Traditional and Heavy-Tailed Self Regularization in Neural Network Models

Charles H. Martin (charles@CalculationConsulting.com) and Michael W. Mahoney (mmahoney@stat.berkeley.edu)

Motivation

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 study regularization

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

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

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

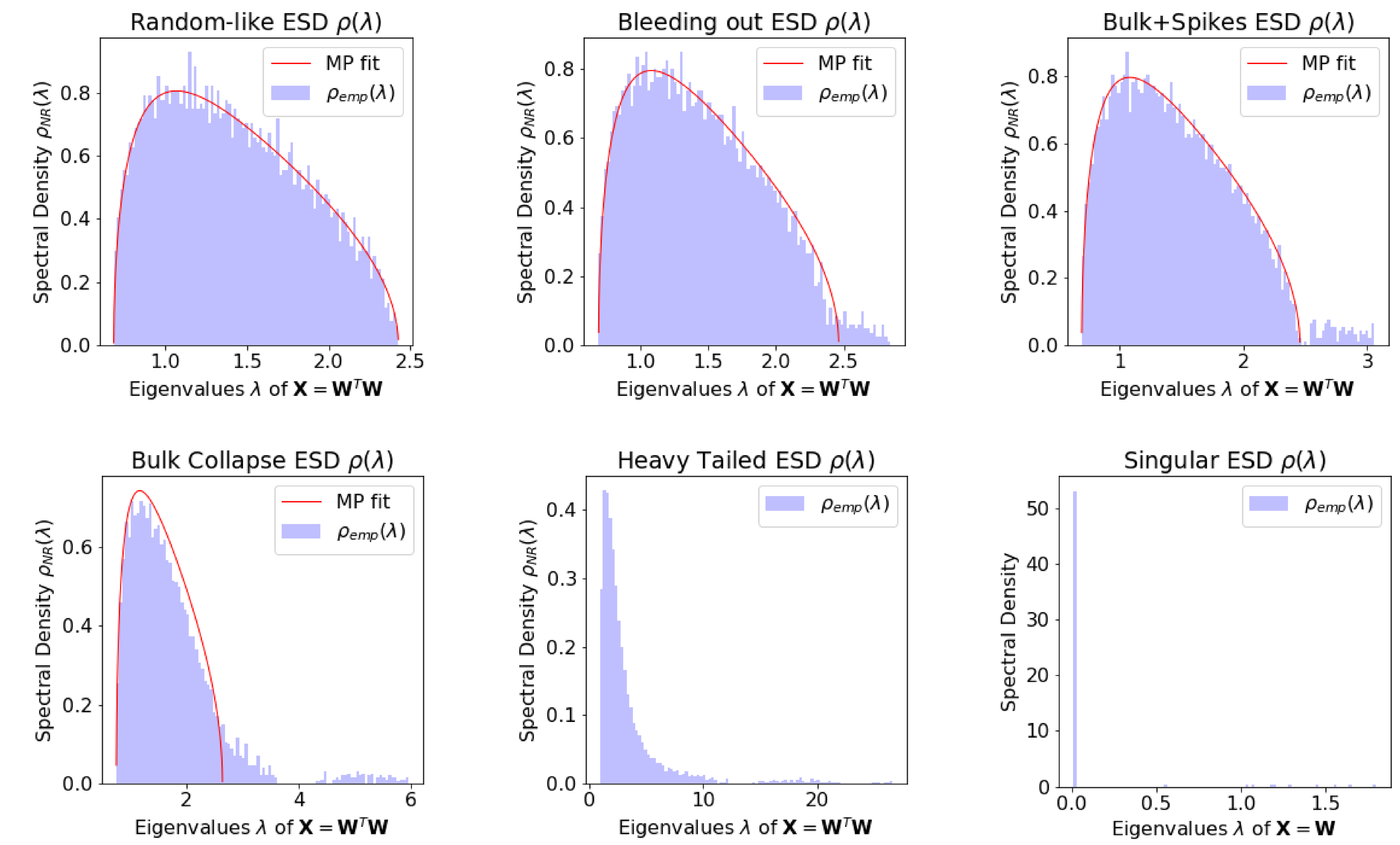
$$\min_{W_l, b_l} \mathcal{L}\left(\sum_i E_{DNN}(d_i) - y_i\right) + \alpha \sum_l \|\mathbf{W}_l\|$$

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 $\mathbf{W}_L$.

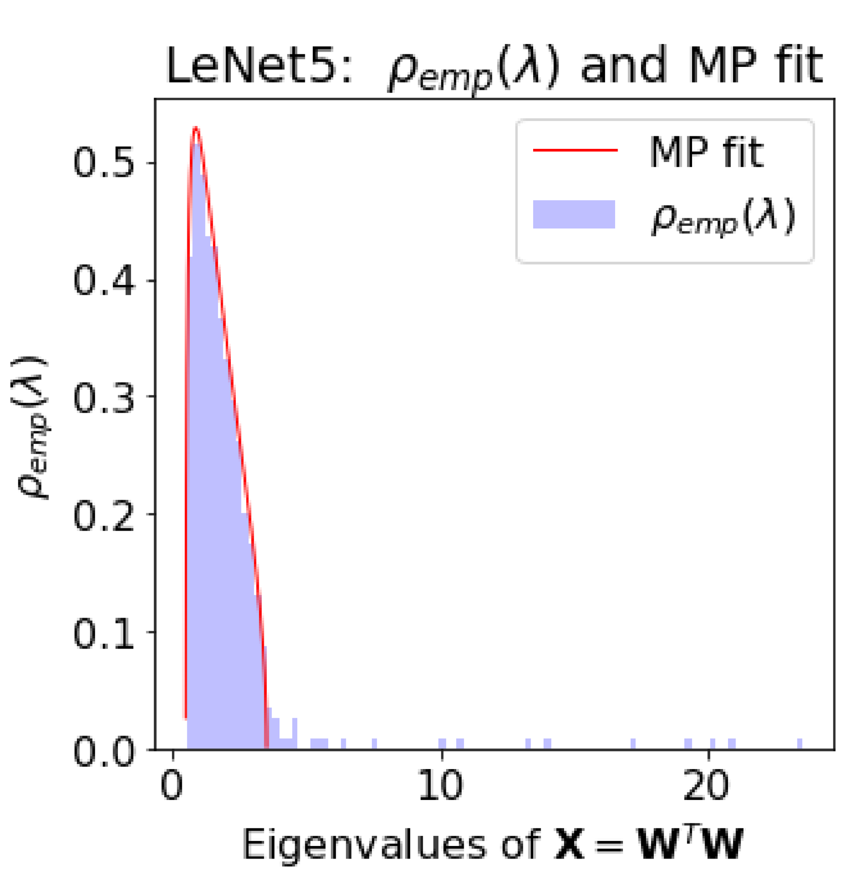
Phenomenological Theory: 5+1 Phases of Training



.

Old/Small Models ...

... exhibit "Bulk+Spike" ~ Tikhonov regularization



Simple scale threshold

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

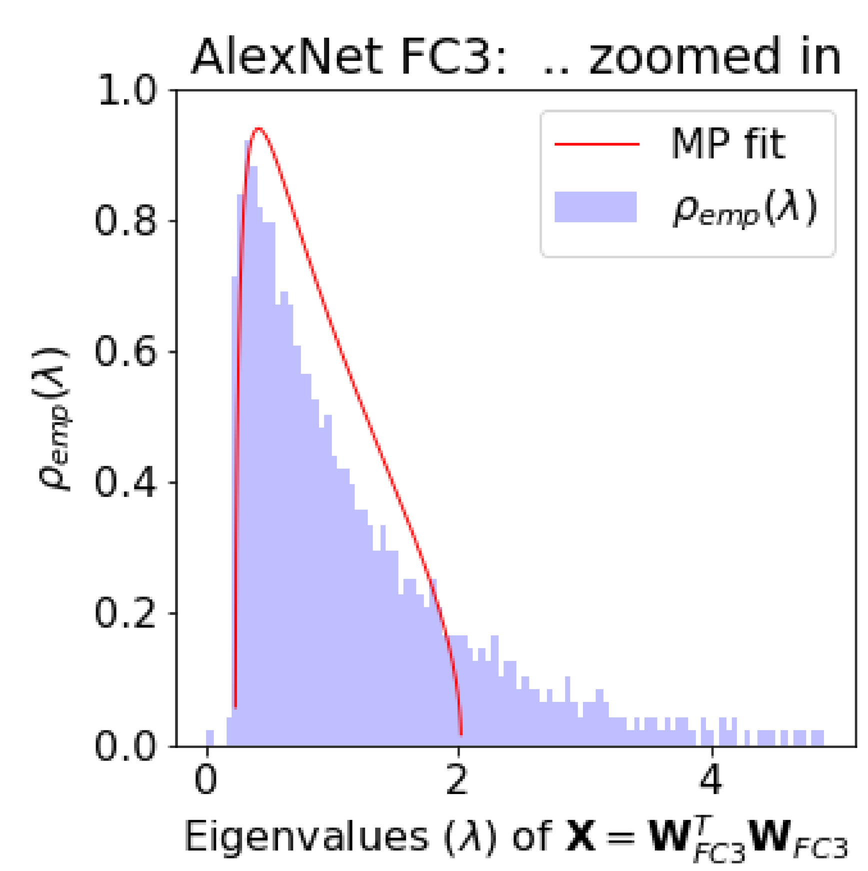
Eigenvalues $> \alpha$ (Spikes) carry most of the signal/information

Corresponds to usual "signal+noise" model

Smaller, older models like LeNet5 exhibit traditional regularization

New/Large Models ...

... exhibit novel Heavy-Tailed Self-Regularization



$\mathbf{W}$ is *strongly-correlated* and highly non-random:

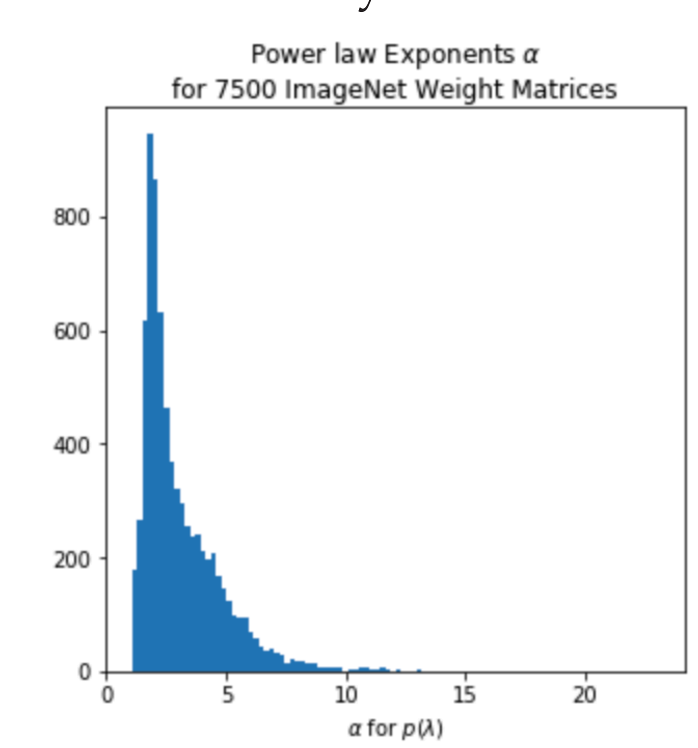
- Can *model* strongly-correlated systems by heavy-tailed random matrices

Use known results from Gaussian Random Matrix Theory, Heavy-Tailed Random Matrix Theory, and Polymer Theory

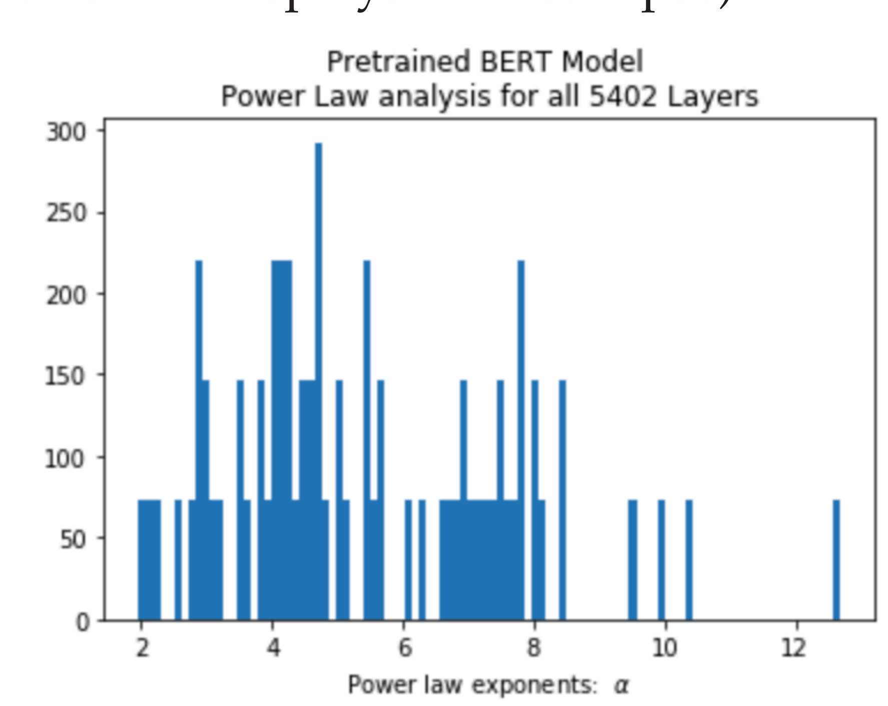
"All" larger, modern DNNs exhibit novel Heavy-tailed self-regularization

Remarkably Universal

All these ImageNet models display remarkable Heavy Tailed Universality:



The pretrained BERT model is *not* optimal (has large exponents and displays rank collapse)



Uses, implications, extensions

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