Traditional and Heavy-Tailed Self Regularization in Neural Network Models

Charles H. Martin & Michael W. Mahoney

ICML, June 2019

(charles@calculationconsulting.com & mmahoney@stat.berkeley.edu)

Motivations: towards a Theory of Deep Learning

Theoretical: deeper insight into *Why Deep Learning Works*?

- convex versus non-convex optimization?
- explicit/implicit regularization?
- is / why is / when is deep better?
- VC theory versus Statistical Mechanics theory?
- ...

Practical: use insights to improve engineering of DNNs?

- when is a network fully optimized?
- can we use labels and/or domain knowledge more efficiently?
- large batch versus small batch in optimization?
- designing better ensembles?
- ...

How we will study regularization

The Energy Landscape is *determined* by layer weight matrices $\mathbf{W}_L$:

$$E_{DNN} = h_L(\mathbf{W}_L \times h_{L-1}(\mathbf{W}_{L-1} \times h_{L-2}(\cdots) + \mathbf{b}_{L-1}) + \mathbf{b}_L)$$

Traditional regularization is applied to $\mathbf{W}_L$:

$$\min_{W_l, b_l} \mathcal{L}\left(\sum_i E_{DNN}(d_i) - y_i\right) + \alpha \sum_l \|\mathbf{W}_l\|$$

Different types of regularization, e.g., different norms $\|\cdot\|$, *leave different empirical signatures on* $\mathbf{W}_L$.

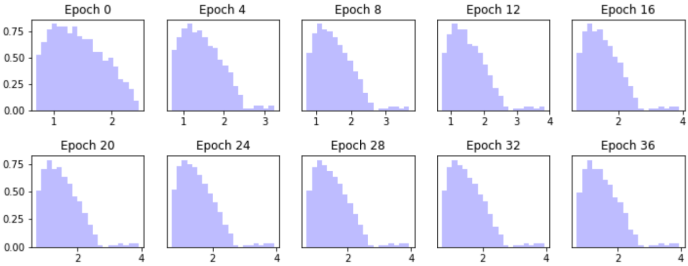
What we do:

- Turn off "all" regularization.
- Systematically turn it back on, explicitly with α or implicitly with knobs/switches.
- Study empirical properties of $\mathbf{W}_L$.

ESD: detailed insight into W_L

Empirical Spectral Density (ESD: eigenvalues of $X = \mathbf{W}_L^T \mathbf{W}_L$)

Eopch 0: Random Matrix



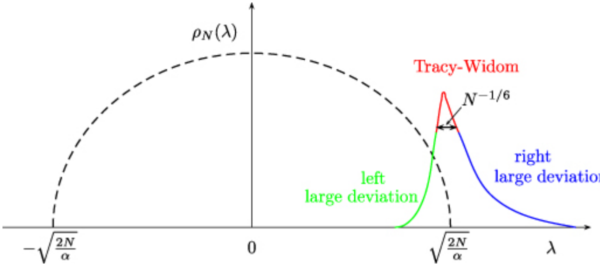
Eopch 36: Random + Spiles

Entropy decrease corresponds to:

- modification (later, breakdown) of random structure and
- onset of a new kind of self-regularization.

Random Matrix Theory 101: Wigner and Tracy-Widom

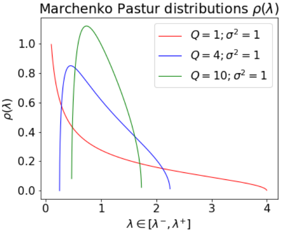
- Wigner: *global bulk statistics* approach universal semi-circular form
- Tracy-Widom: *local edge statistics* fluctuate in universal way



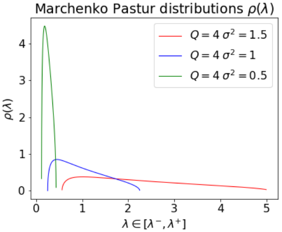
Problems with Wigner and Tracy-Widom:

- Weight matrices usually not square
- Typically do only a single training run

Random Matrix Theory 102': Marchenko-Pastur



(a) Vary aspect ratios



(b) Vary variance parameters

Figure: Marchenko-Pastur (MP) distributions.

Important points:

- *Global bulk stats*: The overall shape is deterministic, fixed by Q and σ.
- *Local edge stats*: The edge λ^+ is very crisp, i.e., $\Delta\lambda_M = |\lambda_{max} - \lambda^+| \sim O(M^{-2/3})$, plus Tracy-Widom fluctuations.

We use both *global bulk statistics* as well as *local edge statistics* in our theory.

Random Matrix Theory 103: Heavy-tailed RMT

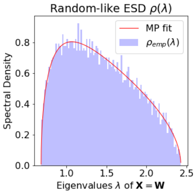
Go beyond the (relatively easy) Gaussian Universality class:

- *model* strongly-correlated systems ("signal") with heavy-tailed random matrices.

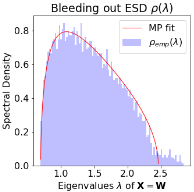
	Generative Model w/ elements from Universality class	Finite-N Global shape $\rho_N(\lambda)$	Limiting Global shape $\rho(\lambda),\ N \to \infty$	Bulk edge Local stats $\lambda \approx \lambda^+$	(far) Tail Local stats $\lambda \approx \lambda_{max}$
Basic MP	Gaussian	MP distribution	MP	TW	No tail.
Spiked-Covariance	Gaussian, + low-rank perturbations	MP + Gaussian spikes	MP	TW	Gaussian
Heavy tail, $4 < \mu$	(Weakly) Heavy-Tailed	MP + PL tail	MP	Heavy-Tailed*	Heavy-Tailed*
Heavy tail, $2 < \mu < 4$	(Moderately) Heavy-Tailed (or "fat tailed")	PL** $\sim \lambda^{-(a\mu+b)}$	PL $\sim \lambda^{-(\frac{1}{2}\mu+1)}$	No edge.	Frechet
Heavy tail, $0 < \mu < 2$	(Very) Heavy-Tailed	PL** $\sim \lambda^{-(\frac{1}{2}\mu+1)}$	PL $\sim \lambda^{-(\frac{1}{2}\mu+1)}$	No edge.	Frechet

Basic MP theory, and the spiked and Heavy-Tailed extensions we use, including known, empirically-observed, and conjectured relations between them. Boxes marked "*" are best described as following "TW with large finite size corrections" that are likely Heavy-Tailed, leading to bulk edge statistics and far tail statistics that are indistinguishable. Boxes marked "**" are phenomenological fits, describing large ($2 < \mu < 4$) or small ($0 < \mu < 2$) finite-size corrections on $N \to \infty$ behavior.

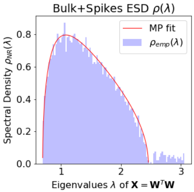
Phenomenological Theory: 5+1 Phases of Training



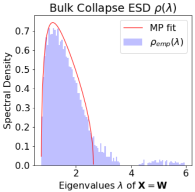
(a) Random-like.



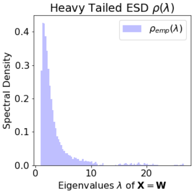
(b) Bleeding-out.



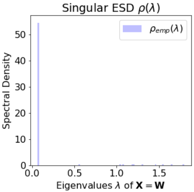
(c) Bulk+Spikes.



(d) Bulk-decay.



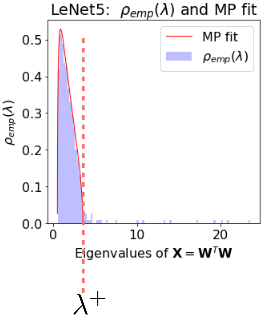
(e) Heavy-Tailed.



(f) Rank-collapse.

Figure: The 5+1 phases of learning we identified in DNN training.

Old/Small Models: Bulk+Spike $\sim$ Tikhonov regularization



simple scale threshold

$$\mathbf{x} = \left(\hat{\mathbf{X}} + \alpha \mathbf{I}\right)^{-1} \hat{\mathbf{W}}^T \mathbf{y}$$

eigenvalues $> \alpha$ (Spikes) carry most of the signal/information

Smaller, older models like LeNet5 exhibit traditional regularization

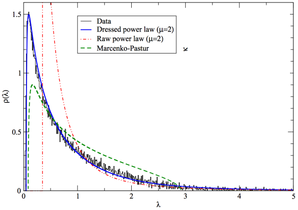
New/Large Models: Heavy-tailed Self-regularization

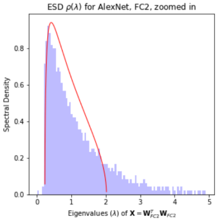
W is *strongly-correlated* and highly non-random

- Can *model* strongly-correlated systems by heavy-tailed random matrices

Then RMT/MP ESD will also have heavy tails

Known results from RMT / polymer theory (Bouchaud, Potters, etc.)





AlexNet
ReseNet50
Inception V3
DenseNet201
...

Larger, modern DNNs exhibit novel Heavy-tailed self-regularization

Uses, implications, and extensions

- Exhibit all phases of training by varying just the batch size ("explaining" the generalization gap)
- A Very Simple Deep Learning (VSDL) model (with load-like parameters α, & temperature-like parameters τ) that exhibits a non-trivial phase diagram
- Connections with minimizing frustration, energy landscape theory, and the spin glass of minimal frustration
- A "rugged convexity" since local minima do *not* concentrate near the ground state of heavy-tailed spin glasses
- A novel capacity control metric (the weighted sum of power law exponents) to predict trends in generalization performance for state-of-the-art models

Use our tool:

- "pip install weightwatcher"

Stop by the poster for more details ...