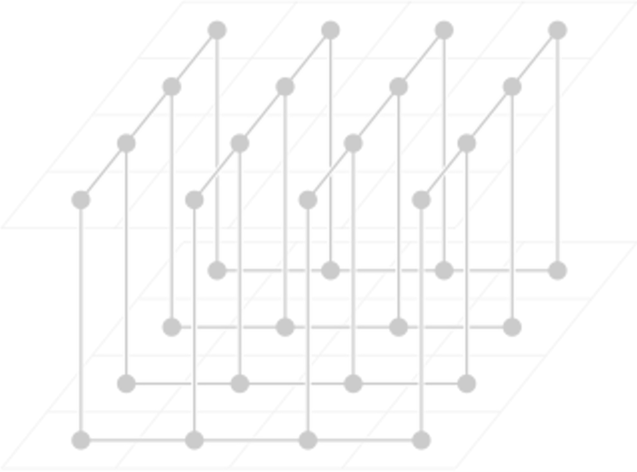
Random matrix theory and modern machine learning*

Michael W. Mahoney

(ICSI, LBNL, and Department of Statistics, UC Berkeley)

November 2025



Short version; a longer version (from June 2025) is on my web site.

Overview

Motivations:

- WeightWatcher, Weight Diagnostics for Analyzing ML Models (with Charles H. Martin)
- Randomized Numerical Linear Algebra for Modern ML (with Michal Derezinski)

Some Theory:

- RMT for NNs: Linear to Nonlinear; Shallow to Deep; etc. (with Zhenyu Liao)

Applications:

- Models of Heavy-Tailed Mechanistic Universality (with Zhichao Wang and Liam Hodgkinson)
- Spectral Estimation with Free Decompression (with Siavash Ameli, Chris van der Heide, and Liam Hodgkinson)
- Determinant Estimation under Memory Constraints and Neural Scaling Laws (with S. Ameli, C. van der Heide, L. Hodgkinson, and F. Roosta)

Overview

Motivations:

- WeightWatcher, Weight Diagnostics for Analyzing ML Models (with Charles H. Martin)
- Randomized Numerical Linear Algebra for Modern ML (with Michal Derezinski)

Some Theory:

- RMT for NNs: Linear to Nonlinear; Shallow to Deep; etc. (with Zhenyu Liao)

Applications:

- Models of Heavy-Tailed Mechanistic Universality (with Zhichao Wang and Liam Hodgkinson)
- Spectral Estimation with Free Decompression (with Siavash Ameli, Chris van der Heide, and Liam Hodgkinson)
- Determinant Estimation under Memory Constraints and Neural Scaling Laws (with S. Ameli, C. van der Heide, L. Hodgkinson, and F. Roosta)

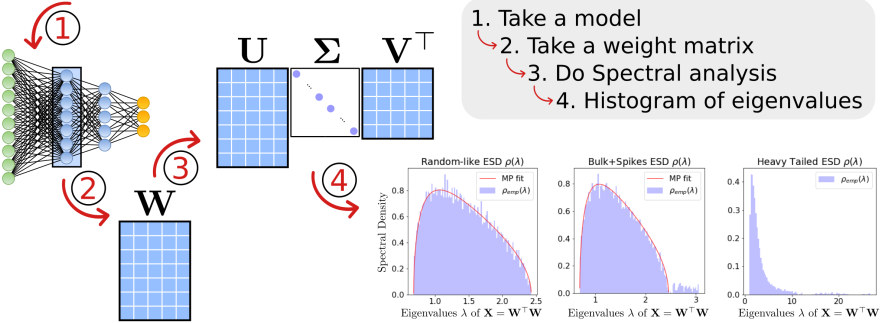
Lots of DNNs analyzed: Look at nearly every publicly-available SOTA model in CV and NLP

- *Don't evaluate your method on one/two/three NNs, evaluate it on:*
  - *dozens (2017)*
  - *hundreds (2019)*
  - *thousands (2021)*
- *Don't use bad/toy models, use SOTA models.*
  - *If you do, don't be surprised if low-quality/toy models are different than high-quality/SOTA models.*
- *Don't train models, instead validate pre-trained models.*
  - *Validating models is harder than training models.*

Watching weights with WeightWatcher

https://github.com/CalculatedContent/WeightWatcher

Analyzing DNN Weight matrices with WeightWatcher



1. Take a model
2. Take a weight matrix
3. Do Spectral analysis
4. Histogram of eigenvalues

1
2
3
4
W
U
Σ
$V^\top$
Random-like ESD $\rho(\lambda)$
Bulk+Spikes ESD $\rho(\lambda)$
Heavy Tailed ESD $\rho(\lambda)$
MP fit
$\rho_{emp}(\lambda)$
Spectral Density
Eigenvalues λ of $\mathbf{X} = \mathbf{W}^\top\mathbf{W}$
0.0
0.2
0.4
0.6
0.8
1.0
1.5
2.0
2.5
1
2
3
0.1
0.3
0
10
20

- Analyze one layer of pre-trained model
- Compare multiple layers of pre-trained model
- Monitor NN properties as you train your own model

“pip install weightwatcher”

Using the theory

Different ways one could *use* a theory.

- Perform diagnostics for model validation, to develop hypotheses, etc.*
- Make predictions about model quality, generalization, transferability, etc.*
- Did post-training modifications damage my model?*
- Will buying more data help?*
- Will training longer help?*
- Will quantizing or distilling help?*
- Construct a regularizer to do model training.**

*Ideally, by peeking at very little or no data.

**If you have lots of data, lots of GPUs, etc.

Overview

Motivations:

- WeightWatcher, Weight Diagnostics for Analyzing ML Models (with Charles H. Martin)
- Randomized Numerical Linear Algebra for Modern ML (with Michal Derezinski)

Some Theory:

- RMT for NNs: Linear to Nonlinear; Shallow to Deep; etc. (with Zhenyu Liao)

Applications:

- Models of Heavy-Tailed Mechanistic Universality (with Zhichao Wang and Liam Hodgkinson)
- Spectral Estimation with Free Decompression (with Siavash Ameli, Chris van der Heide, and Liam Hodgkinson)
- Determinant Estimation under Memory Constraints and Neural Scaling Laws (with S. Ameli, C. van der Heide, L. Hodgkinson, and F. Roosta)

Outline

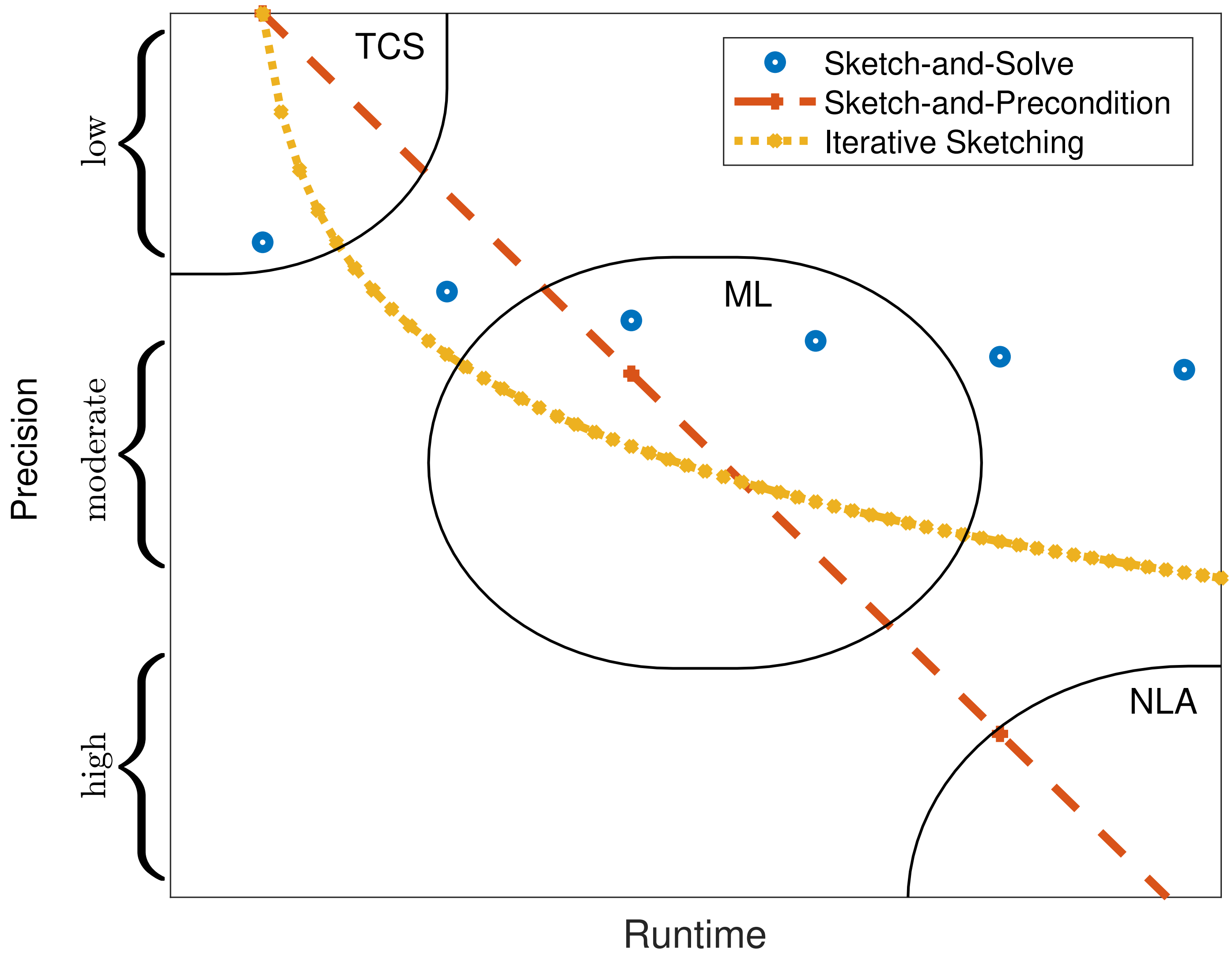
Part I
Foundations of RandNLA

1. Initial thoughts
   - Overview
2. Foundations of "classical" RandNLA
   - Matrix Multiplication
   - Least-squares Approximation
   - Low-rank Approximation
3. Foundations of "modern" RandNLA
   - Algorithmic Gaussianization via Random Matrix Theory
   - RMT for Sampling via DPPs

Part II
Recent and Upcoming Advances

4. Advances in RandNLA for Optimization
   - Gradient Sketch
   - Hessian Sketch
   - Sketch-and-Project
5. Advances in RandNLA for ML
   - Statistical Learning Approaches
   - Statistical Inference Approaches
   - Random Matrix Theory Approaches
6. Putting Randomness into LAPACK
   - RandBLAS/RandLAPACK
7. Concluding thoughts

Using RandNLA methods more generally ...



Landscape of Algorithmic Gaussianization

Sub-gaussian concentration of $x \in \mathbb{R}^d$ w.r.t. a set of functions $\mathcal{F}$

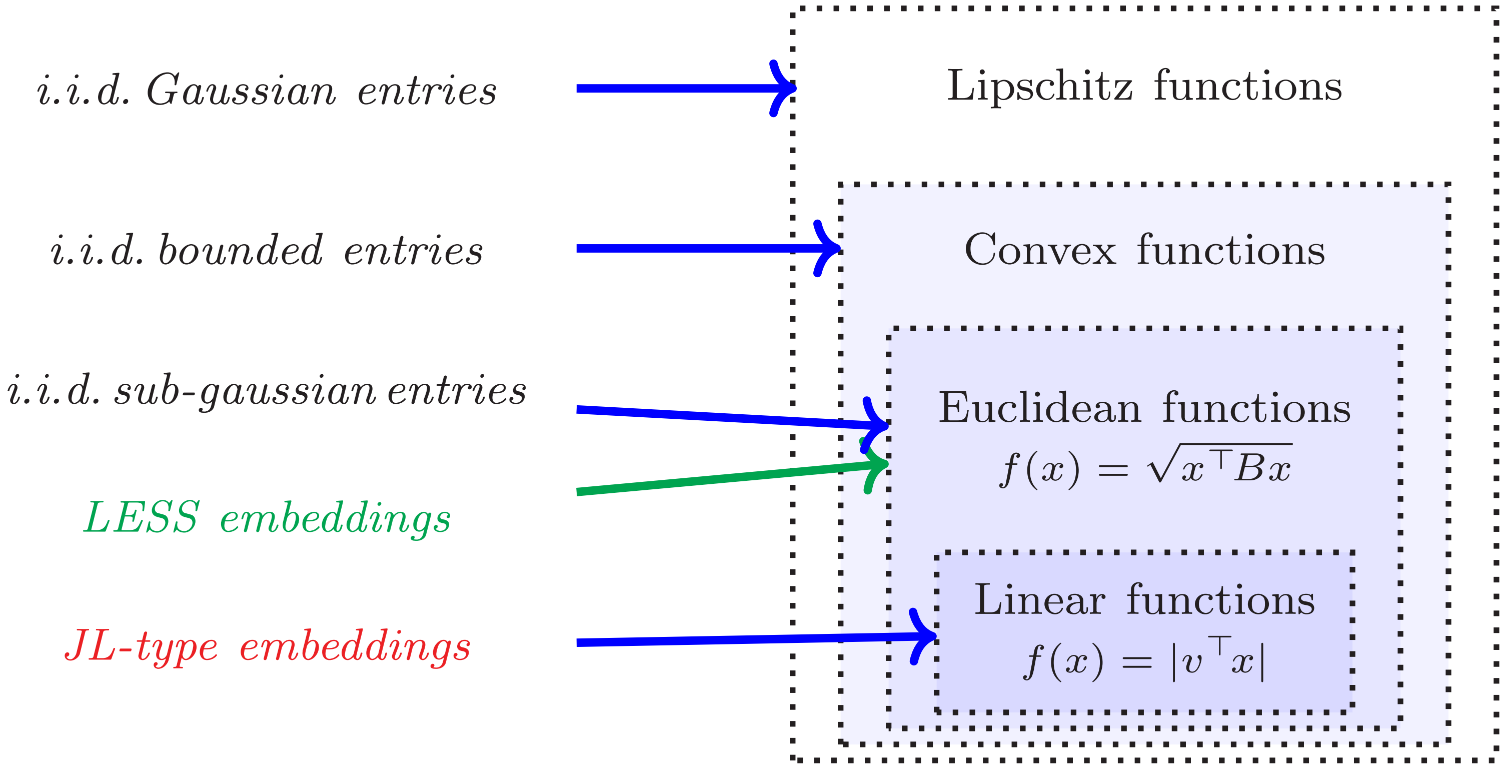
$$\forall f \in \mathcal{F}: \qquad X = f(x) - \mathbb{E}\, f(x) \quad \text{is} \quad \underbrace{O(\|f\|_{\mathrm{Lip}})\text{-sub-gaussian}}_{\mathbb{E}\,\exp(cX^2/\|f\|_{\mathrm{Lip}}) \le 2}$$

Examples

$x \in \mathbb{R}^d$

Concentration

$\mathcal{F} \subseteq \{\mathbb{R}^d \to \mathbb{R}\}$



i.i.d. Gaussian entries

i.i.d. bounded entries

i.i.d. sub-gaussian entries

LESS embeddings

JL-type embeddings

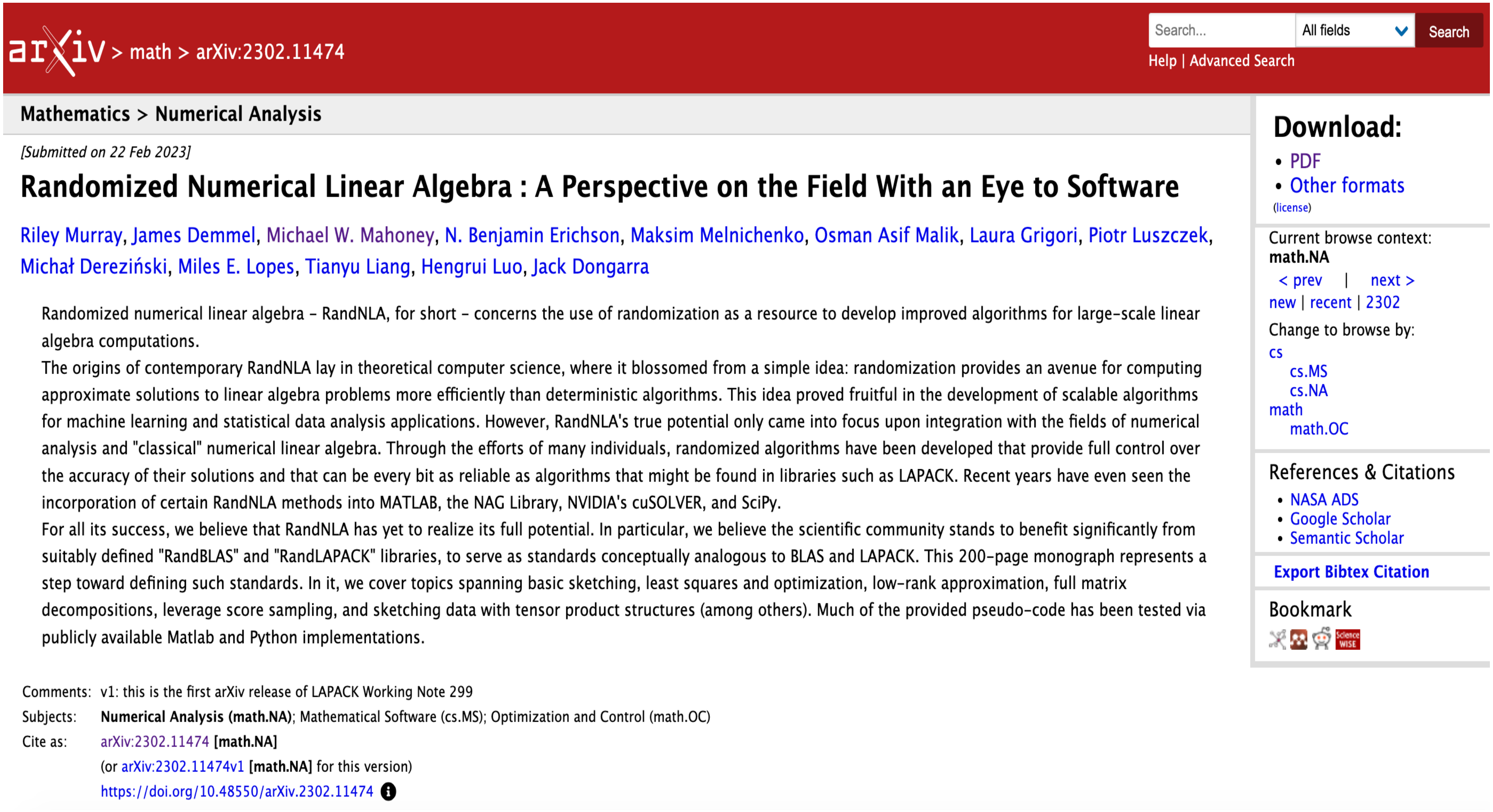
Lipschitz functions

Convex functions

Euclidean functions $f(x) = \sqrt{x^\top B x}$

Linear functions $f(x) = |v^\top x|$

“The RandLAPACK book”


arXiv > math > arXiv:2302.11474
Search...
All fields
Search
Help | Advanced Search
Mathematics > Numerical Analysis

[Submitted on 22 Feb 2023]

Randomized Numerical Linear Algebra : A Perspective on the Field With an Eye to Software

Riley Murray, James Demmel, Michael W. Mahoney, N. Benjamin Erichson, Maksim Melnichenko, Osman Asif Malik, Laura Grigori, Piotr Luszczek, Michał Dereziński, Miles E. Lopes, Tianyu Liang, Hengrui Luo, Jack Dongarra

Randomized numerical linear algebra - RandNLA, for short - concerns the use of randomization as a resource to develop improved algorithms for large-scale linear algebra computations.

The origins of contemporary RandNLA lay in theoretical computer science, where it blossomed from a simple idea: randomization provides an avenue for computing approximate solutions to linear algebra problems more efficiently than deterministic algorithms. This idea proved fruitful in the development of scalable algorithms for machine learning and statistical data analysis applications. However, RandNLA's true potential only came into focus upon integration with the fields of numerical analysis and "classical" numerical linear algebra. Through the efforts of many individuals, randomized algorithms have been developed that provide full control over the accuracy of their solutions and that can be every bit as reliable as algorithms that might be found in libraries such as LAPACK. Recent years have even seen the incorporation of certain RandNLA methods into MATLAB, the NAG Library, NVIDIA's cuSOLVER, and SciPy.

For all its success, we believe that RandNLA has yet to realize its full potential. In particular, we believe the scientific community stands to benefit significantly from suitably defined "RandBLAS" and "RandLAPACK" libraries, to serve as standards conceptually analogous to BLAS and LAPACK. This 200-page monograph represents a step toward defining such standards. In it, we cover topics spanning basic sketching, least squares and optimization, low-rank approximation, full matrix decompositions, leverage score sampling, and sketching data with tensor product structures (among others). Much of the provided pseudo-code has been tested via publicly available Matlab and Python implementations.

Comments:	v1: this is the first arXiv release of LAPACK Working Note 299
Subjects:	**Numerical Analysis (math.NA)**; Mathematical Software (cs.MS); Optimization and Control (math.OC)
Cite as:	arXiv:2302.11474 **[math.NA]**
	(or arXiv:2302.11474v1 **[math.NA]** for this version)
	https://doi.org/10.48550/arXiv.2302.11474 ⓘ

Overview

Motivations:

- WeightWatcher, Weight Diagnostics for Analyzing ML Models
  (with Charles H. Martin)
- Randomized Numerical Linear Algebra for Modern ML
  (with Michal Derezinski)

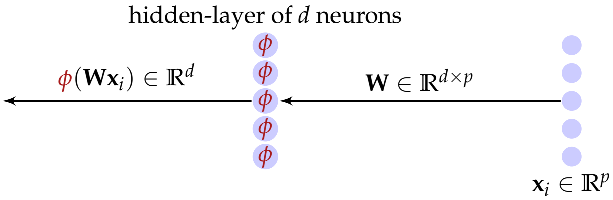
Some Theory:

- RMT for NNs: Linear to Nonlinear; Shallow to Deep; etc.
  (with Zhenyu Liao)

Applications:

- Models of Heavy-Tailed Mechanistic Universality
  (with Zhichao Wang and Liam Hodgkinson)
- Spectral Estimation with Free Decompression
  (with Siavash Ameli, Chris van der Heide, and Liam Hodgkinson)
- Determinant Estimation under Memory Constraints and Neural Scaling Laws
  (with S. Ameli, C. van der Heide, L. Hodgkinson, and F. Roosta)

A deep neural network model

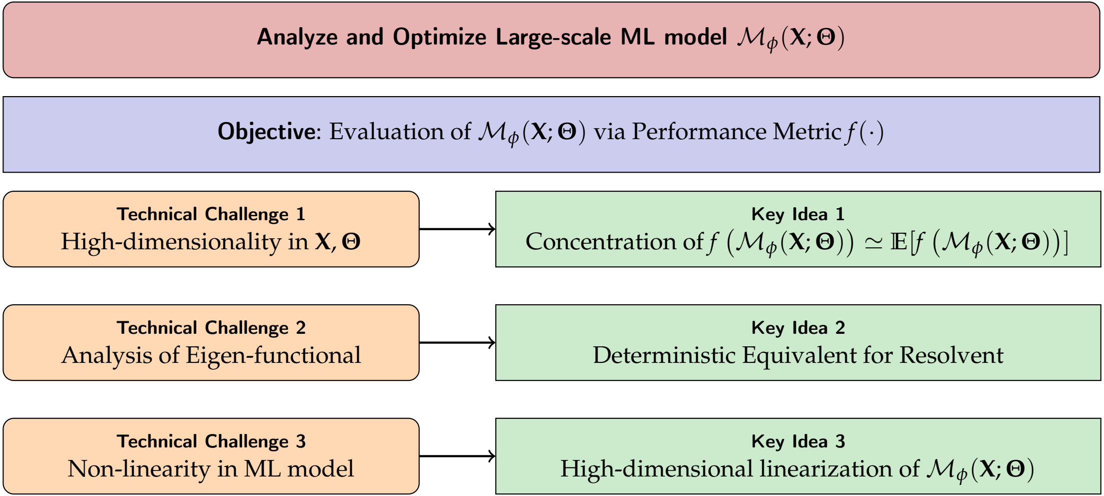

hidden-layer of d neurons
ϕ
ϕ
ϕ
ϕ
ϕ
$\phi(\mathbf{W}\mathbf{x}_i) \in \mathbb{R}^d$
$\mathbf{W} \in \mathbb{R}^{d \times p}$
$\mathbf{x}_i \in \mathbb{R}^p$

- **linear transformation** with first-layer weight matrix $\mathbf{W} \in \mathbb{R}^{d \times p}$
- **nonlinear transformation**: activation function $\phi \colon \mathbb{R} \to \mathbb{R}$ acting entry-wise on $\mathbf{W}\mathbf{x}_i$
- **data representation** at the output of first-layer $\boxed{\mathbf{x}_i \mapsto \phi(\mathbf{W}\mathbf{x}_i)}$
- do the same thing in a layer-by-layer fashion:

$$\frac{1}{\sqrt{d_L}} \mathbf{w}^{\mathsf{T}} \phi_L \left(\frac{1}{\sqrt{d_{L-1}}} \mathbf{W}_L \phi_{L-1} \left(\cdots \frac{1}{\sqrt{d_2}} \phi_2 \left(\frac{1}{\sqrt{d_1}} \mathbf{W}_2 \phi_1 (\mathbf{W}_1 \mathbf{x}_i) \right) \right) \right), \tag{1}$$

for a large number n of input data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$

Technical challenges and key ideas



Analyze and Optimize Large-scale ML model $\mathcal{M}_\phi(\mathbf{X};\boldsymbol{\Theta})$

Objective: Evaluation of $\mathcal{M}_\phi(\mathbf{X};\boldsymbol{\Theta})$ via Performance Metric $f(\cdot)$

Technical Challenge 1
High-dimensionality in $\mathbf{X}, \boldsymbol{\Theta}$

Key Idea 1
Concentration of $f\left(\mathcal{M}_\phi(\mathbf{X};\boldsymbol{\Theta})\right) \simeq \mathbb{E}[f\left(\mathcal{M}_\phi(\mathbf{X};\boldsymbol{\Theta})\right)]$

Technical Challenge 2
Analysis of Eigen-functional

Key Idea 2
Deterministic Equivalent for Resolvent

Technical Challenge 3
Non-linearity in ML model

Key Idea 3
High-dimensional linearization of $\mathcal{M}_\phi(\mathbf{X};\boldsymbol{\Theta})$

High-dimensional Equivalent

Definition (High-dimensional Equivalent)

Let $\mathcal{M}_\phi(\mathbf{X}) \in \mathbb{R}^{p\times n}$ be a (nonlinear) random matrix model that depends on a random matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$ and function $\phi\colon \mathbb{R} \to \mathbb{R}$ (typically applied entrywise). Let $f\left(\mathcal{M}_\phi(\mathbf{X})\right)$ be a scalar observation of $\mathcal{M}_\phi(\mathbf{X})$ for some $f\colon \mathbb{R}^{p\times n} \to \mathbb{R}$. We say that $\tilde{\mathcal{M}}_\phi(\mathbf{X})$ (random or deterministic) is a High-dimensional Equivalent of $\mathcal{M}_\phi(\mathbf{X})$ with respect to $f(\cdot)$ if

$$f(\mathcal{M}_\phi(\mathbf{X})) - f(\tilde{\mathcal{M}}_\phi(\mathbf{X})) \to 0, \tag{2}$$

in probability or almost surely as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$. We denote this relation as

$$\mathcal{M}_\phi(\mathbf{X}) \overset{f}{\leftrightarrow} \tilde{\mathcal{M}}_\phi(\mathbf{X}) \text{ or simply } \mathcal{M}_\phi(\mathbf{X}) \leftrightarrow \tilde{\mathcal{M}}_\phi(\mathbf{X}), \tag{3}$$

when f is clear from context.

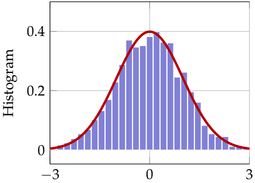
- without (entrywise) nonlinearities, $f(\mathbf{X})$ concentrates around expectation $f(\mathbf{X}) \simeq \mathbb{E}[f(\mathbf{X})]$, and can be assessed through **Deterministic Equivalent** $f(\bar{\mathbf{X}})$;
- for scalar eigenspectral functionals, Deterministic Equivalent for Resolvent framework provides a unified approach to eigenspectral functionals of random matrices;
- for nonlinear models in two different scaling regimes (LLN versus CLT), $\phi(\mathbf{X})$ can be linearized to yield a **Linear Equivalent**.

Concentration versus non-concentration behavior

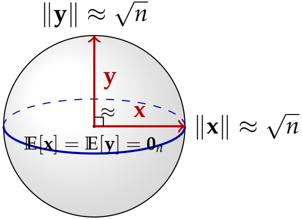
“Concentration” versus “non-concentration” around the mean

Consider two independent random vectors $\mathbf{x} = [x_1, \ldots, x_n]^\top$ and $\mathbf{y} = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$, with i.i.d. entries of zero mean and unit variance. We have the following observations.

1. In the one-dimensional case with $n = 1$, we have $\Pr(|x - 0| > t) \leq t^{-2}$ and $\Pr(|y - 0| > t) \leq t^{-2}$ by Markov’s inequality, so that one-dimensional random variables “concentrate” around their means.
2. In the multi-dimensional case with $n \geq 1$, we have $\mathbb{E}[\|\mathbf{x} - \mathbf{0}\|_2^2] = \mathbb{E}[\mathbf{x}^\top \mathbf{x}] = \operatorname{tr}(\mathbb{E}[\mathbf{x}\mathbf{x}^\top]) = n$ and $\mathbb{E}[\|\mathbf{x} - \mathbf{y}\|_2^2] = \mathbb{E}[\mathbf{x}^\top \mathbf{x} + \mathbf{y}^\top \mathbf{y}] = 2n$. Thus, for $n \gg 1$, the expected Euclidean distance between $\mathbf{x}$ and its mean $\mathbf{0}$ is large: high-dimensional random vectors do not “concentrate” around their means.



(a) “Concentration” around the mean


$\|\mathbf{y}\| \approx \sqrt{n}$
$\mathbf{y}$
$\approx$
$\mathbf{x}$
$\|\mathbf{x}\| \approx \sqrt{n}$
$\mathbb{E}[\mathbf{x}] = \mathbb{E}[\mathbf{y}] = \mathbf{0}_n$

(b) “Non-concentration” around the mean

High-dimensional concentration of scalar observation

- while large random vectors do not "concentrate" round their means, their scalar functionals (often) do
- for a scalar observation map $f\colon \mathbb{R}^n \to \mathbb{R}$ and random vector $\mathbf{x} \in \mathbb{R}^n$, we typically have

$$f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})] \to 0, \tag{4}$$

  with high probability for n large.
- a basic example is the linear function $f(\mathbf{x}) = \mathbf{1}_n^\top \mathbf{x}/n = \frac{1}{n}\sum_{i=1}^n x_i$: By the Large of Large Numbers (LLN) and the Central Limit Theorem (CLT), we have $f(\mathbf{x}) = \mathbb{E}[f(\mathbf{x})] + O(n^{-1/2})$ with high probability
- For a random matrix $\mathbf{X} \in \mathbb{R}^{p\times n}$ in the proportional regime with n, p both large, similar holds:

1. just as for vectors, $\mathbf{X}$ does not concentrate, e.g., in a spectral norm sense; e.g., $\|\mathbf{X} - \mathbb{E}[\mathbf{X}]\| \not\to 0$ as $n, p \to \infty$.
2. at the same time, scalar (e.g., eigenspectral) functionals $f\colon \mathbb{R}^{p\times n} \to \mathbb{R}$ of the random matrix $\mathbf{X}$ do concentrate; i.e., $f(\mathbf{X}) - \mathbb{E}[f(\mathbf{X})] \to 0$ as $n, p \to \infty$. This is the key idea of Deterministic Equivalent.

Definition (Deterministic Equivalent)

A Deterministic Equivalent is a special case of the High-Dimensional Equivalent, applied to a linear model $\mathcal{M}_\phi(\mathbf{X}) = \mathbf{X}$. We denote

$$f(\mathbf{X}) - f(\tilde{\mathbf{X}}) \to 0 \text{ as } n, p \to \infty \quad \Leftrightarrow \quad \mathbf{X} \overset{f}{\leftrightarrow} \tilde{\mathbf{X}} \text{ or simply } \mathbf{X} \leftrightarrow \tilde{\mathbf{X}}. \tag{5}$$

Nonlinear objects in two different scaling regimes

Definition (Two scaling regimes)

Consider a scalar functional $f(\mathbf{x})$ of $\mathbf{x} \in \mathbb{R}^n$, via an observation map $f \colon \mathbb{R}^n \to \mathbb{R}$:

1. **LLN regime**: this holds when $f(\mathbf{x})$ exhibits a LLN-type concentration, strongly concentrating around its mean $\mathbb{E}[f(\mathbf{x})]$, and its distribution function becomes degenerate; that is, it holds when $f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})] \to 0$ in probability or almost surely, as $n \to \infty$.
2. **CLT regime**: this holds when $f(\mathbf{x})$ exhibits a CLT-type concentration, remaining random and maintaining a non-degenerate distribution function; that is, it holds when $\sqrt{n}\,(f(\mathbf{x}) - \mathbb{E}[f(\mathbf{x})]) \to \mathcal{N}(0,1)$ in distribution, as $n \to \infty$.

Nonlinear objects in two scaling regimes

Let $\mathbf{x} \in \mathbb{R}^n$ be a random vector such that $\sqrt{n}\mathbf{x}$ has i.i.d. Gaussian entries $\mathcal{N}(0,1)$ (the $\sqrt{n}$ scaling ensures $\mathbb{E}[\|\mathbf{x}\|^2] = 1$). Let $\mathbf{y} \in \mathbb{R}^n$ be a deterministic vector of unit norm $\|\mathbf{y}\| = 1$. Consider two nonlinear objects:

1. **LLN regime**: random variables $f_{\mathrm{LLN}}(\mathbf{x}) = \|\mathbf{x}\|_2^2$ or $f_{\mathrm{LLN}}(\mathbf{x}) = \mathbf{x}^\top \mathbf{y}$ that both exhibit LLN-type concentration (i.e., nearly deterministic for n large), and we are interested in $\phi(f_{\mathrm{LLN}}(\mathbf{x}))$; and
2. **CLT regime**: random variables $f_{\mathrm{CLT}}(\mathbf{x}) = \sqrt{n}(\|\mathbf{x}\|_2^2 - 1)$ or $f_{\mathrm{CLT}}(\mathbf{x}) = \sqrt{n} \cdot \mathbf{x}^\top \mathbf{y}$ that both exhibit CLT-type concentration (they remain inherently random and have non-degenerate distributions for n large), and we are interested in $\phi(f_{\mathrm{CLT}}(\mathbf{x}))$.

Linearization in the two scaling regimes

Theorem (Taylor's theorem)

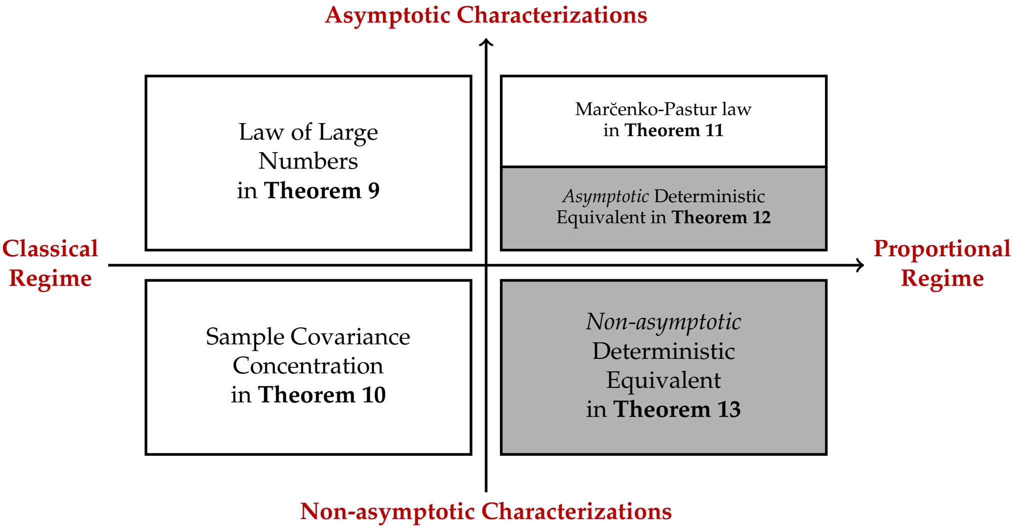
Let $\phi\colon \mathbb{R} \to \mathbb{R}$ be a function that is at least k times continuously differentiable in a neighborhood of some point $\tau \in \mathbb{R}$. Then, there exists $h_k\colon \mathbb{R} \to \mathbb{R}$ such that
$\phi(x) = \phi(\tau) + \phi'(\tau)(x-\tau) + \frac{\phi''(\tau)}{2}(x-\tau)^2 + \ldots + \frac{\phi^{(k)}(\tau)}{k!}(x-\tau)^k + h_k(x)(x-\tau)^k$, *with* $\lim_{x\to\tau} h_k(x) = 0$.
Consequently, $h_k(x)(x-\tau)^k = o(|x-\tau|^k)$ *as* $x \to \tau$.

Theorem (Hermite polynomial expansion)

The i^{th} normalized Hermite polynomial, $\mathrm{He}_i(t)$, is given by $\mathrm{He}_0(t) = 1, \mathrm{He}_i(t) = \frac{(-1)^i}{\sqrt{i!}} e^{\frac{t^2}{2}} \frac{d^i}{dt^i}\left(e^{-\frac{t^2}{2}}\right), i \geq 1$. *The normalized Hermite polynomials*

1. *are orthogonal with respect to Gaussian measure, i.e.,* $\int \mathrm{He}_m(t)\mathrm{He}_n(t)\mu(dt) = \delta_{mn}$ *for* $\mu(dt) = \frac{1}{\sqrt{2\pi}}e^{-\frac{t^2}{2}}dt$*; and*
2. *can be used to formally expand any square-integrable function $\phi \in L^2(\mu)$ as*
   $\phi(\xi) \sim \sum_{i=0}^{\infty} a_{\phi;i}\mathrm{He}_i(\xi), \quad a_{\phi;i} = \int \phi(t)\mathrm{He}_i(t)\mu(dt) = \mathbb{E}[\phi(\xi)\mathrm{He}_i(\xi)]$, *for* $\xi \sim \mathcal{N}(0,1)$. *The coefficients $a_{\phi;i}$s are the Hermite coefficients of ϕ:*

$$a_{\phi;0} = \mathbb{E}[\phi(\xi)],\ a_{\phi;1} = \mathbb{E}[\xi\phi(\xi)],\ \sqrt{2}a_{\phi;2} = \mathbb{E}[\xi^2\phi(\xi)] - a_{\phi;0},\ \nu_\phi = \mathbb{E}[\phi^2(\xi)] = \sum_{i=0} a_{\phi;i}^2. \tag{6}$$



Asymptotic Characterizations

Law of Large Numbers in **Theorem 9**

Marčenko-Pastur law in **Theorem 11**

Asymptotic Deterministic Equivalent in **Theorem 12**

Classical Regime

Proportional Regime

Sample Covariance Concentration in **Theorem 10**

Non-asymptotic Deterministic Equivalent in **Theorem 13**

Non-asymptotic Characterizations

Figure: Taxonomy of four different ways to characterize the sample covariance matrix $\hat{\mathbf{C}} = \frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}$.

Asymptotic behavior of SCM in the classical regime via law of large numbers

Theorem (Asymptotic Law of Large Numbers for SCM)

Let p be fixed, and let $\mathbf{X} \in \mathbb{R}^{p \times n}$ be a random matrix with independent sub-gaussian columns $\mathbf{x}_i \in \mathbb{R}^p$ such that $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}$ and $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^{\mathsf{T}}] = \mathbf{I}_p$. Then one has,

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \to 0, \tag{10}$$

almost surely, as $n \to \infty$.

- LLN is "parameterized" to hold only in the **classical limit**, not the **proportional limit**
- many variants and extensions of the LLN exist, but become vacuous when applied to the **proportional regime** $n, p \to \infty$ and $p/n \to c \in (0, \infty)$, see below for an example

Non-asymptotic behavior of SCM in the classical regime via matrix concentration

Theorem (Non-asymptotic matrix concentration for SCM, [Ver18, Theorem 4.6.1])

Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ be a random matrix with independent sub-gaussian columns $\mathbf{x}_i \in \mathbb{R}^p$ such that $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}$ and $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\mathsf{T}] = \mathbf{I}_p$. Then, one has, with probability at least $1 - 2\exp(-t^2)$, for any $t \geq 0$, that

$$\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \leq C_1 \max(\delta, \delta^2), \quad \delta = C_2(\sqrt{p/n} + t/\sqrt{n}), \tag{11}$$

for some constants $C_1, C_2 > 0$, independent of n, p.

Proof: combines Bernstein's concentration inequality with ϵ-net argument, see [Ver18] for details.

1. can reproduce the LLN asymptotic result by taking $n \to \infty$ with Borel–Cantelli lemma
2. **Classical regime.** Here, $n \gg p$, **say that** $n \sim p^2$. Then with high probability, that $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 = O(n^{-1/4})$ and conveys a similar intuition to the asymptotic LLN result
3. **Proportional regime.** Here, n, p **are both large and** $n \sim p$. Then, with high probability, that $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 = O(\sqrt{p/n}) = O(1)$, and qualitatively different LLN with a vacuous $\sim 100\%$ relative error, e.g., as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$.

Proportional regime: eigenvalues via traditional RMT and the Marčenko-Pastur law

Theorem (Limiting spectral distribution for SCM: Marčenko-Pastur law, [MP67])

Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ be a random matrix with i.i.d. sub-gaussian columns $\mathbf{x}_i \in \mathbb{R}^p$ such that $\mathbb{E}[\mathbf{x}_i] = \mathbf{0}$ and $\mathbb{E}[\mathbf{x}_i \mathbf{x}_i^\mathsf{T}] = \mathbf{I}_p$. Then, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, with probability one, the empirical spectral measure (ESD) $\mu_{\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}}$ of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ converges weakly to a probability measure μ given explicitly by

$$\boxed{\mu(dx) = (1 - c^{-1})^+ \delta_0(x) + \frac{1}{2\pi c x} \sqrt{(x - E_-)^+ \, (E_+ - x)^+} \, dx,} \tag{12}$$

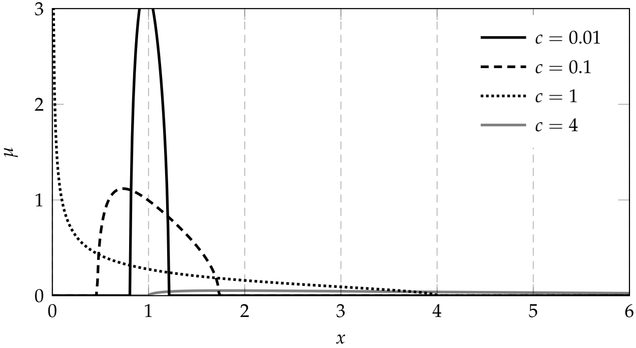
where $E_\pm = (1 \pm \sqrt{c})^2$ and $(x)^+ = \max(0, x)$, which is known as the Marčenko-Pastur distribution.

- provides a more refined characterization of the eigenspectrum of $\hat{\mathbf{C}}$ (than, e.g., matrix concentration):

(i) **Classical regime.** Here, $n \gg p$ **so that** $c = p/n \to 0$, the Marčenko-Pastur law in Equation (12) shrinks to a Dirac mass, in agreement with $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \sim 0$

(ii) **Proportional regime.** Here, $n \sim p \gg 1$, and by the (true but vacuous) matrix concentration result $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 = O(p/n) = O(1)$, and, depending on the ratio $c = p/n$, the eigenvalues of $\hat{\mathbf{C}}$ can be very different from one, and takes the form of the Marčenko-Pastur law

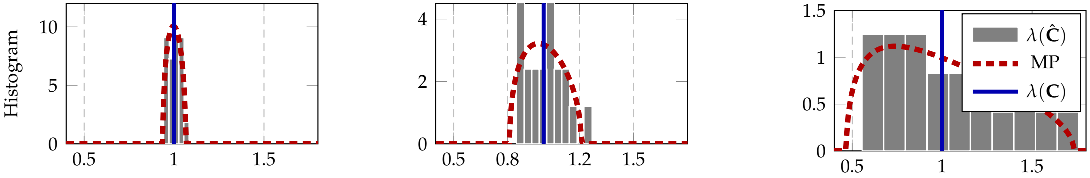
- we have in fact $\|\hat{\mathbf{C}} - \mathbf{I}_p\|_2 \simeq c + 2\sqrt{c}$ as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$



- averaged amount of eigenvalues of $\hat{\mathbf{C}}$ lying within the interval $[1-\delta, 1+\delta]$, for $\delta \ll 1$, as

$$\mu([1-\delta,1+\delta]) = \int_{1-\delta}^{1+\delta} \frac{1}{2\pi c x}\sqrt{\left(x-(1-\sqrt{c})^2\right)^+\left((1+\sqrt{c})^2-x\right)^+}\,dx$$
$$= \frac{1}{2\pi c}\int_{-\delta}^{\delta}\left(\sqrt{4c-c^2}+O(\varepsilon)\right)\,d\varepsilon = \frac{\sqrt{4c^{-1}-1}}{\pi}\delta + O(\delta^2).$$

- for $p \approx 4n$ there is asymptotically no eigenvalue of $\hat{\mathbf{C}}$ close to one!
- in accordance with the shape of the limiting Marčenko-Pastur law with $c = 4$ above

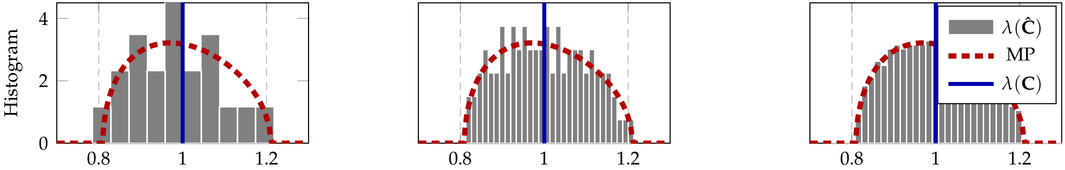


(a) $n = 1\,000p$

(b) $n = 100p$

(c) $n = 10p$

Figure: **Varying n and $c = p/n$ for fixed p.** Histogram of the eigenvalues of $\hat{\mathbf{C}}$ versus the limiting Marčenko-Pastur law in Theorem 11, for $\mathbf{X}$ having standard Gaussian entries with $p = 20$ and different $n = 1\,000p, 100p, 10p$ from left to right.



(a) $p = 20$

(b) $p = 100$

(c) $p = 500$

Figure: **Varying n and p for fixed $c = p/n$.** Histogram of the eigenvalues of $\hat{\mathbf{C}}$ versus the Marčenko-Pastur law, for $\mathbf{X}$ having standard Gaussian entries with $n = 100p$ and different $p = 20, 100, 500$ from left to right.

An asymptotic Deterministic Equivalent for resolvent

Theorem (An asymptotic Deterministic Equivalent for resolvent, [CL22, Theorem 2.4])

Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ be a random matrix having i.i.d. sub-gaussian entries of zero mean and unit variance, and denote $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T} - z\mathbf{I}_p)^{-1}$ the resolvent of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$ for $z \in \mathbb{C}$ not an eigenvalue of $\frac{1}{n}\mathbf{X}\mathbf{X}^\mathsf{T}$. Then, as $n, p \to \infty$ with $p/n \to c \in (0, \infty)$, the deterministic matrix $\bar{\mathbf{Q}}(z)$ is a Deterministic Equivalent of the random resolvent matrix $\mathbf{Q}(z)$ with

$$\mathbf{Q}(z) \leftrightarrow \bar{\mathbf{Q}}(z), \quad \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p, \tag{13}$$

with $m(z)$ the unique valid Stieltjes transform as solution to

$$czm^2(z) - (1 - c - z)m(z) + 1 = 0. \tag{14}$$

- The equation of $m(z)$ is quadratic and has two solutions defined via the complex square root
- **only one** satisfies $\Im[z] \cdot \Im[m(z)] > 0$ as a “valid” Stieltjes transform, and leads to the Marčenko-Pastur law

$$\boxed{\mu(dx) = (1 - c^{-1})^+ \delta_0(x) + \frac{1}{2\pi c x}\sqrt{(x - E_-)^+ (E_+ - x)^+}\, dx,} \tag{15}$$

for $E_\pm = (1 \pm \sqrt{c})^2$ and $(x)^+ = \max(0, x)$.

A non-asymptotic Deterministic Equivalent for resolvent

Theorem (A non-asymptotic Deterministic Equivalent for resolvent)

Let $\mathbf{X} \in \mathbb{R}^{p \times n}$ *be a random matrix having i.i.d. sub-gaussian entries with zero mean and unit variance, and denote* $\mathbf{Q}(z) = (\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}} - z\mathbf{I}_p)^{-1}$ *the resolvent of* $\frac{1}{n}\mathbf{X}\mathbf{X}^{\mathsf{T}}$ *for* $z < 0$*. Then, there exists universal constants* $C_1, C_2 > 0$ *depending only on the sub-gaussian norm of the entries of* $\mathbf{X}$ *and* $|z|$*, such that for any* $\varepsilon \in (0,1)$*, if* $n \geq (C_1 + \varepsilon)p$*, one has*

$$\|\mathbb{E}[\mathbf{Q}(z)] - \bar{\mathbf{Q}}(z)\|_2 \leq \frac{C_2}{\varepsilon} \cdot n^{-\frac{1}{2}}, \quad \bar{\mathbf{Q}}(z) = m(z)\mathbf{I}_p, \tag{16}$$

for $m(z)$ *the unique positive solution to the Marčenko-Pastur equation* $czm^2(z) - (1 - c - z)m(z) + 1 = 0, c = p/n$*.*

- this is a deterministic characterization of the **expected resolvent**
- to get DE, it remains to show **concentration** results for trace and bilinear forms: more or less standard

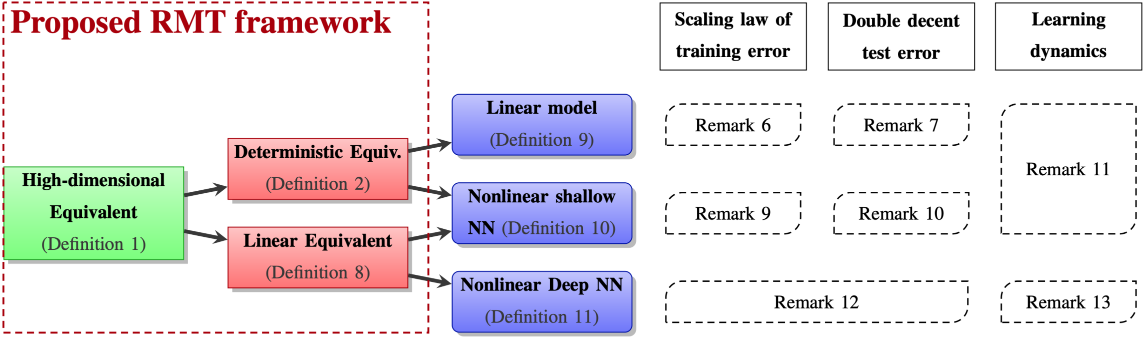
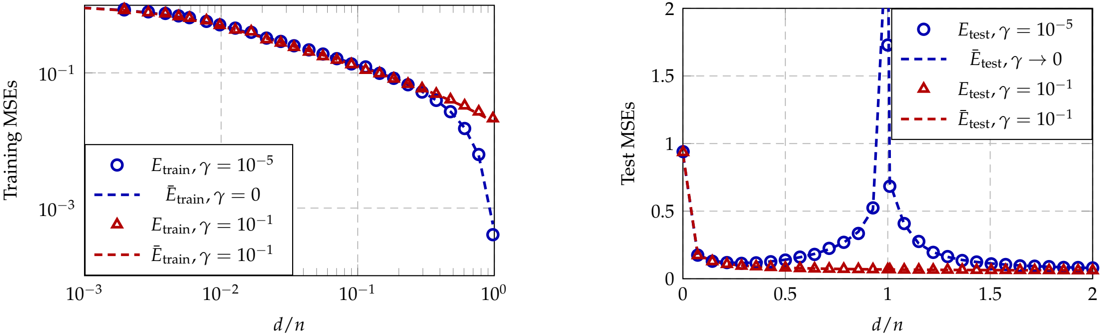

Proposed RMT framework
High-dimensional Equivalent (Definition 1)
Deterministic Equiv. (Definition 2)
Linear Equivalent (Definition 8)
Linear model (Definition 9)
Nonlinear shallow NN (Definition 10)
Nonlinear Deep NN (Definition 11)
Scaling law of training error
Double decent test error
Learning dynamics
Remark 6
Remark 7
Remark 11
Remark 9
Remark 10
Remark 12
Remark 13

Figure: Overview of [LM25], summarizing major concepts and results and where to find them.

Numerical results



(a) Training MSE

(b) Test MSE

Figure: Empirical and theoretical training and test MSEs of single-hidden-layer NN model, as a function of d/n, for $\gamma = 10^{-1}$ and $\gamma = 10^{-5}$, with Gaussian $\mathbf{W}$ and ReLU activation $\phi(t) = \max(t, 0)$, $n = 1\,024$ training samples and $n' = 1\,024$ test samples from the MNIST dataset (number 1 and 2).**Figure 7a**: log-log plot of training MSEs averaged over 30 runs. **Figure 7b**: test MSEs averaged over 30 runs on independent test sets of size $\hat{n} = 2\,048$.

Overview

Motivations:

- WeightWatcher, Weight Diagnostics for Analyzing ML Models
  (with Charles H. Martin)
- Randomized Numerical Linear Algebra for Modern ML
  (with Michal Derezinski)

Some Theory:

- RMT for NNs: Linear to Nonlinear; Shallow to Deep; etc.
  (with Zhenyu Liao)

Applications:

- Models of Heavy-Tailed Mechanistic Universality
  (with Zhichao Wang and Liam Hodgkinson)
- Spectral Estimation with Free Decompression
  (with Siavash Ameli, Chris van der Heide, and Liam Hodgkinson)
- Determinant Estimation under Memory Constraints and Neural Scaling Laws
  (with S. Ameli, C. van der Heide, L. Hodgkinson, and F. Roosta)

Motivation: Heavy-Tailed Phenomena in Modern Models

- Gradient norms (Simsekli et al., 2019) and loss curves (Hestness et al., 2017; Kaplan et al., 2020; Hoffmann et al., 2022).
- Eigenvalues of Gram matrices in neural nets: data covariance (Sorscher et al., 2022; Zhang et al., 2023), activation (conjugate kernel) (Pillaud-Vivien et al., 2018; Agrawal et al., 2022; Wang et al., 2023), Hessian (Xie et al., 2023), Jacobian (Wang et al., 2023).
- Strong correlation between heavy-tailed trained weight matrices & model performance: Heavy-Tailed Self-Regularization (HT-SR) Theory (Martin and Mahoney, 2021b) and Layer-wise Diagnostics (Zhou et al., 2023; Lu et al., 2024).
- Power law appears in neural scaling laws (Kaplan et al., 2020; Wei et al., 2022; Defilippis et al., 2024; Paquette et al., 2024; Lin et al., 2024).

Need new RMT for **Heavy-Tailed Mechanistic Universality (HT-MU)**.

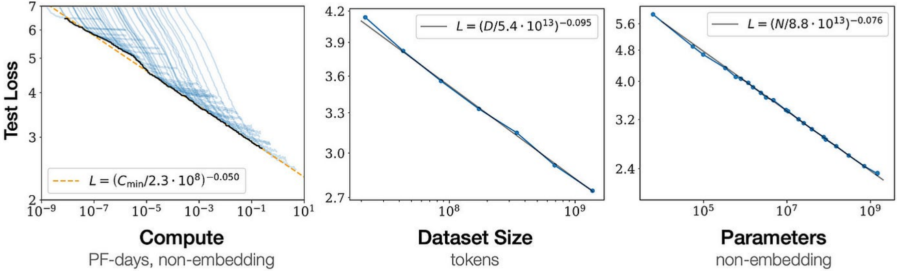
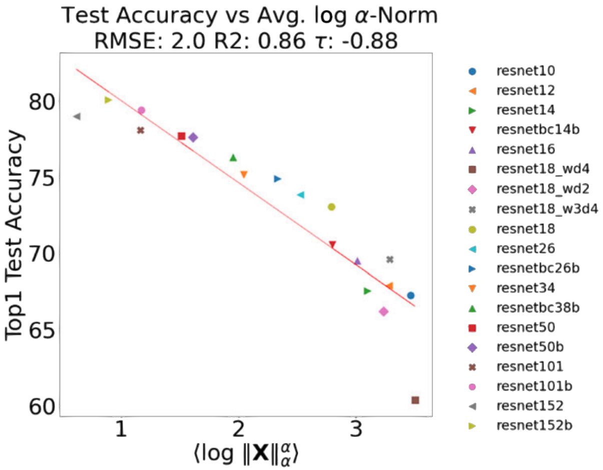


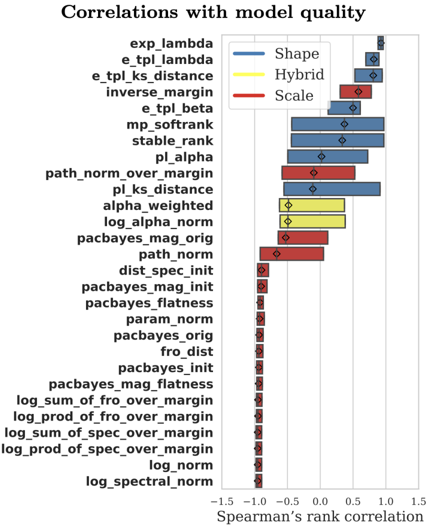
Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute[2] used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Kaplan et al. (2020). Scaling laws for neural language models.

Hoffmann et al. (2022). Training compute-optimal large language models.



Martin, C. H., Peng, T., & Mahoney, M. W. (2021). Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data. Nature Communications, 12(1), 4122.



Yang, Y., Theisen, R., Hodgkinson, L., Gonzalez, J. E., Ramchandran, K., Martin, C. H., & Mahoney, M. W. (2023). *Test accuracy vs. generalization gap: model selection in NLP without accessing training or testing data.*

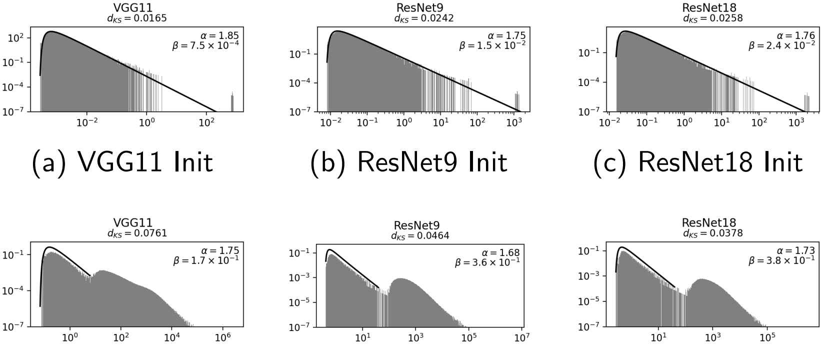
Heavy-Tailed Mechanistic Universality

What might constitute "universality" in neural network weights?

- In RMT:
  - it denotes the emergence of system-independent properties derivable from a few global parameters defining an ensemble.
- In statistical physics:
  - it arises in systems with very strong correlations, at or near a critical point or phase transition;
  - it is characterized by measuring experimentally "observables" that display heavy-tailed behavior, with (universal) power law exponents.

Although trained weight matrices are *not* random, but rather strongly correlated through training, RMT provides a useful descriptive framework.

NTK Spectra at Initialization vs. Post-Training



(a) VGG11 Init

(b) ResNet9 Init

(c) ResNet18 Init

(d) VGG11 Trained

(e) ResNet9 Trained

(f) ResNet18 Trained

Figure: NTK eigenvalue histograms and inverse-Gamma fits near zero. Initialization: mild inverse-Gamma behavior. Post-Training: pronounced heavy-tail

Heavy-Tailed Mechanistic Universality

Definition

Heavy-tailed distributions (informally): densities decaying slower than exponential, often exhibiting power-law tails

$$f(x) \sim c\,x^{-\alpha}, \quad x \to \infty,$$

or inverse-Gamma behavior near zero $f(x) \sim c\,x^{\alpha} e^{-\beta/x}, \quad x \to 0^{+}$.

Possible Approaches for Describing HT-MU:

- iid Heavy-Tailed Elements: (Arous and Guionnet, 2008) Elements of feature matrices are not independent and heavy-tailed.
- Kesten Phenomenon: (Hodgkinson and Mahoney, 2021; Vladimirova et al., 2018; Hanin and Nica, 2020) a mechanism discovered by Kesten (1973) for recursive systems.
- Population Covariance: power-law in, power-law out (PIPO) principle.

Comparison of Possible Mechanisms

	Power Law		Inverse
Mechanism	**Elements**	**Spectrum**	**Gamma**
iid Heavy-Tailed Elements	✓	✓	×
Kesten Phenomenon	✓	✓	✓/×
Population Covariance	✓/×	✓	✓/×
Structured Matrices (Ours)	×	✓	✓
Empirical Observations (Features)	×	✓	✓
Empirical Observations (Weights)	×	✓	×

Table: Comparison of various mechanisms: capacity to yield power laws, in feature matrix **elements** and feature matrix **spectral densities**; capacity to yield an inverse Gamma law for the spectral density in a neighborhood of zero.

Modeling Framework

Entropic Regularization Setup

- **Stochastic Minimization Operator**

$$\underset{\Theta}{\overset{\pi_\Theta,\tau}{\mathrm{smin}}}\, f(\Theta) \;:=\; \min_{q\in\mathcal{P}} \left[\mathbb{E}_{q(\Theta)}[f(\Theta)] \;+\; \tau\,\mathrm{KL}(q \,\|\, \pi_\Theta)\right],$$

  where $\mathcal{P}$ is the set of probability densities on the support of π_Θ, and

  - π_Θ is the initial prior ($\Theta =$ model coefficients).
  - $\tau > 0$ is the "temperature" (controls early stopping).

- Stochastic optimization models (Mandt et al., 2016; Chaudhari and Soatto, 2018) have strong links to Bayesian inference (Germain et al., 2016) and statistical physics of generalization (Mezard and Montanari, 2009).

- Applying to the training loss optimizes a PAC-Bayes bound on the test error (Xie et al., 2023). As τ decreases during training, optimizer smoothly interpolates between π_Θ and the final optimal density.

Entropic Regularization Setup

Feature Learning Setup: Stochastic minimization in *two stages*

$$\overset{\pi_\Theta,\tau}{\underset{\Theta}{\text{smin}}}\, L(\Theta,\Phi) \quad \text{and} \quad q(\Phi) \;=\; \overset{\pi_\Phi,\eta}{\underset{\Phi}{\text{argsmin}}}\left[\overset{\pi_\Theta,\tau}{\underset{\Theta}{\text{smin}}}\, L(\Theta,\Phi)\right].$$

- π_Θ, π_Φ: initial densities of model coefficients Θ and features Φ.
- $\tau, \eta > 0$: "temperatures" control coefficient vs. feature learning rates.

Proposition (Optimal Feature Density)

$$q(\Phi) \;\propto\; \left[\mathcal{Z}_\tau(\Phi)\right]^{\tau/\eta} \pi_\Phi(\Phi), \quad \mathcal{Z}_\tau(\Phi) \;=\; \mathbb{E}_{\Theta\sim\pi_\Theta} \exp\left(-L(\Theta,\Phi)/\tau\right).$$

Of particular interest: late stage of training, $\tau, \eta \to 0^+$ with $\tau/\eta \to \rho > 0$.

Master Model Ansatz

- **Ansatz:** for trained feature matrices, with parameters $\alpha, \beta > 0$ and initial density π:

$$q(M) \ \propto \ (\det M)^{-\alpha} \exp\big(-\beta \, \mathrm{tr}(\Sigma \, M^{-1})\big) \, \pi(M)$$

  - $\alpha, \beta > 0$ depend on model/optimizer hyperparameters.
  - Σ is label/covariance-related (e.g., $Y \, Y^\top$).
  - $\pi(M)$ is the prior "initialization" density of the feature matrix.

- **Key Observation:** The trained feature matrix M generally follows an *inverse-Wishart-type density* (Mardia et al., 2024).
  1. First consider $\Sigma = I$ to remove the effect of Σ, the density π of feature matrices M at initialization completely determines the density $q(M)$. Change of variables $M \mapsto Q \Lambda Q^\top$ for orthogonal Q and diagonal Λ; so we only need to study the spectral distribution Λ.
  2. Second, we will consider a general Σ to get spectral densities of trained feature/weight matrices.

RMT for Heavy-Tailed Spectral Behavior

Eigenvector Structure and Beta-Ensembles

- To derive a *spectral density* from the Master Model Ansatz, diagonalize $M = Q\,\mathrm{diag}(\lambda)\,Q^{\top}$ and set $\Sigma = I$.
- **Key Assumption:** *Distribution of eigenvectors Q is not uniform!* (non-Haar) due to implicit model biases.
- Use **Beta-Ensemble** (Dumitriu and Edelman, 2002; Forrester, 2010) with parameter $\kappa \in [0, \infty]$ to capture the Master Model Ansatz:

$$q_{\kappa}(\lambda_1, \ldots, \lambda_N) \;\propto\; \prod_{i=1}^{N} V(\lambda_i)\; \prod_{i<j} |\lambda_i - \lambda_j|^{\kappa/N}$$

- Take $V(\lambda) = \lambda^{-\alpha}\,\exp(-\beta\,\lambda^{-1})$ to match *Master Model Ansatz*.
- The $1/N$ "high temperature" scaling has also been examined (Forrester and Mazzuca, 2021), but with a different application.
- Although $\pi(M)$ could be complicated, we argue that much of the behavior of π is captured by the extent of the eigenvalue repulsions. κ controls *eigenvalue repulsion*.

Main Theorem: HTMP Distribution

Theorem (Generalized Marchenko–Pastur)

Let M_N follow $q_\kappa(\lambda_1, \ldots, \lambda_N) \propto \prod_{i=1}^{N} \lambda_i^{-\alpha} e^{-\beta \lambda_i^{-1}} \prod_{i<j} |\lambda_i - \lambda_j|^{\frac{\kappa(N)}{N}}$ with parameter $\kappa(N)$. Define

$$\gamma(N) = \frac{\kappa(N)/2}{\alpha - \kappa(N)/2 - 1} \to \gamma \in (0, 1) \quad \textit{as } N \to \infty.$$

Then the empirical spectral distribution of $\frac{2\gamma(N)\beta}{\kappa(N)} M_N^{-1}$ converges to:

1. **MP**$_\gamma$ *(Marchenko-Pastur distribution) if $\kappa(N) \to \infty$;*
2. **HTMP**$_{\gamma,\kappa}$ *(High-Temperature MP) if $\kappa(N) \to \kappa \in (0, \infty)$.*

This beta-ensemble result is derived from a sequence of random matrix theory from Dumitriu and Edelman (2006); Dung and Duy (2021).

Main Theorem: Tail Behavior for Trained Features

Theorem (Spectral Density of Trained Feature Matrix)

Let ρ_N be the ESD of a trained feature matrix M_N, and μ_Σ the spectral measure of label covariance Σ. Then

$$\boxed{\rho_N(\lambda) \xrightarrow[N\to\infty]{} \big(\mu_\Sigma \boxtimes \rho\big)(\lambda),}$$

where $\boxtimes$ is multiplicative free convolution, ρ is either $\lambda^{-2}\,\rho_{\mathrm{MP}}(\lambda^{-1})$ (if $\kappa = \infty$) or $\lambda^{-2}\,\rho_{\mathrm{HTMP}}(\lambda^{-1})$ (if $\kappa < \infty$). Additionally,

- *Bounded vs. Heavy-Tailed: $\kappa = \infty \implies$ bounded support; $\kappa < \infty \implies$ power-law tail.*
- *Inverse-Gamma near zero: If $\kappa < \infty$, density $\rho(x) \sim x^{-\frac{\kappa}{2\gamma}-1-\frac{\kappa}{2}} \exp\left(-\frac{\beta_-}{x}\right)$ as $x \to 0^+$.*
- *Power-law Tail: $\rho(x) \sim x^{-\frac{\kappa}{2\gamma}-1+\frac{\kappa}{2}}$ for $x \to \infty$.*

Remarks

- The power law for the limiting density ρ contains a tail exponent that gets heavier as κ decreases: i.e., as the structure of the underlying matrix becomes more rigid.
- Decreasing κ increases implicit model bias, consistent with Martin and Mahoney (2021b) and Simsekli et al. (2019), who claim heavier tails imply stronger model biases and better model quality and generalization ability.[1]
- HTMP model represents the first RMT ensemble that captures key empirical properties of (strongly-correlated) modern state-of-the-art neural networks (Martin and Mahoney, 2020, 2021a,b; Yang et al., 2023).

[1] Very important: these models' elements need *not* have heavy-tailed behavior.

Application 1: Neural Scaling Laws

- **Setup:** Ridge regression on activation matrix $\Phi \in \mathbb{R}^{n \times d}$, $m = 1$:

$$\hat{w} = \operatorname*{argmin}_{w} L(w) \; = \; \frac{1}{n}\|\Phi w - Y\|^2 + \frac{\mu}{n}\|w\|^2.$$

  Assume $y_i = w_*^\top \varphi(x_i)$, and $\mathbb{E}_x[\varphi(x)\varphi(x)^\top] = I$.

- **Spectral Assumption:** $\Phi\Phi^\top$ follows $\textbf{HTMP}_{\gamma,\kappa}$ (Master Model).
- *Data-Free Scaling Law:* Predicts test loss decay solely from spectral tail; no access to held-out data required. Previous scaling law works focus on power laws in the dataset (e.g., Wei et al., 2022; Defilippis et al., 2024; Paquette et al., 2024; Lin et al., 2024)

Proposition

Let $\mu = n^{-\ell}$ with $\ell \in (0, 1)$. Then, the **Generalization Error** *satisfies*

$$\mathcal{L} := \mathbb{E}_{x,w_*}[(\varphi(x)^\top \hat{w} - y)^2] \; \asymp \; n^{-\ell\left(2 + \frac{\kappa}{2\gamma} - \frac{\kappa}{2}\right)}, \quad n \to \infty$$

with high probability.

Application 2: Optimizer Trajectories

- Empirical observation (Mandt et al., 2016; Simsekli et al., 2019; Hodgkinson et al., 2022): *Lower* and *Upper* power-law tails in the distribution of stochastic gradient norms $\|\widehat{\nabla} L_N\|$ during training:

$$\Pr(\|\widehat{\nabla} L_N\| \leq x) \sim C_- \, x^{\alpha}, \quad x \to 0^+,$$
$$\Pr(\|\widehat{\nabla} L_N\| > x) \sim C_+ \, x^{-\beta}, \quad x \to \infty.$$

- **Model:** Assume residuals $\bar{Y}$ are Gaussian, NTK matrix $J \sim$ inverse-Wishart (or **HTMP**) independent of $\bar{Y}$.
- **Application:** Under these assumptions, $\|\widehat{\nabla} L_N\|$ exhibits both lower and upper power-law tails.
- There has been significant theoretical justification for the upper power law in terms of the Kesten mechanism (Hodgkinson and Mahoney, 2021; Gurbuzbalaban et al., 2021, 2022), but there has been little justification for the lower power law before.

Application 3: 5+1 Phases of Trained Weight Matrices

- **Empirical Observation** (Martin and Mahoney, 2019, 2020, 2021b; Yang et al., 2023; Zhou et al., 2023): Trained weight matrices can exhibit 5+1 Phases of Training:
  1. Random-Like (MP bulk, no outliers).
  2. Bleeding-Out (MP bulk with emerging spikes).
  3. Bulk+Spikes (distinct spikes outside bulk).
  4. Bulk-Decay (bulk extends, no finite support).
  5. Heavy-Tailed (power-law tail).
  6. Rank-Collapse (mass at zero eigenvalue).
- **Application:** Consider $A = W^\top W$ with trained weight W, then $\frac{\beta}{\alpha - \kappa/2 - 1} A$ converges to $\mathbf{HTMP}_{\gamma,\kappa}$.
- Decreasing κ across training $\Rightarrow$ transition from bounded support to heavy tail. Power law exponents in the spectrum of weight matrices are strongly predictive of model performance.

5+1 Phases for Trained Weight: HTMP Fits

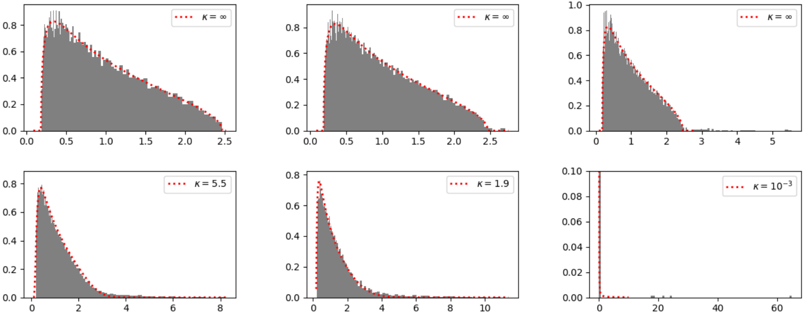


Figure: Weight spectral densities for MiniAlexNet trained on CIFAR-10 with batch sizes 1000, 800, 250, 100, 50, 5 (top to bottom). *Fitted MP/HTMP curves shown in red dashed* with different κ.

As batch size decreases, κ decreases $\Rightarrow$ heavier tail.
(a)–(c): $\kappa = \infty$ for MP or MP+spike behavior.
(d)–(f): Finite κ for heavy tail plus eventual rank collapse.

Conclusions

- **Master Model**: A unified RMT framework (Master Model Ansatz) that captures heavy-tailed spectral behavior of trained feature matrices from a Bayesian perspective.
- **HTMP Ensemble**: High-temperature MP ($\textbf{HTMP}_{\gamma,\kappa}$) arises when eigenvector entropy $\propto \kappa$ is finite; interpolates between MP ($\kappa \to \infty$) and heavy-tailed regimes ($\kappa \to 0^+$).
- **Key Insights**
  1. *Data Contribution*: Heavy-tailed population covariance $\Sigma \implies$ heavy-tailed trained spectra (PIPO).
  2. *Eigenvector Structure*: More architectural bias (smaller κ) $\implies$ heavier tails.
  3. *Training Dynamics*: As $\tau, \eta \to 0$, HTMP hyperparameters α, β, κ evolve, explaining transitions (5+1 phases).
- **Applications**
  - Neural scaling laws (ridge regression) predicted by HTMP exponents.
  - Lower/upper power-law tails in SGD trajectories explained.
  - 5+1 training phases fit by tuning κ for HTMP.

Overview

Motivations:

- WeightWatcher, Weight Diagnostics for Analyzing ML Models (with Charles H. Martin)
- Randomized Numerical Linear Algebra for Modern ML (with Michal Derezinski)

Some Theory:

- RMT for NNs: Linear to Nonlinear; Shallow to Deep; etc. (with Zhenyu Liao)

Applications:

- Models of Heavy-Tailed Mechanistic Universality (with Zhichao Wang and Liam Hodgkinson)
- Spectral Estimation with Free Decompression (with Siavash Ameli, Chris van der Heide, and Liam Hodgkinson)
- Determinant Estimation under Memory Constraints and Neural Scaling Laws (with S. Ameli, C. van der Heide, L. Hodgkinson, and F. Roosta)

Tiers of Matrix Difficulty

Explicit: the whole matrix fits in memory

Implicit: can make use of matrix-vector products (e.g. CG, SLQ)

Out-of-core: parts of the matrix can be loaded into memory a piece at a time

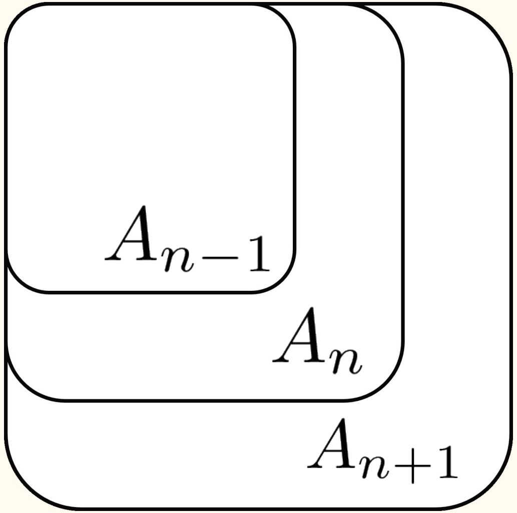
Impalpable: most matrix entries are inaccessible, matrix-vector products are unavailable (e.g. distributed or enormous datasets)

	Access in Memory		
		Matrix–Vector	Any
Type	Matrix	Product	Subblock
Explicit	✓	✓	✓
Implicit	✗	✓	✗
Out-of-core	✗	∼	✓
Impalpable	✗	✗	✗

Extrapolating Matrices

Suppose our matrix of interest is embedded in an infinite sequence of nested matrices

$$A_1, A_2, A_3, \ldots \qquad A_n \in \mathbb{R}^{n \times n}$$



so that $(A_n)_{ij} = (A_{n+1})_{ij}$

Objective: Find eigenspectrum of A_n using eigenspectrum of A_{n_s} for $n_s \ll n$

Free Probability

How do we ensure the eigenvalues of submatrices represent the whole matrix?

An important topic in random matrix theory involving random matrices with uniformly random eigenvectors, so that probability distributions of matrix dependents (including submatrices) *depend only on the eigenspectra*.

Theorem (Nica, 1993): Any sequence of matrices can be turned into an (asymptotically) free sequence of random matrices by applying random permutations σ to the rows and columns:

$$\tilde{A}_{ij} = A_{\sigma(i)\sigma(j)}$$

Free Decompression

Let $m(t, \cdot)$ be the Stieltjes transform of the enlargement of A by a factor of e^t
Under the large matrix limit, $m(t, \cdot)$ satisfies the *partial differential equation:*

$$\frac{\partial m}{\partial t} = -m + \frac{1}{m} \frac{\partial m}{\partial z}$$

Proof: Random matrix theory arguments involving the R-transform and the celebrated theorem of (Nica & Speicher, 1996).

To our knowledge, this operation has always been considered in reverse (*free compression*), finding eigenspectra of submatrices, given the eigenspectrum of the full matrix. We are the first to attempt ***free decompression***.

Free decompression of a random submatrix $\mathbf{A}_n$ to a larger matrix $\mathbf{A}$ requires:

1. **estimation** of its Stieltjes transform $m_{\mathbf{A}_n}$;
2. **evolution** of $m_{\mathbf{A}_n}$ in n using PDE;
3. **evaluation** of the spectral distribution of $\mathbf{A}$.

An Engineering Challenge

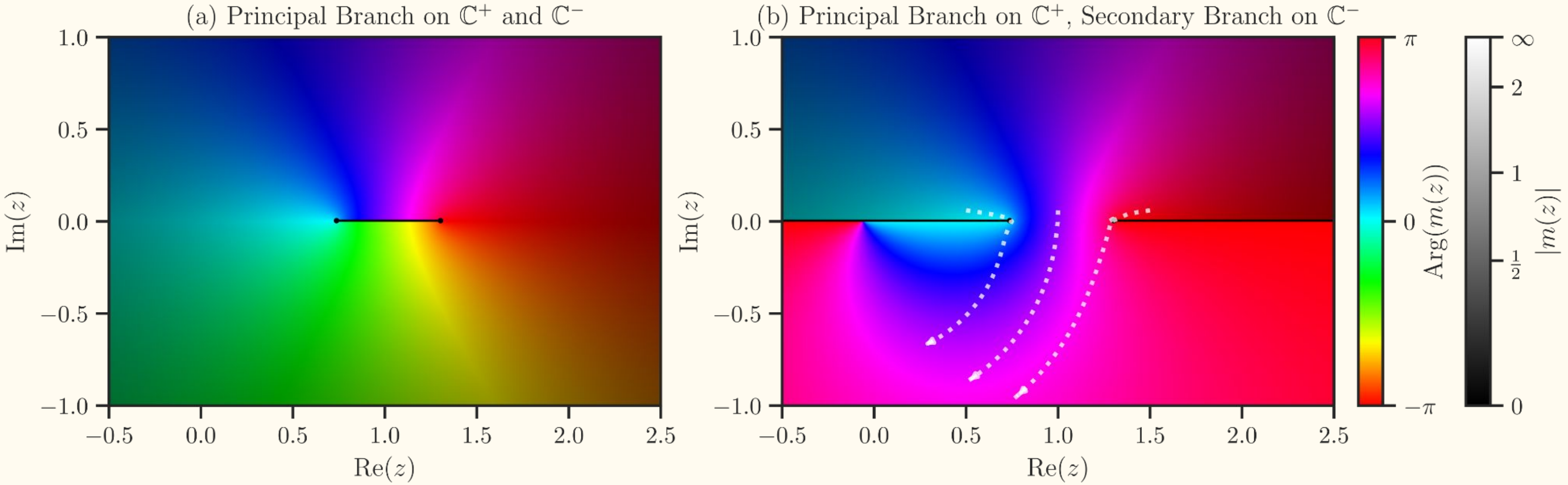
This is a very difficult equation to solve!

Solve the PDE using method of characteristics in the complex plane. But…

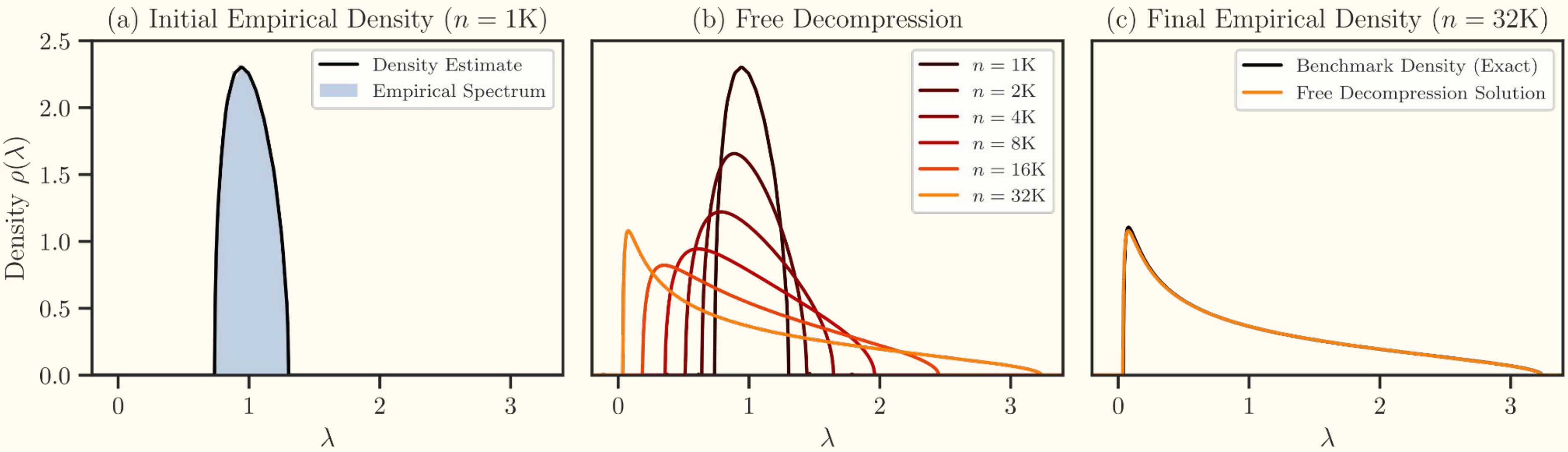
Proposition: All characteristic curves pass through the (discontinuous) branch cut for the principal branch of the Stieltjes transform.

- ➢ To solve the characteristic equations, a new *secondary* branch is required.
- ➢ Tantamount to (ill-posed) numerical analytic continuation.
- ➢ Naively solving the PDE fails: we need to directly tackle the analytic continuation problem.

Analytic Continuation of Stieltjes Transform



Wishart Matrices (Marchenko-Pastur Law)

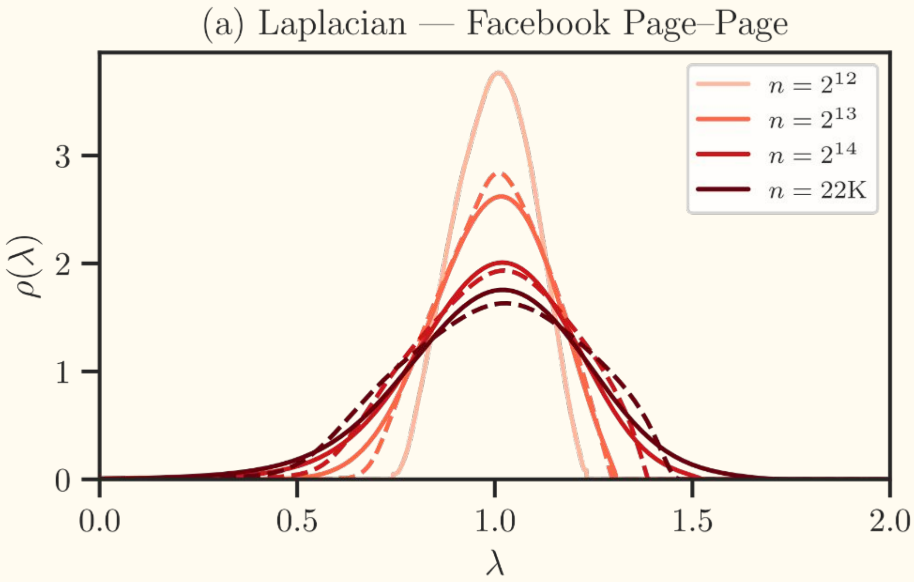


Histogram of eigenvalues of small matrix & density estimate

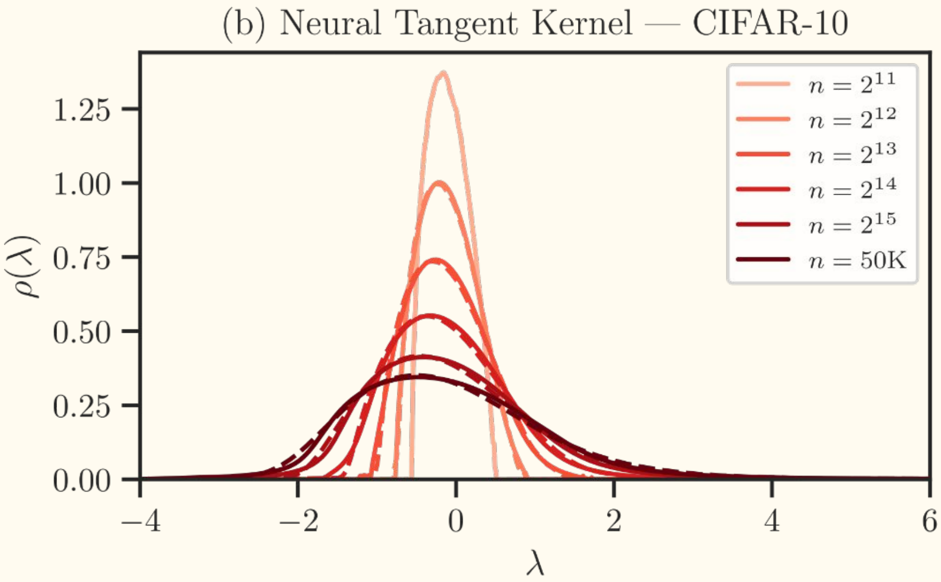
Densities under free decompression

Expected density & solution from free decompression

Experiments with Real Data



Symmetrically normalized Laplacian matrix of the SNAP Facebook dataset



log-NTK matrix computed from the CIFAR-10 dataset using a ResNet-50 model

Empirical spectral density (solid) vs. free decompression estimate from $n = 2^{11}$ (dashed)

freealg

freealg is our Python package that implements free decompression for estimating eigenspectra.

`pip install freealg`

(work in progress!)

Listing 1: A minimal usage example of the `freealg` package.

```
# Install freealg with "pip install freealg"
import freealg as fa

# Create an object for the Marchenko--Pastur distribution with the parameter $\lambda = \frac{1}{50}$
mp = fa.distributions.MarchenkoPastur(1/50)

# Generate a matrix of size $n_s = 1000$ corresponding to this distribution
A = mp.matrix(size=1000)

# Create a free-form object for the matrix within the support $I = [\lambda_-, \lambda_+]$
ff = fa.FreeForm(A, support=(mp.lam_m, mp.lam_p))

# Fit the distribution using Jacobi polynomials of degree $K = 20$, with $\alpha = \beta = \frac{1}{2}$
# Also fit the glue function via Pade of degree $[(p+q)/q]$ with $p = 0$, $q = 1$.
psi = ff.fit(method='jacobi', K=20, alpha=0.5, beta=0.5, reg=0.0, damp='jackson',
             pade_p=0, pade_q=1, optimizer='ls', plot=True)

# Decompress the spectral density corresponding to a larger matrix of size $n = 2^5 \times n_s$,
rho_large = ff.decompress(size=32_000, plot=True)
```

arXiv

Siavash Ameli, Chris van der Heide, Liam Hodgkinson, Michael W. Mahoney. (2025) *Spectral Estimation with Free Decompression*. arxiv: 2506.11994

Overview

Motivations:

- WeightWatcher, Weight Diagnostics for Analyzing ML Models (with Charles H. Martin)
- Randomized Numerical Linear Algebra for Modern ML (with Michal Derezinski)

Some Theory:

- RMT for NNs: Linear to Nonlinear; Shallow to Deep; etc. (with Zhenyu Liao)

Applications:

- Models of Heavy-Tailed Mechanistic Universality (with Zhichao Wang and Liam Hodgkinson)
- Spectral Estimation with Free Decompression (with Siavash Ameli, Chris van der Heide, and Liam Hodgkinson)
- Determinant Estimation under Memory Constraints and Neural Scaling Laws (with S. Ameli, C. van der Heide, L. Hodgkinson, and F. Roosta)

Overview

Log-determinant is widely encountered in linear algebra and statistics:

- Gaussian process (kernel methods)
- Determinantal point process
- Volume form (Bayesian computation)

Challenges

- It is often **the most difficult term** to compute in these applications.
- **Memory-wall** (time complexity isn't the only bottleneck)

Outline

I. Large Matrices

- Neural Tangent Kernels
- Arithmetic Precision

II. MEMDET

- Compute exact log-det
- Out-of-core

III. FLODANCE

- Approximate log-det
- Utilize scale law

IIII. Results

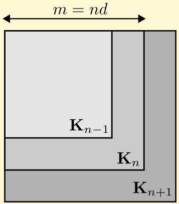
- NTK matrices
- Matérn kernel

Neural Tangent Kernel Sizes

			Matrix Size		
Dataset	Training Set	Classes	float16	float32	float64
CIFAR-10	50,000	10	0.5 TB	1.0 TB	2.0 TB
MNIST	60,000	10	0.72 TB	1.5 TB	2.9 TB
SVHN	73,257	10	1.1 TB	2.2 TB	4.2 TB
ImageNet-1k	1,281,167	1000	3,282,778 TB	6,565,556 TB	13,131,111 TB[*]

[*] 13.1 exabytes is an order of magnitude larger than CERN's current data storage capacity.

Scale Law



$$\frac{\det\left(\mathbf{K}_n\right)}{\det\left(\mathbf{K}_{n-1}\right)} \sim n^{\nu}$$

- n: num dataset
- d: num classes
- $m = nd$: matrix size

Lemma

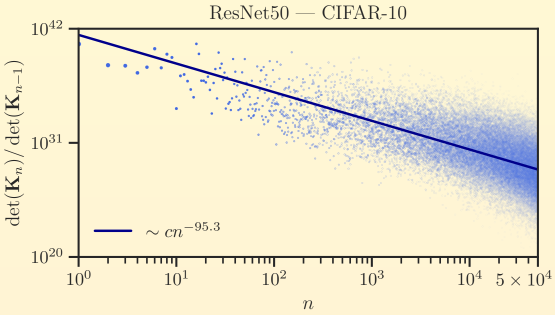
Let $f : \mathcal{X} \to \mathbb{R}^d$ be a zero-mean vector-valued m-dimensional Gaussian process with covariance kernel κ. For each $n \geq 2$, let

$$E(n) := \mathbb{E}[d^{-\frac{1}{2}} \|f(x_n)\|^2 \mid f(x_i) = 0$$

denote the mean-squared error of fitting the f to the zero function using $x_1, \ldots, x_{n-1}$. Then

$$\frac{\mathrm{pdet}(\mathbf{K}_n)}{\mathrm{pdet}(\mathbf{K}_{n-1})} \leq E(n)^d, \quad \forall n > 1,$$

with equality if $d = 1$.



- NTK of ResNet50 on CIFAR-10
- Number of classes: $d = 10$
- Dataset images: $n = 50\text{K}$
- Matrix size: $m = 500\text{K}$.

Comparison of Methods

Method			Rel.	Est.	Wall
Name	Settings	TFLOPs	Error	Cost	Time
SLQ	$l = 100$, $s = 104$	5203	55%	\$83	1.8 days
MEMDET	LDL, $n_b = 32$	41,667	**0%**	\$601	13.8 days
FLODANCE	$n_s = 500$, $q = 0$	**0.04**	4%	\$0.04	1 min
FLODANCE	$n_s = 5000$, $q = 4$	41.7	**0.02%**	\$4	1.5 hr

- **Largest NTK formation** and **exact logdet computation** to our knowledge
- ResNet50, full CIFAR-10 with all $n = 50\text{K}$ images
- Matrix size $m = 500{,}000$ dense matrix, **double precision**, **2TB** size.
- MEMDET computes the **exact** log-determinant, serves as **benchmark**.
- Costs and wall time are based on an NVIDIA H100 GPU (\$2/hour).
- Wall time include NTK formation.

Resources

Reference

Ameli, S., van der Heide, C., Hodgkinson, L., Roosta, F., Mahoney, M.W., (2025). Determinant Estimation under Memory Constraints and Neural Scaling Laws, *The 42nd International Conference on Machine Learning*.

Related Work

Ameli, S., van der Heide, C., Hodgkinson, L., Mahoney, M.W., (2025). Spectral Estimation with Free Decompression. *arXiv: 2506.11994*

Software

Package	Documentation	Install	Implements
detkit	`ameli.github.io/detkit`	`pip install detkit`	MEMDET FLODANCE
imate	`ameli.github.io/imate`	`pip install imate`	SLQ
freealg	`ameli.github.io/freealg`	`pip install freealg`	(Related work)

Overview

Motivations:

- WeightWatcher, Weight Diagnostics for Analyzing ML Models (with Charles H. Martin)
- Randomized Numerical Linear Algebra for Modern ML (with Michal Derezinski)

Some Theory:

- RMT for NNs: Linear to Nonlinear; Shallow to Deep; etc. (with Zhenyu Liao)

Applications:

- Models of Heavy-Tailed Mechanistic Universality (with Zhichao Wang and Liam Hodgkinson)
- Spectral Estimation with Free Decompression (with Siavash Ameli, Chris van der Heide, and Liam Hodgkinson)
- Determinant Estimation under Memory Constraints and Neural Scaling Laws (with S. Ameli, C. van der Heide, L. Hodgkinson, and F. Roosta)