A Statistical Learning Theory for Models with High Complexity

Lijia Zhou

Department of Statistics, University of Chicago

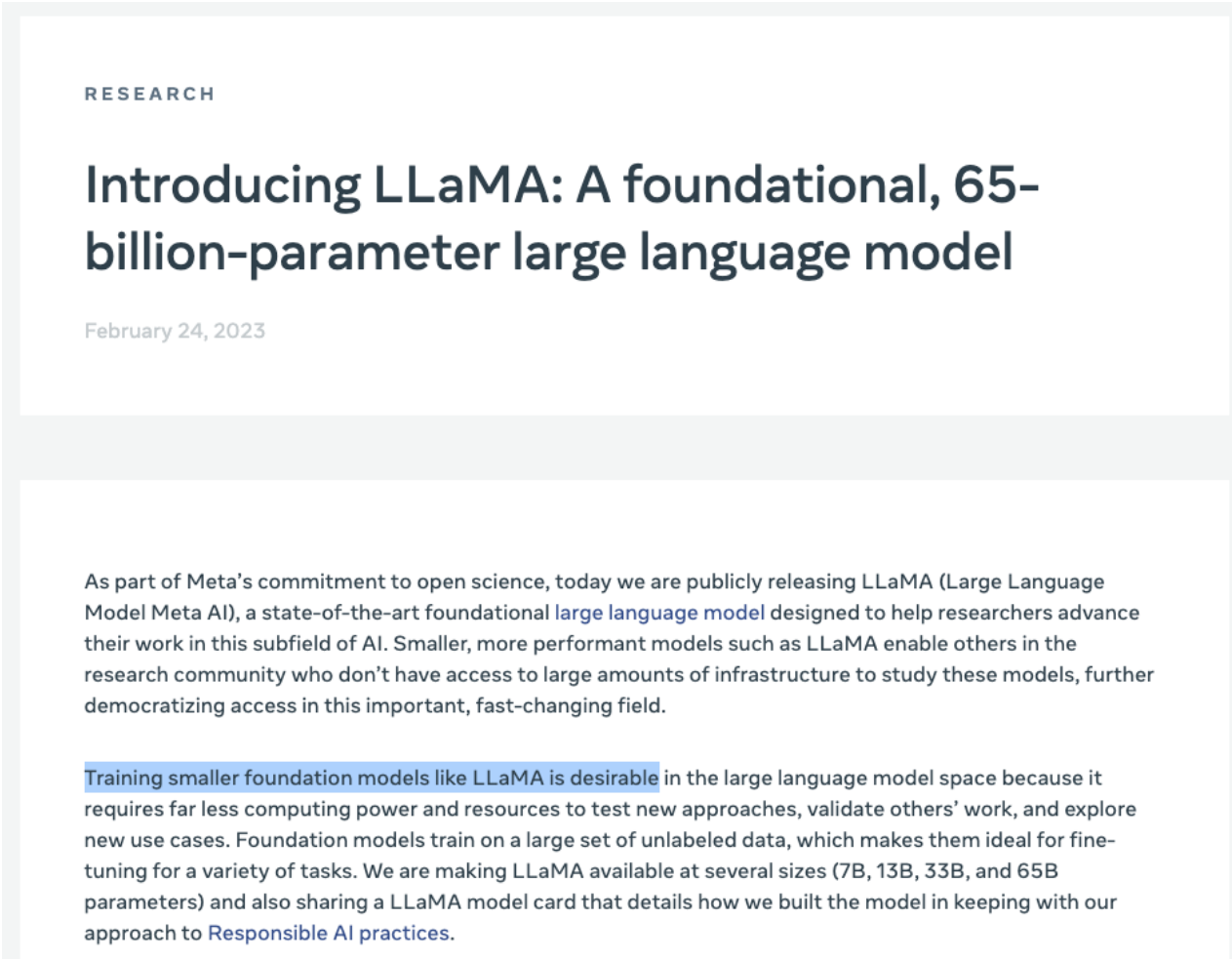
Advisor: Nati Srebro (TTIC)
Committee: Chao Gao (UChicago), Rina Foygel Barber (UChicago)

Joint work with Frederic Koehler (Stanford), Danica Sutherland (UBC), Pragya Sur (Harvard), Zhen Dai (UChicago), Jamie Simon (UC Berkeley), Gal Vardi (TTIC/Hebrew University)

April 17, 2023

Motivation

- modern ML models are becoming incredibly larger!
- 7-billion-parameter model is small???



RESEARCH

Introducing LLaMA: A foundational, 65-billion-parameter large language model

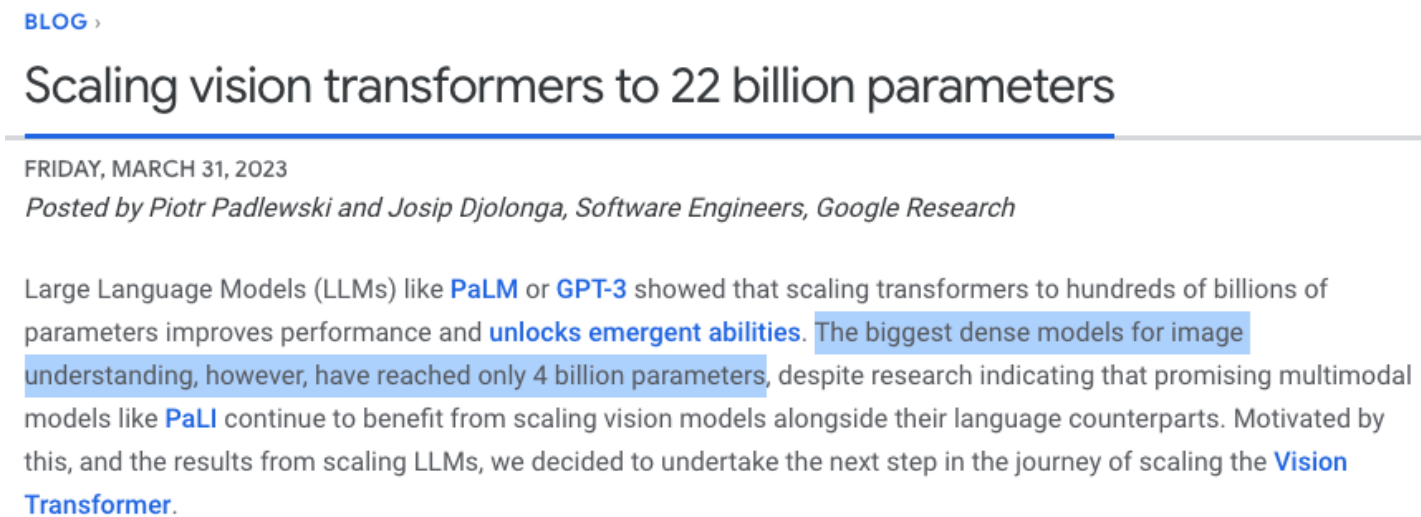
February 24, 2023

As part of Meta's commitment to open science, today we are publicly releasing LLaMA (Large Language Model Meta AI), a state-of-the-art foundational large language model designed to help researchers advance their work in this subfield of AI. Smaller, more performant models such as LLaMA enable others in the research community who don't have access to large amounts of infrastructure to study these models, further democratizing access in this important, fast-changing field.

Training smaller foundation models like LLaMA is desirable in the large language model space because it requires far less computing power and resources to test new approaches, validate others' work, and explore new use cases. Foundation models train on a large set of unlabeled data, which makes them ideal for fine-tuning for a variety of tasks. We are making LLaMA available at several sizes (7B, 13B, 33B, and 65B parameters) and also sharing a LLaMA model card that details how we built the model in keeping with our approach to Responsible AI practices.

Motivation

- number of parameters in GPT
  - 2018: 117 Million, 2019: 1.5 Billion, 2020: 175 Billion
  - GPT-4 $>>$ 200 Billions
- vision transformers



BLOG ›

Scaling vision transformers to 22 billion parameters

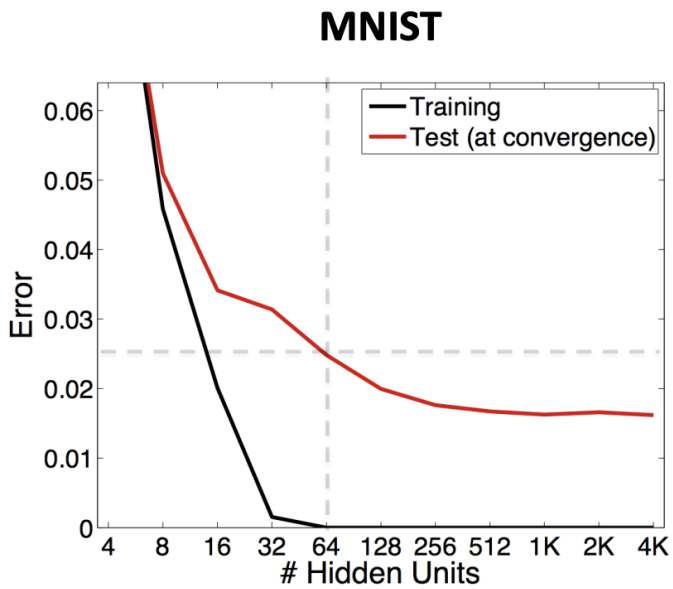
FRIDAY, MARCH 31, 2023

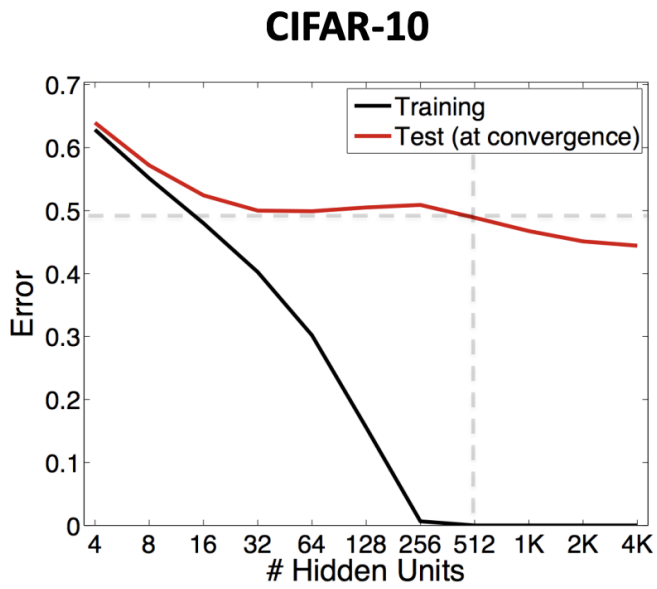
Posted by Piotr Padlewski and Josip Djolonga, Software Engineers, Google Research

Large Language Models (LLMs) like PaLM or GPT-3 showed that scaling transformers to hundreds of billions of parameters improves performance and unlocks emergent abilities. The biggest dense models for image understanding, however, have reached only 4 billion parameters, despite research indicating that promising multimodal models like PaLI continue to benefit from scaling vision models alongside their language counterparts. Motivated by this, and the results from scaling LLMs, we decided to undertake the next step in the journey of scaling the Vision Transformer.

Bigger = Better!

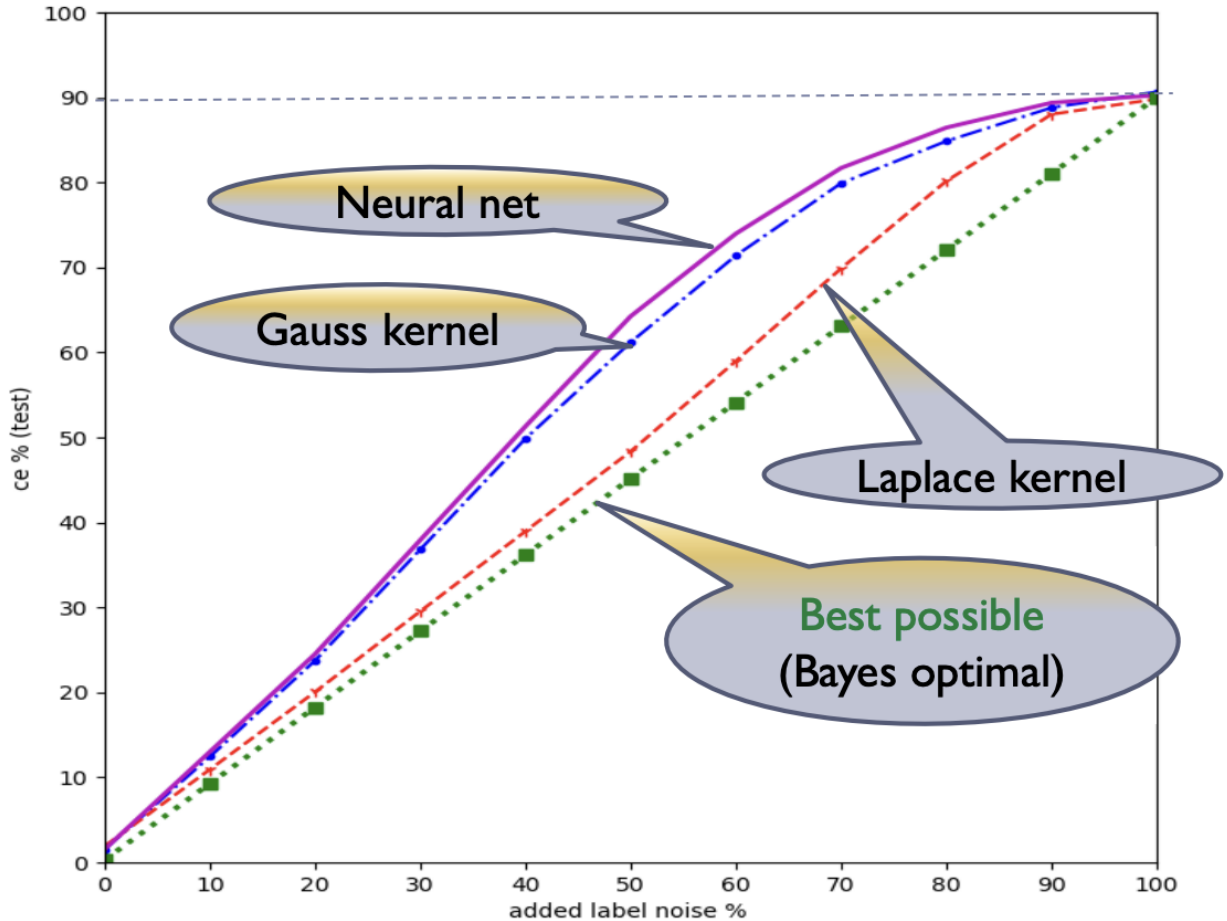
- 2-layer NNs with an increasing number of hidden units [Neyshabur et al. 2015]:





Interpolation learning

- it is possible for a model to generalize well while interpolating noisy training labels [Belkin et al. 2018]



A Challenge for Statistical Learning Theory

- Occam's razor: simpler models generalize better!
- **uniform convergence**: training error is close to the test error for *all* low-complexity models
- However, models that interpolate noisy training labels often have **very high complexity** (even in terms of dimension-free measures such as ℓ_2 norm)

A Challenge for Statistical Learning Theory

- Why do high-dimensional interpolants generalize?
- How can we analyze models with high complexity?
  - uniform convergence!

Plan

- Setting
- Moreau envelope generalization theory
- Applications
  - linear regression
  - max-margin classification
  - phase retrieval, relu regression
  - matrix sensing
  - (single-index) neural networks
- Universality

Setting

We receive i.i.d. sample pairs (x_i, y_i) from some data distribution $\mathcal{D}$ over $\mathbb{R}^d \times \mathcal{Y}$.

Fix any continuous loss function $f : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}$, we can fit a linear model $(\hat{w}, \hat{b})$ by minimizing the empirical loss $\hat{L}_f$:

$$\hat{L}_f(w, b) = \frac{1}{n} \sum_{i=1}^{n} f(\langle w, x_i \rangle + b, y_i),$$

with the goal of achieving small population loss L_f:

$$L_f(w, b) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[f(\langle w, x \rangle + b, y)]$$

Goal

We want to prove high probability bounds on $L_f(\hat{w}, \hat{b})$ with *weak assumptions* on $\mathcal{D}$. For example,

$$L_f(\hat{w}, \hat{b}) \leq \inf_{w,b} L_f(w, b) + \epsilon_{\mathcal{D},n} \tag{0.1}$$

- confidence interval for prediction error
- understand what properties of the **optimization algorithm** and **data distribution** can allow us to learn in high-dimensional settings?
- theoretical tools to predict generalization behaviors and guide the design of better algorithms and model selection procedures in practice

Implicit Bias of GD

If $d > n$ and we run gradient descent on $\hat{L}_f(w, b)$, then

$$\hat{w}_t, \hat{b}_t \to \underset{w,b:\hat{L}_f(w,b)=0}{\arg\min} \|w\|_2^2 + \|b\|_2^2 \tag{0.2}$$

Proof: $\hat{w}_t$ stays in the $\{w = X^T v \mid v \in \mathbb{R}^n\}$ and so the limit satisfies the KKT condition.

Examples:

1. Linear regression
2. Logistic regression: converge in direction to the max-margin solution
3. Matrix Factorization: if $X = UU^T$ and we run GD on U, then $\hat{X}_t$ converge to the minimal nuclear norm solution

Examples

- minimal ℓ_2 norm interpolant:

$$\begin{aligned} \arg\min \quad & \|w\|_2 \\ \text{s.t.} \quad & \forall i \in [n],\ \langle w, x_i \rangle + b = y_i \end{aligned} \tag{3.3}$$

- max-margin classification:

$$\begin{aligned} \arg\min \quad & \|w\|_2 \\ \text{s.t.} \quad & \forall i \in [n],\ (\langle w, x_i \rangle + b) y_i \geq 1 \end{aligned} \tag{3.4}$$

- phase retrieval:

$$\begin{aligned} \arg\min \quad & \|w\|_2 \\ \text{s.t.} \quad & \forall i \in [n],\ \langle w, x_i \rangle^2 = y_i^2 \end{aligned} \tag{3.5}$$

- ReLU regression:

$$\begin{aligned} \arg\min \quad & \|w\|_2 \\ \text{s.t.} \quad & \forall i \in [n],\ \sigma(\langle w, x_i \rangle) = y_i \end{aligned} \tag{3.6}$$

Examples

We also consider the problem of matrix sensing:

$$\begin{aligned} \arg\min \quad & \|X\|_* \\ \text{s.t.} \quad & \forall i \in [n], \langle A_i, X \rangle = y_i \end{aligned} \tag{3.7}$$

These estimators satisfy $\hat{L}_f = 0$ with

- $f(\hat{y}, y) = (\hat{y} - y)^2$
- $f(\hat{y}, y) = (1 - \hat{y}y)_+^2$
- $f(\hat{y}, y) = (|\hat{y}| - y)^2$.
- $f(\hat{y}, y) = \begin{cases} (\hat{y} - y)^2 & \text{if} \quad y > 0 \\ \sigma(\hat{y})^2 & \text{if} \quad y = 0 \end{cases}$

Gaussian Multi-Index Model

The distribution $\mathcal{D}$ over $\mathbb{R}^d \times \mathcal{Y}$ is given by

(A) $x \sim \mathcal{N}(\mu, \Sigma)$,

(B) there exist $w_1^*, ..., w_k^* \in \mathbb{R}^d$, a random variable $\xi \sim \mathcal{D}_\xi$ independent of x (not necessarily Gaussian), and an unknown link function $g : \mathbb{R}^{k+1} \to \mathcal{Y}$ such that

$$\eta_i = \langle w_i^*, x \rangle, \quad y = g(\eta_1, ..., \eta_k, \xi). \tag{0.8}$$

Examples

1. $\mathcal{Y} = \mathbb{R}$ and $y = \langle w^*, x \rangle + \xi$
2. $\mathcal{Y} = \mathbb{R}$ and

$$y = \underbrace{\langle w^*, x \rangle}_{\text{linear signal}} + \underbrace{|x_1| \cdot \cos x_2}_{\text{non-linear term}} + \underbrace{x_3 \cdot \xi}_{\text{heteroscedasticity}}$$

3. $\mathcal{Y} = \{-1, 1\}$ and

$$\Pr(y = 1) = \text{sigmoid}(\langle w^*, x \rangle)$$

4. g is a neural network with k hidden units in the first layer

Model Complexity

Let $W = [w_1^*, ..., w_k^*] \in \mathbb{R}^{d \times k}$ and $Q = I - W(W^T \Sigma W)^{-1} W^T \Sigma$.

$x^\perp = Q^T x$ is the components of x that is independent of y

- $Q^2 = Q$: Q is a projection
- $Q^T \Sigma W = 0$: $Q^T x$ and $W^T x$ are independent
- y only depends on x through $W^T x$

Let C_δ be a continuous function such that with probability at least $1 - \delta/4$ over $x^\perp \sim \mathcal{N}\left(0, \Sigma^\perp\right)$, uniformly over all $w \in \mathbb{R}^d$,

$$\langle w, x^\perp \rangle \leq C_\delta(w). \tag{4.9}$$

Moreau Envelope Generalization Theory

Moreau envelope of f with parameter $\lambda \geq 0$ is defined as

$$f_\lambda(\hat{y}, y) = \inf_u \ f(u, y) + \lambda(u - \hat{y})^2$$

and is usually viewed as a smooth approximation to the original function f.

Theorem

Under some mild conditions on f, there exists $\epsilon = \tilde{O}(\sqrt{k/n})$ such that w.p. at least $1 - \delta$, for all $w \in \mathbb{R}^d$, $b \in \mathbb{R}$ and $\lambda \geq 0$

$$L_{f_\lambda}(w, b) \leq (1 + \epsilon) \left(\hat{L}_f(w, b) + \frac{\lambda C_\delta(w)^2}{n} \right). \tag{4.10}$$

Properties of the Moreau Envelope

- if for each $y \in \mathcal{Y}$, f is M-Lipschitz with respect to the first argument, then for any $\lambda \geq 0$

$$f_\lambda(\hat{y}, y) \geq f(\hat{y}, y) - \frac{M^2}{4\lambda}. \tag{4.11}$$

- if f is non-negative and for each $y \in \mathcal{Y}$, $\sqrt{f}$ is $\sqrt{H}$-Lipschitz with respect to the first argument, then for any $\lambda \geq 0$

$$f_\lambda(\hat{y}, y) \geq \frac{\lambda}{\lambda + H} f(\hat{y}, y). \tag{4.12}$$

- for the square loss, squared hinge loss, phase retrieval loss, and the ReLU loss, it holds that for any $\lambda \geq 0$

$$f_\lambda(\hat{y}, y) = \frac{\lambda}{\lambda + 1} f(\hat{y}, y) \tag{4.13}$$

Lipschitz loss

Combining (4.10) with (4.11)

$$\begin{aligned} L_f(w,b) &\leq (1+\epsilon)\left(\hat{L}_f(w,b) + \inf_{\lambda \geq 0} \frac{\lambda C_\delta(w)^2}{n} + \frac{M^2}{4\lambda}\right) \\ &= (1+\epsilon)\left(\hat{L}_f(w,b) + M\sqrt{\frac{C_\delta(w)^2}{n}}\right) \end{aligned} \tag{4.14}$$

Examples:

- Absolute loss, Logistic loss / Binomial GLM, Hinge loss
- Huber's loss, modified Huber's loss

Square-root Lipschitz loss

Combining (4.10) with (4.12)

$$L_f(w,b) \le (1+\epsilon)\left(\inf_{\lambda\ge 0}\frac{\lambda+H}{\lambda}\hat{L}_f(w,b)+\frac{(\lambda+H)C_\delta(w)^2}{n}\right)$$
$$= (1+\epsilon)\left(\sqrt{\hat{L}_f(w,b)}+\sqrt{\frac{H\,C_\delta(w)^2}{n}}\right)^2 \tag{4.15}$$

For any 1-Lipschitz $\sigma : \mathbb{R} \to \mathbb{R}$ (e.g., identity, absolute value, ReLU, sigmoid, tanh), the following loss functions are 1 square-root Lipschitz:

- $f(\hat{y}, y) = (\sigma(\hat{y}) - y)^2$
- $f(\hat{y}, y) = (1 - \sigma(\hat{y})y)_+^2$

ℓ_2 Benign Overfitting for Linear regression, Max-Margin Classification, ReLU regression and Phase Retrieval

By Cauchy-Schwarz, we have

$$\langle w, x^{\perp} \rangle \leq \|w\|_2 \|x^{\perp}\|_2 \approx \|w\|_2 \sqrt{\mathrm{tr}(\Sigma^{\perp})}$$

and so uniformly over all $(w, b) \in \mathbb{R}^{d+1}$

$$L_f(w, b) \leq (1 + \epsilon) \left(\sqrt{\hat{L}_f(w, b)} + \|w\|_2 \sqrt{\frac{\mathrm{tr}(\Sigma^{\perp})}{n}} \right)^2$$

Norm bound

Defintion

The effective rank of a covariance matrix Σ is defined as

$$R(\Sigma) = \frac{\mathrm{tr}(\Sigma)^2}{\mathrm{tr}(\Sigma^2)}.$$

Theorem

Fix any $(w^\sharp, b^\sharp) \in \mathbb{R}^{d+1}$. There exists

$$\rho \lesssim \sqrt{\frac{k \log(n/k)}{n}} + \log\left(\frac{1}{\delta}\right)\left(\frac{1}{\sqrt{n}} + \frac{1}{\sqrt{R(\Sigma^\perp)}} + \frac{n}{R(\Sigma^\perp)}\right),$$

such that with probability at least $1-\delta$, it holds that

$$\min_{(w,b)\in\mathbb{R}^{d+1}:\hat{L}_f(w,b)=0} \|w\|_2 \leq \|w^\sharp\|_2 + (1+\rho)\sqrt{\frac{nL_f(w^\sharp, b^\sharp)}{\mathrm{tr}(\Sigma^\perp)}}. \quad (4.16)$$

Generalization Bound

Corollary

Let $\hat{w}, \hat{b} = \arg\min_{(w,b):\hat{L}_f(w,b)=0} \|w\|_2$ *be the minimal norm interpolant. With probability at least* $1-\delta$*, it holds that*

$$L_f(\hat{w}, \hat{b}) \leq (1+\rho)\left(\sqrt{L_f(w^\sharp, b^\sharp)} + \|w^\sharp\|_2 \sqrt{\frac{\mathrm{tr}(\Sigma^\perp)}{n}}\right)^2. \quad (4.17)$$

We establish consistency when

$$\|w^\sharp\|_2 \sqrt{\frac{\mathrm{tr}(\Sigma^\perp)}{n}} \to 0, \quad \frac{k}{n} \to 0, \quad \frac{n}{R(\Sigma^\perp)} \to 0 \quad (4.18)$$

Examples:

- Junk feature:

$$\Sigma = \begin{pmatrix} I_k & 0 \\ 0 & \frac{1}{d} I_d \end{pmatrix} \implies \Sigma^{\perp} = \frac{1}{d} I_d$$

- Gaussian kernels on the hypersphere: by rotation invariance, Σ has a block diagonal structure and the i-th block has dimension $O_d(d^i)$
  - Consider the scaling $n = d^l$ where $l \notin \mathbb{N}$, then we can take $k = O(d^{\lfloor l \rfloor}) = o(n)$. Moreover, it holds that $R(\Sigma^{\perp}) = \Omega(d^{\lceil l \rceil}) = \omega(n)$

Minimal Nuclear Norm in Matrix Sensing

Given (i.i.d. standard Gaussian) random measurement matrices $A_1, ..., A_n$ and measurements $y_1, ..., y_n$ given by $y_i = \langle A_i, X^* \rangle + \xi_i$ where ξ is independent of A_i, and $\mathbb{E}\xi = 0$ and $\mathbb{E}\xi^2 = \sigma^2$, we hope to reconstruct the matrix $X^* \in \mathbb{R}^{d_1 \times d_2}$ with sample size $n \ll d_1 d_2$.

We assume X^* has rank r and we consider the minimal nuclear norm solution

$$\hat{X} = \underset{X \in \mathbb{R}^{d_1 \times d_2} : \langle A_i, X \rangle = y_i}{\arg\min} \|X\|_*. \tag{4.19}$$

The dual norm of nuclear norm is the spectral norm and so

$$\sigma^2 + \|\hat{X} - X^*\|_F^2 = L(\hat{X}) \leq (1 + o(1)) \frac{\|\hat{X}\|_*^2 (\mathbb{E}\|A\|)^2}{n}$$

Nuclear norm bound

Similar analysis shows that

$$\|\hat{X}\|_* \leq \|X^*\|_* + \left(1 - \frac{n}{d_1 d_2}\right)^{-1/2} \frac{\mathbb{E}\|A\|_*}{\mathbb{E}\|A\|_F^2} \sqrt{n}\sigma$$

and so

$$\|\hat{X} - X^*\|_F^2 + \sigma^2$$

$$\leq (1 + o(1)) \left(\frac{\|X^*\|_* \mathbb{E}\|A\|}{\sqrt{n}} + \left(1 - \frac{n}{d_1 d_2}\right)^{-1/2} \frac{\mathbb{E}\|A\|_* \mathbb{E}\|A\|}{\mathbb{E}\|A\|_F^2} \sigma \right)^2 .$$

Low-Rank Matrix Recovery

Importantly, since X^* has rank r, we have $\|X^*\|_* \leq \sqrt{r}\|X^*\|_F$. Moreover, it is well-known that $\mathbb{E}\|A\| \approx \sqrt{d_1} + \sqrt{d_2}$. Rearranging the uniform convergence bound, we obtain

$$\mathbb{E}\left[\frac{\|\hat{X} - X^*\|_F^2}{\|X^*\|_F^2}\right] \leq \left(\left(1 - \frac{n}{d_1 d_2}\right)^{-1} \left(\frac{\mathbb{E}\|A\|_* \mathbb{E}\|A\|}{\mathbb{E}\|A\|_F^2}\right)^2 - 1\right) \frac{\sigma^2}{\|X^*\|_F^2} + O\left(\frac{r(d_1+d_2)}{n} + \sqrt{\frac{r(d_1+d_2)}{n}} \frac{\sigma}{\|X^*\|_F}\right).$$

In the noiseless case $\sigma = 0$, we recover the classical rate

$$\mathbb{E}\left[\frac{\|\hat{X} - X^*\|_F^2}{\|X^*\|_F^2}\right] \lesssim \frac{r(d_1+d_2)}{n}$$

Benign Matrix Sensing

The singular values of A concentrates to the square root of the Marchenko-Pastur law if d_1/d_2 converge to a constant. However, by Holder's inequality, $\frac{\mathbb{E}\|A\|_* \mathbb{E}\|A\|}{\mathbb{E}\|A\|_F^2}$ can only converge to 1 when the singular values are all the same. This happens if and only if $d_1/d_2 \to \{0, \infty\}$. Therefore, when the signal to noise ratio $\frac{\|X^*\|_F^2}{\sigma^2}$ is constant, we obtain consistency when the following holds

- $r(d_1 + d_2) = o(n)$
- $d_1 d_2 = \omega(n)$
- $d_1/d_2 \to \{0, \infty\}$

This can happen for example when

$$r = \Theta(1), d_1 = \Theta(n^{1/2}), d_2 = \Theta(n^{2/3}).$$

Local Gaussian width

We define a mapping $\phi : \mathbb{R}^d \to \mathbb{R}^{k+1}$ by

$$\phi(w) := (w^T \Sigma W, \|w\|_{\Sigma^\perp}). \tag{4.20}$$

It is obvious that for $w \in \mathcal{K}$

$$\begin{aligned} \langle w, x^\perp \rangle &\leq \max_{w' \in \mathcal{K}: \phi(w') = \phi(w)} \langle w', x^\perp \rangle \\ &\approx \mathbb{E}\left[\max_{w' \in \mathcal{K}: \phi(w') = \phi(w)} \langle w', x^\perp \rangle \right] := C(\phi(w)) \end{aligned} \tag{4.21}$$

Optimal Complexity Control

Suppose that f is convex. If $\hat{w}, \hat{b}$ is the constrained ERM for any bounded convex set, then it holds that

$$\hat{L}_f(\hat{w}, \hat{b}) \approx \max_{\lambda \geq 0} L_{f_\lambda}(\hat{w}, \hat{b}) - \frac{\lambda C(\phi(\hat{w}))^2}{n}.$$

In particular, if $f_\lambda = \frac{\lambda}{1+\lambda}$, then

$$\hat{L}_f(\hat{w}, \hat{b}) \approx \left(\sqrt{L_f(\hat{w}, \hat{b})} - \frac{C(\phi(\hat{w}))}{\sqrt{n}} \right)_+^2$$

and so

$$\hat{L}_f(\hat{w}, \hat{b}) \approx 0 \implies L_f(\hat{w}, \hat{b}) \approx \frac{C(\phi(\hat{w}))^2}{n}.$$

Single-Index Neural Network

Let $\sigma(x) = \max(x, 0)$ be the ReLU activation function, N be the number of hidden units, and $\theta = (w, a, b) \in \mathbb{R}^{d+2N}$ parameterize the class of simple neural nets:

$$h_\theta(x) := \sum_{i=1}^{N} a_i \sigma(\langle w, x \rangle - b_i).$$

If we use the square loss or the squared hinge loss, we are essentially minimizing

$$f(\hat{y}, y, \theta) = \left(\sum_{i=1}^{N} a_i \sigma(\hat{y} - b_i) - y \right)^2 \text{ or } \left(1 - \sum_{i=1}^{N} a_i \sigma(\hat{y} - b_i) y \right)_+^2$$

and $\sqrt{f}$ is $\max_{j \in [N]} \left| \sum_{i=1}^{j} a_i \right|$ Lipschitz if we sort the b_i's.

Generalization Bound

The square-root Lipschitz parameter depends on the model θ, but we can still show that for some $\epsilon = \tilde{O}\left(\frac{k+N}{n}\right)$, it holds that

$$L(\theta) \leq (1+\epsilon)\left(\sqrt{\hat{L}(\theta)} + \frac{\max_{j\in[N]}\left|\sum_{i=1}^{j} a_i\right| \, \|w\| \, \mathbb{E}\|x^{\perp}\|_*}{\sqrt{n}}\right)^2.$$

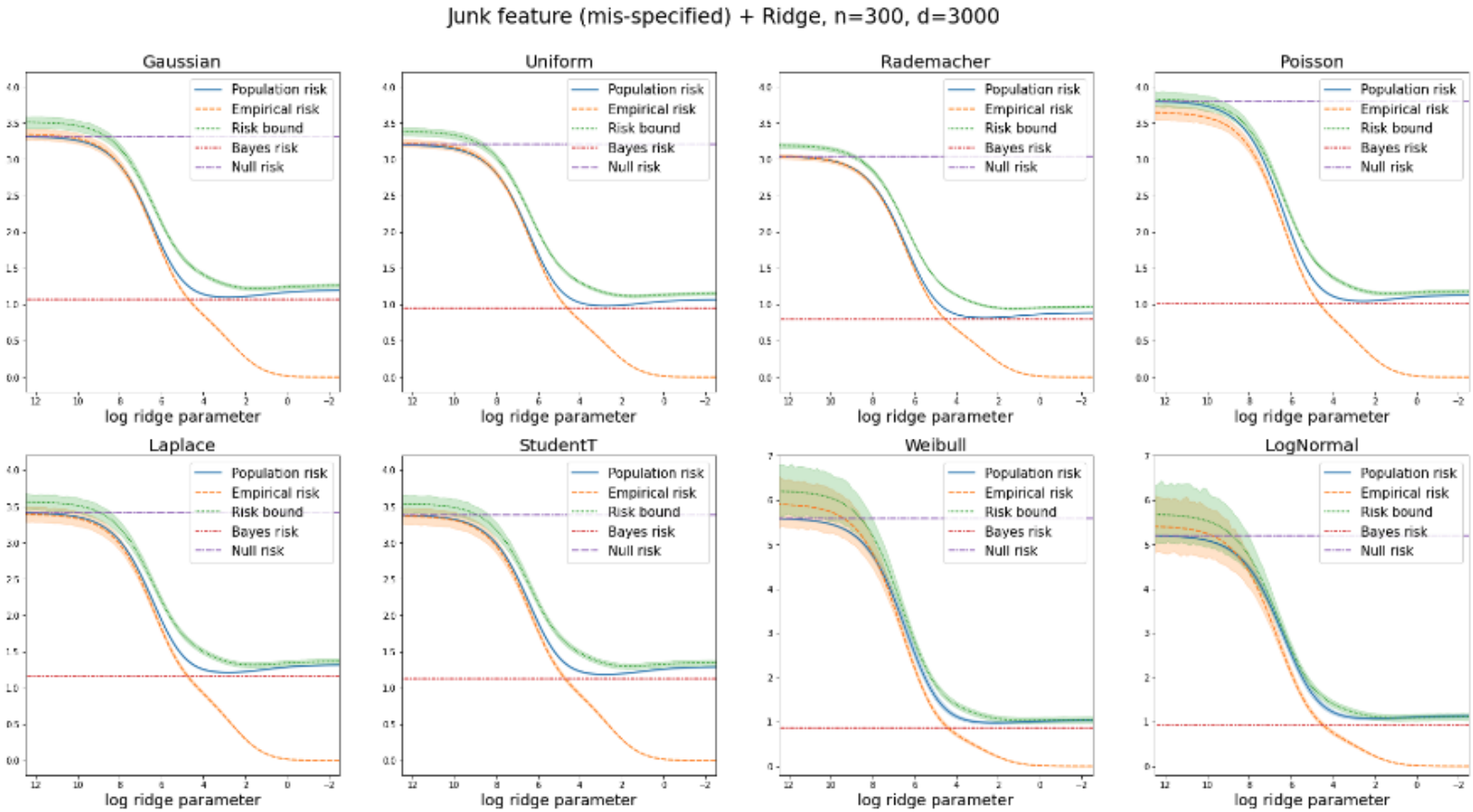
For linear models, we measure the complexity by the norm of the coefficients. For (single-index) neural net, we can measure by

$$\max_{j\in[N]}\left|\sum_{i=1}^{j} a_i\right| \, \|w\|.$$

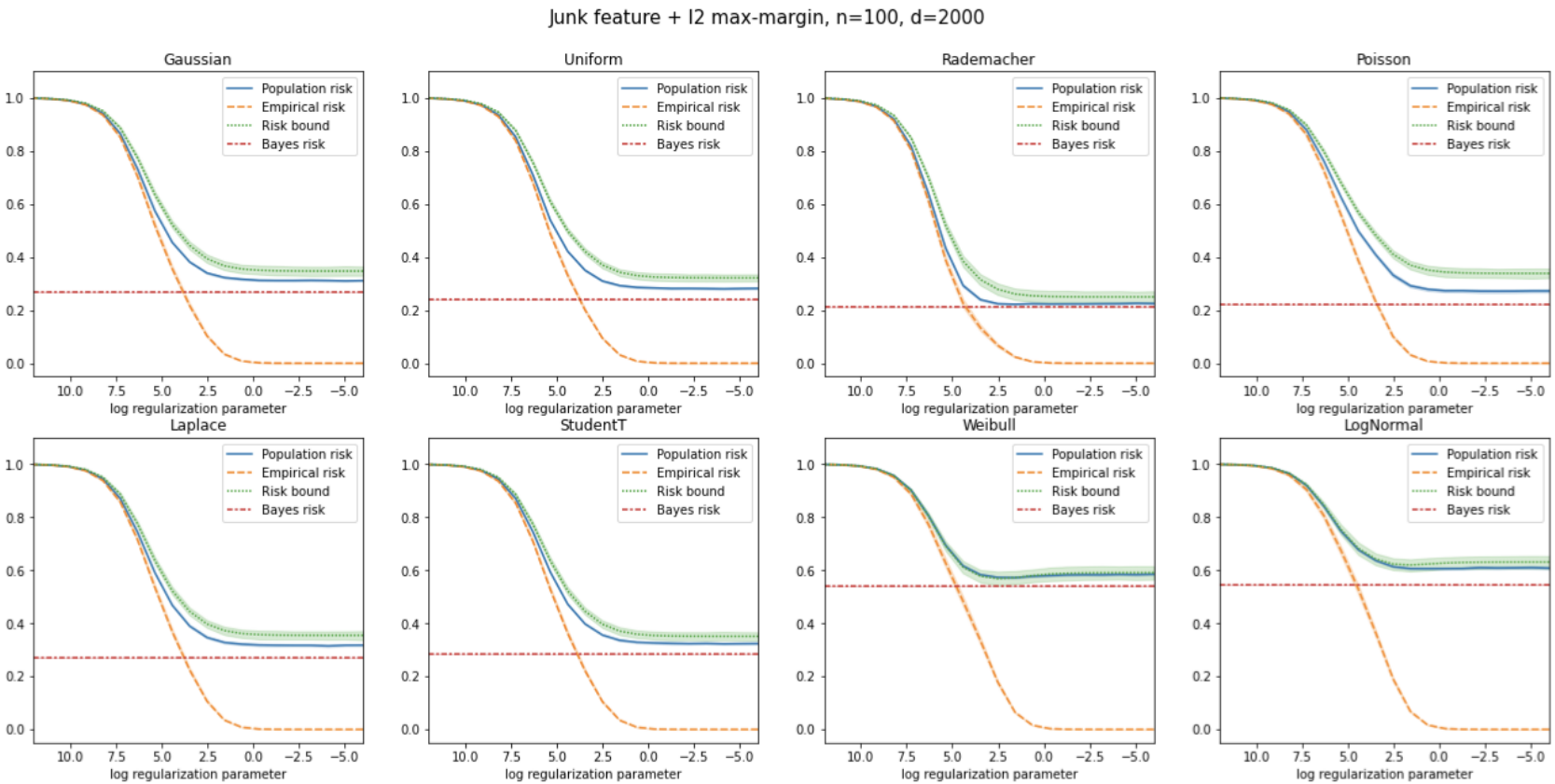
Universality

- omniscient risk estimator: $x_i = \Sigma^{1/2} z_i$ where
  - z_i has i.i.d. coordinates with zero mean, unit variance, and bounded 12th moments, as in Wu and Xu (2020)
  - z_i has independent coordinates with zero mean, unit variance, finite moments of all order, as in Hastie et al (2020)
- universality of Gaussian Minimax Theorem
  - can be proven in similar settings using Lindeberg's method, as in Han and Shen (2022)
- Gaussian equivalence
  - random features: $x_i = \sigma(Wz_i)$ where W is a randomly initialized matrix and z_i is standard Gaussian, as in Hu and Lu (2022)
  - kernel regression with $K(x, x') = h(\langle x, x' \rangle)$ on uniform distribution and x is uniform on hypersphere or boolean hypercube, as in Misiakiewicz (2022)

Experiments: Regression



Experiments: Classification



Failure of Universality

Suppose that $\mathcal{D}$ is given by

(A) $x = (x_{|k}, x_{|d-k})$ where $x_{|k} \sim \mathcal{N}(0, \Sigma_{|k})$ and there exists a function $h : \mathbb{R}^k \to \mathbb{R}$ such that

$$x_{|d-k} = h(x_{|k}) \cdot z \tag{0.22}$$

where $z \sim \mathcal{N}\left(0, \Sigma_{|d-k}\right)$ is independent of $x_{|k}$.

(B) there exists a function $g : \mathbb{R}^{k+1} \to \mathbb{R}$ such that

$$y = g(x_{|k}, \xi) \tag{0.23}$$

where $\xi \sim \mathcal{D}_\xi$ is independent of x.

Failure of Universality

It can be shown that

$$\frac{\|\hat{w}\|_2^2 \cdot \mathbb{E}\|x_{|d-k}\|_2^2}{n} \to \inf_w \mathbb{E}[h(x_{|k})^2] \cdot \mathbb{E}\left[\left(\frac{y - \langle w, x\rangle}{h(x_{|k})}\right)^2\right]$$

However, if we choose $\Sigma_{|d-k}$ to be benign, then the Gaussian theory would predict

$$\inf_w \mathbb{E}[(\langle w, x\rangle - y)^2] \leq \inf_w \mathbb{E}[h(x_{|k})^2] \cdot \mathbb{E}\left[\left(\frac{y - \langle w, x\rangle}{h(x_{|k})}\right)^2\right]$$

Counterexample

Consider $k = 1$, $x_1 \sim \mathcal{N}(0, 1)$, $h(|x_1|) = 1 + |x_1|$ and $y = h(|x_1|)^2$. Then it holds that

$$\inf_w \mathbb{E}[(\langle w, x \rangle - y)^2] - \inf_w \mathbb{E}[h(x_{|k})^2] \cdot \mathbb{E}\left[\left(\frac{y - \langle w, x \rangle}{h(x_{|k})}\right)^2\right]$$

$$= \mathrm{var}(h(|x_1|)) > 0.$$

In this case, we cannot pretend that x is Gaussian because

- the complexity function C_δ is defined like a **worst-case** Rademacher complexity
- for Gaussian data it's not that different from the **expected** Rademacher complexity because the norm of the tail of x concentrates
- the tail of x **does not concentrate** in this counterexample!

Summary

We propose a statistical learning theory for models with high complexity. We use it to explains benign overfitting in:

- linear regression (or kernel regression),
- max-margin classification,
- phase retrieval, ReLU regression,
- matrix sensing.

This theory is non-asymptotic, requires very mild assumptions, can be easily adapted to different complexity measure (ℓ_1, ℓ_2, nuclear norm, etc) and has the potential to explain even more complex models such as deep neural nets.

However, this theory requires a Gaussian feature assumption and universality can fail in unexpected ways.

Acknowledgement

- Nati
- Nati's group:
  - Danica Sutherland, Gal Vardi, Sam Buchanan, Lingxiao Wang, Kavya Ravichandran, Owen Melia, Anmol Kabra, Gene Li, Kumar Kshitij Patel, Omar Montasser, Donya Saless, Nirmit Joshi, Akilesh Tangella, Blake Woodworth, Xiaoxia Wu, Suriya Gunasekar, Pritish Kamath, Brian Bullins
  - MLO reading group
- Frederic Koehler, Pragya Sur, Zhen Dai, Jamie Simon
- Haochen Wang
- thesis committee:
  - Chao Gao, Rina Foygel Barber
- reading and research:
  - Per Mykland, Lek-Heng Lim, Yali Amit

Acknowledgement

- undergraduate research advisor and recommender:
  - Michael Stein, Zheng (Tracy) Ke, Weibiao Wu
- courses at uchicago and TTIC
- TA and statistics consulting program
- many more stats & CAM students...
- administrative staffs
- Qi Zhu
- You Zhou, Jieling Li