# POWER LAW COVARIANCE IS GREAT


Elliot Paquette

McGill

*Based on joint work with:*

McGill: *Courtney Paquette*, *Keliang Xiao*

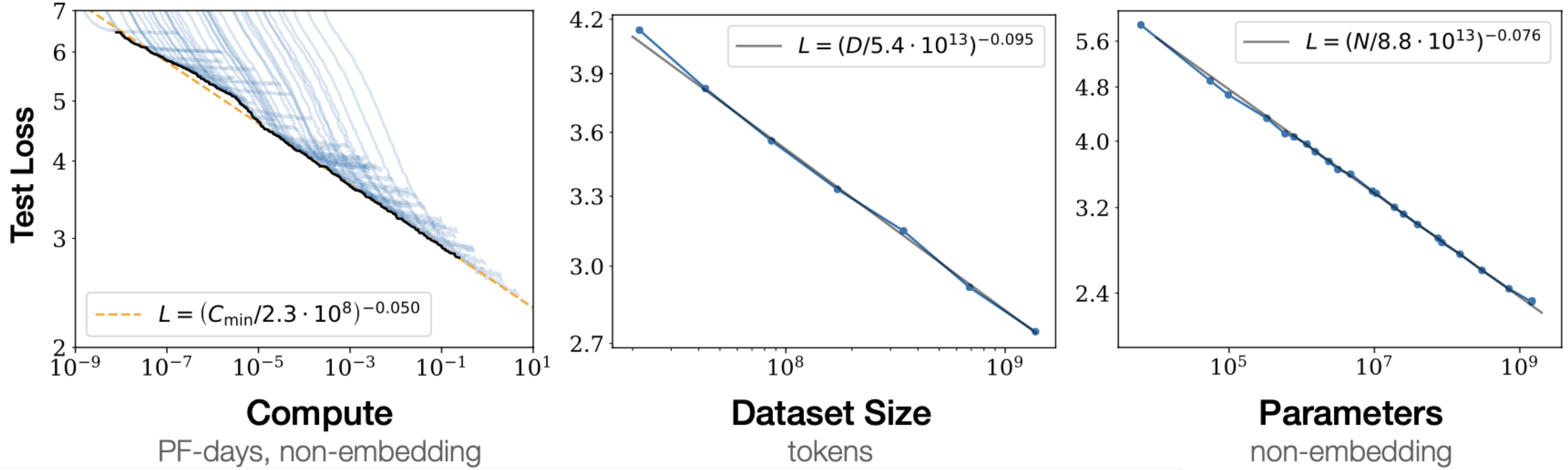
USC University of Southern California: Yizhe Zhu

Mila: *Gauthier Gidel*, *Damien Ferbach*

DeepMind: *Katie Everett*, *Lechao Xiao*, *Jeffrey Pennington*

# THE SCALING HYPOTHESIS

Kaplan et al. 2020

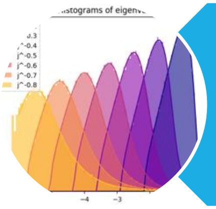



**Smooth power laws:** Performance has a power-law relationship with each of the three scale factors $N, D, C$ when not bottlen... (see Figure 1). We observ... must flatten out eventually...

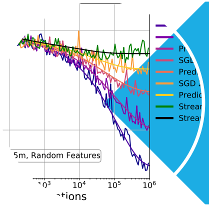
**Convergence is inefficient:** When working within a fixed compute budget $C$ but without any other restrictions on the model size $N$ or available data $D$, we attain optimal performance by training *very large models* and stopping *significantly short of convergence* (see Figure 3). Maximally compute-efficient training would therefore be far more sample efficient than one might expect based on training small models to convergence, with data requirements growing very slowly as $D \sim C^{0.27}$ with training compute. (Section 6)

See also Hoffman et al. (Chinchilla), which has $n \propto f^{0.5}$,
$n$ =number of samples. $f$ =number of flops

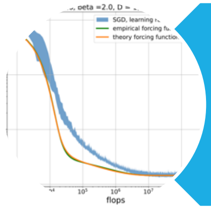
# TALK PLAN



Part 1: The Power law Random features model

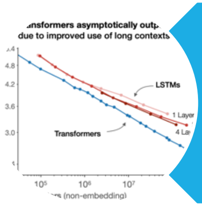



Part 2: The role of the nonlinearity




Part 3: Scaling laws for the linear model

- *In which we can see many different behaviors of SGD*




Part 4: What can change a scaling law?

# DESIGNING A MODEL...

Suppose $X$ is a latent data vector in $\mathbb{R}^v$.

We access an embedding of $X$ into $\mathbb{R}^d$, by $\sigma(W^\top X)$ for $W \sim N(0, I_v \otimes I_d/d)$.

The targets are computed in the latent space: $\hat{\sigma}(\langle X, \hat{\beta} \rangle)$.

We fit the linear model $\langle \theta, \sigma(W^\top X) \rangle$ with 1-pass SGD, MSE loss.

# Power law data-geometry

$X \sim N(0, \Sigma)$ with $\Sigma_{jj} = j^{-2\alpha}$ **SOURCE**

- Observable
- Real-world

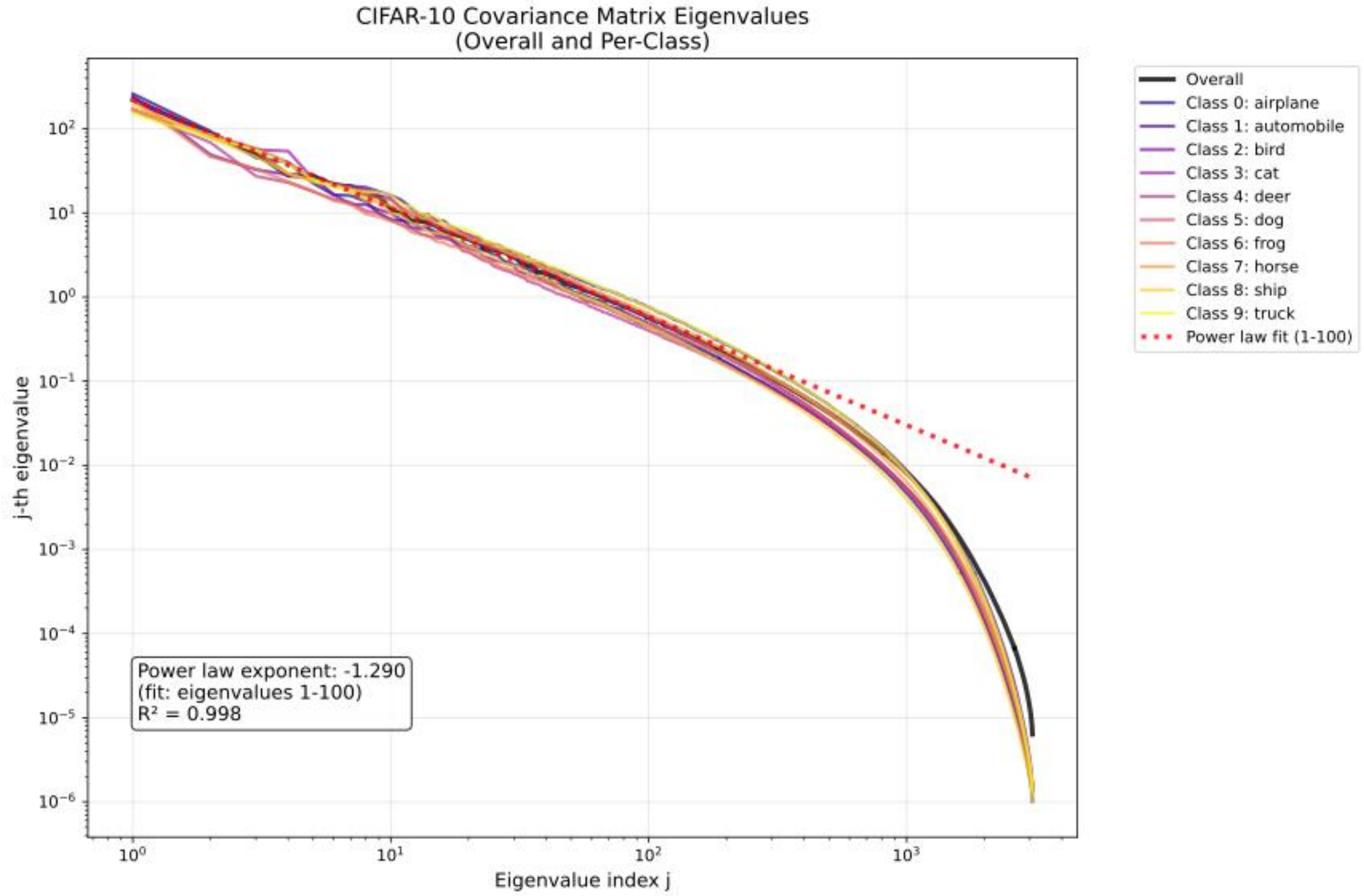
$\hat{\beta}_j = j^{-\beta}$. **Capacity**

- Artificial
- Phenomenological

*Larger $\alpha, \beta \Rightarrow$ Lower complexity*

Ren, Nichani, Wu, Lee '25 Arous, Erdogdu, Vural, Wu '25

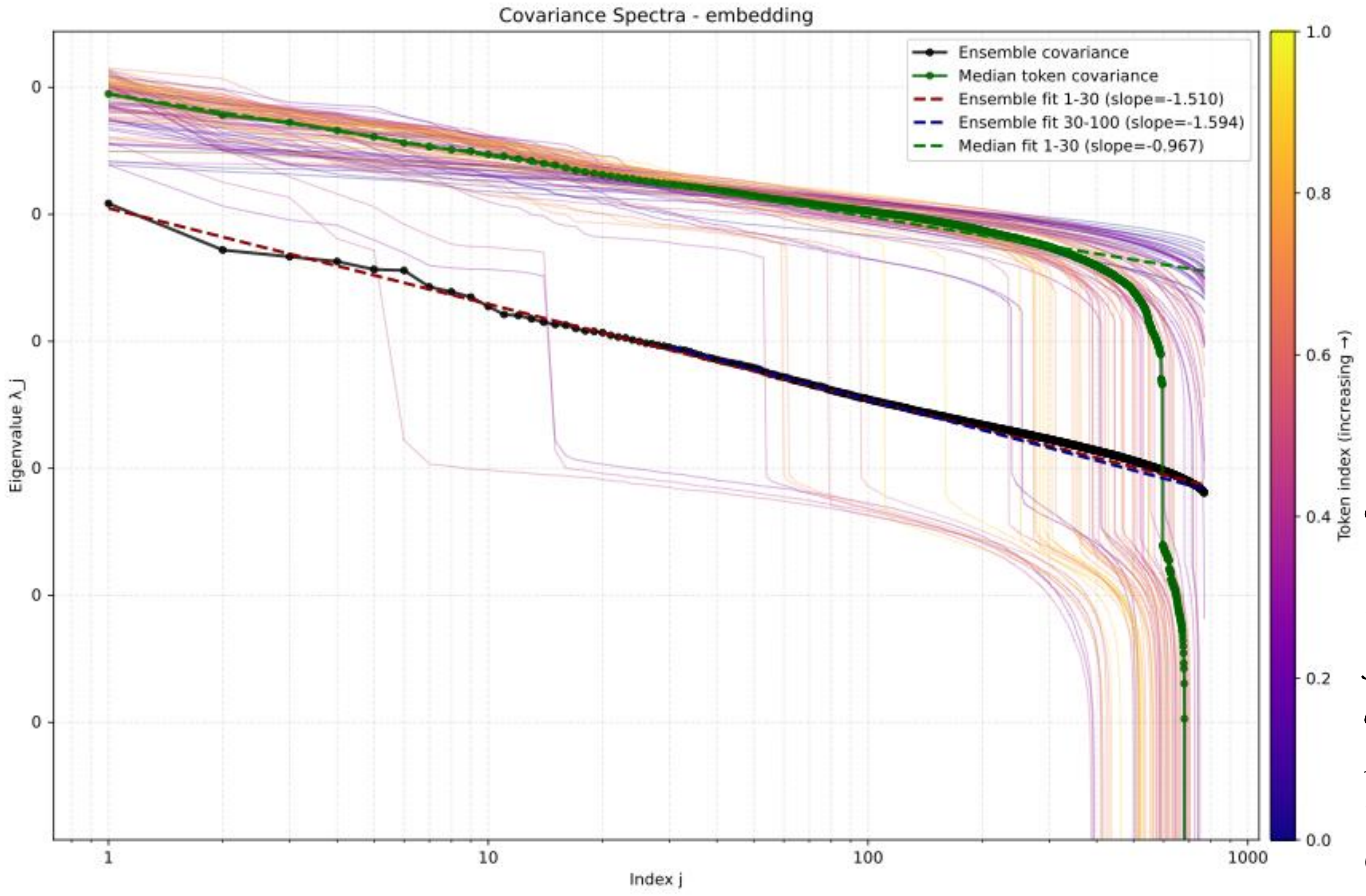
# CIFAR-10


CIFAR-10 Covariance Matrix Eigenvalues
(Overall and Per-Class)
j-th eigenvalue
Eigenvalue index j
Power law exponent: -1.290
(fit: eigenvalues 1-100)
R² = 0.998
Overall
Class 0: airplane
Class 1: automobile
Class 2: bird
Class 3: cat
Class 4: deer
Class 5: dog
Class 6: frog
Class 7: horse
Class 8: ship
Class 9: truck
Power law fit (1-100)






# FINEWEB-EDU-(UNTRAINED-EMBEDDING LAYER)




768-dimensional (GPT2-small)

Gaussian embedding

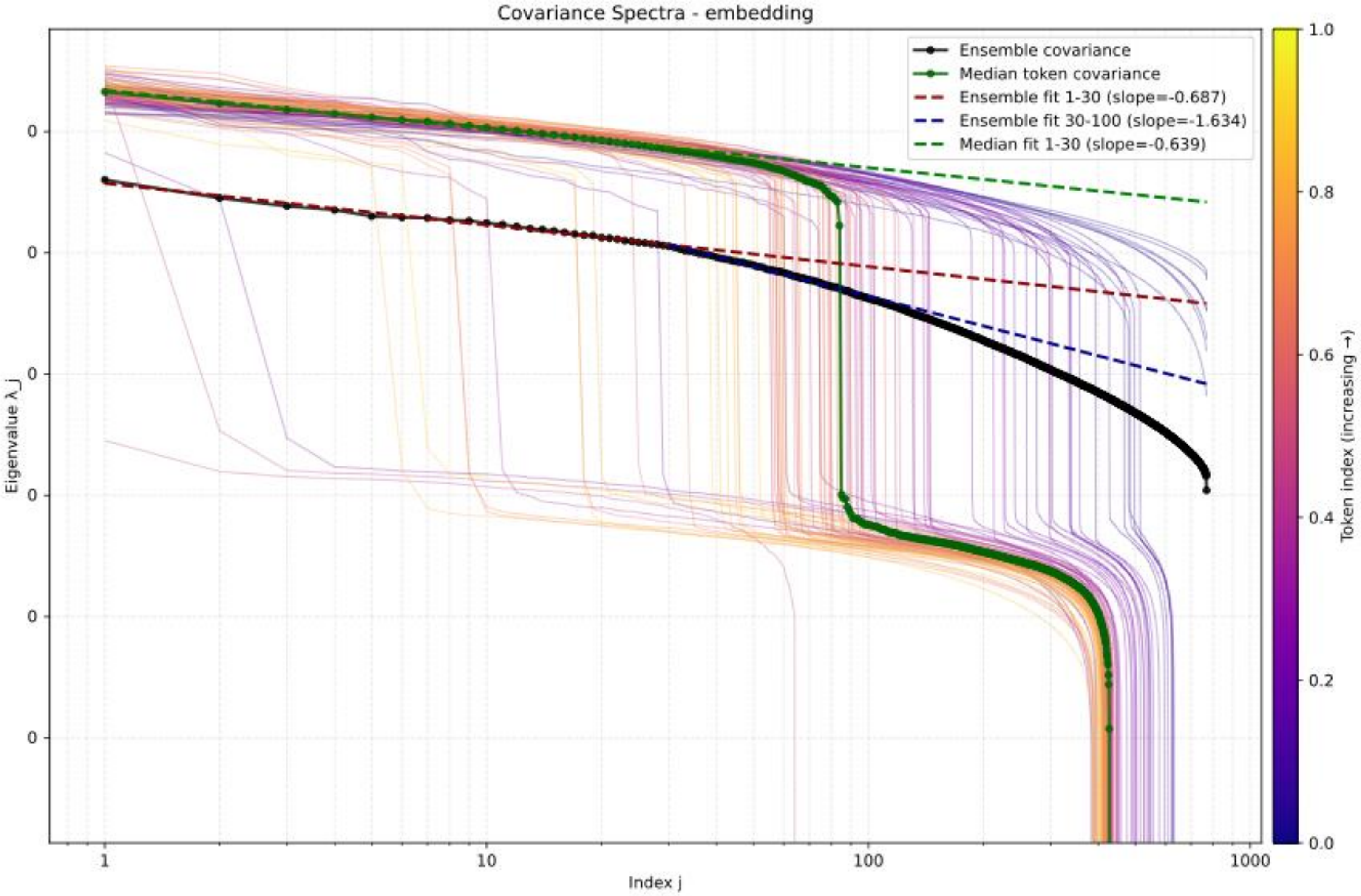
Condition on the token T

Return embedding of token (T-1)

*The Independent Jane:*

*For all the love, romance and scandal in Jane Austen's books, what they are really about is freedom and independence. Independence of thought and the freedom to choose. Elizabeth's refusal of Mr. Collins offer of marriage showed an independence seldom seen in heroines of*

# FINEWEB-EDU-(TRAINED-EMBEDDING LAYER)




768-dimensional (GPT2-small)

Gaussian embedding

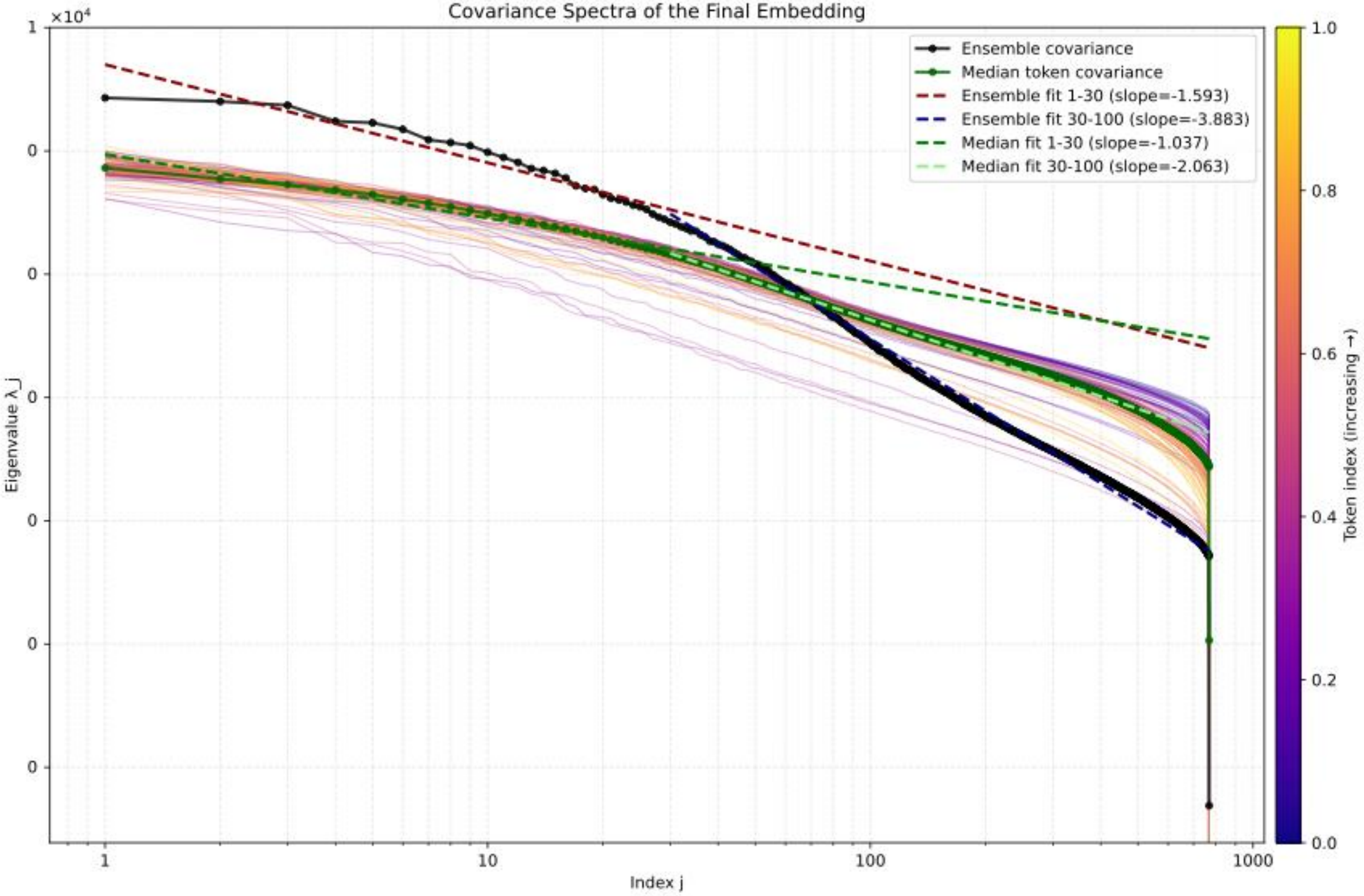
Condition on the token T

Return embedding of token (T-1)

*The Independent Jane:*

*For all the love, romance and scandal in Jane Austen's books, what they are really about is freedom and independence. Independence of thought and the freedom to choose. Elizabeth's refusal of Mr. Collins offer of marriage showed an independence seldom seen in heroines of*

# FINEWEB-EDU-(UNTRAINED-READOUT LAYER)




768-dimensional (GPT2-small)

Gaussian embedding

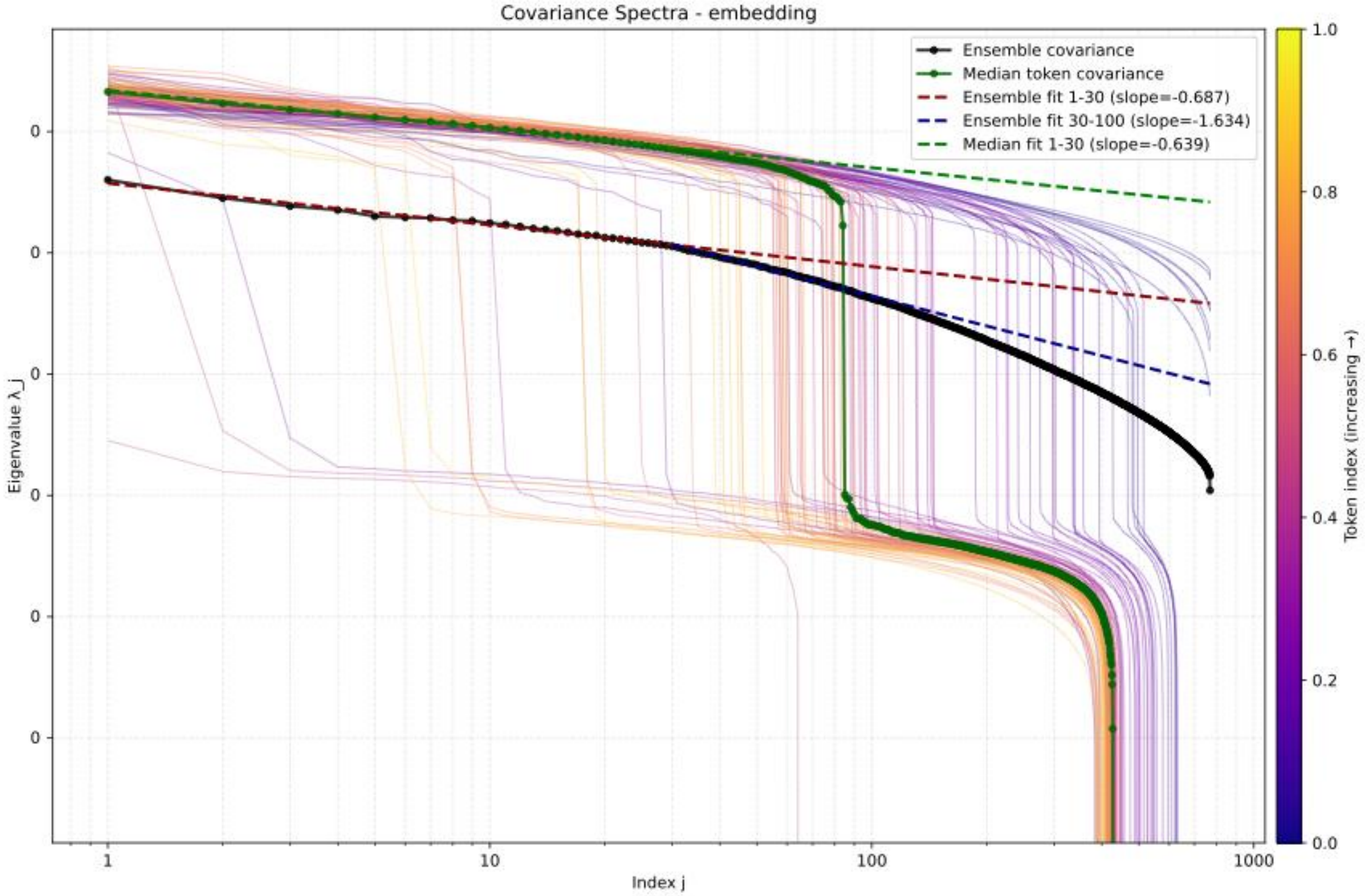
Condition on the token T

Return embedding of token (T-1)

*The Independent Jane:*

*For all the love, romance and scandal in Jane Austen's books, what they are really about is freedom and independence. Independence of thought and the freedom to choose. Elizabeth's refusal of Mr. Collins offer of marriage showed an independence seldom seen in heroines of*

# FINEWEB-EDU-(TRAINED-READOUT LAYER)




768-dimensional (GPT2-small)

Gaussian embedding

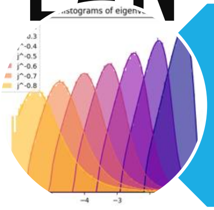
Condition on the token T

Return embedding of token (T-1)

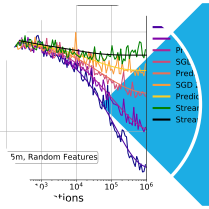
*The Independent Jane:*

*For all the love, romance and scandal in Jane Austen's books, what they are really about is freedom and independence. Independence of thought and the freedom to choose. Elizabeth's refusal of Mr. Collins offer of marriage showed an independence seldom seen in heroines of*

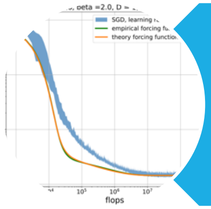
# TALK PLAN



Part 1: The Power law Random features model

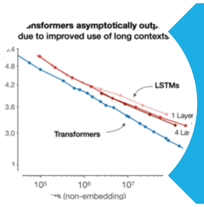



Part 2: The role of the nonlinearity




Part 3: Scaling laws for the linear model

- *In which we can see many different behaviors of SGD*




Part 4: What can change a scaling law?

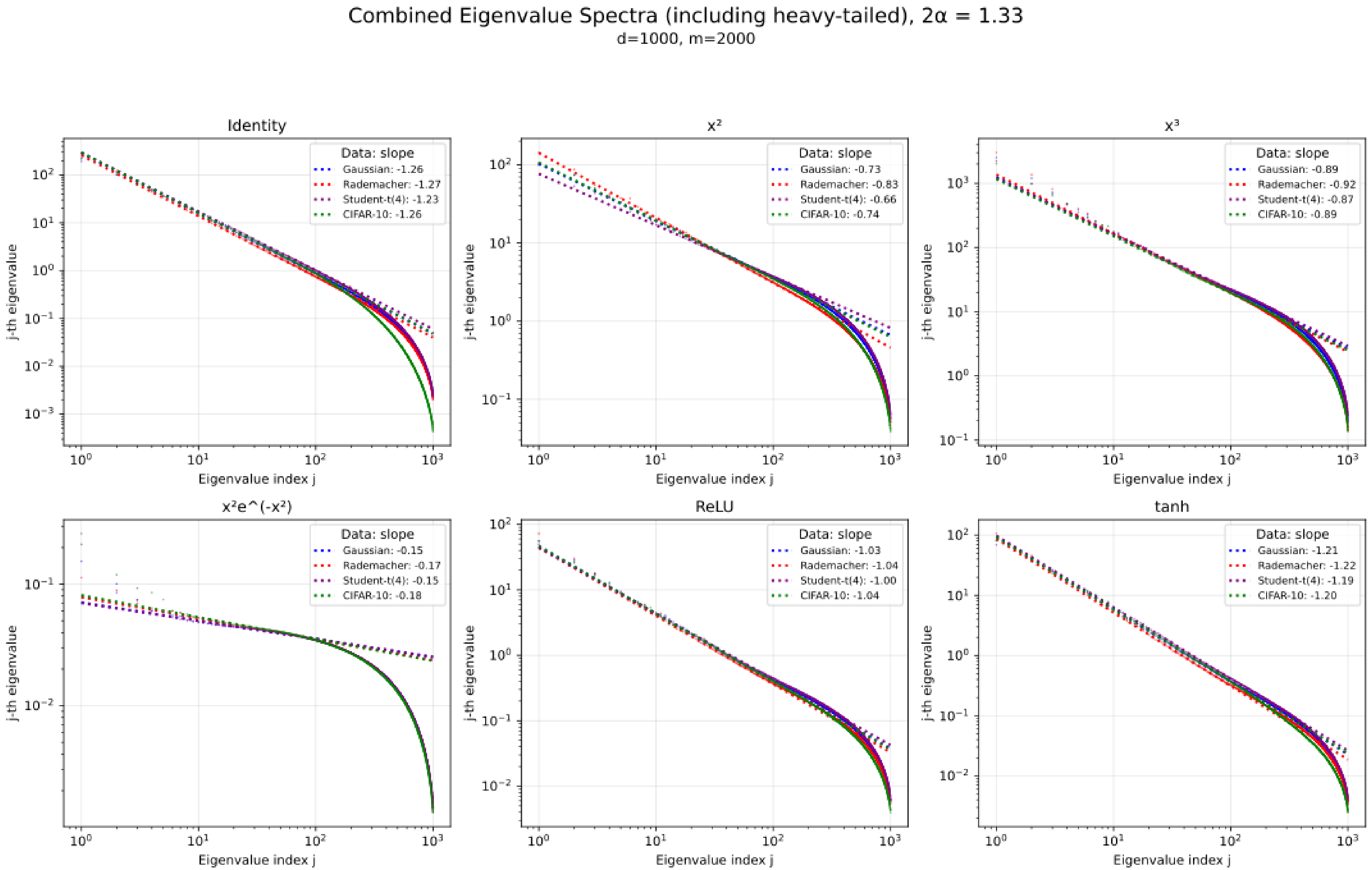
# QUESTION

Suppose your data $X \in \mathbb{R}^v$ has power law covariance.

*Capponetto & De Vito '05, '07 (and therein); Bach '17; Bahri '21,*

*'Zipf-law' of word distribution (1920s-30s)*

What is the distribution of $\sigma(W^\top X)$ for a Gaussian matrix $W$?



Combined Eigenvalue Spectra (including heavy-tailed), 2α = 1.33
d=1000, m=2000
Identity
Data: slope
Gaussian: -1.26
Rademacher: -1.27
Student-t(4): -1.23
CIFAR-10: -1.26
x²
Data: slope
Gaussian: -0.73
Rademacher: -0.83
Student-t(4): -0.66
CIFAR-10: -0.74
x³
Data: slope
Gaussian: -0.89
Rademacher: -0.92
Student-t(4): -0.87
CIFAR-10: -0.89
x²e^(-x²)
Data: slope
Gaussian: -0.15
Rademacher: -0.17
Student-t(4): -0.15
CIFAR-10: -0.18
ReLU
Data: slope
Gaussian: -1.03
Rademacher: -1.04
Student-t(4): -1.00
CIFAR-10: -1.04
tanh
Data: slope
Gaussian: -1.21
Rademacher: -1.22
Student-t(4): -1.19
CIFAR-10: -1.20
j-th eigenvalue
Eigenvalue index j


*2510.xxxxx*

With Yizhe Zhu (USC)
And Keliang Xiao (McGill)

- $\sigma(x) = x^p, p \in \mathbb{N}$.
- $2\alpha > 1$.
- $X$ is normally distributed with variances $j^{-2\alpha}$ in $\mathbb{R}^v$ with $v > d$.
  ($v = \infty$ *is allowed*).
- $W$ is $N(0, I \otimes I/d) \in \mathbb{R}^{d \times v}$

The eigenvalues of

$$K(W) = \mathbb{E}_X(\sigma(W^\top X) \otimes \sigma(W^\top X))$$

satisfy for all $1 \le j \le c_0 d$

$$c_1 \left( \frac{\log^{p-1}(j+1)}{j} \right)^{2\alpha} \le \lambda_j\big(K(W)\big) \le c_2 \left( \frac{\log^{p-1}(j+1)}{j} \right)^{2\alpha}$$

With probability tending to 1 as $d \to \infty$.

*c_j are nonrandom constants depending on* $\alpha, p$.

# PROOF IDEA
# $\sigma(x) = x^2$

1. Reduce to dominant kernel terms

$$K(W) = \mathbb{E}_X(\ \sigma(W^\top X) \otimes \sigma(W^\top X))$$
$$K(W)_{ij} \approx \langle W_i, \Sigma_X W_j \rangle^2 = \langle W_i^{\otimes 2}, \Sigma_X^{\otimes 2} W_j^{\otimes 2} \rangle$$

2. Do head-tail decomposition

$$\Sigma_X^{\otimes 2} = H^\epsilon + T^\epsilon$$

*Bartlett, Long, Lugosi, Tsigler '20*
*Lin, Wu, Kakade, Bartlett, Lee '24*

$H^\epsilon$ keeps all directions larger than $\epsilon$

$$T^\epsilon \perp H^\epsilon$$

3. In the head, we can reverse

$$W^{\pi\top} H^\epsilon W^\pi \to \sqrt{H} W^{\pi\top} W^\pi \sqrt{H}$$

Then $W^{\pi\top} W^\pi \approx I_\pi d$

4. Bound the tail in norm like $O(\epsilon)$

5. Spectrum matches that of $\Sigma_X^{\otimes 2}$ up to multiplicative constants.

# CONCLUSION

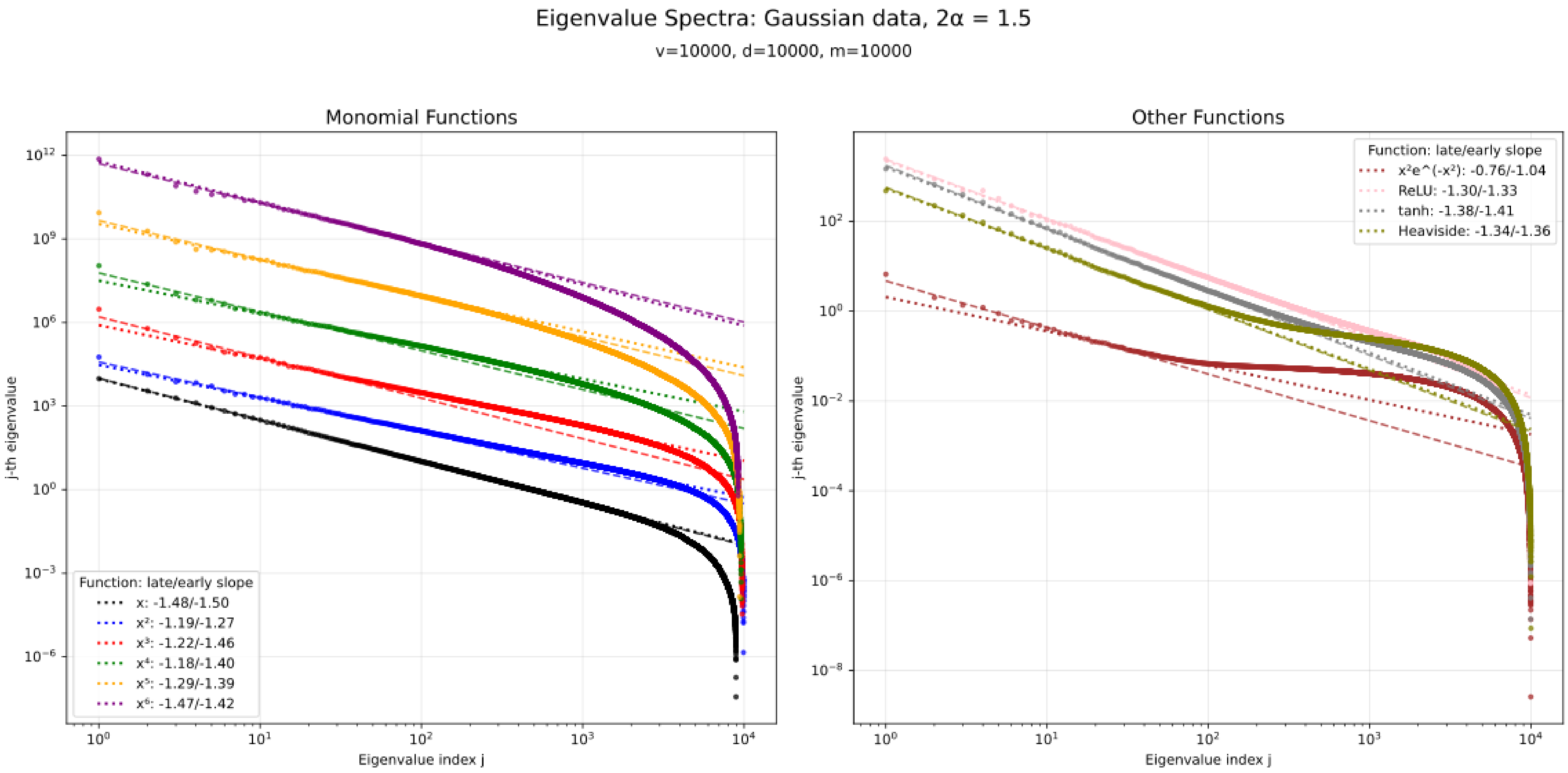
Spectrally, *polynomial* nonlinearity does very little.

Open Qs: so so many

Learning theory..

Non-polynomial..

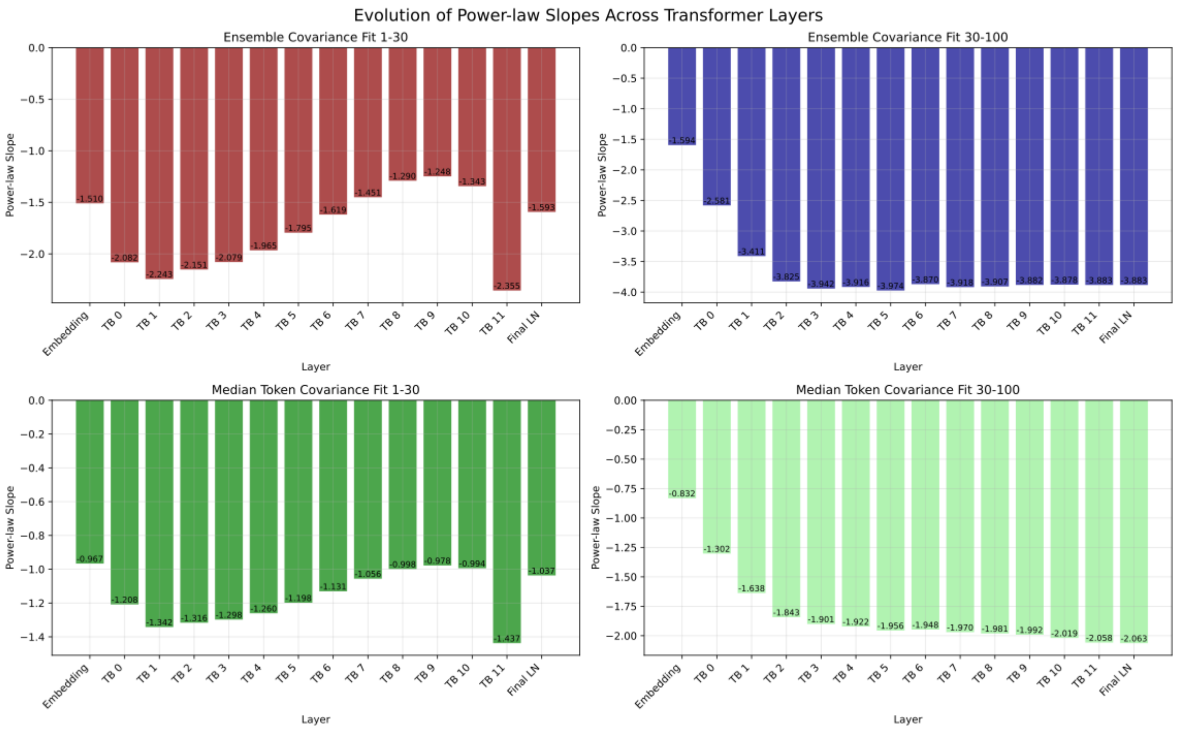
Universality..



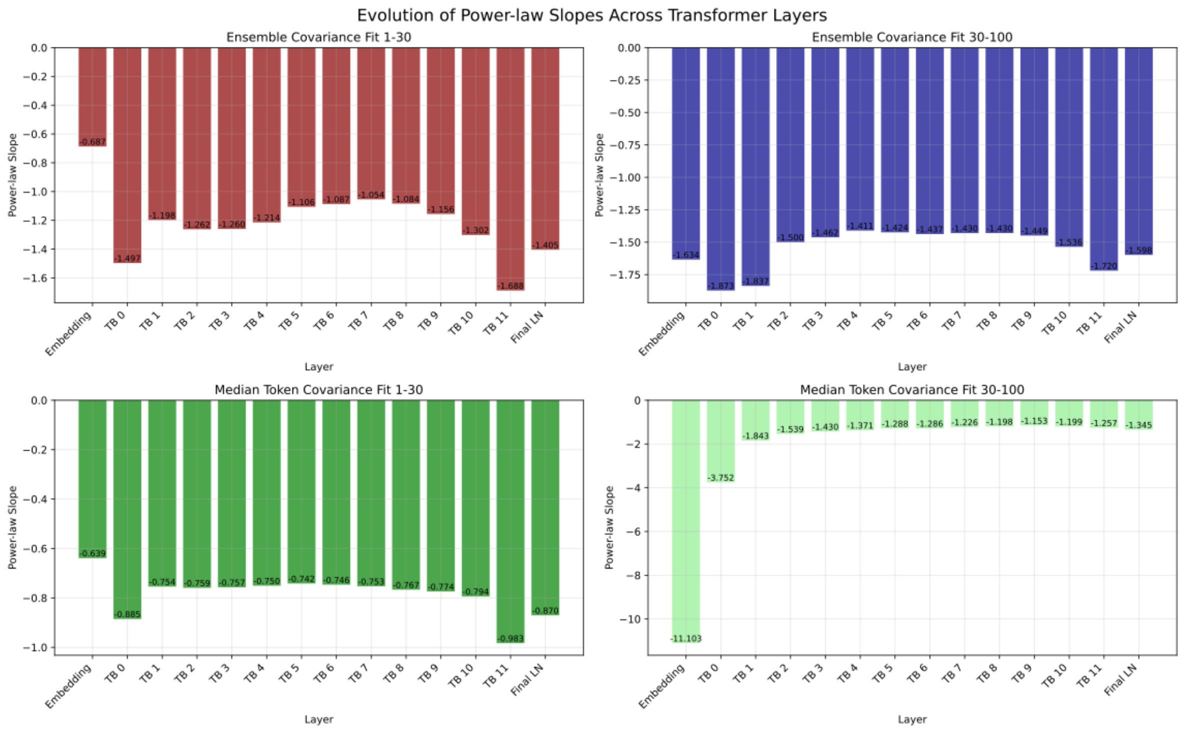
Eigenvalue Spectra: Gaussian data, 2α = 1.5
v=10000, d=10000, m=10000
Monomial Functions
j-th eigenvalue
Eigenvalue index j
Function: late/early slope
x: -1.48/-1.50
x²: -1.19/-1.27
x³: -1.22/-1.46
x⁴: -1.18/-1.40
x⁵: -1.29/-1.39
x⁶: -1.47/-1.42
Other Functions
Function: late/early slope
x²e^(-x²): -0.76/-1.04
ReLU: -1.30/-1.33
tanh: -1.38/-1.41
Heaviside: -1.34/-1.36

# SLOPE EVOLUTIONS ACROSS LAYERS (GPT2)

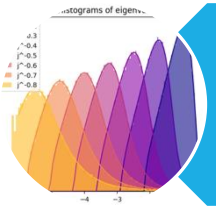
## Init



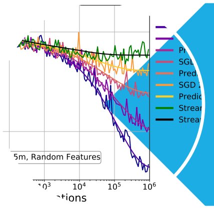

## After training



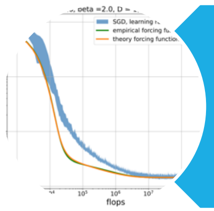
# TALK PLAN



Part 1: The Power law Random features model

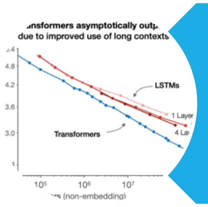



Part 2: The role of the nonlinearity




Part 3: Scaling laws for the linear model

- *In which we can see many different behaviors of SGD*




Part 4: What can change a scaling law?

# THE POWER LAW RANDOM FEATURES MODEL

$$\min_{\Theta}\{R(\Theta) := \frac{1}{2}\mathbb{E}\Big[\big(\langle\Theta, W^{\top}X\rangle - \langle\hat{\beta}, X\rangle\big)^2\Big]\}.$$
$$W \sim N(0, (I_v \otimes I_d)/d).$$

$$X \sim N(0,\Sigma) \text{ with } \Sigma_{jj} = j^{-2\alpha}$$

$$\hat{\beta}_j = j^{-\beta}.$$

Parameters

Latent Dim. $v$

Embedding Dim. $d$

Data ('Source') Complexity. $1/\alpha$

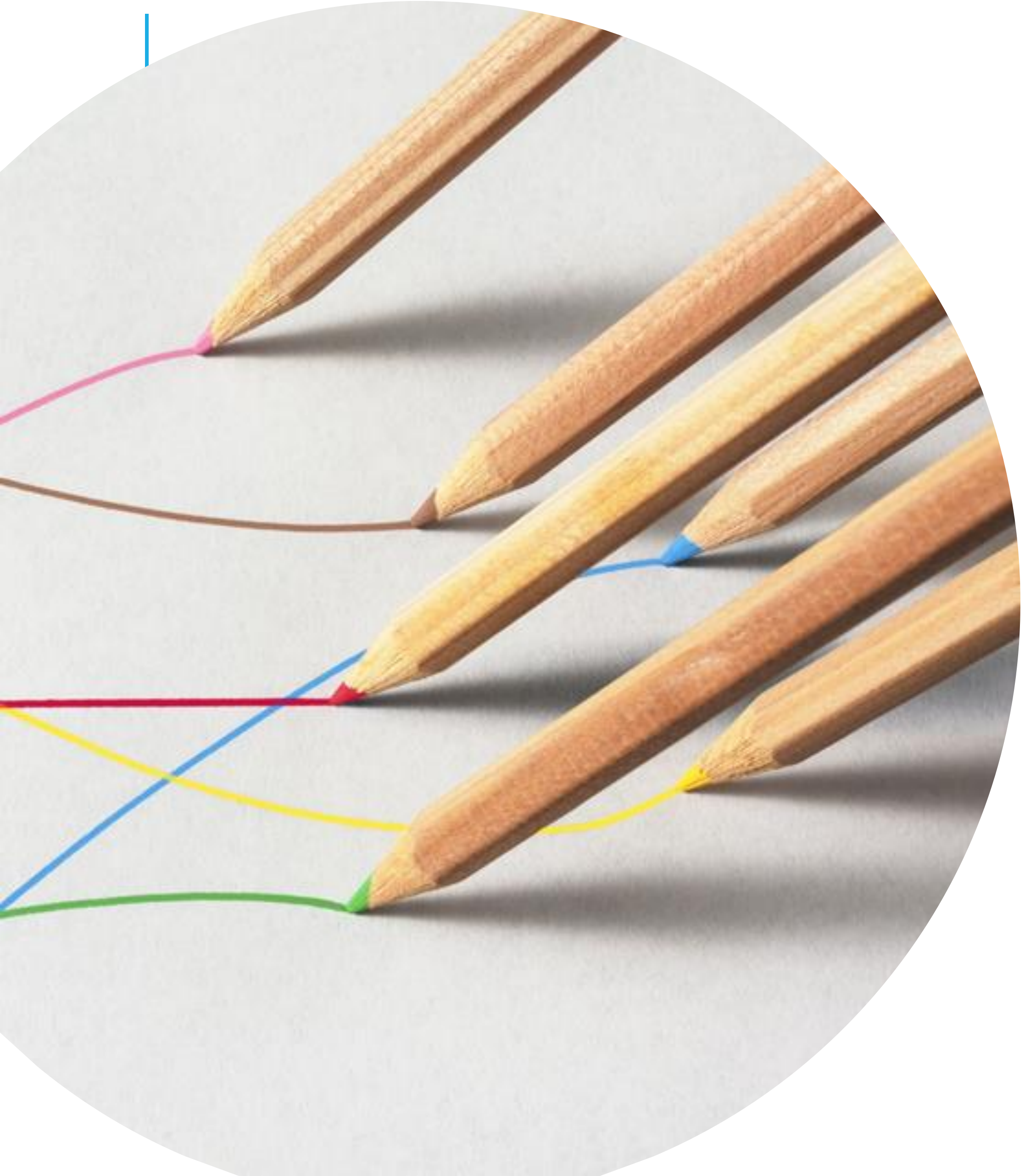
Target ('Capacity') Complexity. $1/\beta$

*Maloney, Roberts, Sully '22*

*Bahri et al. '21*

*Defillipis, Loureiro, Misiakiewicz '24*

*This work:*
$Paquette^{\otimes 2}, Xiao, Pennington\ '24$



Q

How do the power law exponents effect the loss curves in the linear model in one-pass SGD?

# LOSS CURVES OF ONE-PASS GAUSSIAN SGD

1. Streaming SGD: For iid samples:

$$\Theta_{k+1} = \Theta_k - \gamma W^\top X_{k+1}(\langle \Theta_{\mathrm{k}}, W^\top X_{k+1}\rangle - \langle \hat{\beta}, X_{k+1}\rangle)$$

2. Our main object of interest:

$$\psi(k) \coloneqq \mathbb{E}_X\big(R(\Theta_k)\big)$$

$$R(\Theta) \coloneqq \frac{1}{2}\mathbb{E}[(\langle \Theta, W^\top X\rangle - \langle \beta, X\rangle)^2]\}.$$

3. Gaussian SGD for LR satisfies:

$$\psi(k) \coloneqq \mathbb{E}_X\big(R(\Theta_k)\big)$$

$$\psi(k) = F(k) + \mathcal{K} * \psi(k)$$

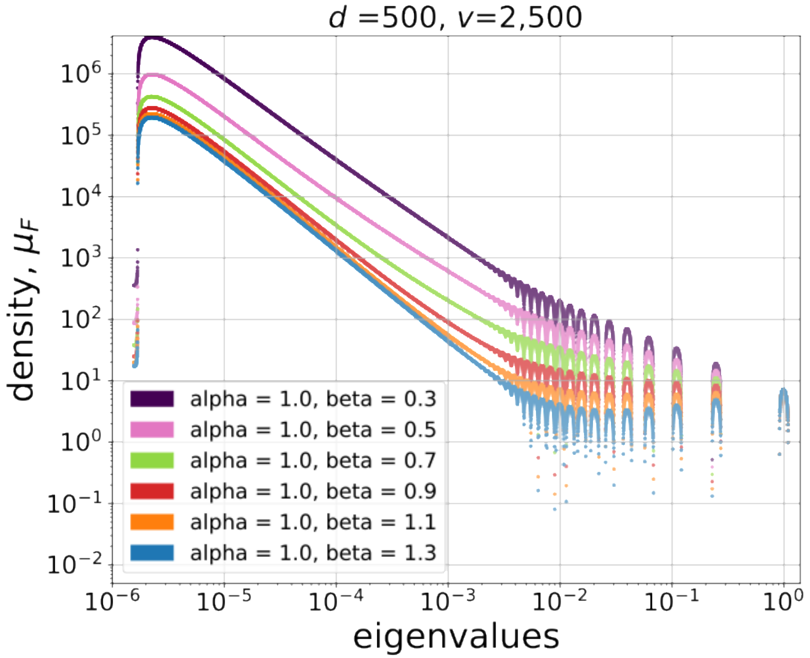
- $F(k)$ is (approximately) the loss under mean gradient descent
- $\mathcal{K}(k)$ is the risk curve of 1 unit of variance of SGD noise
- $\mathcal{K} * \psi$ is the convolution:
  - $\mathcal{K} * \psi(k) = \sum_r \mathcal{K}(k - r - 1)\psi(r)$

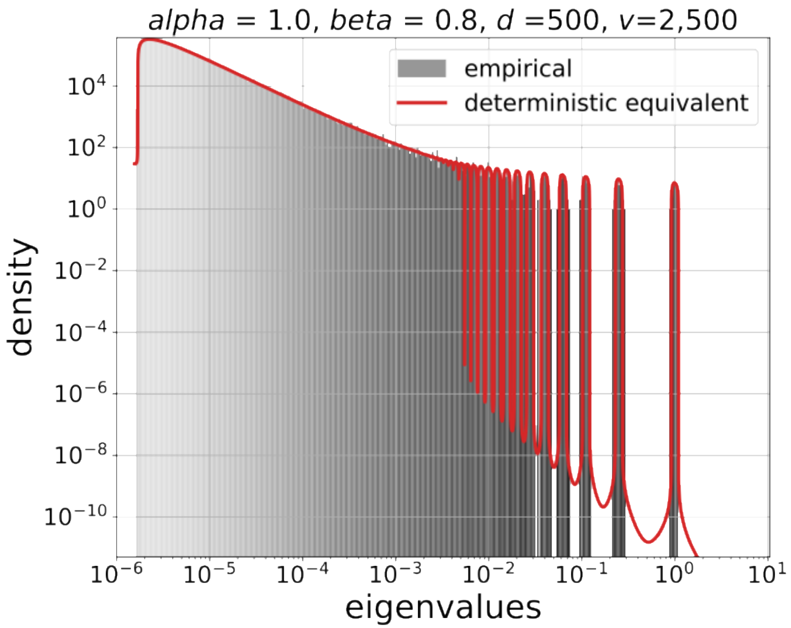
# RELATING $F, \mathcal{K}$ TO THE DETERMINISTIC EQUIVALENT

Two weighted deterministic equivalents $\mu_F, \mu_{\mathcal{K}}$ so that

$$F(k) \approx \int_0^\infty (1 - 2\gamma z + 2\gamma^2 z^2)^k \ \mu_F(dz)$$

$$\mathcal{K}(k) \approx \int_0^\infty \gamma^2 (1 - 2\gamma z + 2\gamma^2 z^2)^k \ \mu_{\mathcal{K}}(dz)$$




$$\mu_F(dz) = \lim_{\epsilon \to 0} \frac{\Im}{\pi} \langle \hat{\beta}, \left( \frac{\Sigma}{\Sigma\, m(z + i\epsilon) - z} \right) \hat{\beta} \rangle$$

$$\mu_{\mathcal{K}}(dz) = z^2 \lim_{\epsilon \to 0} \frac{\Im}{\pi} Tr \left( \frac{1}{\Sigma\, m(z + i\epsilon) - z} \right)$$

## Theorem

*FOR THE DETERMINISTIC EQUIVALENT:*

$$\psi(k) = F(k) + (\mathcal{K} * \psi)(k).$$

*"Kesten's Lemma"*

Suppose $\gamma$ is at most half the convergence threshold.

There is a constant $C = C(\alpha, \beta)$ so that for all $k$

$$F(k) + (\mathcal{K} * F)(k) \leq \psi(k) \leq F(k) + C(\mathcal{K} * F)(k).$$

*Hence it suffices to understand the rates of decay of F, K.*

$$F(k) \asymp_\gamma F_0(k) + F_{pp}(k) + F_{ac}(k)$$
$$\mathcal{K}(k) \asymp_\gamma K_{pp}(k)$$

Phase diagram determined by disappearance of terms.

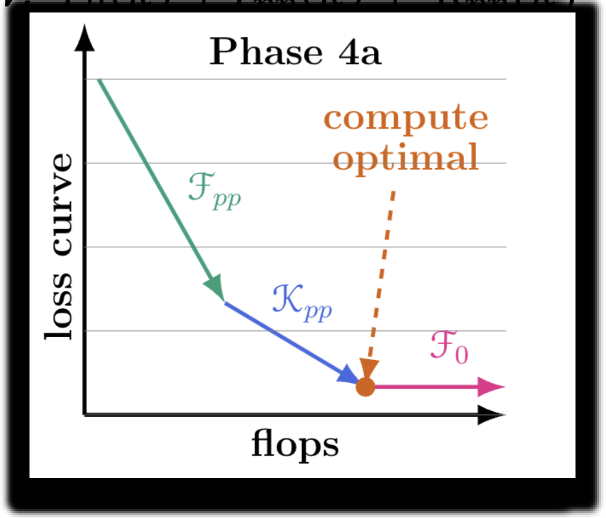
$$\psi(k) \asymp$$

*I.* $F_0(k) + F_{pp}(k)$

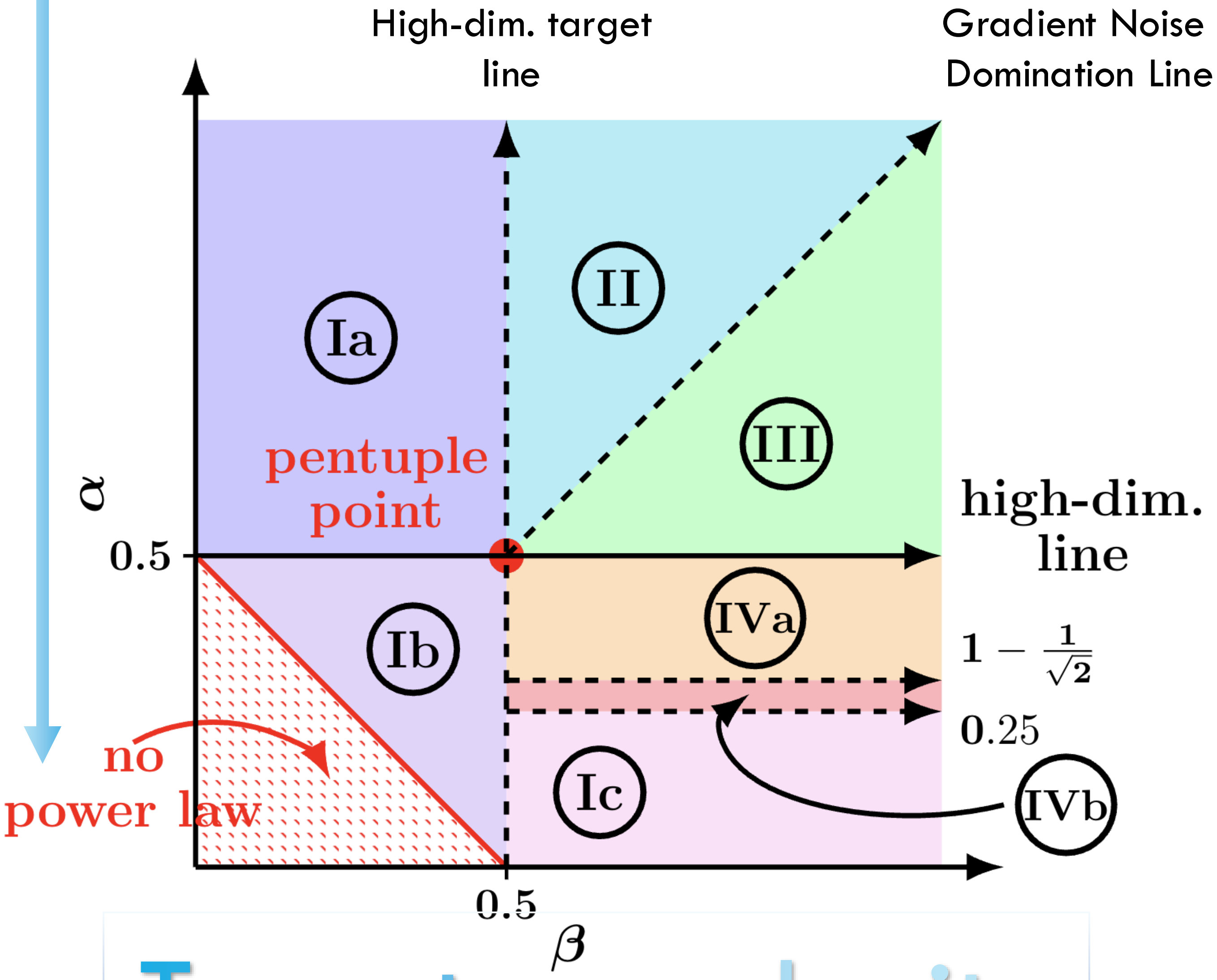
*II.* $F_0(k) + F_{pp}(k) + F_{ac}(k)$

*III.* $F_0(k) + F_{pp}(k) + F_{ac}(k) + K_{pp}(k)$

*IV.* $F_0(k) + F_{pp}(k) + K_{pp}(k)$




Data complexity




Target complexity

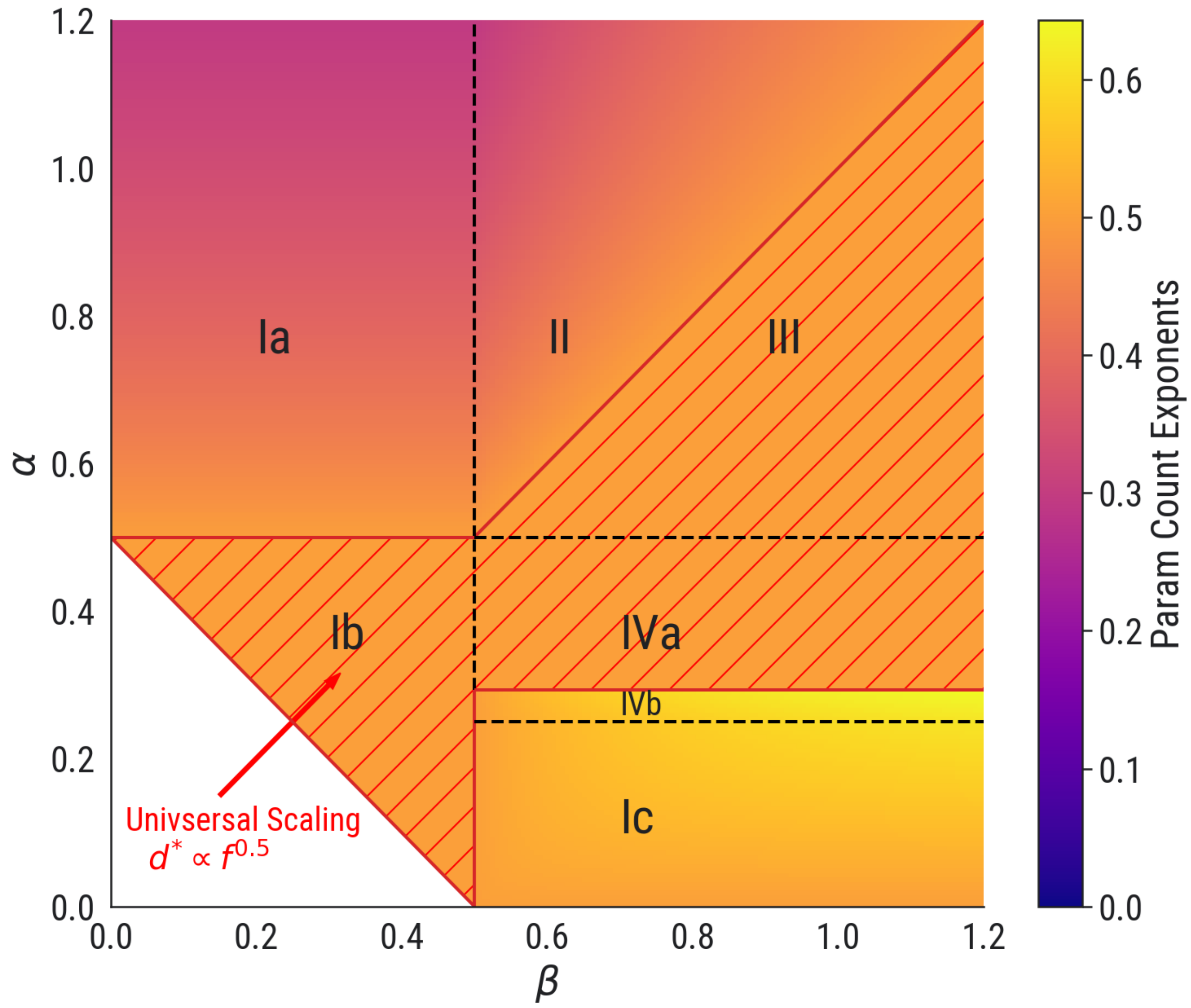
# UNIVERSAL SCALING REGIME

*Compute optimal* $d^*$

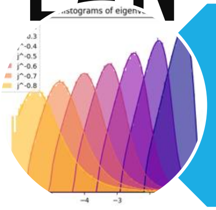
$$\underset{d}{\operatorname{argmin}}\, \psi\left(\frac{f}{d}; d, v, \alpha, \beta\right)$$

$$d^* \propto \sqrt{f} \Leftrightarrow n \propto d$$

Derived empirically for language models in Hoffman et al. '22 (Chinchilla)

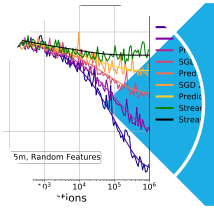


# TALK PLAN



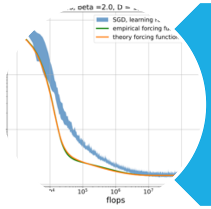
## Part 1: The Powerlaw Random features model

- *Phenomenological model of scaling laws*



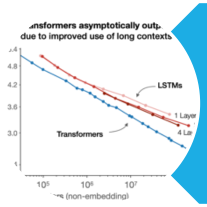

## Part 2: Theory for loss curves (Volterra equations, SDEs, and more)

- *Quadratic models*




## Part 3: Compute optimal scaling laws for streaming SGD on random features

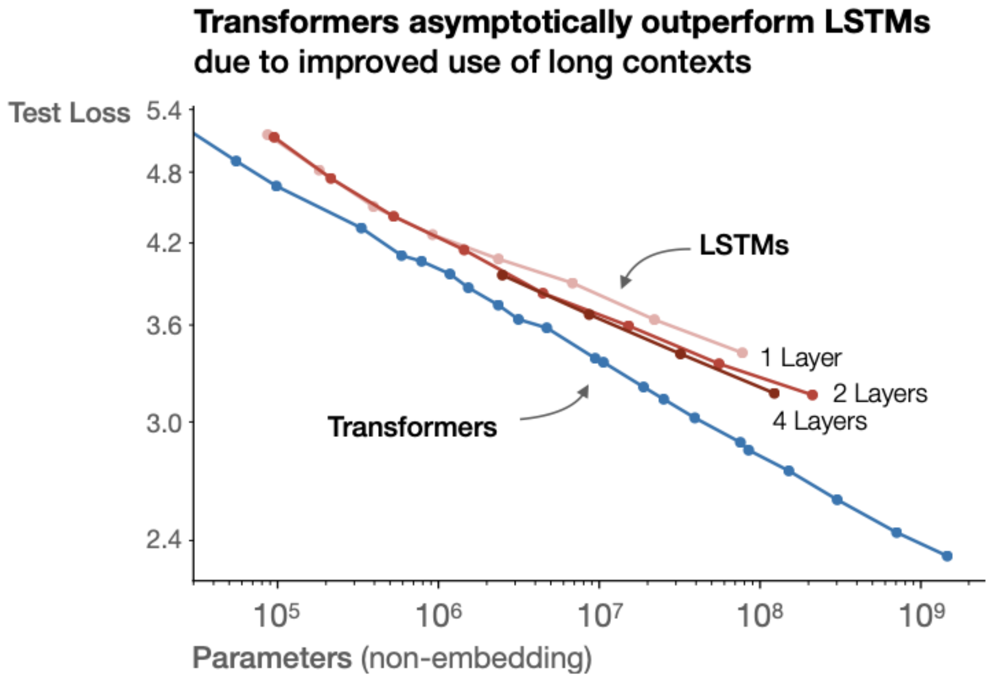
- *In which we can see many different behaviors of SGD*



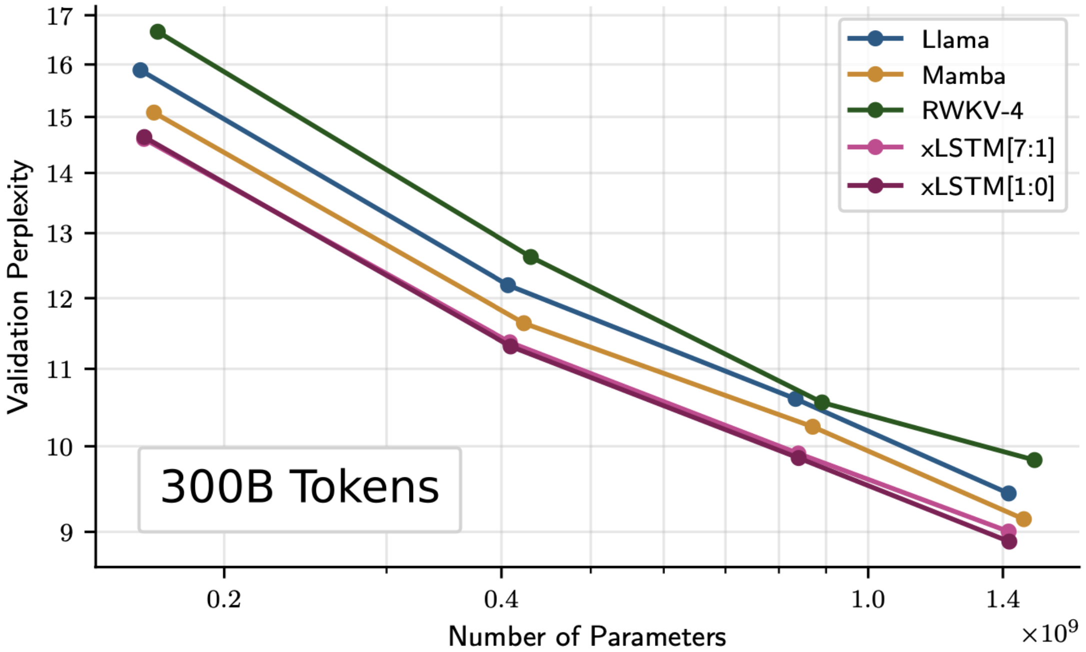

## Part 4: What can change a scaling law?

# DOES THE MODEL MATTER FOR THE SCALING LAW?

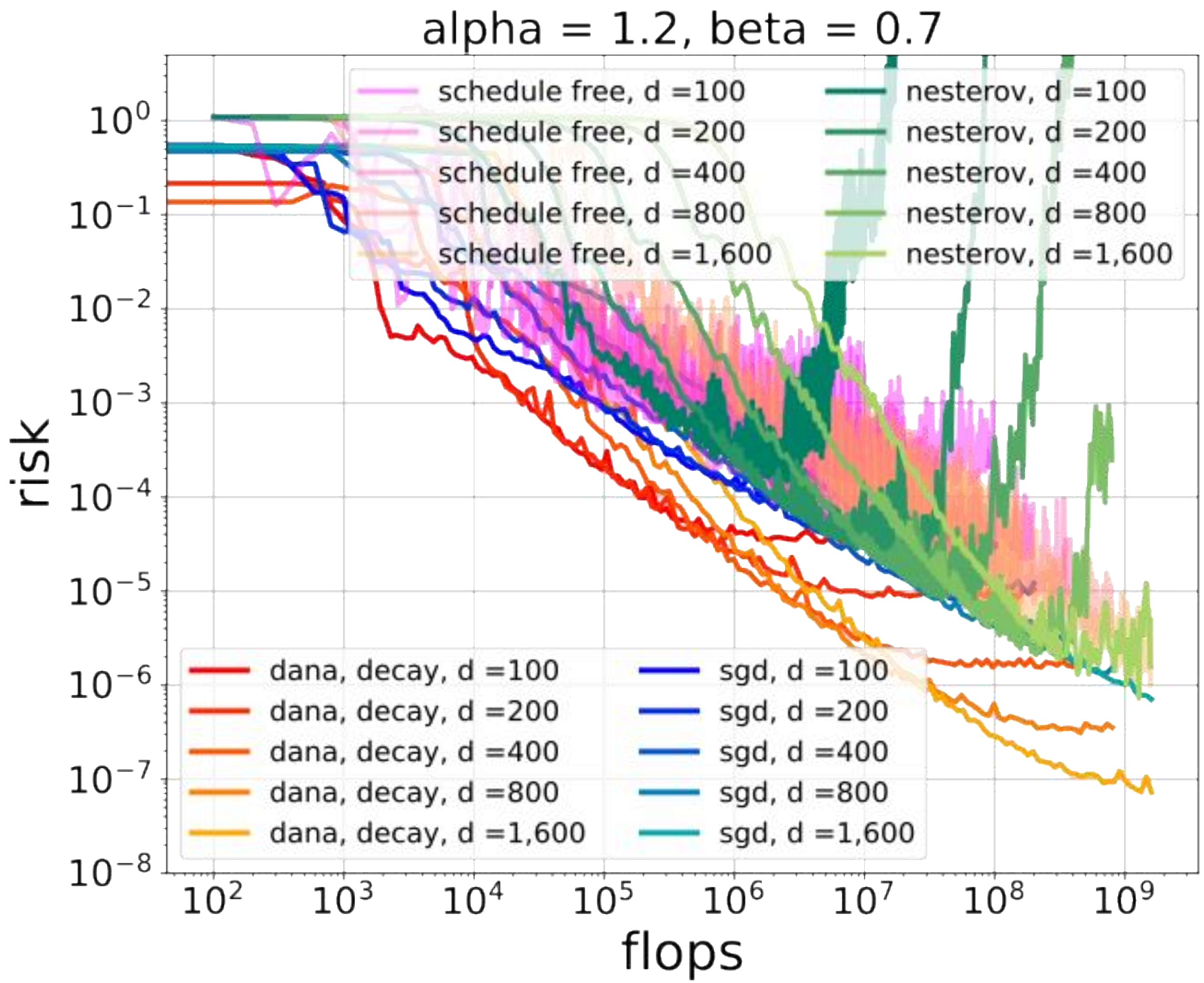
Yes, but less than you might expect..



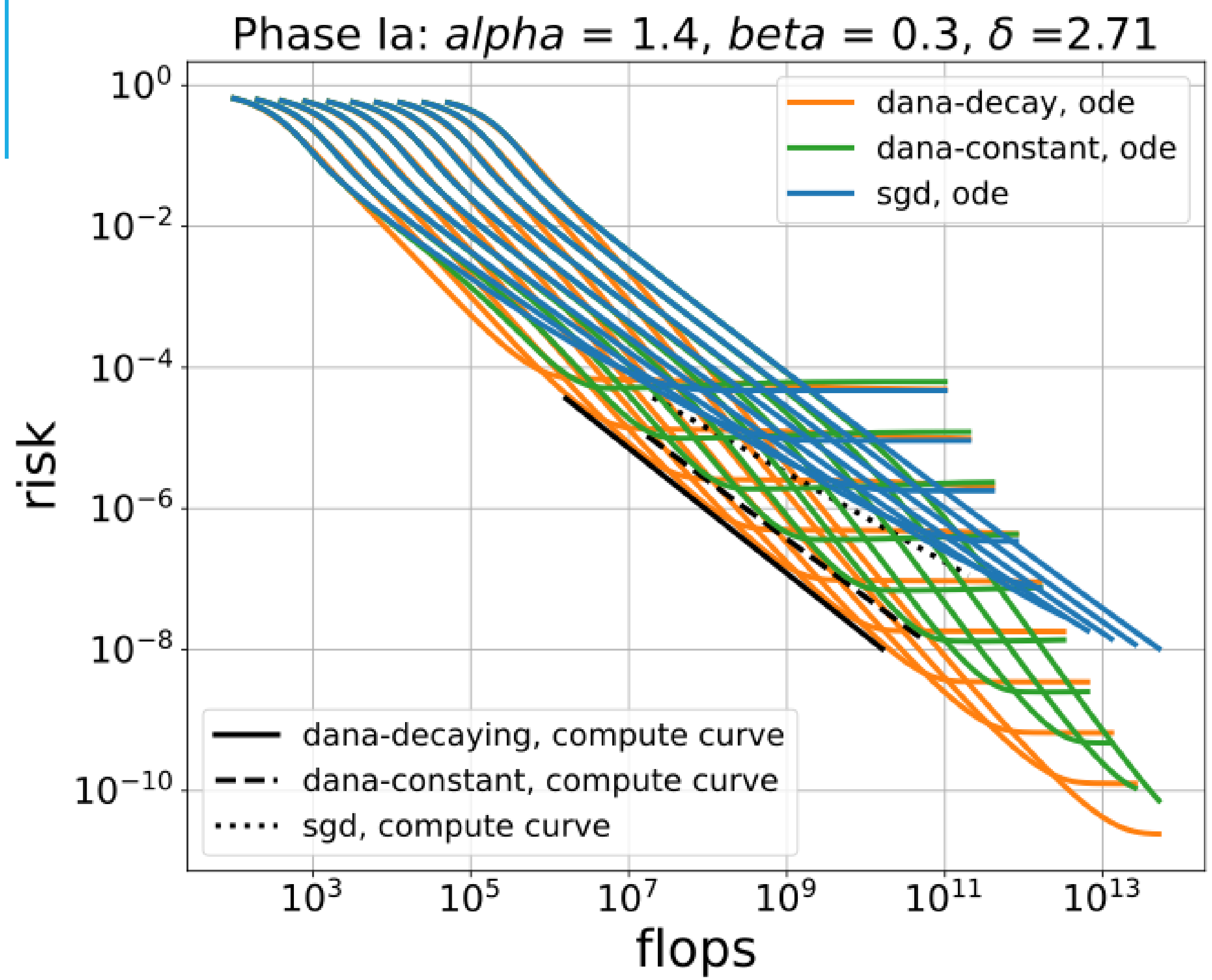

Kaplan et al. 2020




Beck et al. 2024 “xLSTM”




Yes, but all out-of-the-box algorithms are the same..




*Joint with: Ferbach, Everett, Paquette, Gidel. ‘25*

Momentum can change the scaling laws, but only if the hyperparameters are chosen problem-aware.

Stay tuned…

# INTERMISSION