# Heavy-Tailed Self-Regularization in Deep Neural Networks

NeurIPS 2023 Workshop on Heavy Tails

charles@calculationconsulting.com

# Research: Implicit Self-Regularization in Deep Learning

*Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning.*

**(JMLR 2021)**

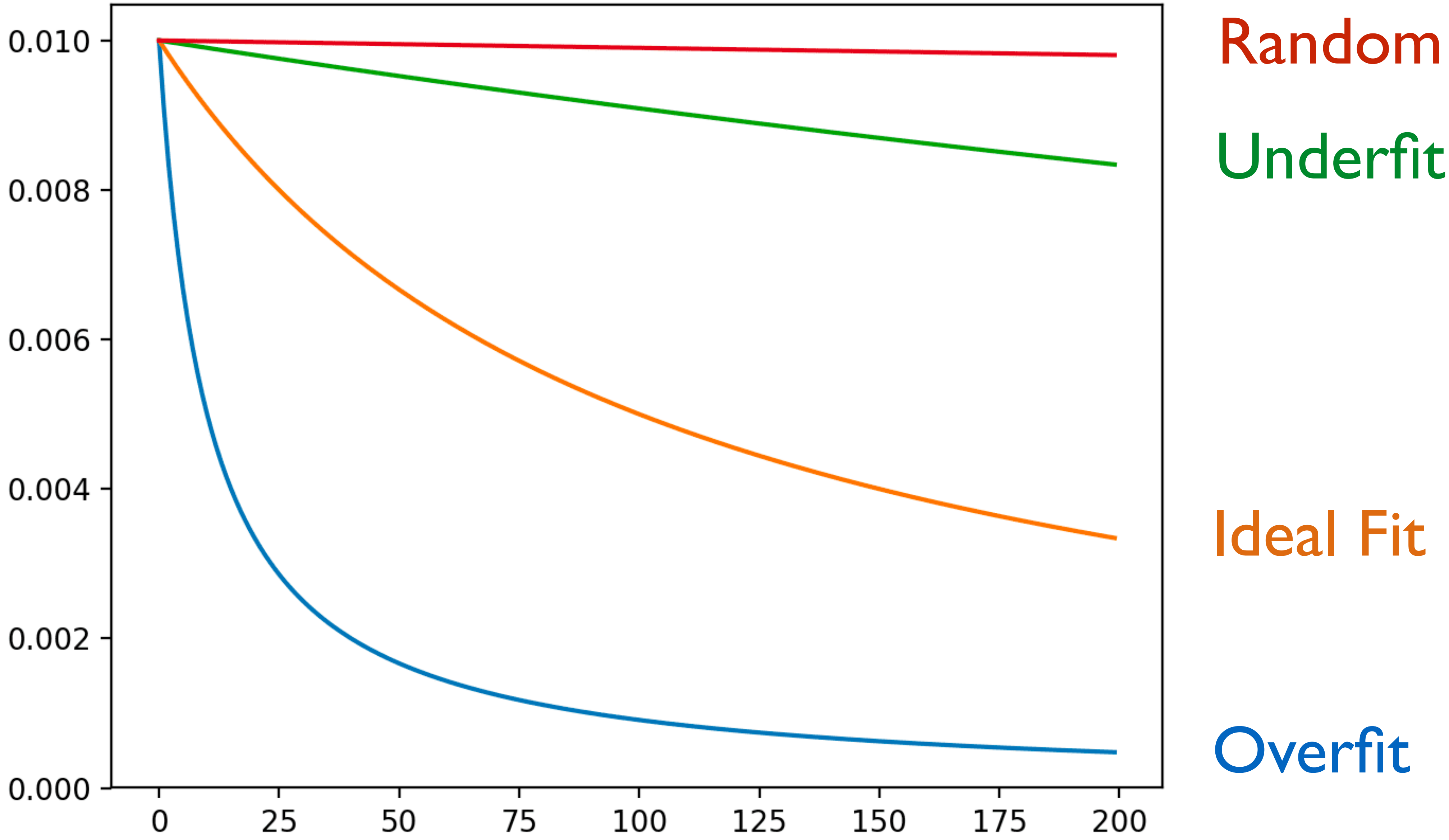
*Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data*

**(Nature Communications 2021)**

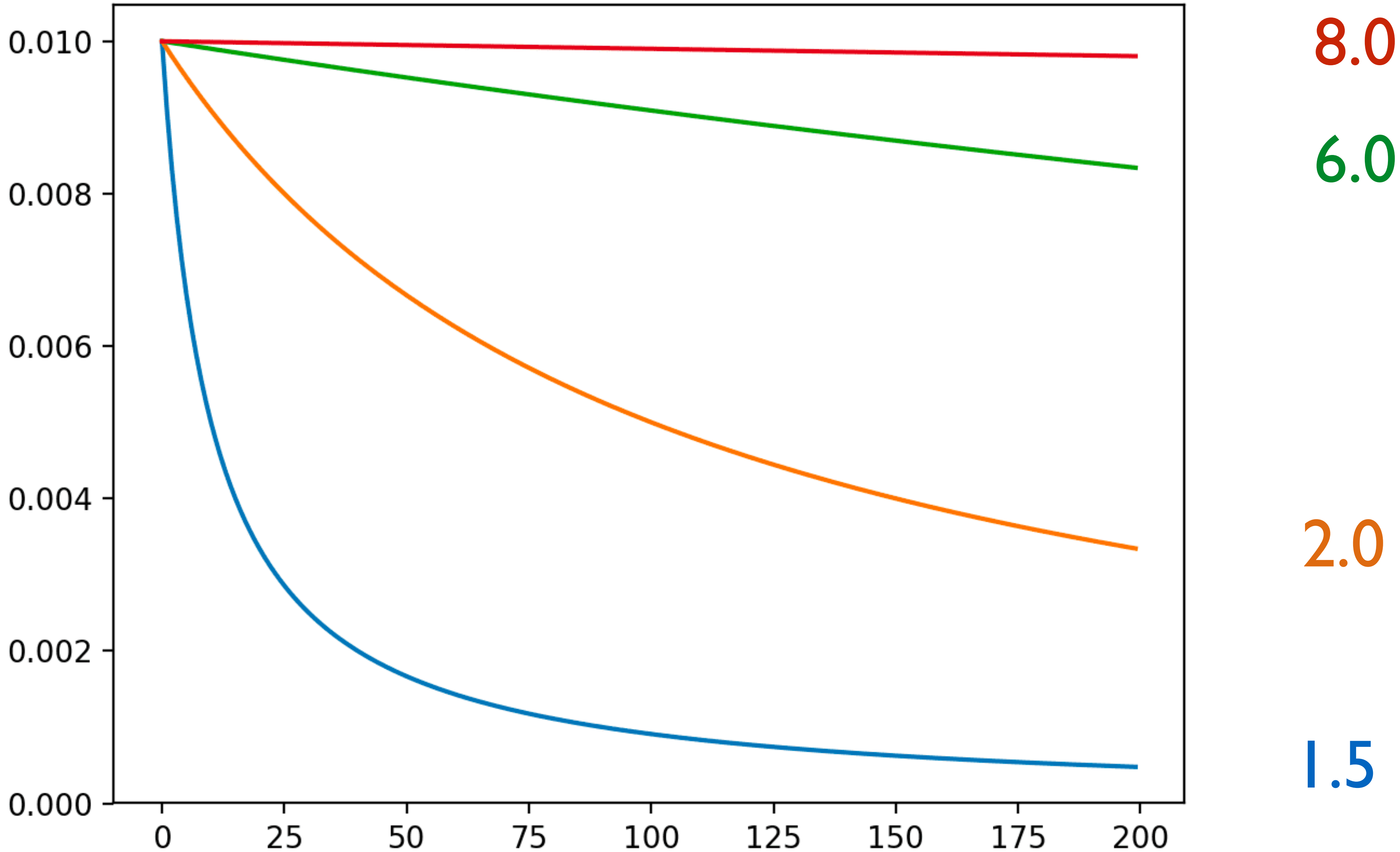
*SETOL: SemiEmpirical Theory of Learning*

**(Invited submission, Philosophical Magazine)**

# Problem: How do NN layers converge ?

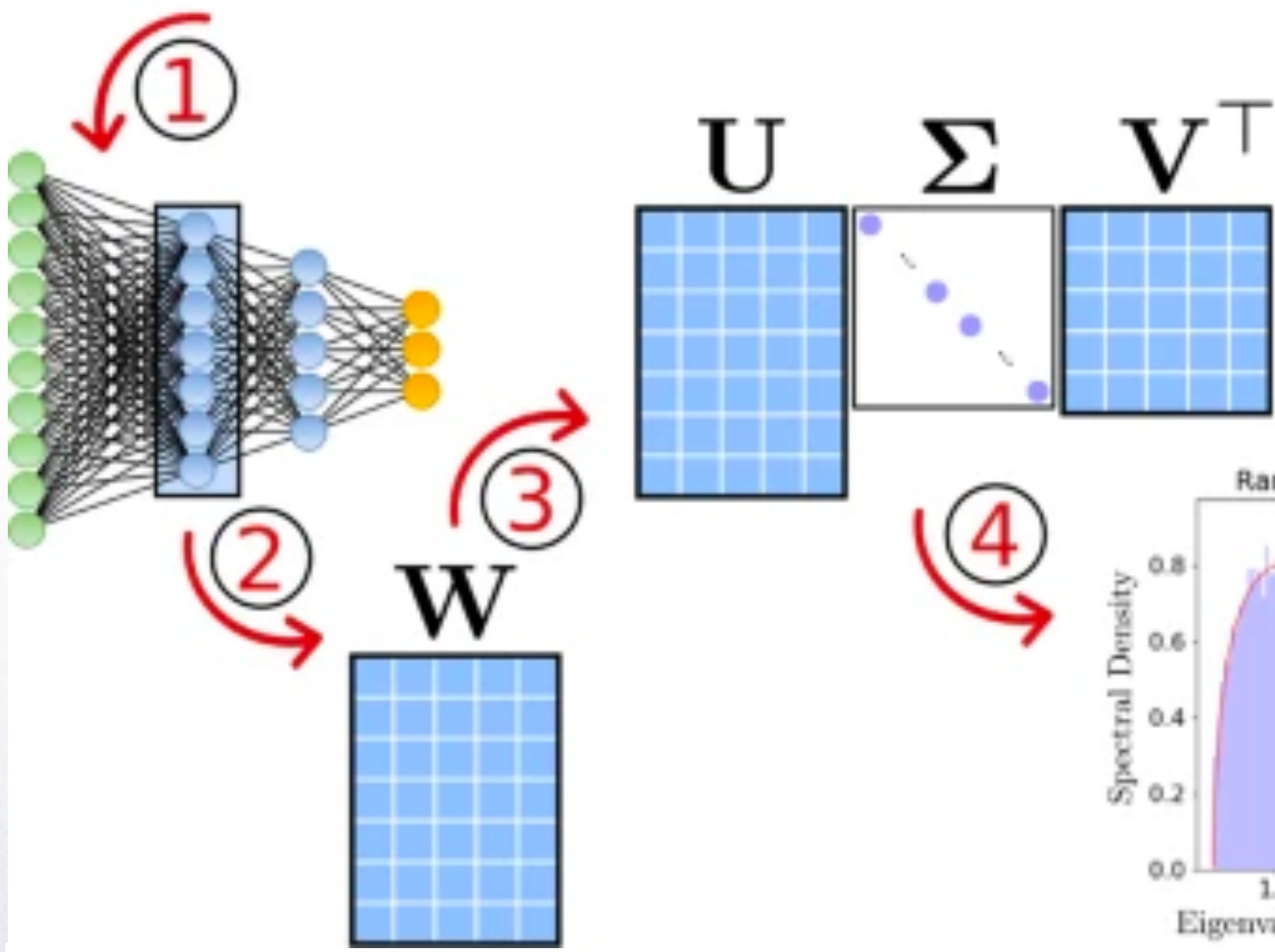


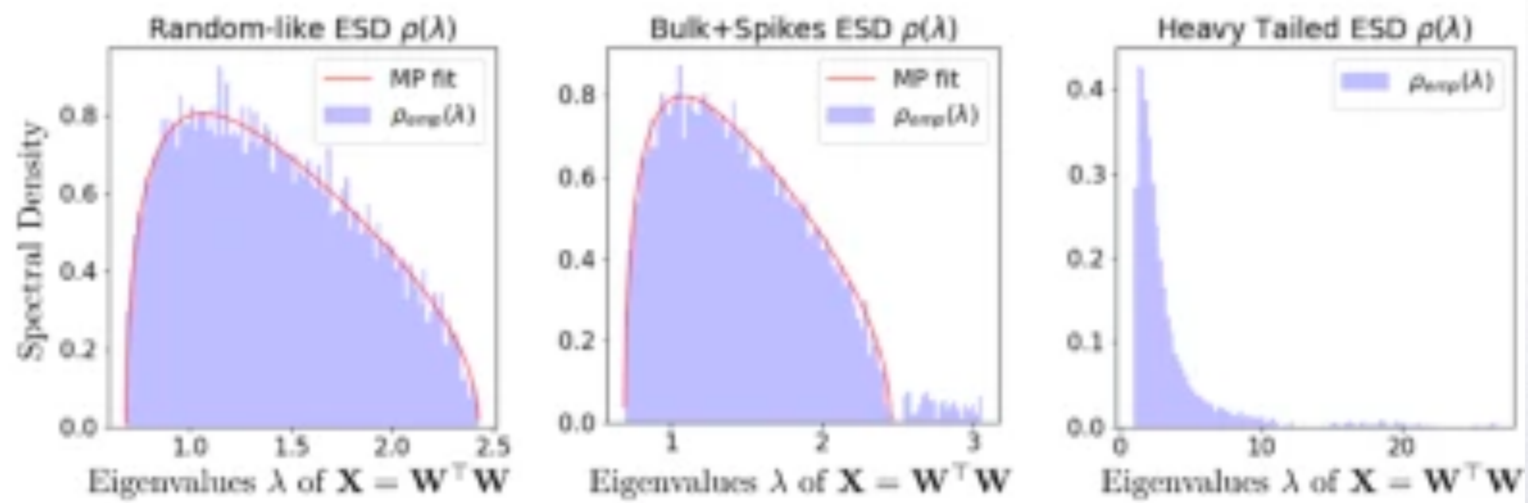
# Solution: weightwatcher layer quality metric



# Analyzing DNN Weight matrices with WeightWatcher

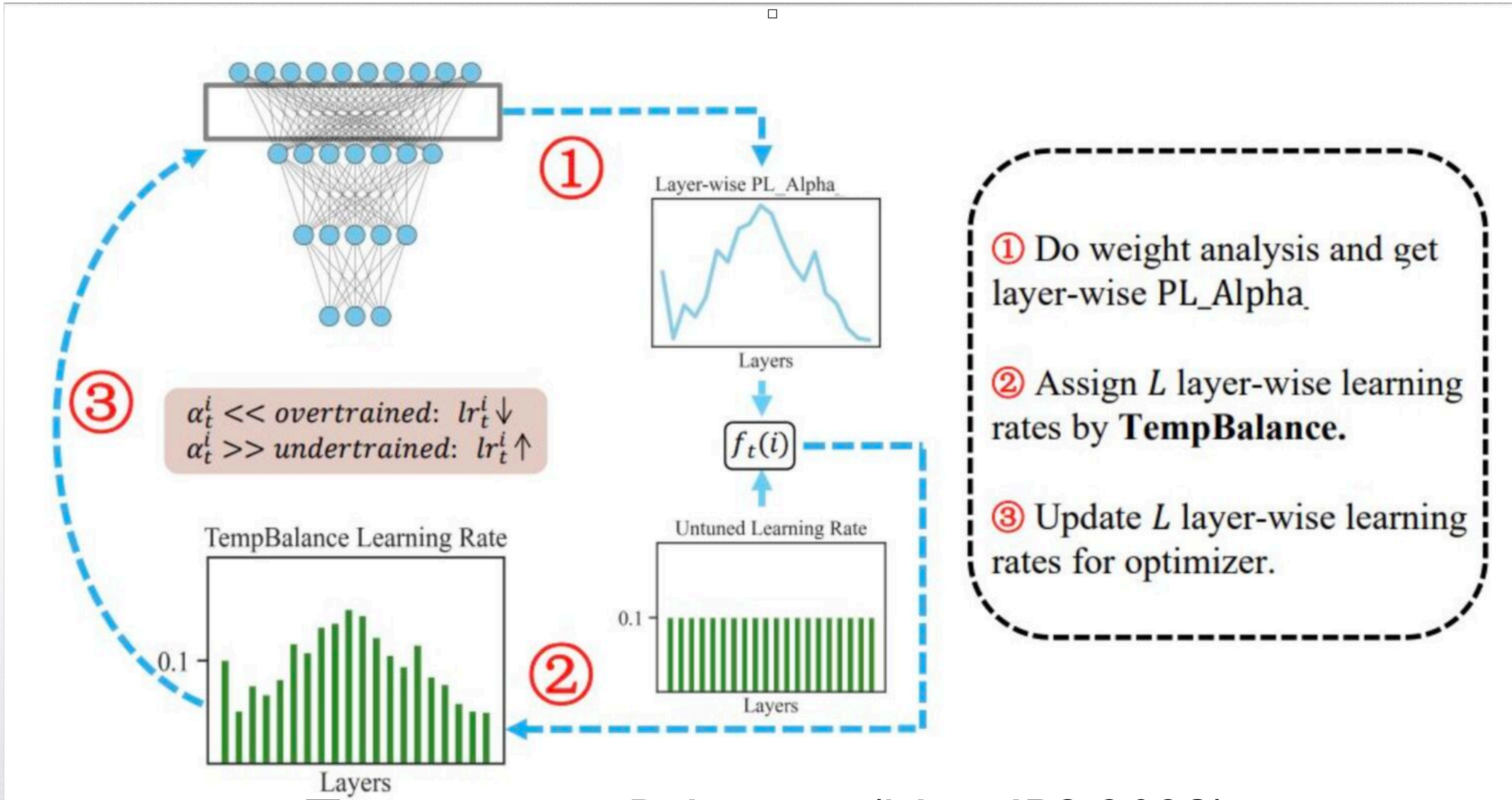
1. Take a model
2. Take a weight matrix
3. Do Spectral analysis
4. Histogram of eigenvalues



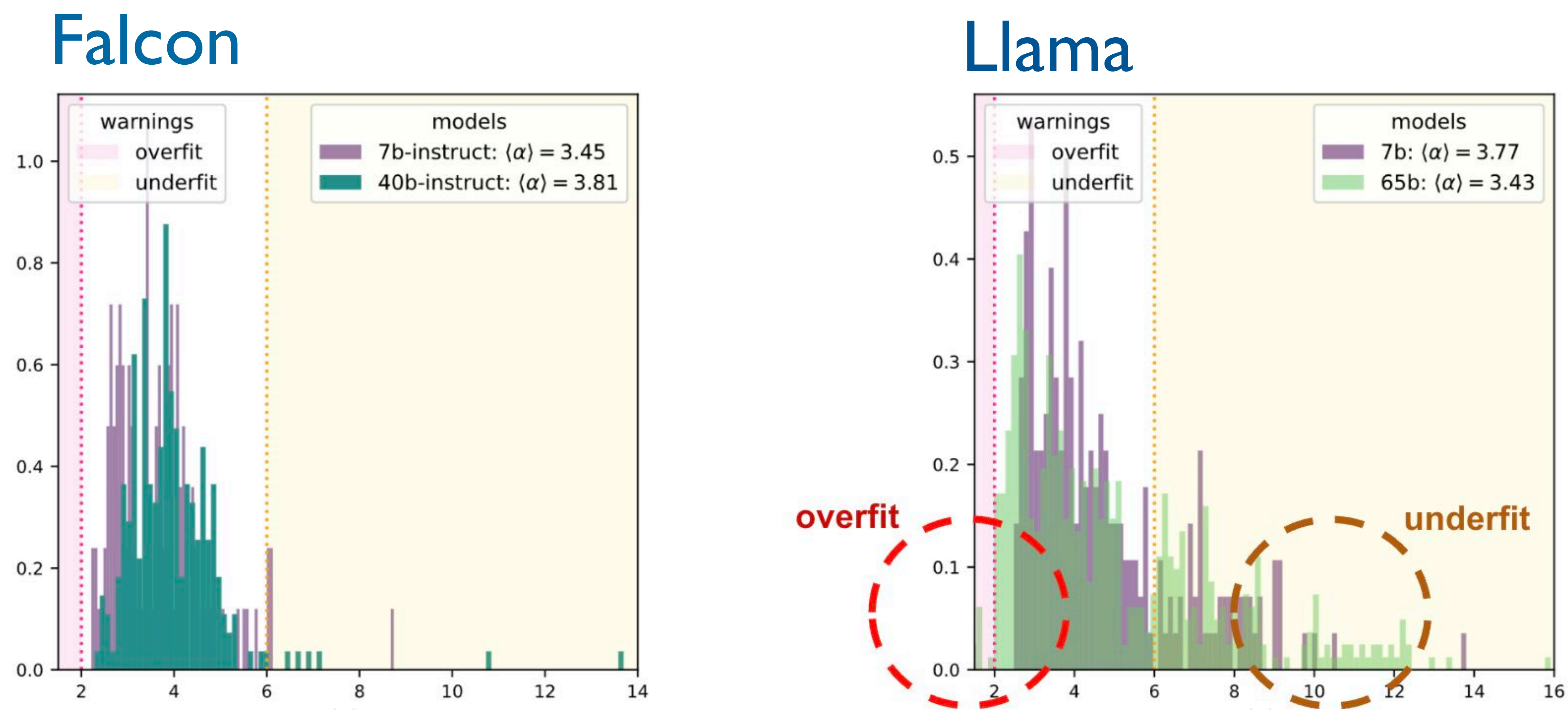

weightwatcher layer quality metric: $\rho_{emp}(\lambda) \sim \lambda^{-\alpha}$.

# Accelerate Training: Adjust Layer Learning Rates




Temperature Balancing (NeurIPS 2023)

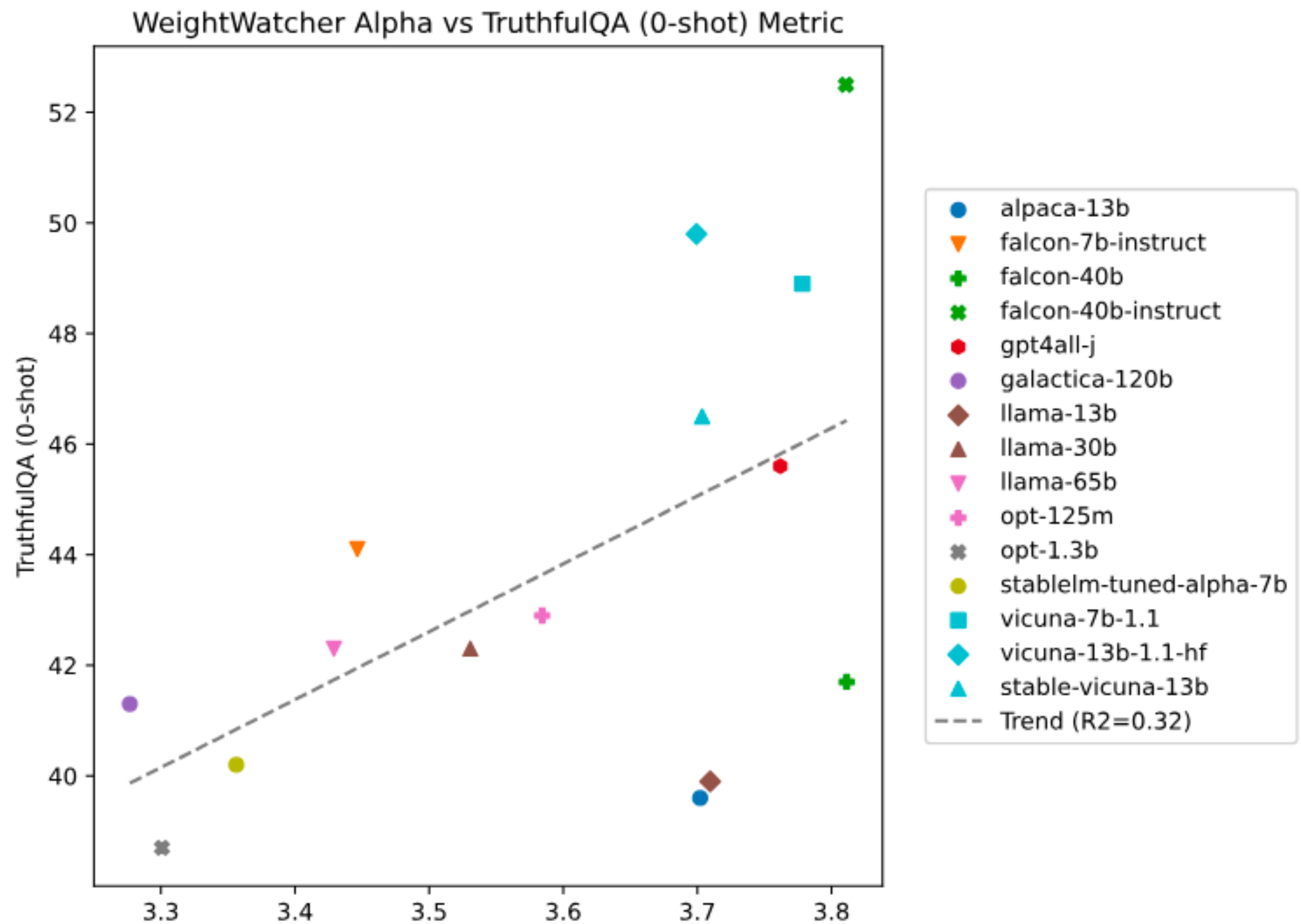
# Comparing LLMs: Detect 'Bad' Layers




weightwatcher layer quality metric

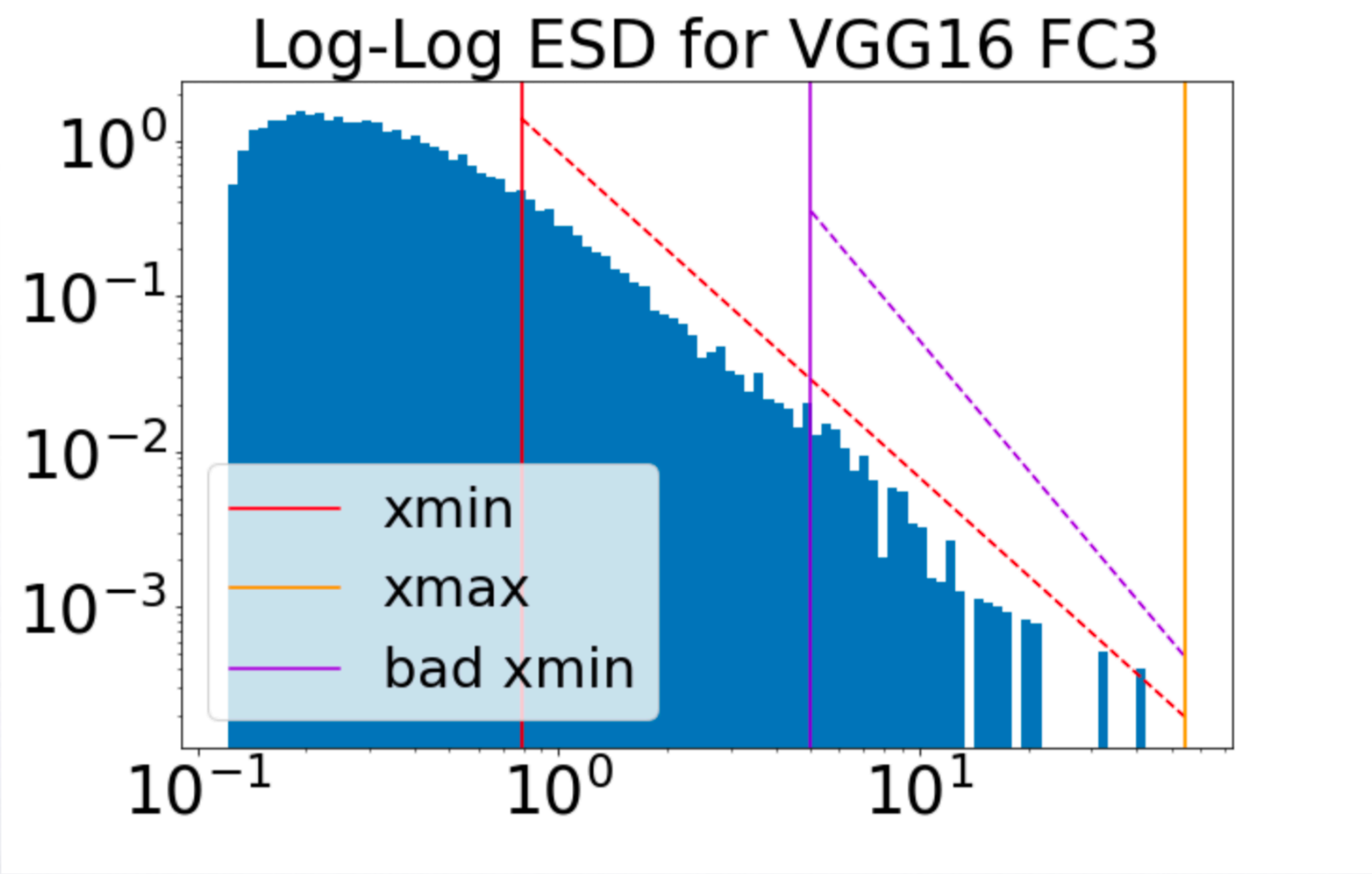
# LLM Hallucinations: Gauge Truthfulness of Base Models

TruthfulQA metric




weightwatcher quality metric

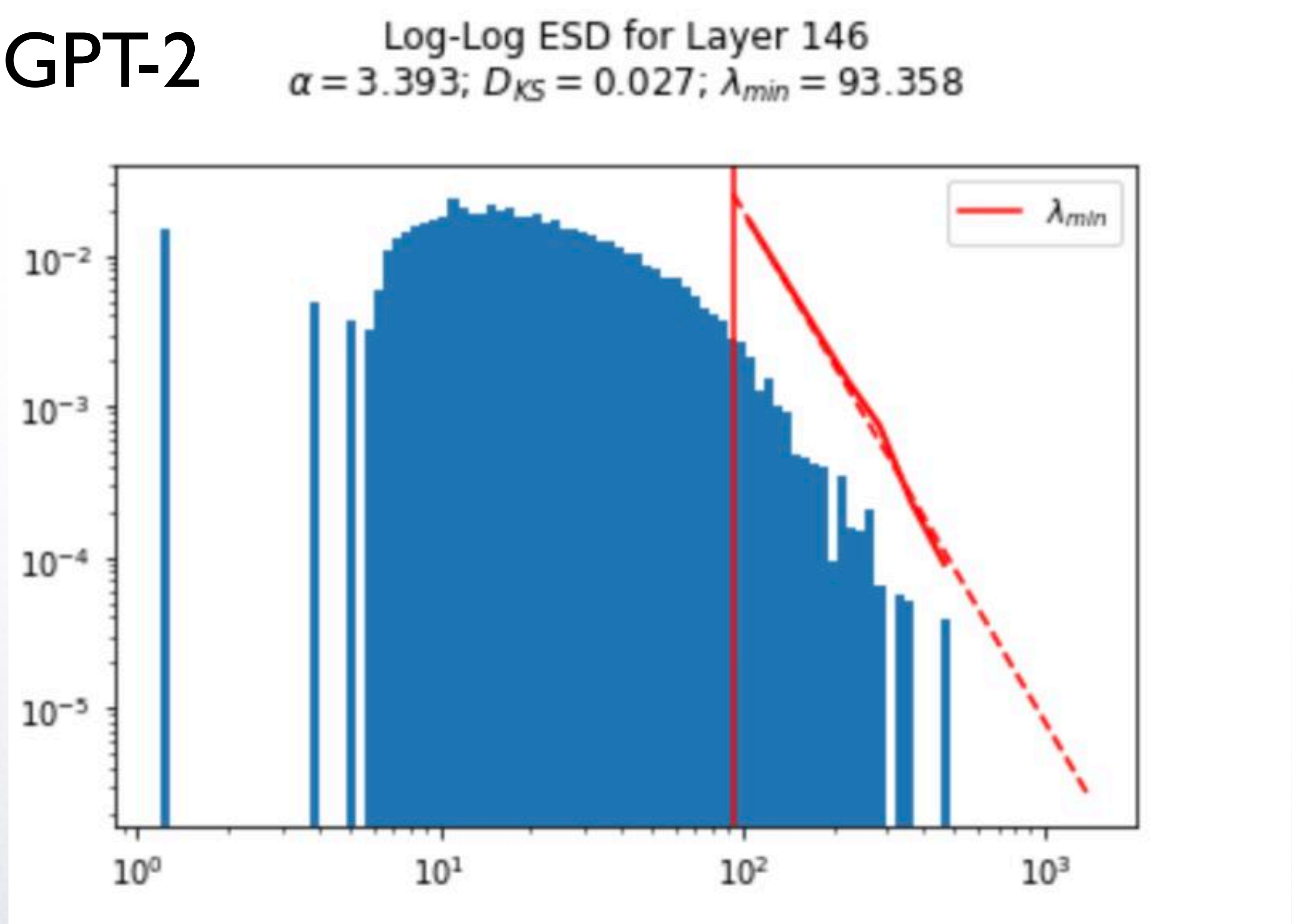
# WeightWatcher: analyzes the ESD (eigenvalues) of the layer weight matrices




The tail of the ESD contains the information

# WeightWatcher: analyzes the ESD (eigenvalues) of the layer weight matrices

GPT-2




Fits a Power Law
(or Truncated Power Law)

$$\rho_{emp}(\lambda) \sim \lambda^{-\alpha}.$$

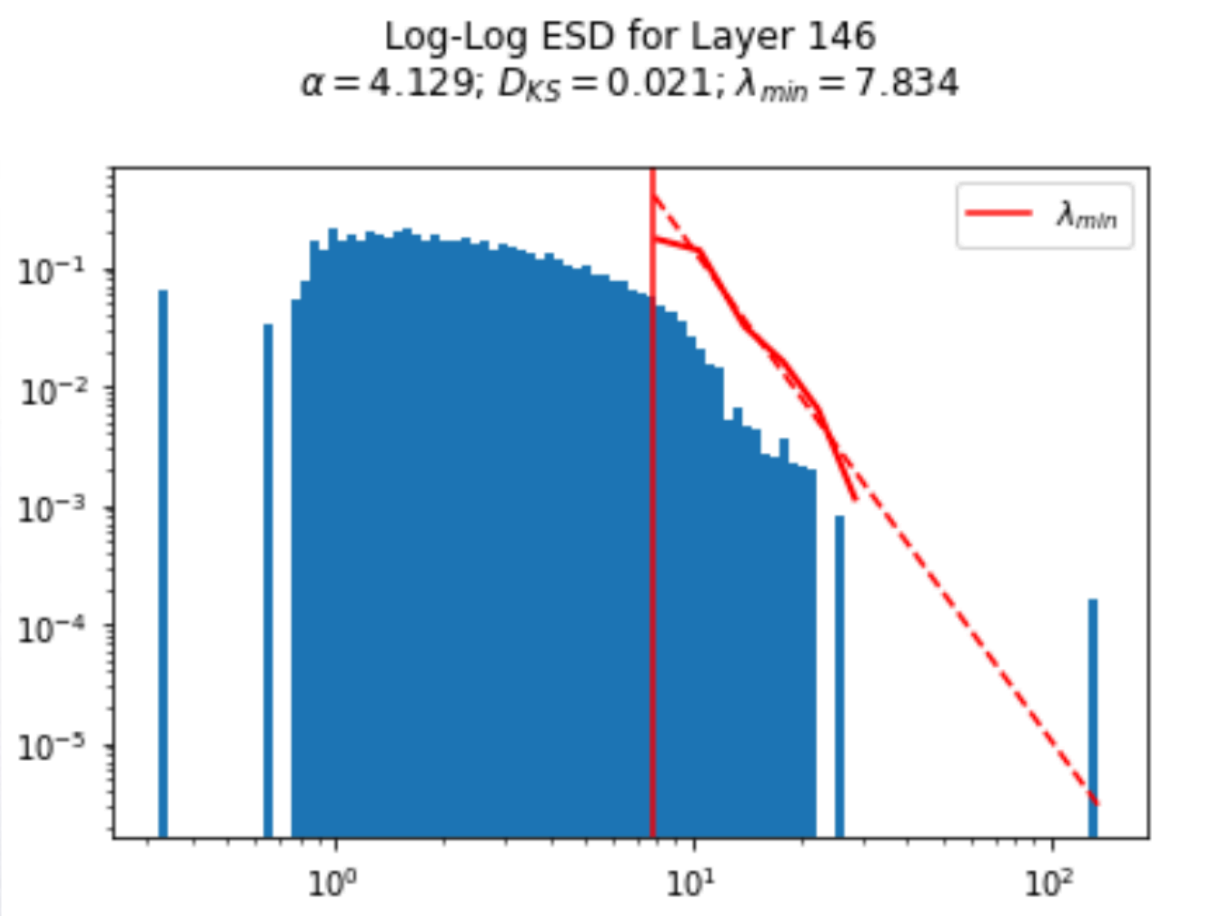
alpha in [2, 6]

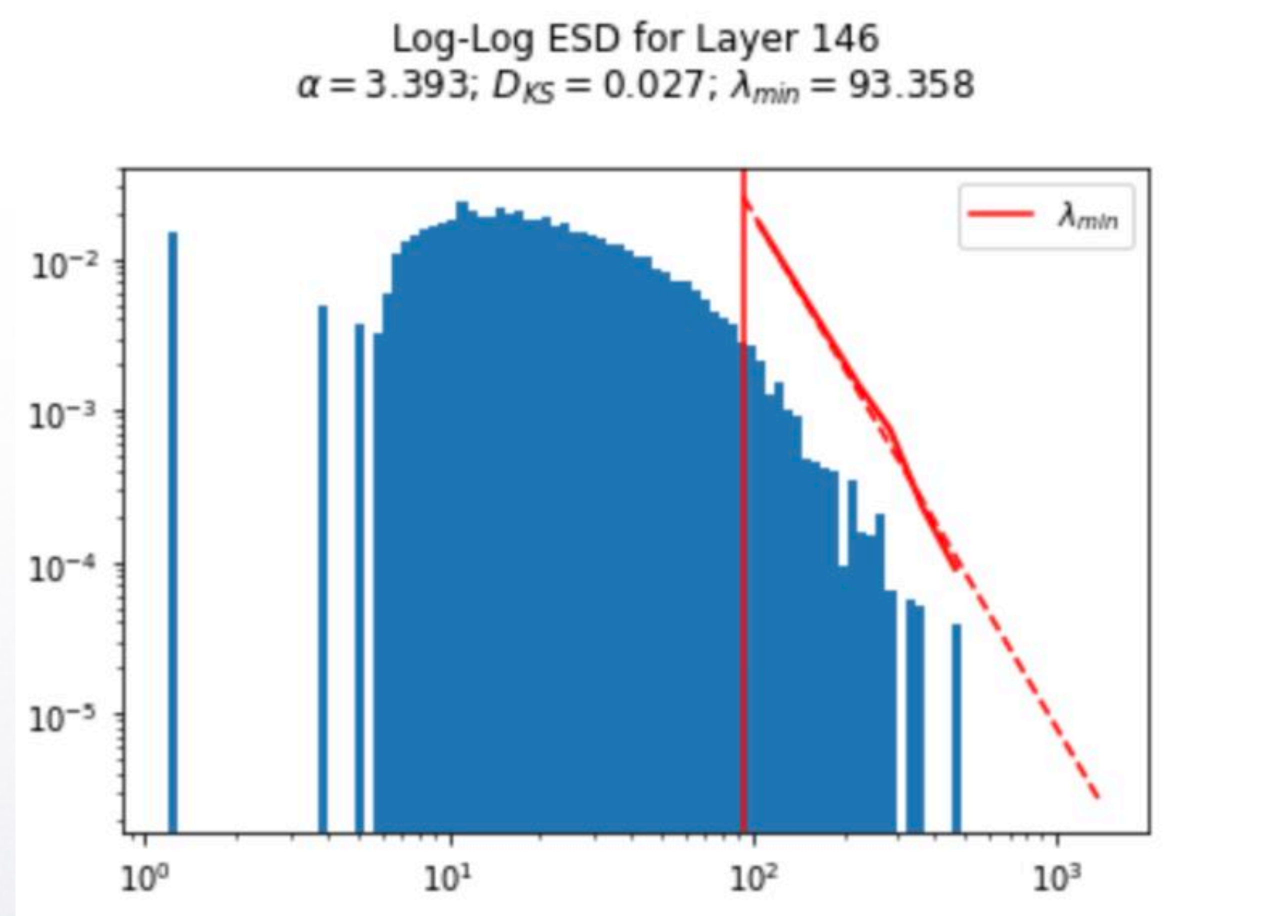
Good quality of fit (D is small)

```
watcher.analyze(plot=True)
```

Well trained layers are heavy-tailed and well shaped

# WeightWatcher: analyzes the ESD (eigenvalues) of the layer weight matrices

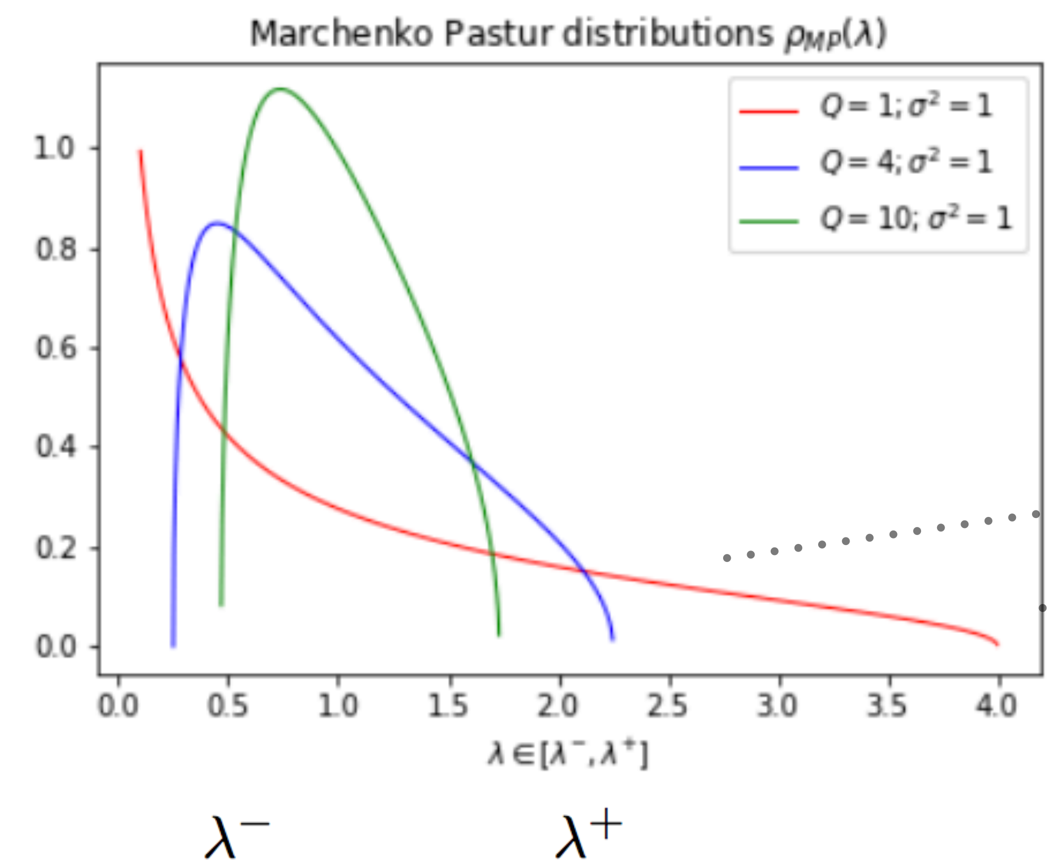



Better trained layers are more heavy-tailed and better shaped

# Random Matrix Theory: Marcenko Pastur

RMT says if W is a simple random Gaussian matrix,
then the ESD will have a very simple , known form




Shape depends on Q=N/M
(and variance ~ 1)

Eigenvalues tightly bounded

very crisp edges

$\Delta\lambda_M = |\lambda_{max} - \lambda^+| \sim \mathcal{O}(M^{-\frac{2}{3}})$.

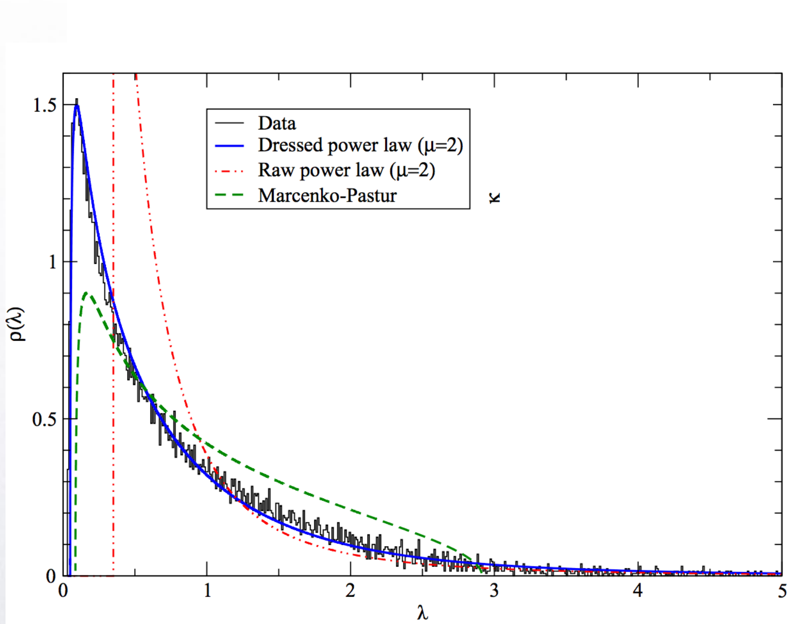
plus Tracy-Widom fluctuations

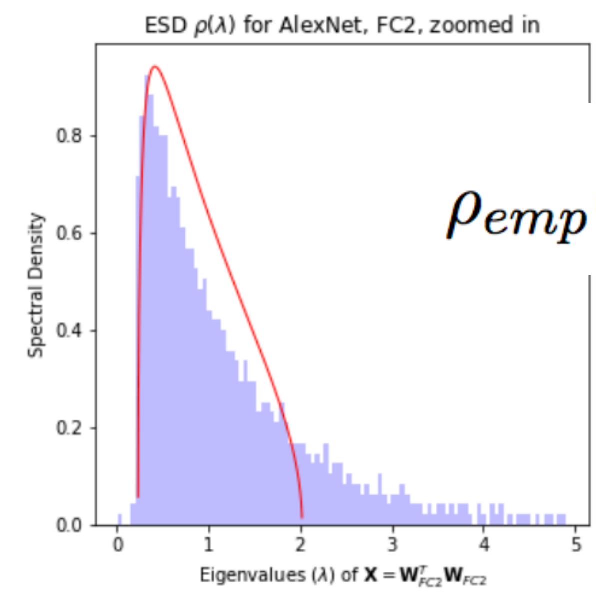
a few spikes may appear

# Random Matrix Theory: Heavy Tailed

But if W is heavy tailed, the ESD will also have heavy tails
(i.e. its all spikes, bulk vanishes)

If *W* is strongly correlated , then the ESD can be *modeled* as if W is drawn from a heavy tailed distribution

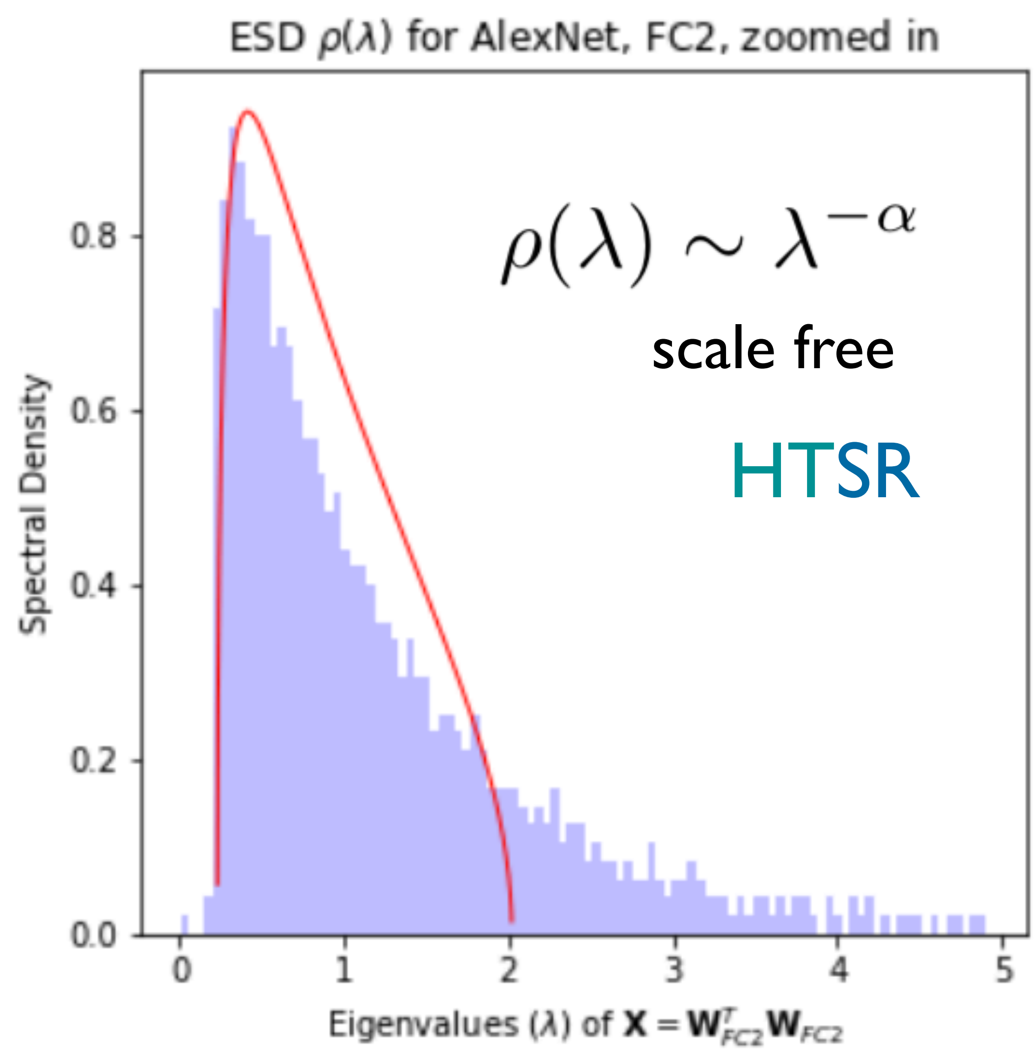



$$\rho_{emp}(\lambda) \sim \lambda^{-\alpha}.$$

Nearly all pre-trained DNNs display heavy tails…as shall soon see

# Heavy-Tailed: Self-Regularization

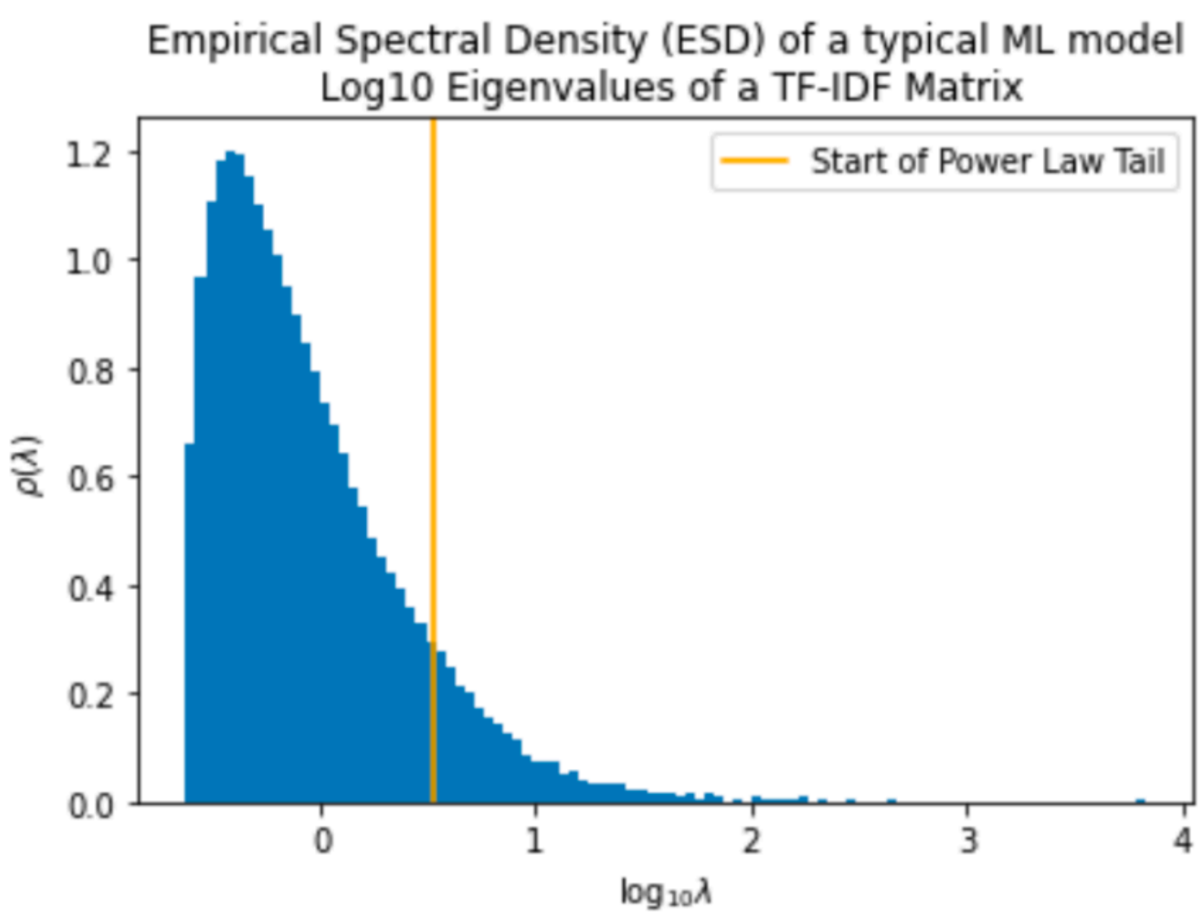


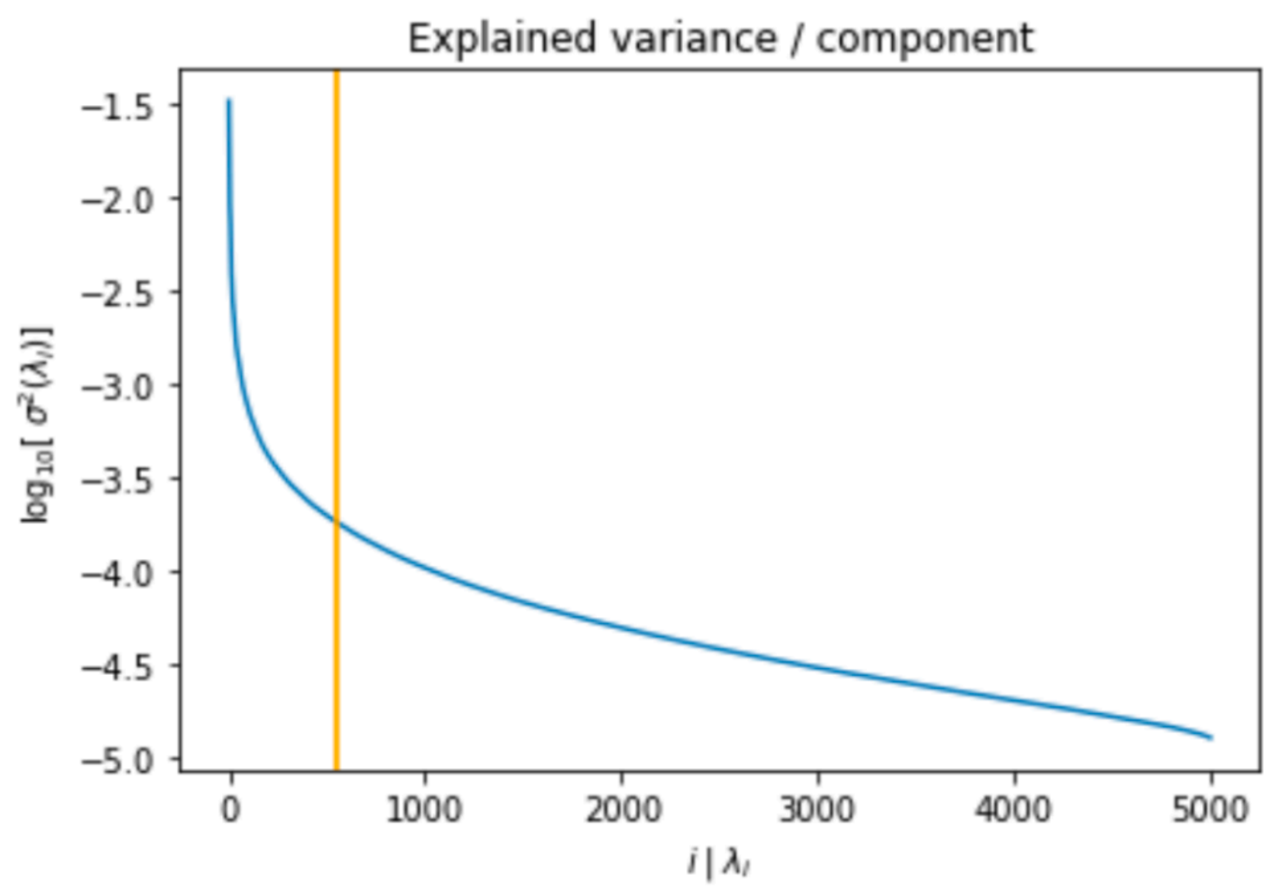

AlexNet,
VGG11, VGG13, …
ResNet, …
DenseNet,
BERT, RoBERT, …
GPT, GPT2, …
Flan-T5
Bloom
Llama
Falcon
…

All large, well trained, modern DNNs exhibit *heavy tailed self-regularization*

# Heavy-Tails: in Latent Semantic Analysis

LSA on 20newsgroups; great PL fit; alpha = 2.2



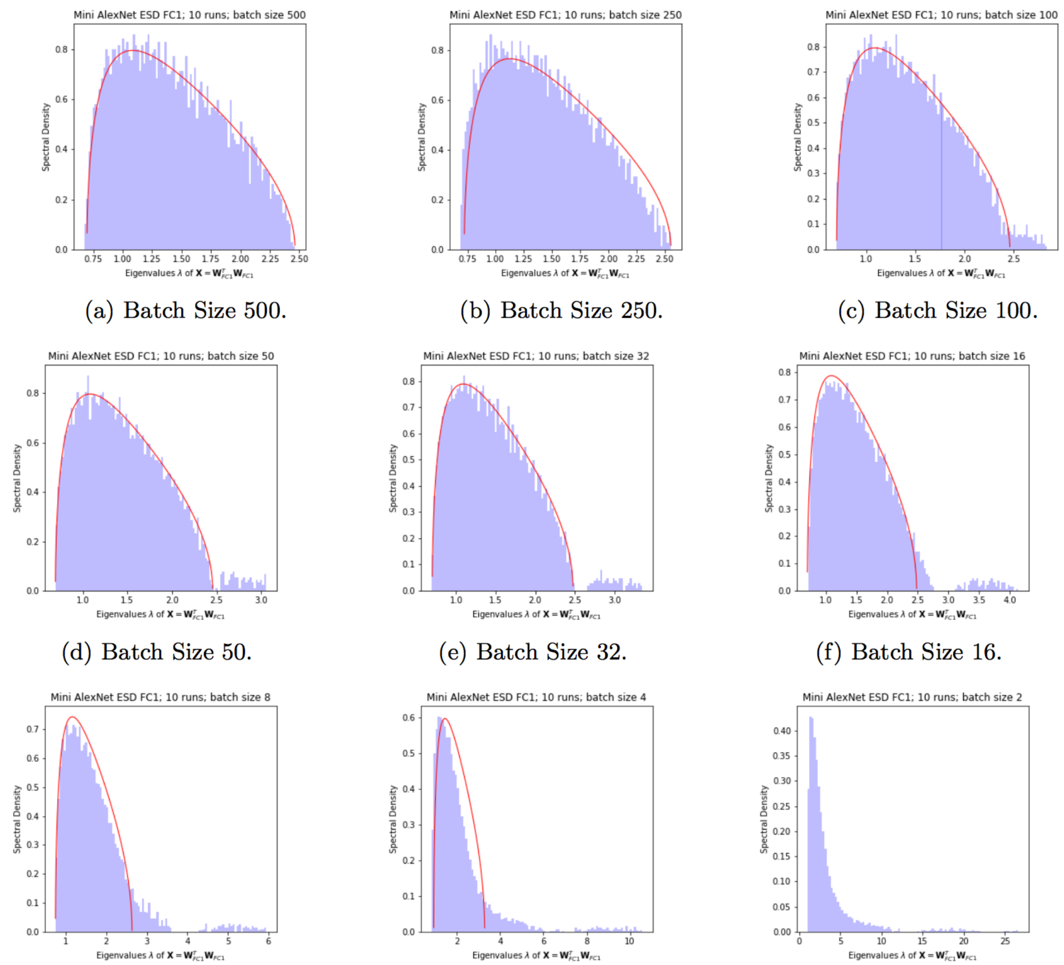

*Heavy Tails do not solely come from of SGD training*

# HTSR Theory: Heavy Tailed Self Regularization

Gaussian random matrix

Bulk+ Spikes

Heavy Tailed




(a) Batch Size 500. (b) Batch Size 250. (c) Batch Size 100.

(d) Batch Size 50. (e) Batch Size 32. (f) Batch Size 16.

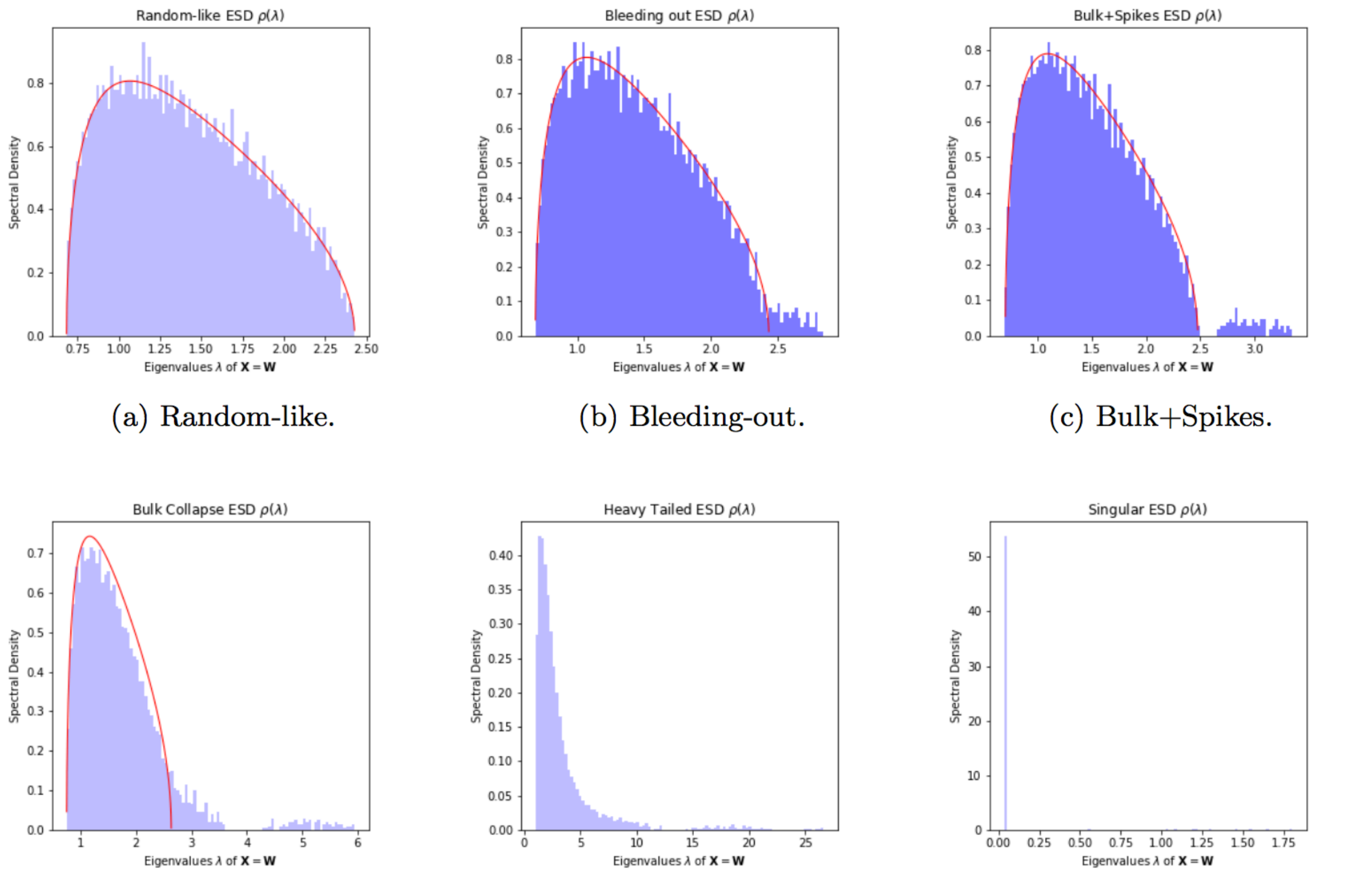
(g) Batch Size 8. (h) Batch Size 4. (i) Batch Size 2.

Small, older NNs

Large, modern DNNs
and/or
Small batch sizes

*DNN training induces breakdown of Gaussian random structure and the onset of a new kind of heavy tailed self-regularization*

# HTSR Theory: 5+1 Phases of Training




(a) Random-like. (b) Bleeding-out. (c) Bulk+Spikes.

(d) Bulk-decay. (e) Heavy-tailed. (f) Singularity.

**Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning**

*Charles H. Martin, Michael W. Mahoney*; JMLR 22(165):1–73, 2021.

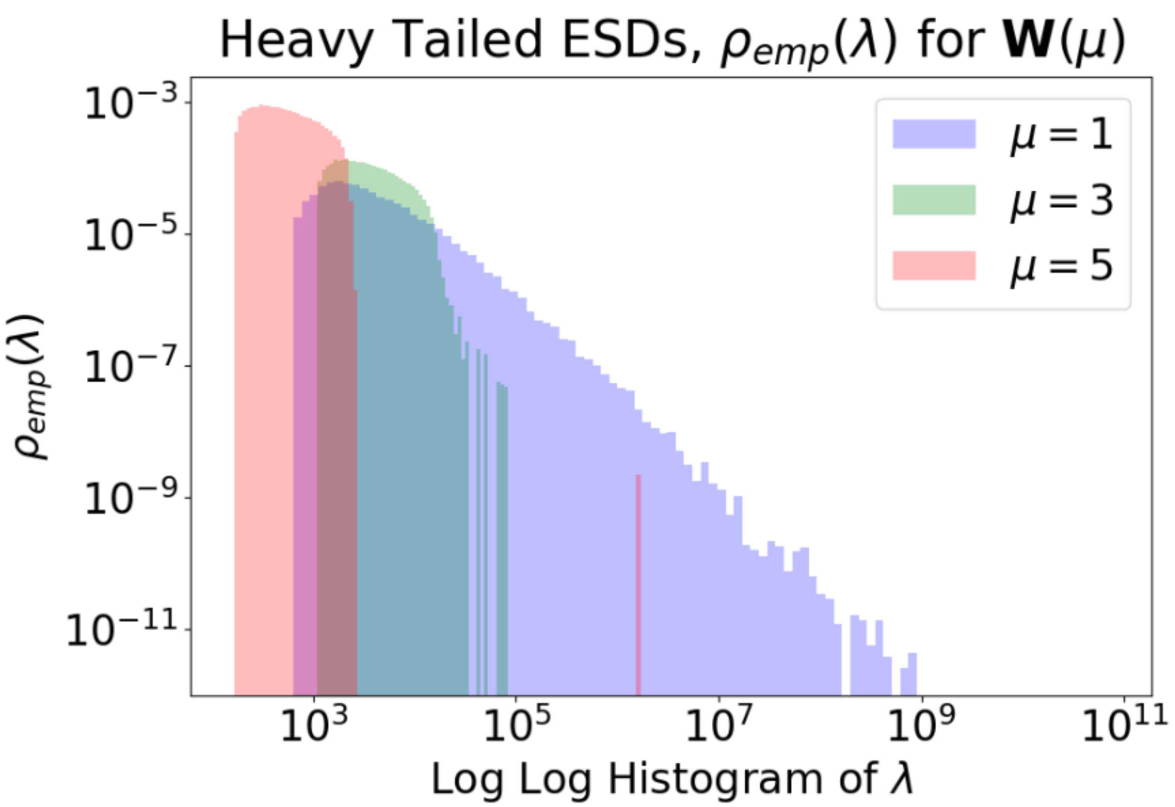
# Heavy Tailed RMT: Universality Classes

| | 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, i.e., Eqn. (8) | 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 |

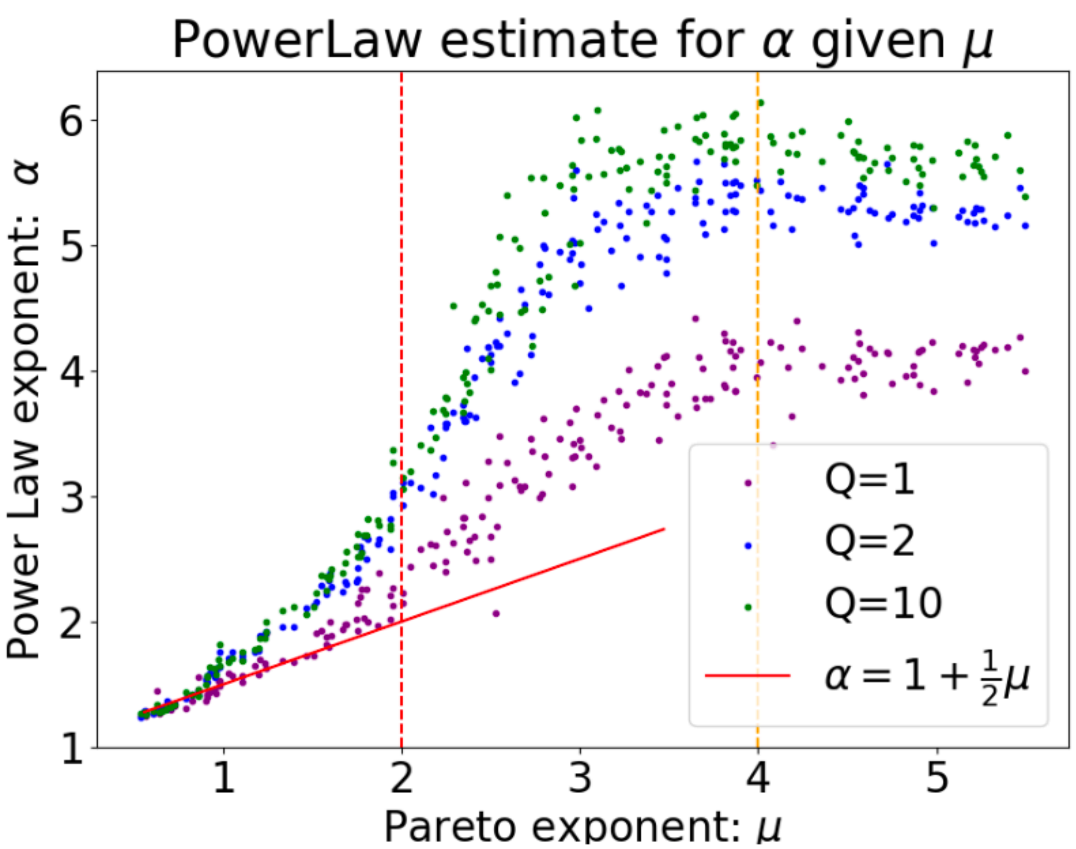
*Charles H. Martin, Michael W. Mahoney*; JMLR 22(165):1–73, 2021.

The familiar Wigner/MP Gaussian class is not the only Universality class in RMT

# Heavy Tailed RMT: Universality Classes




(a) Heavy Tailed ESDs




(b) PL $\alpha$ vs HT $\mu$ exponent

$$W_{ij} \sim P(x) \sim \frac{1}{x^{1+\mu}}, \quad \mu > 0.$$

# Statistical Mechanics
## Why this works…

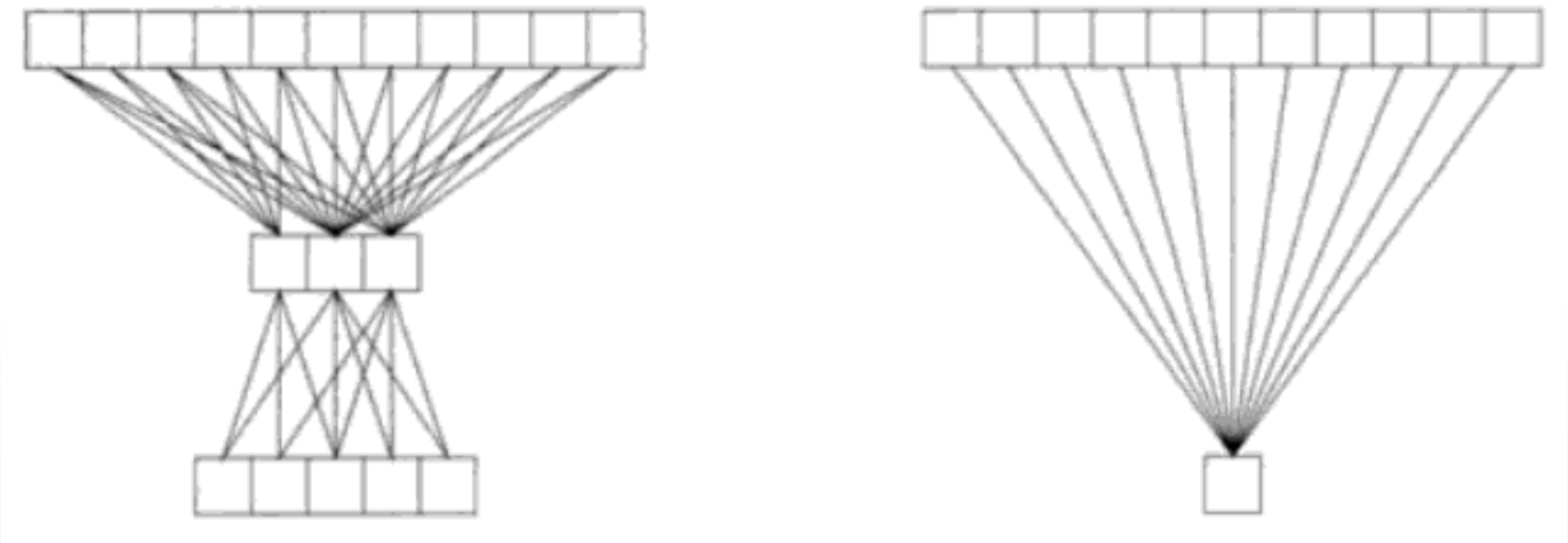
ORIGINAL ARTICLE

**SETOL: A Semi-Empirical Theory of (Deep) Learning**

Charles H Martin[a] and Christopher Hinrichs and Michael W Mahoney[b]

[a]Calculation Consulting, San Francisco, CA 94122; [b] UC Berkeley

# Classic Set Up: Student-Teacher model

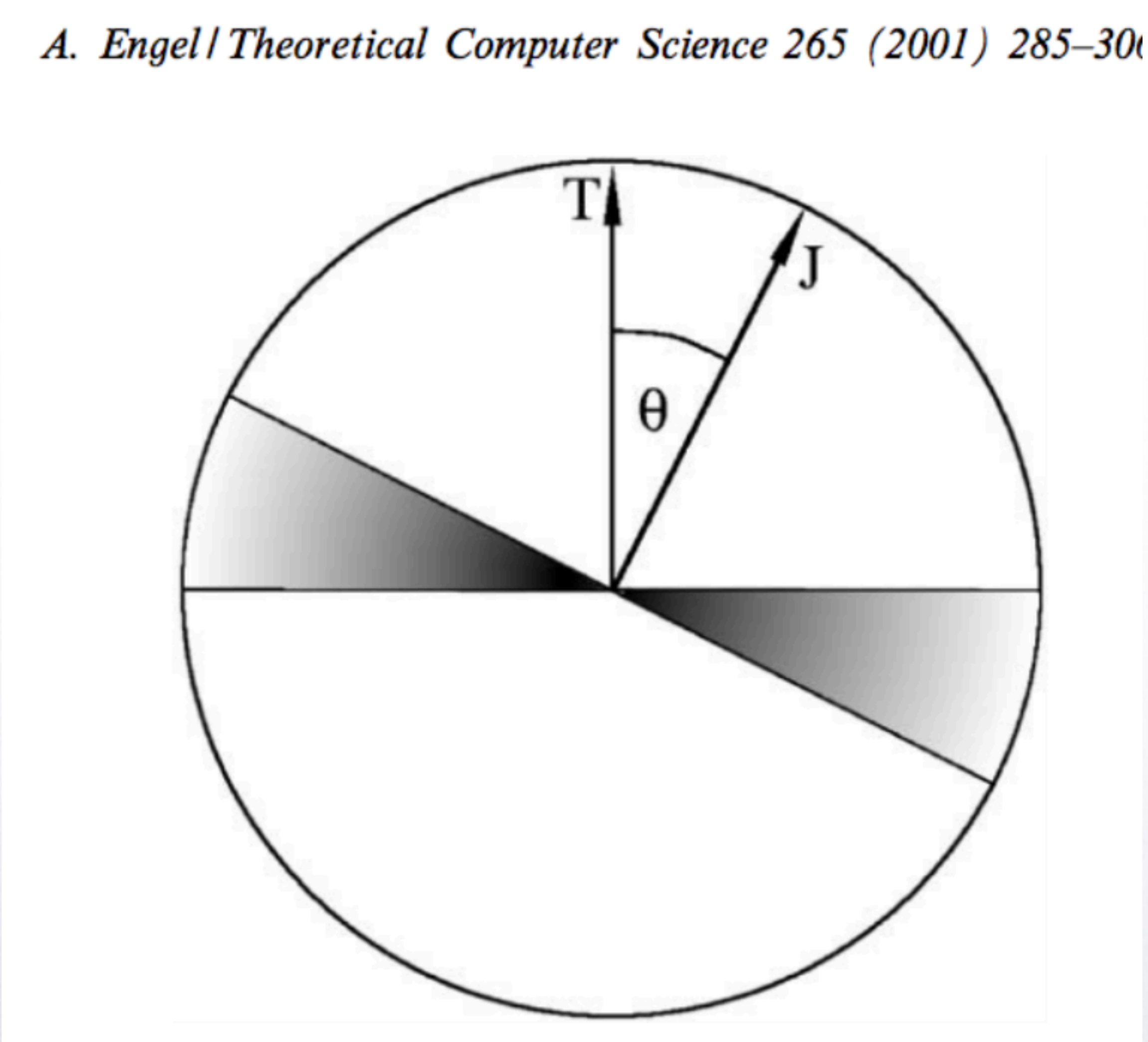


MultiLayer Feed Forward Network

Perceptron

Statistical Mechanics of Learning from the 1990s

# Classic Set Up: Student-Teacher model




Average overlap over random students J

$$\Omega_0(R) = \int d\mu(\mathbf{J}) \left\langle \delta\left(R - \frac{\mathbf{J}\cdot\mathbf{T}}{n}\right) \right\rangle_{\mathbf{T}}.$$

Linear Perceptron, High-T limit

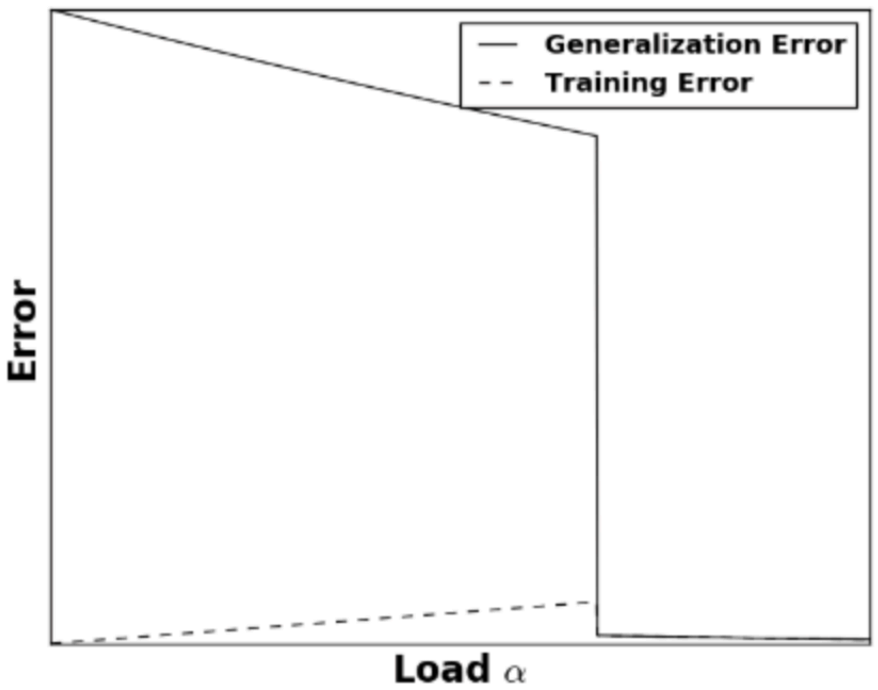
Generalization Error ~ 1 - R

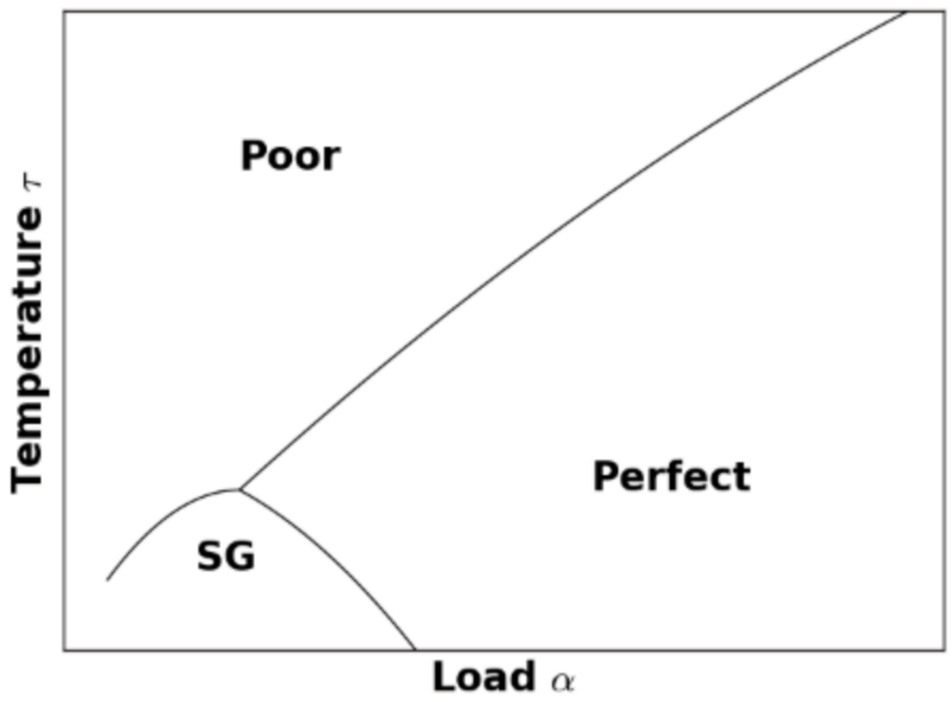
Exhibits phase behavior when overfit

# Key Results: Complex Phase Behavior

**Rethinking generalization requires revisiting old ideas: statistical mechanics approaches and complex learning behavior**

Charles H. Martin, Michael W. Mahoney (2017)




overfit models behave like glassy / meta-stable spin glasses

# New Set Up: Matrix-generalized Student-Teacher

Student vector —> weight matrix

$$\mathbf{J} \rightarrow \mathbf{S} \in \mathcal{R}^{N \times M}$$

Heavy-Tailed correlation matrices

$$\mathbf{A} := \frac{1}{N}\mathbf{S}^T\mathbf{S} \qquad \mathbf{X} := \frac{1}{N}\mathbf{W}^T\mathbf{W}$$

Student-Teacher matrix overlap

$$\mathbf{R} := \frac{1}{N}\mathbf{S}^T\mathbf{W} \qquad \mathbf{R}^T\mathbf{R} := \frac{1}{N}\mathbf{W}^T\mathbf{A}\mathbf{W}$$

(IZ) Free Energy associated with the generalization error

$$\mathbf{F}^{IZ} := -\frac{1}{\beta}\ln\int d\mu(\mathbf{S})\exp\left[-\beta(1 - \frac{1}{N}(\mathrm{Tr}\left[\mathbf{R}^T\mathbf{R}\right])^{1/2})\right]$$

# Layer Quality Metrics : SemiEmpirical Theory

$$\lim_{N\to\infty} \frac{1}{N} \log \; \mathbb{E}_A \left[ exp\left( \frac{\beta}{2} Tr\left[\mathbf{W}^T\mathbf{A}\mathbf{W}\right] \right) \right] = \frac{\beta}{2} \sum_{i=1}^{M} G_A(\lambda_i)$$

"Asymptotics of HCZI integrals …" Tanaka (2008)

"Generalized Norm"
simple, functional form
can infer from empirical fit

Eigenvalues of Teacher
empirical fit: R-transform:

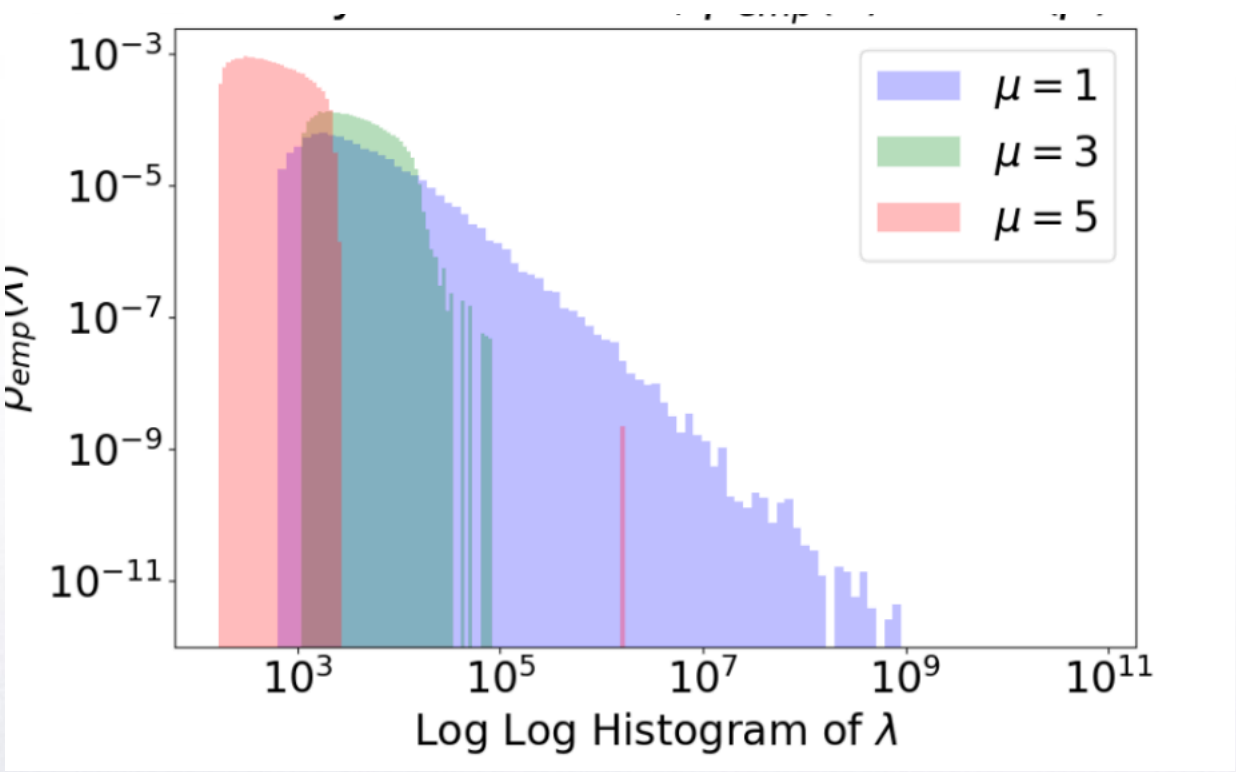
$$G_A(\lambda) := \int_0^\lambda R_A(z)dz \longleftarrow R(z) := \kappa_1 + \kappa_2 z + \kappa_3 z^2 + \ldots$$

free cumulants

WeightWatcher
Layer Quality metric

$$log \sum_{i=1}^{M} G_A(\lambda_i)$$

# R-Transform : Heavy Tails




$$R(z) := \kappa_1 + \kappa_2 z + \kappa_3 z^2 + \ldots$$

Alpha smaller =>

Heavy tail = >

Larger higher-order cumulants =>

Better generalization…

(down to alpha=2)

# New Principle of Learning: Volume Preserving Transformation

As the correlations concentrate into the tail $\mathbf{X} \rightarrow \mathbf{X}^{eff}$,

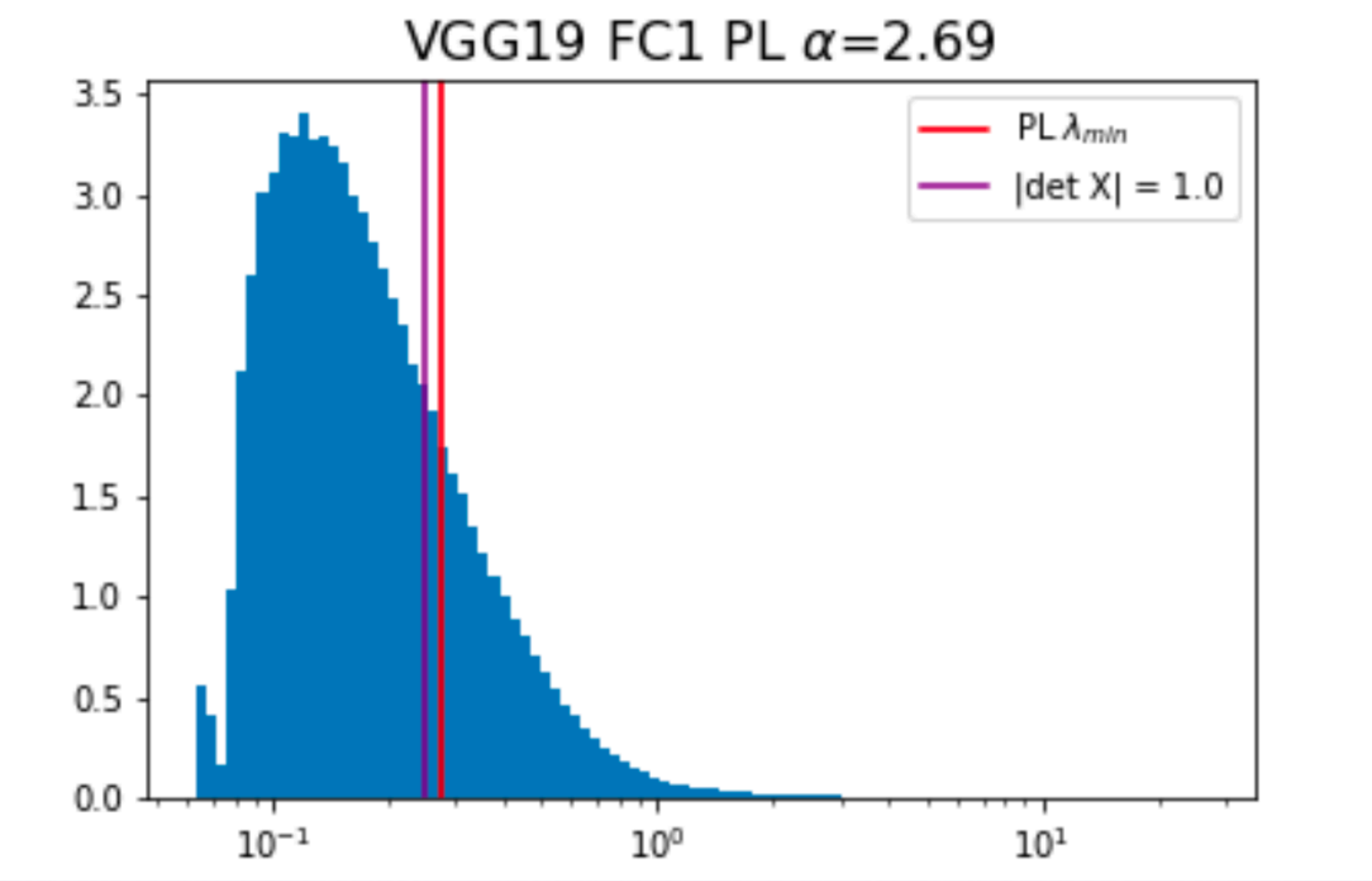
The change of measure must satisfy a volume preserving transformation $d\mu(\mathbf{S}) \rightarrow d\mu(\mathbf{A})$

$$\Big\langle \det \mathbf{A} \Big\rangle_{\mathbf{A}} \simeq \det \mathbf{X} = \prod_{t} \lambda_t \;\; \forall \lambda_t \in \rho_{tail}(\lambda),$$

Which sets a condition on the eigenvalues in the tail

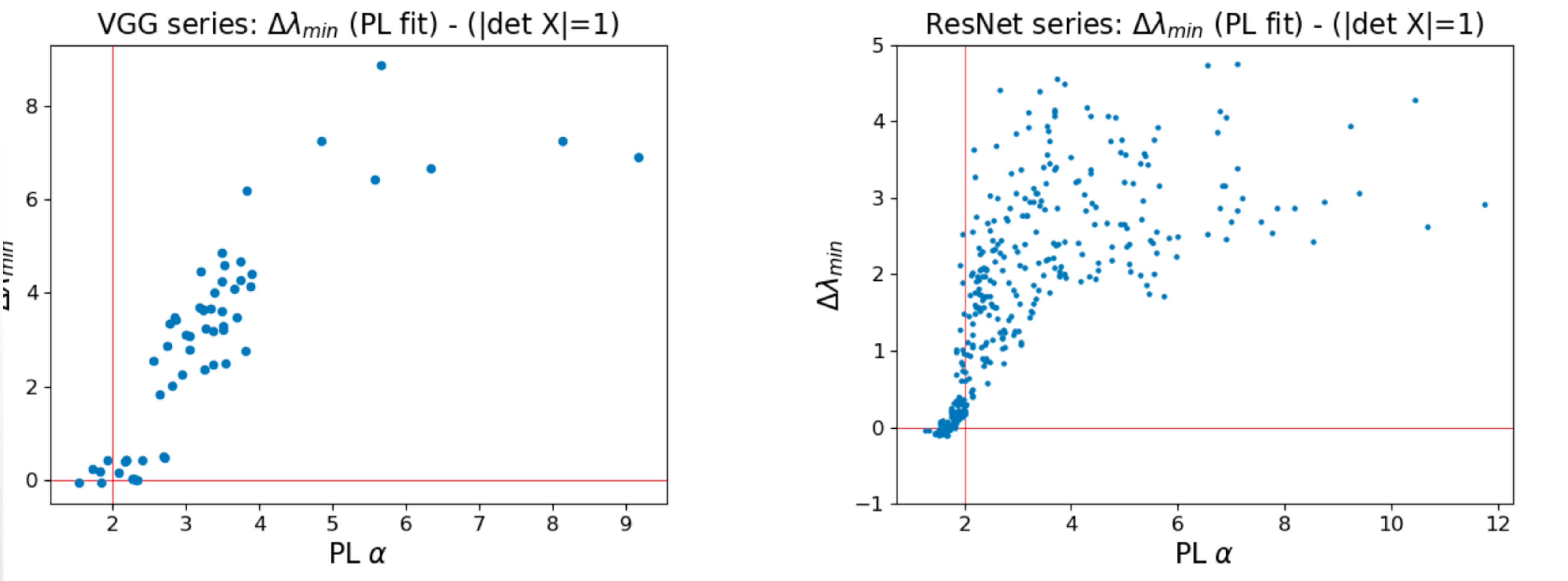
$$|\det \mathbf{X}| \simeq 1; \quad \mathrm{Tr}\left[\log \mathbf{X}\right] = \log|\det \mathbf{X}| \simeq 0.$$

# New Principle of Ideal Learning: Volume Preserving Transformation



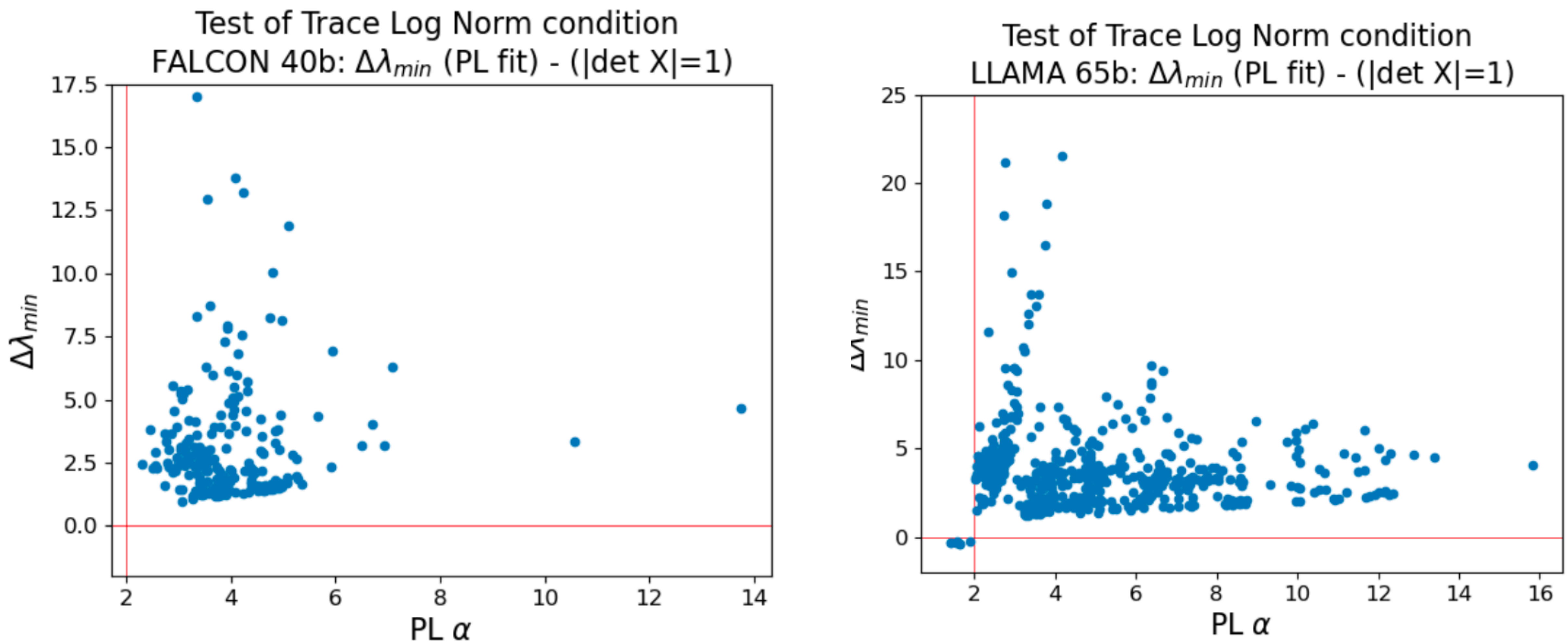

As alpha -> 2, the |det X|=1 condition holds!

# New Principle of Ideal Learning: Volume Preserving Transformation




As alpha -> 2, the |det X|=1 condition holds!

# New Principle of Ideal Learning:
# Volume Preserving Transformation

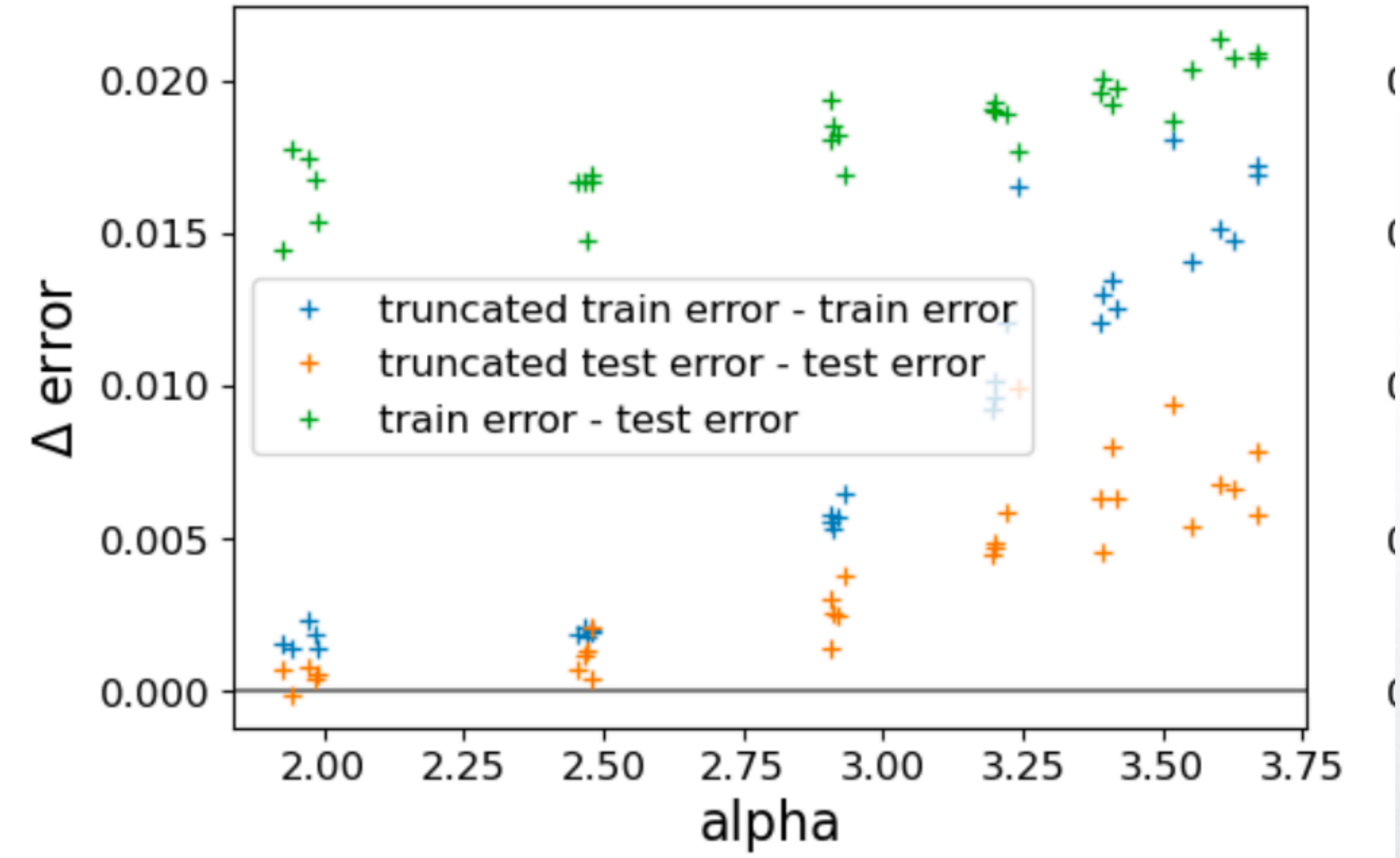



(b) Falcon 40B

(d) Llama 65B

# New Principle of Ideal Learning:

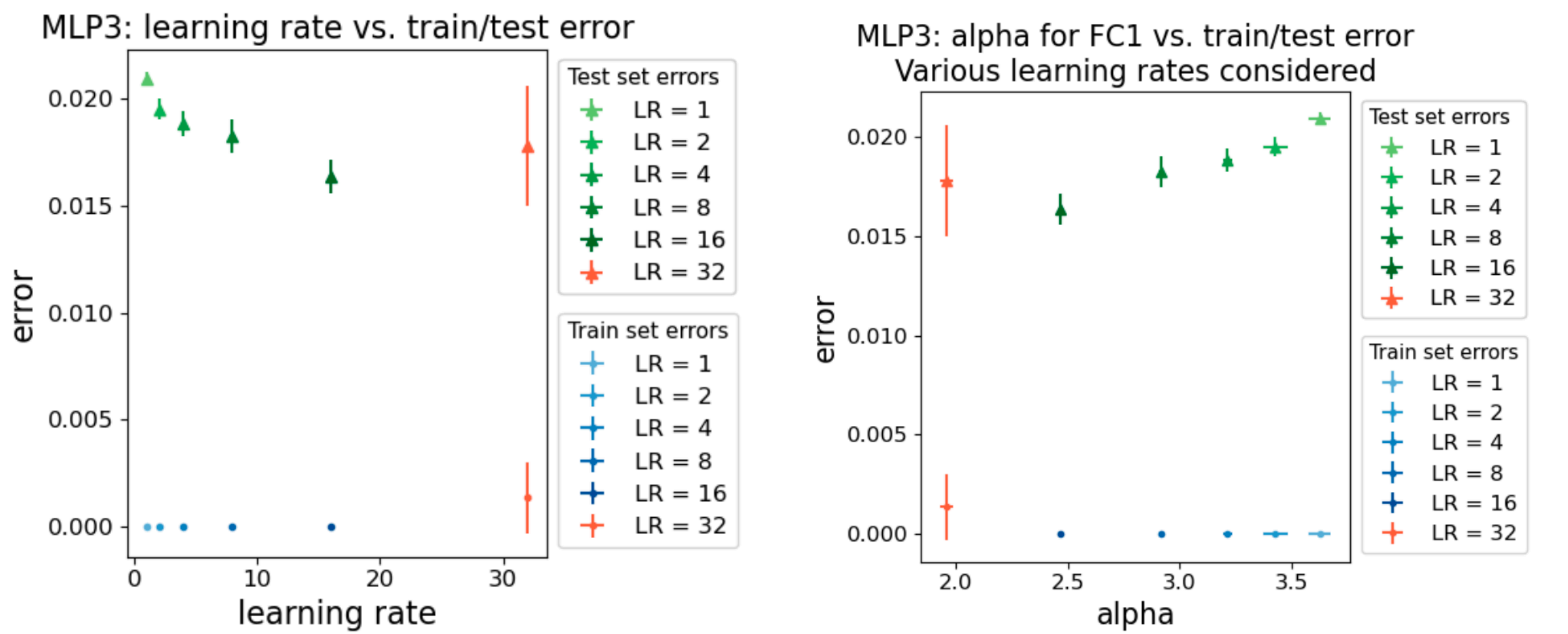
MLP3 on MNIST, varying LR
Truncated SVD on FC1




As alpha -> 2, SVD generalizing components concentrate in the tail

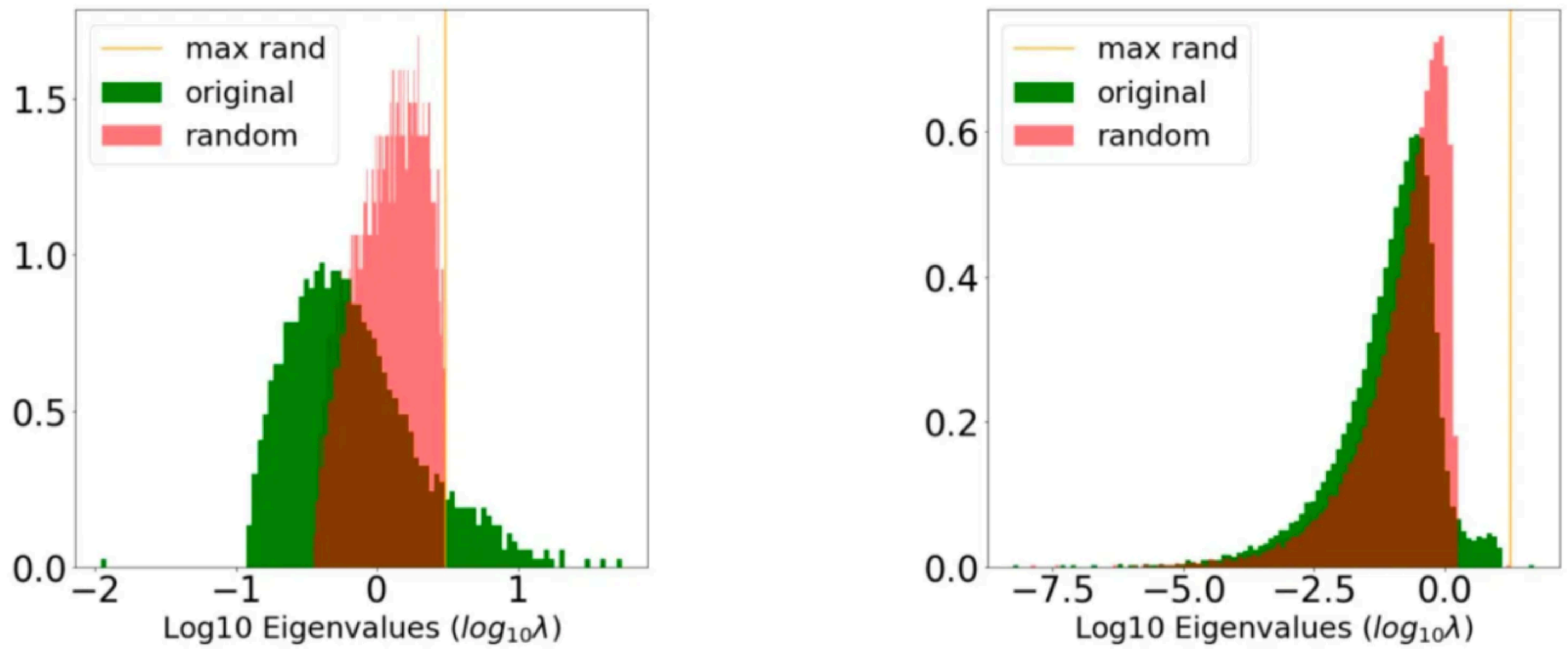
# Layer Quality Metrics : Detecting Overfitting

## MLP3 on MNIST, varying LR




As alpha < 2, layers appear to be overfit

# Layer Quality Metrics : Correlation Traps



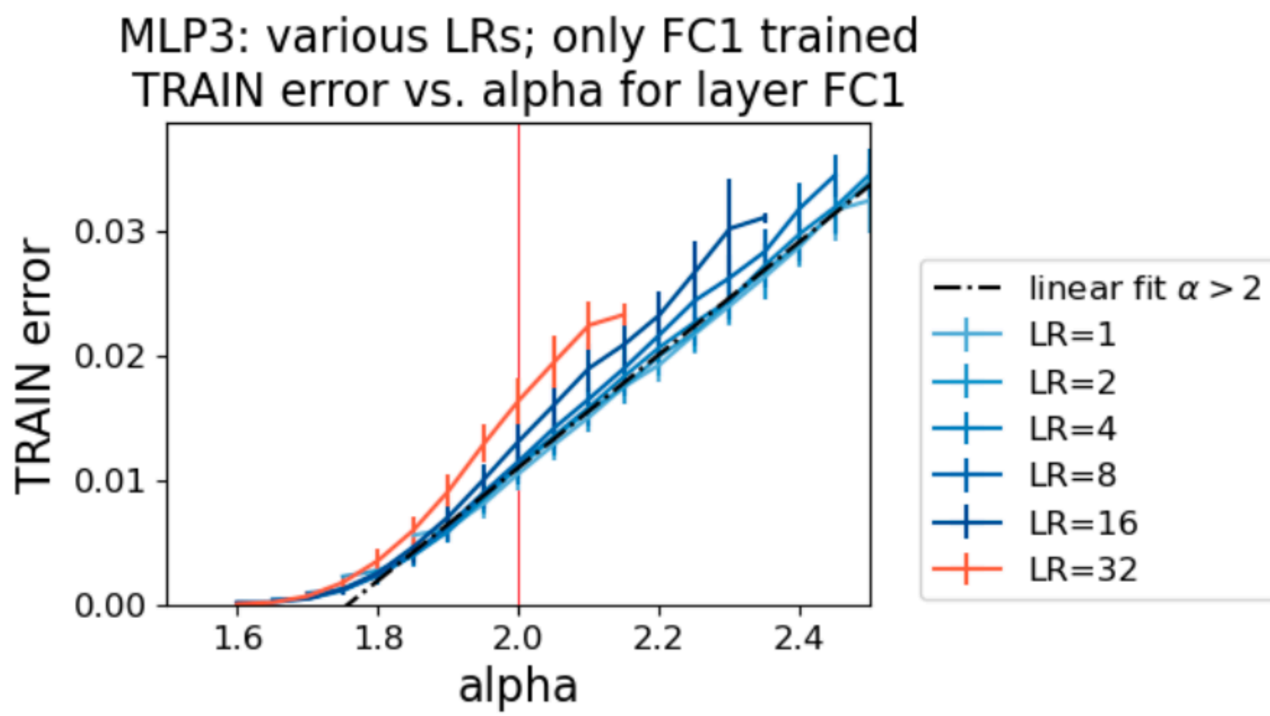

(a) ESD of **W** and randomized **W**.

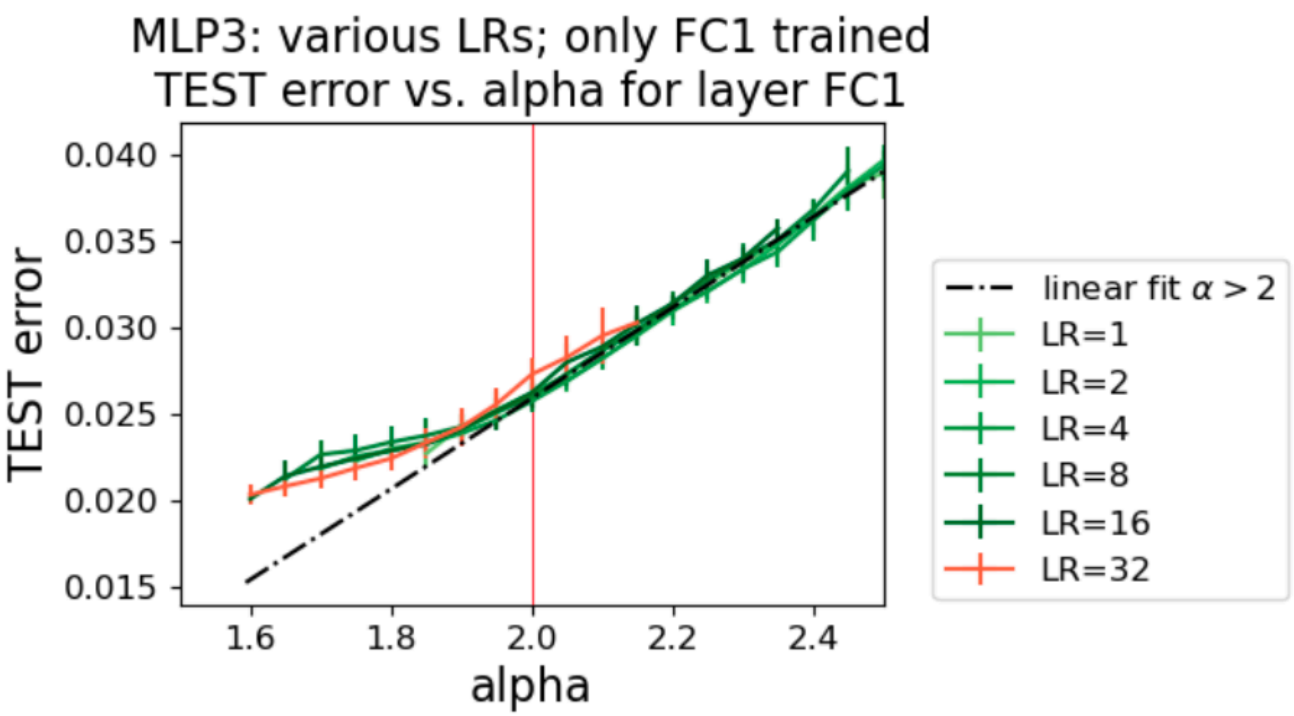
(b) ESD of **W** and randomized **W**.

A result of spuriously large learning rates

# Layer Quality Metrics : Detecting Overfitting

MLP3 on MNIST: Reduced capacity
Train FC1, freeze FC2, FC3

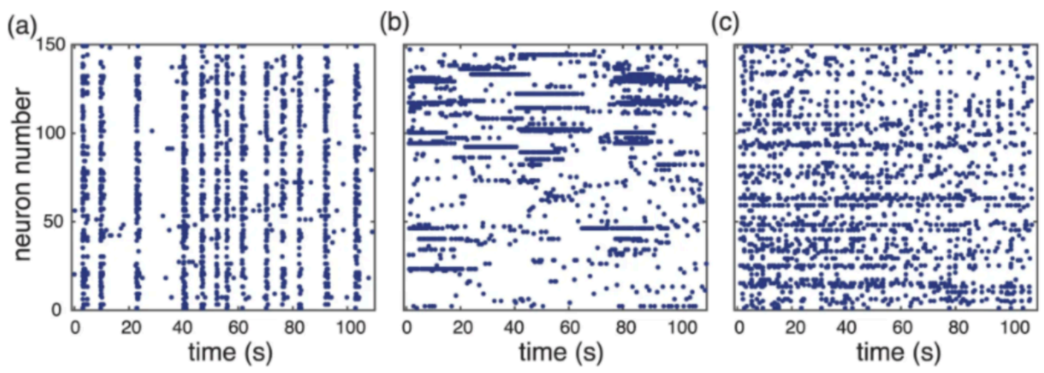


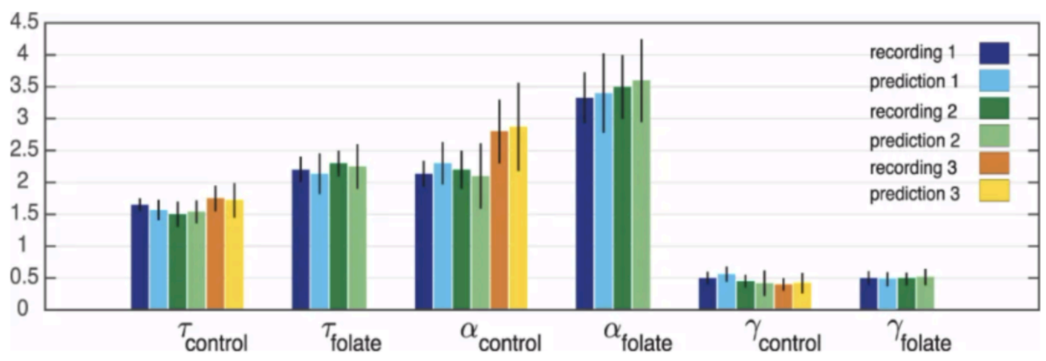

As alpha < 2, layer FC1 appears to be overfit
Why ?: IZ Free Energy becomes non-extensive

# Heavy Tailed Correlations: in Neuroscience

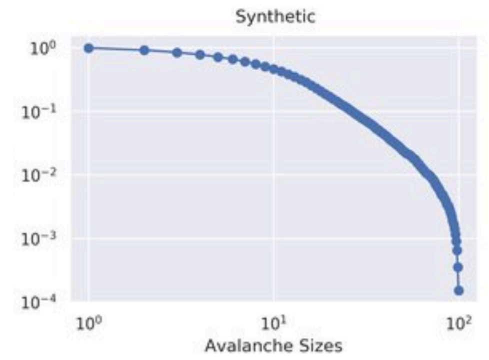
From: Neuronal avalanche dynamics indicates different universality classes in neuronal cultures

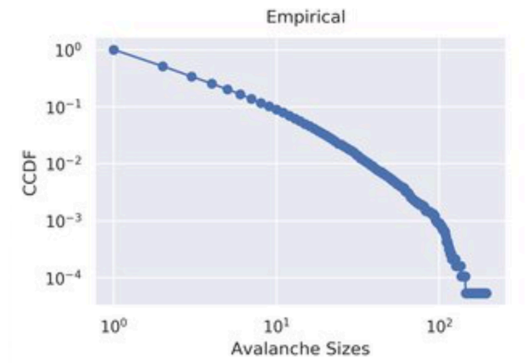



(c) Spiking activity of cultured neurons




(d) Critical exponents, fit to scalaing model




weightwatcher supports several PL fits
from experimental neuroscience

plus totally new shape metrics
we have invented (and published)

Spiking (i.e real) neurons exhibit power law behavior

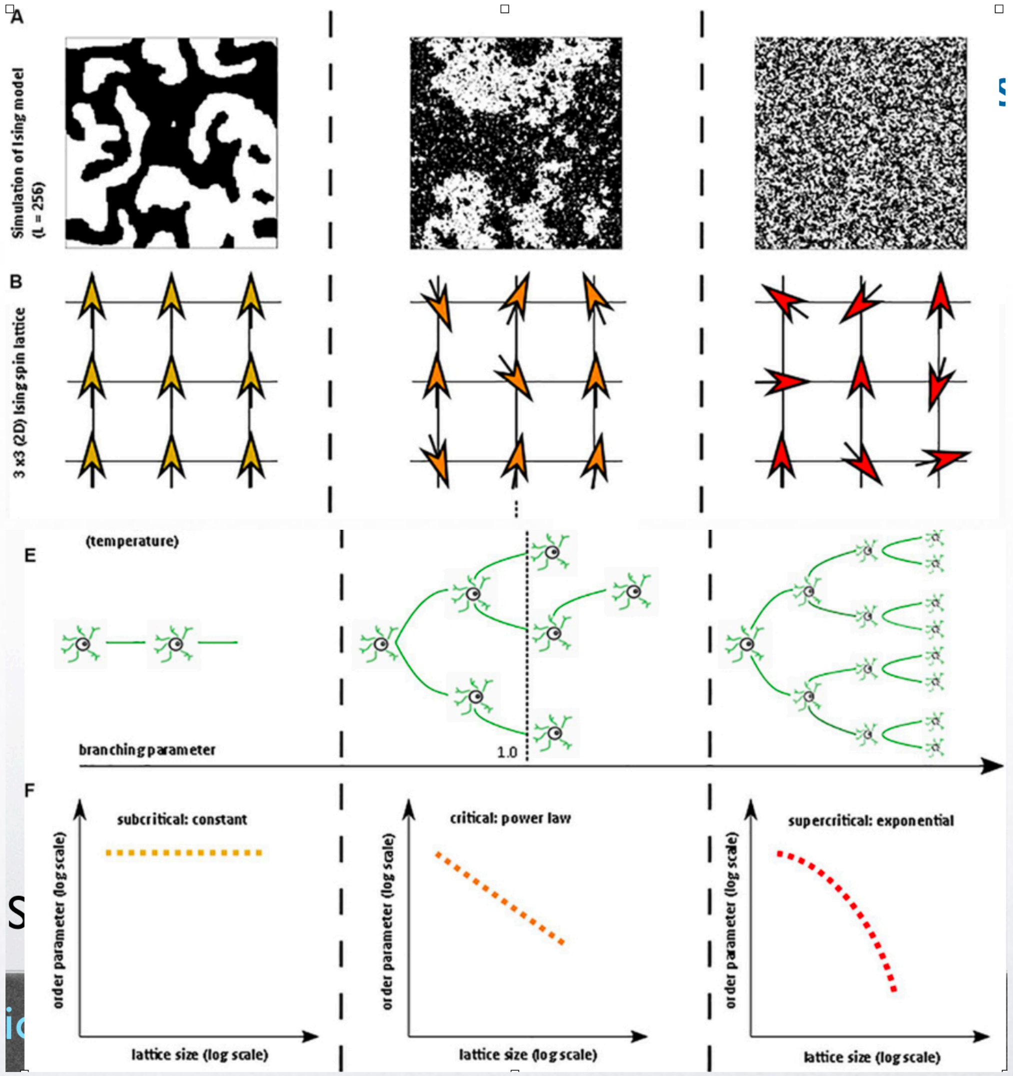
# WeightWatcher: why Power Law fits ?

*The Critical Brain Hypothesis*

Evidence of Self-Organized Criticality (SOC)
Per Bak (How Nature Works)

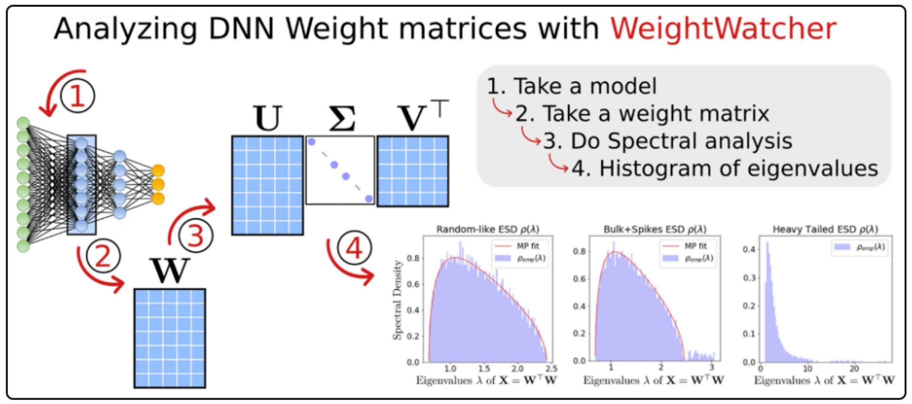
As neural systems become more complex
they exhibit power law behavior
and then truncated power law behavior

We see *exactly* this behavior in DNNs
and it is predictive of learning capacity




Spiking (i.e real) neurons exhibit (truncated) power law behavior w|w weight|watcher.ai

HOME USAGE RESEARCH PRESENTATIONS CODE EXAMPLES COMMUNITY ABOUT




# weight | watcher

## Data-Free Diagnostics for Deep Learning

WeightWatcher (w|w) is an open-source, diagnostic tool for analyzing Deep Neural Networks (DNN), without needing access to training or even test data. It is based on theoretical research into Why Deep Learning Works, using the new Theory of Heavy-Tailed Self-Regularization (HT-SR), [published in JMLR and Nature Communications](#).

WeightWatcher is a one-of-a-kind must-have tool for anyone training, deploying, or monitoring Deep Neural Networks (DNNs).

pip install weightwatcher

[and check out our latest LLM Leaderboard!](#)

https://weightwatcher.ai 115K+ downloads; 1000 stars

downloads 104k | release v0.7.1.5 | license Apache-2.0 | Published in Nature | Video Tutorial | discord 14 online | LinkedIn | Blog

We are looking for early adopters and collaborators


charles@calculationconsulting.com