P_Λ is the projector on the space spanned by the eigenvectors associated to the d largest eigenvalues of $\Phi_{RF}\,\Phi_{RF}^T$.

*) $\| P_\Lambda (\hat{\theta}(T) - \theta^*) \|$ negligible

in that subspace $\hat{\theta}(T)$ already close to convergence despite early stopping

*) $\mathbb{E}_x\left[\left(\phi^T(Vx) \; P_\Lambda^\perp (\hat{\theta}(T) - \theta^*)\right)^2 \right]$

$$\le 2\,\underbrace{\mathbb{E}_x\left[\left((Vx)^T \; P_\Lambda^\perp (\hat{\theta}(T) - \theta^*)\right)^2\right]}_{\tilde{O}\left(d/n + n/d^{3/2}\right) \text{ by bounding } \| V^T P_\Lambda^\perp \Phi_{RF}^+ \|} + 2\,\underbrace{\mathbb{E}_x\left[\left(\tilde{\phi}^T(Vx) \; P_\Lambda^\perp (\hat{\theta}(T) - \theta^*)\right)^2\right]}_{\tilde{O}\left(d/n + n/d^{3/2}\right) \text{ by bounding } \| \mathbb{E}_x [\tilde{\phi}(Vx)\tilde{\phi}^T(Vx)] \|_{op}}$$

$\phi(z) = z + \tilde{\phi}(z)$

(2.A) + (2.B) give that

$$|\underbrace{\hat{R}}_{\text{test error of } \hat{\Theta}(T)} - \underbrace{R^*}_{\text{test error of } \theta^*}| = \tilde{O}\left(\frac{d}{n\varepsilon} + \sqrt{\frac{d}{n}} + \sqrt{\frac{n}{d^{3/2}}} \right)$$

Step 1 gives that $\hat{\Theta}(T) = \Theta(T)$ concluding the proof. ☒

EXTRA :

Sudakov- Fernique's inequality

<u>THEOREM</u> Let $(X_t)_{t\in T}$ and $(Y_t)_{t\in T}$ be two mean zero Gaussian processes. Assume that for all $s,t \in T$, $\mathbb{E}(X_t - X_s)^2 \leq \mathbb{E}(Y_t - Y_s)^2$. Then,

$$\mathbb{E} \sup_{t\in T} X_t \leq \mathbb{E} \sup_{t\in T} Y_t$$

Tsirelson, Ibragimov, Sudakov

Borell -TIS inequality

<u>THEOREM</u> Let $(X_t)_{t\in T}$ be a mean zero Gaussian process which is almost surely finite. Let $\sigma_T^2 := \sup_{t\in T} \mathbb{E} X_t^2$. Then, σ_T^2 and $\mathbb{E} \sup_{t\in T} |X_t|$ are finite and, for all $u>0$,

$$\mathbb{P}\left(\sup_{t\in T} |X_t| > \mathbb{E} \sup_{t\in T} |X_t| + u \right) \leq \exp\left(- \frac{u^2}{2\sigma_T^2} \right)$$

DP-SGD WITH CLIPPING : SETTING

→ So far, clipping treated as a nuisance (we set C_{clip} so that clipping never happens with high probability)

→ Empirical evidence that aggressive clipping helps performance

We discuss a sharp analysis that captures the effect of clipping

*) Linear regression

$$y_i = x_i^T \theta^* + z_i \ , \qquad x_i \sim N(0, \Sigma) \ , \qquad z_i \sim N(0, \gamma^2) \ , \qquad z_i \perp x_i$$

→ input dimension d = # parameters p

→ $Tr(\Sigma) = d$, $\quad \lambda_{max}(\Sigma) / \lambda_{min}(\Sigma) = \Theta(1)$ (well-conditioned covariance)

→ $\|\theta^*\| = \Theta(1)$

*) Proportional regime

→ n, d large with $d/n \to \gamma \in (0, \infty)$

*) (One-pass) Differentially-private stochastic gradient descent (DP-SGD)

$$\theta^{k+1} = \theta^k - \eta_k \nabla_\theta \, \ell(y_{k+1}, f(x_{k+1}; \theta^k)) \Big/ \max\left(1, \frac{\|\nabla_\theta \ell(y_{k+1}, f(x_{k+1}; \theta^k))\|_2}{C_{clip}}\right)$$
$$+ 2 \, C_{clip} \, \sigma_k \, N(0, I_d)$$

Square loss : $\ell(y, \hat{y}) = \frac{1}{2} (y - \hat{y})^2$

Linear model : $f(x; \theta) = x^T \theta$

Test error : $R(\theta) := R(f(x;\theta)) = \frac{1}{2} \mathbb{E}_{(x,y)} | y - f(x;\theta)|^2 = \frac{1}{2} \mathbb{E}_{x} | x^T (\theta^* - \theta) + z |^2$

$$= \frac{1}{2} \left(\|\Sigma^{1/2} (\theta^* - \theta)\|^2 + \gamma^2 \right)$$

→ # samples n = # iterations of the algorithm T

→ Each data point used only once

→ $C_{clip} = c\sqrt{d}$ ← constant independent of n, d

Before, $C_{clip} = \sqrt{p}\log^2 n = \sqrt{d}\log^2 n$ so that clipping never happens w.h.p.

$p = d$ in linear regression

Here, $C_{clip} \sim \sqrt{d}$ and $\| \nabla_\vartheta \ell(y, f(x;\vartheta))\| = \|x\| \, |x^T\vartheta - y| \sim \sqrt{d}$

so that clipping is frequent.

*) <u>zero-concentrated DP (zCDP)</u>

<u>DEFINITION</u> (ρ-zCDP) [Bun, Steinke, 2016] Given $\alpha \in (1, +\infty)$ and two random variables X and X' with laws p_X and $p_{X'}$, their α-Renyi divergence is

$$D_\alpha(X \| X') = \frac{1}{\alpha - 1} \ln \int \left(\frac{p_X(\vartheta)}{p_{X'}(\vartheta)} \right)^\alpha p_{X'}(\vartheta)\, d\vartheta .$$

Then, a randomized algorithm A satisfies <u>ρ-zero concentrated differential privacy</u> if, for any pair of adjacent datasets D, D' and any $\alpha \in (1, \infty)$, we have $D_\alpha(A(D) \| A(D')) \leq \alpha\rho$.

$$\rho^2/2\text{-zCDP} \implies \left(\rho^2/2 + \rho\sqrt{2\ln(1/\delta)}, \delta\right)\text{-DP} \quad \text{[Bun, Steinke, 2016]}$$

$\rho = \Theta(1) \not\Rightarrow \varepsilon = \Theta(1), \ \delta \ll 1/n$ so ρ-zCDP with $\rho = \Theta(1)$ is a relaxed notion of privacy

<u>PROPOSITION</u> (following [Feldman, Koren, Talwar, 2020]) The output ϑ_n of DP-SGD is $(\rho^2/2)$-zCDP, where

$$\rho = \max_{k \in \{1, \dots, n\}} \frac{\eta_k}{\sqrt{\sum_{j=k}^{n} \sigma_j^2}}$$

$\longrightarrow$ Each sample x_k protected by overall noise introduced in next updates $\sum_{j=k}^{n} \sigma_j^2$

$\longrightarrow$ Proof based on privacy amplification by iteration (rather than advanced composition)

$\longrightarrow$ Minimize noise introduced by the algorithm via the schedule:

$$\eta_k = \rho \sqrt{\sum_{j=k}^{n} \sigma_j^2} \iff \rho^2 \sigma_k^2 = \begin{cases} \eta_k^2 - \eta_{k+1}^2 & k \in \{1, \dots, n-1\} \\ \eta_k^2 & k = n \end{cases}$$

Idea: Track the test error of DP-SGD via an SDE in parameter space. This gives a <u>deterministic equivalent</u> of the (stochastic) dynamics since the SDE is equivalent to a system of d coupled ODEs.

<u>DEFINITION</u> (Homogeneized DP-SGD). For any $t \in [0,1)$, we define the homogeneized DP-SGD (H-DP-SGD) as the solution of the SDE:

$$d\Theta_t = -\tilde{\eta}(t)\, \mu_c(\Theta_t) \nabla R(\Theta_t)\, dt + \tilde{\eta}(t) \sqrt{\frac{2\nu_c(\Theta_t) R(\Theta_t) \Sigma}{n}}\, dB_t^s + 2\frac{\sqrt{d}}{n} c\, \tilde{\sigma}(t)\, dB_t^p$$

$\longrightarrow$ $\Theta_0 = 0$

$\longrightarrow$ B_t^s, B_t^p are two independent Wiener processes

$\longrightarrow$ $\tilde{\eta}(t)$ is the continuous version of the discrete step size η_k $\left(\eta_k = \frac{\tilde{\eta}(k/n)}{n}\right)$

(both $\tilde{\eta}^2(\cdot)$ and the absolute value of its first and second derivative are required to be uniformly bounded from above by a constant independent of n, d)

$\longrightarrow$ $\tilde{\sigma}(t)$ is the continuous version of the noise standard deviation σ_k so that $\rho^2 \sigma_k^2 = \eta_k^2 - \eta_{k+1}^2$ becomes $\rho^2 \tilde{\sigma}^2(t) := -\frac{d}{dt} \tilde{\eta}^2(t)$

$\longrightarrow$ $R(\theta) = \frac{1}{2} \mathbb{E} \,|x^T \theta - y|^2$ is the risk

$\longrightarrow \quad \mu_c(\vartheta) = \dfrac{\left\| \mathbb{E}_{(x,y)} \left[r_c(\vartheta, x, y)\, x \right] \right\|}{\left\| \mathbb{E}_{(x,y)} \left[r(\vartheta, x, y)\, x \right] \right\|}$ is the descent reduction factor

effect of clipping

with $r(\vartheta, x, y) = x^T \vartheta - y$ the residual in ϑ and

$r_c(\vartheta, x, y) = r(\vartheta, x, y) \,/\, \max\left(1, \dfrac{|r(\vartheta, x, y)|}{c}\right)$ the clipped residual

$\longrightarrow \quad \nu_c(\vartheta) = \dfrac{\mathbb{E}\left[r_c^2(\vartheta, x, y) \right]}{\mathbb{E}\left[r^2(\vartheta, x, y) \right]}$ is the variance reduction factor

effect of clipping

DP-SGD with Clipping: Sharp Analysis

Theorem [Bombari, Seroussi, M., 2025] Assume $\rho = \Theta(1)$ and $\sup_{t \in [0,1]} \tilde{\eta}(t) < 2/\gamma$.
Let Θ_t and θ_k be independent realizations of H-DP-SGD and DP-SGD.
Then, with high probability,

$$\sup_{t \in [0,1)} \left| R(\Theta_t) - R(\theta^{\lfloor tn \rfloor}) \right| = O\left(\frac{\log^2 n}{\sqrt{n}} \right).$$

*) $\sup_{t \in [0,1]} \tilde{\eta}(t) < 2/\gamma$ is a stability condition for SGD (can be relaxed)

Interpretation The test error of DP-SGD is well approximated by

$$d\Theta_t = \underbrace{- \tilde{\eta}(t) \overbrace{\mu_c(\Theta_t)}^{\text{effect of clipping}} \nabla R(\Theta_t)\, dt}_{(\mathrm{I})} + \underbrace{\tilde{\eta}(t) \sqrt{\frac{2 \overbrace{\nu_c(\Theta_t)}^{\text{effect of clipping}} R(\Theta_t)\, \Sigma}{n}}\, dB_t^s}_{(\mathrm{II})} + \underbrace{2 \frac{\sqrt{d}}{n} c\, \tilde{\sigma}(t)\, dB_t^p}_{(\mathrm{III})}$$

(I) = term corresponding to descent towards the minimizer of $R(\cdot)$

(II) = term corresponding to noise inherent in SGD

(III) = term corresponding to noise needed to enforce DP

SDE $\longrightarrow$ System of ODEs

SDE equivalent to the following deterministic system of d coupled ODEs:

$$dD_i = -2\lambda_i \tilde{\eta}(t)\mu_c(R(t)) D_i \, dt + \lambda_i \tilde{\eta}^2(t)\nu_c(R(t)) \left(R(t) + \zeta^2/2\right)\gamma \, dt + 2c^2 \tilde{\sigma}^2(t)\gamma^2 dt \qquad (\star)$$

where $R(t) = \frac{1}{d}\sum_{i=1}^{d} \lambda_i D_i(t)$ and $\{\lambda_i\}_{i=1}^{d}$ are the eigenvalues of Σ,

so that

$$\sup_{t\in[0,1)} \left| R(t) - R^{exc}(\theta^{\lfloor tn\rfloor}) \right| = O\left(\frac{\log^2 n}{\sqrt{n}}\right).$$

$$R^{exc}(\theta) = R(\theta) - \zeta^2/2$$

excess error / noiseless test risk such that $R^{exc}(\theta^*) = 0$.

It is easier (and already insightful!) to work with an upper and a lower bound:

$\underline{R}(t) \leq R(t) \leq \overline{R}(t)$, where $\overline{R}(t), \underline{R}(t) : [0,1] \longrightarrow \mathbb{R}$ are the unique solutions to the following two (decoupled) ODEs:

$$d\overline{R}(t) = -2\lambda_{min}\tilde{\eta}(t)\mu_c(\overline{R})\overline{R}\, dt + \lambda_{max}\tilde{\eta}^2(t)\nu_c(\overline{R})(\overline{R} + \zeta^2/2)\gamma\, dt + 2c^2\tilde{\sigma}^2(t)\gamma^2 dt$$

$$d\underline{R}(t) = -2\lambda_{max}\tilde{\eta}(t)\mu_c(\underline{R})\underline{R}\, dt + \tilde{\eta}^2(t)\nu_c(\underline{R})(\underline{R} + \zeta^2/2)\gamma\, dt + 2c^2\tilde{\sigma}^2(t)\gamma^2 dt$$

where $\overline{R}(0) = \underline{R}(0) = \|\Sigma^{1/2}\theta^*\|^2/2$ and $\lambda_{max/min} := \lambda_{max/min}(\Sigma)$.

REMARK: $\overline{R}(t) = \underline{R}(t)$ when $\Sigma = I$ (all d ODEs in $(\star)$ decouple and coincide)

Idea of the argument : Doob's decomposition of the stochastic processes given by DP-SGD and H-DP-SGD allows to write any quadratic function of the parameters as sum of predictable component + (vanishing in n) martingale

Equivalence between high-dimensional dynamics of SGD and a "homogenized" SDE proved in a sequence of works :

→ Least squares (no clipping, no private noise)
[Paquette, Paquette, Adlam, Pennington, 2022 ;
Paquette, Paquette, Adlam, Pennington, 2024]

→ GLMs and multi-index models (no clipping, no private noise)
[Collins-Woodfin, Paquette, Paquette, Seroussi, 2024]

→ Clipped dynamics (no private noise)
[Marshall, Xiao, Agarwala, Paquette, 2025]

To give some intuition about the terms (I) and (II) in the <u>INTERPRETATION</u>, we follow Example 9 in [Paquette, 2023] and consider the simplest possible setting :

*) no DP noise ($\sigma_n = 0$)

*) no clipping ($\mu_c \equiv 1$, $\nu_c \equiv 1$)

*) isotropic data ($\Sigma = I$)

In this case, the one-pass SGD iteration reads :

$$\theta^{k+1} = \theta^k - \eta_k \nabla_{\theta^k} \frac{1}{2}\left(x_{k+1}^T \theta^k - y_{k+1}\right)^2$$

$$= \theta^k - \eta_k \nabla_{\theta^k} \frac{1}{2}\left(x_{k+1}^T (\theta^k - \theta^*) - z_{k+1}\right)^2$$

$y_{k+1} = x_{k+1}^T \theta^* + z_{k+1}$

$$= \theta^k - \eta_k \left(x_{k+1}^T (\theta^k - \theta^*) - z_{k+1}\right) x_{k+1}$$

We now compute the first two moments of the gradient $\left(x_{k+1}^T(\theta^k-\theta^*) - z_{k+1}\right) x_{k+1}$:

$$*)\quad \mathbb{E}_{(x_{k+1},y_{k+1})}\left[\left(x_{k+1}^T(\theta^k-\theta^*) - z_{k+1}\right) x_{k+1}\right] = \mathbb{E}\left[x_{k+1}\, x_{k+1}^T(\theta^k-\theta^*)\right] = \theta^k-\theta^* = \nabla R(\theta^k)$$

$z_{k+1} \perp x_{k+1}$ and $\mathbb{E}[z_{k+1}] = 0$ $\qquad$ $\mathbb{E}\, x_{k+1} x_{k+1}^T = \Sigma = I$

Thus, the mean of the gradient gives (I).

$$*)\quad \mathbb{E}_{(x_{k+1},y_{k+1})}\left[\left(x_{k+1}^T(\theta^k-\theta^*) - z_{k+1}\right)^2 x_{k+1}\, x_{k+1}^T\right]$$

$$= \left(\mathbb{E}_{(x_{k+1},y_{k+1})}\left[\left(x_{k+1}^T(\theta^k-\theta^*) - z_{k+1}\right)^2\right] + \sigma^2\right) I + 2(\theta^k-\theta^*)(\theta^k-\theta^*)^T$$

some manipulations using Wick rule

$$= 2R(\theta^k)\, I + \underbrace{2(\theta^k-\theta^*)(\theta^k-\theta^*)^T}_{\text{lower order correction}} \approx 2R(\theta^k)\, I$$

Thus, the covariance of the gradient gives something that looks like (II).

Let us now compute

$$\mathbb{E}_{(x_{k+1},y_{k+1})}\left[R^{exc}(\theta^{k+1}) - R^{exc}(\theta^k)\right] = \mathbb{E}_{(x_{k+1},y_{k+1})}\left[\frac{1}{2}\|\theta^{k+1}-\theta^*\|^2 - \frac{1}{2}\|\theta^k-\theta^*\|^2\right]$$

$$= \mathbb{E}_{(x_{k+1},y_{k+1})}\left[\frac{1}{2}\left(\|\theta^{k+1}-\theta^k\|^2 + 2\langle\theta^{k+1}-\theta^k, \theta^k-\theta^*\rangle\right)\right]$$

$$= \frac{1}{2}\gamma_k^2 \overbrace{\mathbb{E}_{(x_{k+1},y_{k+1})}\left[\left(x_{k+1}^T(\theta^k-\theta^*) - z_{k+1}\right)^2 \|x_{k+1}\|^2\right]}^{Tr\left[2R(\theta^k) I + 2(\theta^k-\theta^*)(\theta^k-\theta^*)^T\right]}$$

$$+ \underbrace{\mathbb{E}_{(x_{k+1},y_{k+1})}\left[-\gamma_k\left(x_{k+1}^T(\theta^k-\theta^*) - z_{k+1}\right) x_{k+1}^T\right]}_{(\theta^k-\theta^*)^T}(\theta^k-\theta^*)$$

$$= -2\eta_k R^{exc}(\vartheta^k) + \eta_k^2 \left(d R^{exc}(\vartheta^k) + d \xi^2/2 + 2 R^{exc}(\vartheta^k) \right)$$

$$= -2\eta_k R^{exc}(\vartheta^k) + \eta_k^2 \gamma \left(n R^{exc}(\vartheta^k) + n \xi^2/2 + \underbrace{\frac{2}{\gamma} R^{exc}(\vartheta^k)}_{} \right)$$

can be neglected as the other terms in the parenthesis are of order n

Recalling $\eta_k = \frac{\tilde{\eta}(k/n)}{n}$ and setting $\rho(t) = \lim_{n \to +\infty} \mathbb{E}\left[R^{exc}(\vartheta^{\lfloor tn \rfloor}) \right]$, we have that the equation above is the Euler-Maruyama discretization of

$$d\rho = -2\tilde{\eta}(t)\rho\, dt + \tilde{\eta}^2(t)\left(\rho + \xi^2/2\right)\gamma\, dt$$

this corresponds to the ODE for $R(t)$ when $\Sigma = I$, $\mu_c \equiv 1$, $\nu_c \equiv 1$, $\tilde{\nabla} \equiv 0$

Already when $\Sigma \neq I$ (still no clipping and no DP noise), the SGD dynamics depends on unboundedly many statistics — and not just the first two moments as above (see Example 13 in [Paquette, 2023]). Thus, we need the framework based on homogenized SGD (and Doob's decomposition).

Noise in the last iteration

REMARK: sup is taken over $t \in [0,1)$ so equivalence does not hold for the very last iterate ϑ^n which gives the private output of the algorithm.

For $k < n$,

$$\rho^2 \sigma_k^2 = -\frac{d}{dk} \eta_k^2 \approx -\frac{1}{n} \frac{d}{d(k/n)} \frac{\tilde{\eta}^2(k/n)}{n^2} = \frac{\rho^2}{n^3} \tilde{\sigma}^2(k/n)$$

For $k = n$,

$$\rho^2 \sigma_n^2 = \eta_n^2 = \frac{\tilde{\eta}^2(1)}{n^2}$$

additional n factor at the denominator on top

This implies that, if $\hat{\eta}(1) > 0$, the noise added to the last iterate is much larger. After a few calculations, the final loss can be shown to be given by

$$\left| R(\vartheta^n) - R(\vartheta^{n-1}) - \frac{2c^2 \tilde{\eta}^2(1)\gamma^2}{\rho^2} \right| = O\left(\frac{\log n}{\sqrt{n}}\right)$$

OPTIMAL RATES & BENEFIT OF CLIPPING

To study DP-SGD, we can now analyze H-DP-SGD or, even easier, the ODEs.

*) We stick to $\Sigma = I$ (see [Bombari, Serusi, M., 2025] for tracking the dependence on $\lambda_{max}/\lambda_{min}$ in the bounds).

*) We work with $\gamma = \frac{d}{n} \longrightarrow 0$ and use $o_\gamma(\cdot)$, $O_\gamma(\cdot)$, $\Omega_\gamma(\cdot)$.

*) We consider three schedules (corresponding to different choices of learning rate $\tilde{\eta}(t)$ and noise standard deviation $\tilde{\sigma}(t)$):

ⓐ Output perturbation $\quad \tilde{\eta}(t) = \tilde{\eta}(0)\,, \quad \tilde{\sigma}(t) = 0$

Learning rate is fixed during all training and private noise added only at the end of the algorithm.

→ Corresponding DP-SGD output denoted by θ^n_{OUT}

ⓑ Constant noise $\quad \tilde{\eta}(t) = \tilde{\eta}(0)\sqrt{1-t}\,, \quad \tilde{\sigma}(t) = \frac{1}{\rho^2}\tilde{\eta}^2(0)$

Linearly decaying $\tilde{\eta}^2(t)$ corresponds to constant level of noise

→ Corresponding DP-SGD output denoted by θ^n_{CNST}

ⓒ Optimal schedule $\quad \tilde{\eta}(t) = \dfrac{c_1}{t + \gamma \max(c_2, c_3/\rho)}$

(For some explicit constants c_1, c_2, c_3 independent of γ, ρ)

This is OPTIMAL in the sense that it gives the optimal decay rate of the excess error (as a function of γ).

→ Corresponding DP-SGD output denoted by θ^n_{OPT}

on arXiv soon...

THEOREM [Bombari, Seroussi, M., 2026] Let $\rho = \Omega_\gamma(\gamma^{1-h})$ for some $h>0$.

① Pick $c = O_\gamma(1)$ and $\tilde{\eta}(0)c = C\log(1/\gamma)$ for a large enough constant C (independent of γ). Then, with high probability,

$$R^{exc}(\theta^n_{OUT}) = O_\gamma\left(\gamma\log(1/\gamma) + \frac{\gamma^2\log^2(1/\gamma)}{\rho^2}\right),$$

$$R^{exc}(\theta^n_{CNST}) = O_\gamma\left(\gamma\log^{2/3}(1/\gamma) + \frac{\gamma^2\log^{4/3}(1/\gamma)}{\rho^2}\right).$$

Furthermore, a lower bound of the same order holds for any choice of hyperparameters $c, \tilde{\eta}(0)$.

② Pick c to be a suitable constant (independent of γ). Then, with high probability,

$$R^{exc}(\theta^n_{OPT}) = O_\gamma\left(\gamma + \frac{\gamma^2}{\rho^2}\right).$$

Furthermore, a minimax lower bound of the same order holds: let θ^n be the solution found by a generic algorithm $\mathcal{M}$ belonging to the set of all $\rho^2/2$-zCDP algorithms $\mathbb{M}$; then, we have

$$\inf_{\mathcal{M}\in\mathbb{M}} \sup_{\|\theta^*\|<1} \mathbb{E}\left[R^{exc}(\theta^n)\right] = \Omega_\gamma\left(\gamma + \frac{\gamma^2}{\rho^2}\right).$$

INTERPRETATION

*) By properly tuning the learning rate $\tilde{\eta}(t)$ (and, consequently, the noise schedule $\tilde{\sigma}(t)$), DP-SGD achieves minimax optimal rates.

*) The sharp analysis of DP-SGD allows to identify optimal hyperparameters. In particular, aggressive clipping ($c = O_\gamma(1)$) leads to good performance: if $c = \omega_\gamma(1)$, then

$$R^{exc}(\theta^n_{OUT}) = \Omega_\gamma\left(\gamma \log(1/\gamma) + \underbrace{\max(1, c^2)} \frac{\gamma^2 \log^2(1/\gamma)}{\rho^2} \right)$$

penalty paid for using large clipping constant

$$R^{exc}(\theta^n_{CNST}) = \Omega_\gamma\left(\gamma \log^{2/3}(1/\gamma) + \overbrace{\max(1, c^2)} \frac{\gamma^2 \log^{4/3}(1/\gamma)}{\rho^2} \right)$$

PROOF IDEAS

*) Bounds on θ^n_{OUT}, θ^n_{CNST} proved by analyzing the corresponding ODEs for $R(t)$

*) Minimax lower bound obtained by extending the analysis of [Cai, Wang, Zhang, 2021] for (ε, δ)-DP to zCDP

*) Optimal schedule obtained by crafting $\hat{\eta}(t)$ in order to match the minimax lower bound

[OPEN PROBLEM] Sharp analysis of clipped dynamics for non-linear models

References

[Abadi, Chu, Goodfellow, McMahan, Mironov, Talwar, Zhang, 2016] Martin Abadi, Andy Chu, Ian Goodfellow, H. Brendan McMahan, Ilya Mironov, Kunal Talwar, and Li Zhang. “Deep Learning with Differential Privacy”. In: *ACM SIGSAC Conference on Computer and Communications Security.* 2016.

[Adamczak, Litvak, Pajor, Tomczak-Jaegermann, 2011] Radoslaw Adamczak, Alexander E Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. “Restricted isometry property of matrices with independent columns and neighborly polytopes by random sampling”. In: *Constructive Approximation*, 2011.

[Bartlett, Montanari, Rakhlin, 2021] Peter L Bartlett, Andrea Montanari, and Alexander Rakhlin. “Deep learning: a statistical viewpoint”. In: *Acta numerica*, 2021.

[Bassily, Feldman, Talwar, Thakurta, 2019] Raef Bassily, Vitaly Feldman, Kunal Talwar, and Abhradeep Thakurta. “Private Stochastic Convex Optimization with Optimal Rates”. In: *Advances in Neural Information Processing Systems.* 2019.

[Bassily, Smith, Thakurta, 2014] Raef Bassily, Adam Smith, and Abhradeep Thakurta. “Private Empirical Risk Minimization: Efficient Algorithms and Tight Error Bounds”. In: *2014 IEEE 55th Annual Symposium on Foundations of Computer Science.* 2014.

[Bombari, Amani, Mondelli, 2022] Simone Bombari, Mohammad Hossein Amani, and Marco Mondelli. “Memorization and optimization in deep neural networks with minimum over-parameterization”. In: *Advances in Neural Information Processing Systems*, 2022.

[Bombari, Kiyani, Mondelli, 2023] Simone Bombari, Shayan Kiyani, and Marco Mondelli. “Beyond the Universal Law of Robustness: Sharper Laws for Random Features and Neural Tangent Kernels”. In: *International Conference on Machine Learning (ICML).* 2023.

[Bombari, Mondelli, 2025] Simone Bombari and Marco Mondelli. “Privacy for free in the overparameterized regime”. In: *Proceedings of the National Academy of Sciences*, 2025.

[Bombari, Seroussi, Mondelli, 2025] Simone Bombari, Inbar Seroussi, and Marco Mondelli. “Better rates for private linear regression in the proportional regime via aggressive clipping”. In: *arXiv preprint arXiv:2505.16329*, 2025.

[Bubeck, Sellke, 2021] Sebastien Bubeck and Mark Sellke. “A Universal Law of Robustness via Isoperimetry”. In: *Advances in Neural Information Processing Systems (NeurIPS).* 2021.

[Bun, Steinke, 2016] Mark Bun and Thomas Steinke. “Concentrated differential privacy: Simplifications, extensions, and lower bounds”. In: *Theory of cryptography conference*. Springer. 2016, pp. 635–658.

[Cai, Wang, Zhang, 2021] T. Tony Cai, Yichen Wang, and Linjun Zhang. “The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy”. In: *The Annals of Statistics*, 2021.

[Chizat, Oyallon, Bach, 2019] Lenaic Chizat, Edouard Oyallon, and Francis Bach. “On lazy training in differentiable programming”. In: *Neural Information Processing Systems (NeurIPS)*. Vol. 32. 2019.

[Collins-Woodfin, Paquette, Paquette, Seroussi, 2024] Elizabeth Collins-Woodfin, Courtney Paquette, Elliot Paquette, and Inbar Seroussi. “Hitting the High-dimensional notes: an ODE for SGD learning dynamics on GLMs and multi-index models”. In: *Information and Inference: A Journal of the IMA*, 2024.

[Cover, 1965] Thomas M. Cover. “Geometrical and statistical properties of systems of linear inequalities with applications in pattern recognition”. In: *IEEE Transactions on Electronic Computers*, 1965.

[Du, Zhai, Poczos, Singh, 2019] Simon S. Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. “Gradient Descent Provably Optimizes Over-parameterized Neural Networks”. In: *International Conference on Learning Representations (ICLR)*. 2019.

[Dwork, McSherry, Nissim, Smith, 2006] Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. “Calibrating Noise to Sensitivity in Private Data Analysis”. In: *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006*. 2006.

[Dwork, Roth, 2014] Cynthia Dwork and Aaron Roth. “The algorithmic foundations of differential privacy”. In: *Foundations and Trends in Theoretical Computer Science*, 2014.

[Feldman, Koren, Talwar, 2020] Vitaly Feldman, Tomer Koren, and Kunal Talwar. “Private stochastic convex optimization: optimal rates in linear time”. In: *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*. STOC 2020. 2020. ISBN: 9781450369794.

[Ghorbani, Mei, Misiakiewicz, Montanari, 2021] Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. “Linearized two-layers neural networks in high dimension”. In: *The Annals of Statistics*, 2021.

[Iurada, Bombari, Tommasi, Mondelli, 2026] Leonardo Iurada, Simone Bombari, Tatiana Tommasi, and Marco Mondelli. “A Law of Data Reconstruction for Random Features (and Beyond)”. In: 2026.

[Jacot, Gabriel, Hongler, 2018] Arthur Jacot, Franck Gabriel, and Clément Hongler. “Neural tangent kernel: Convergence and generalization in neural networks”. In: *Neural Information Processing Systems (NeurIPS)*. 2018.

[Kifer, Smith, Thakurta, 2012] Daniel Kifer, Adam Smith, and Abhradeep Thakurta. "Private Convex Empirical Risk Minimization and High-dimensional Regression". In: *Conference on Learning Theory*. 2012.

[Madry, Makelov, Schmidt, Tsipras, Vladu, 2018] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *International Conference on Learning Representations (ICLR)*. 2018.

[Marshall, Xiao, Agarwala, Paquette, 2025] Noah Marshall, Ke Liang Xiao, Atish Agarwala, and Elliot Paquette. "To Clip or not to Clip: the Dynamics of SGD with Gradient Clipping in High-Dimensions". In: *The Thirteenth International Conference on Learning Representations*. 2025.

[Mei, Misiakiewicz, Montanari, 2022] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. "Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration". In: *Applied and Computational Harmonic Analysis*, 2022.

[Misiakiewicz, Montanari, 2024] Theodor Misiakiewicz and Andrea Montanari. "Six lectures on linearized neural networks". In: *Journal of Statistical Mechanics: Theory and Experiment*, 2024.

[Montanari, Zhong, 2022] Andrea Montanari and Yiqiao Zhong. "The interpolation phase transition in neural networks: Memorization and generalization under lazy training". In: *The Annals of Statistics*, 2022.

[Nguyen, Mondelli, Montufar, 2021] Quynh Nguyen, Marco Mondelli, and Guido Montufar. "Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep ReLU networks". In: *International Conference on Machine Learning (ICML)*. 2021.

[Oymak, Soltanolkotabi, 2019] Samet Oymak and Mahdi Soltanolkotabi. "Overparameterized nonlinear learning: Gradient descent takes the shortest path?" In: *International Conference on Machine Learning (ICML)*. 2019.

[Paquette, Paquette, Adlam, Pennington, 2024] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. "Homogenization of SGD in high-dimensions: exact dynamics and generalization properties". In: *Mathematical Programming*, 2024.

[Paquette, Paquette, Adlam, Pennington, 2022] — . "Implicit Regularization or Implicit Conditioning? Exact Risk Trajectories of SGD in High Dimensions". In: *Advances in Neural Information Processing Systems*. 2022.

[Paquette, 2023] Elliot Paquette. *High-dimensional limits of stochastic gradient descent*. Lecture notes. 2023.

[Rahimi, Recht, 2007] Ali Rahimi and Benjamin Recht. "Random features for large-scale kernel machines". In: *Advances in Neural Information Processing Systems*, 2007.

[Soltanolkotabi, Javanmard, Lee, 2018] Mahdi Soltanolkotabi, Adel Javanmard, and Jason D Lee. "Theoretical insights into the optimization landscape of over-parameterized shallow neural networks". In: *IEEE Transactions on Information Theory*, 2018.

[Szegedy, Zaremba, Sutskever, Bruna, Erhan, Goodfellow, Fergus, 2014] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. "Intriguing properties of neural networks". In: *International Conference on Learning Representations (ICLR)*. 2014.

[Xiao, Hu, Misiakiewicz, Lu, Pennington, 2022] Lechao Xiao, Hong Hu, Theodor Misiakiewicz, Yue Lu, and Jeffrey Pennington. "Precise learning curves and higher-order scalings for dot-product kernel regression". In: *Advances in Neural Information Processing Systems*, 2022.