Compute-Optimal Scaling Laws under Low-Rank Feature Map Perturbations

Konstantinos Christopher Tsiolis

Department of Statistical Sciences, University of Toronto

Vector Institute

Wednesday November 20, 2024

Background: Scaling Laws in Large Language Models

- Large language models (LLMs) contain billions of parameters and are trained to predict masked tokens on large text corpora from the Internet [AAA+23, ABA+23].

Background: Scaling Laws in Large Language Models

- Large language models (LLMs) contain billions of parameters and are trained to predict masked tokens on large text corpora from the Internet [AAA+23, ABA+23].
- Recent empirical work has found that their test error follows a power law in each of:
  - d: The number of parameters ("model size");
  - r: The number of stochastic gradient descent (SGD) iterations.

Background: Scaling Laws in Large Language Models

- Large language models (LLMs) contain billions of parameters and are trained to predict masked tokens on large text corpora from the Internet [AAA+23, ABA+23].
- Recent empirical work has found that their test error follows a power law in each of:
  - d: The number of parameters ("model size");
  - r: The number of stochastic gradient descent (SGD) iterations.

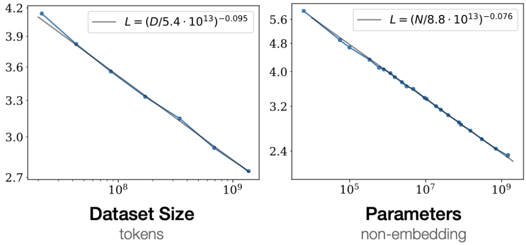


Image source: Kaplan et al. [KMH+20]

Compute-Optimality

- Suppose that the amount of compute to train the model is $C \propto dr$.

Compute-Optimality

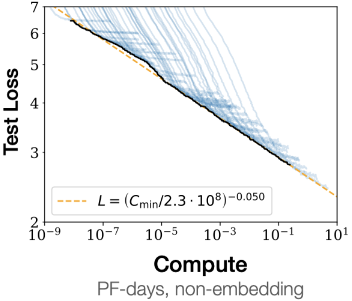
- Suppose that the amount of compute to train the model is $C \propto dr$.
- Key question: **For fixed C, how should we choose d and r?**

Compute-Optimality

- Suppose that the amount of compute to train the model is $C \propto dr$.
- Key question: **For fixed C, how should we choose d and r?**
- [KMH+20, HBM+22] find that the optimal d approximately follows a power law in C, i.e., $d^* \approx sC^\xi$ for constants $s, \xi > 0$.

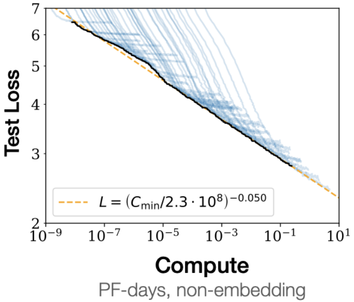
Compute-Optimality

- Suppose that the amount of compute to train the model is $C \propto dr$.
- Key question: **For fixed C, how should we choose d and r?**
- [KMH+20, HBM+22] find that the optimal d approximately follows a power law in C, i.e., $d^* \approx sC^\xi$ for constants $s, \xi > 0$.
  - Moreover, the risk under this choice of d^* follows a power law in C.



Compute-Optimality

- Suppose that the amount of compute to train the model is $C \propto dr$.
- Key question: **For fixed C, how should we choose d and r?**
- [KMH+20, HBM+22] find that the optimal d approximately follows a power law in C, i.e., $d^* \approx sC^{\xi}$ for constants $s, \xi > 0$.
    - Moreover, the risk under this choice of d^* follows a power law in C.



- **Goal**: Rigorously characterize these *compute-optimal scaling laws* in a setting where theoretical analysis is tractable.

A Tractable Setting

- We introduce the setup in the paper "4+3 Phases of Compute-Optimal Neural Scaling Laws" by Paquette et al. [PPXP24].

A Tractable Setting

- We introduce the setup in the paper "4+3 Phases of Compute-Optimal Neural Scaling Laws" by Paquette et al. [PPXP24].
- **Input:** $\mathbf{x} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{D})$, where $\mathbf{D} := \mathrm{diag}(j^{-2\alpha} : j \in [m])$ for some $\alpha > 0$.

A Tractable Setting

- We introduce the setup in the paper "4+3 Phases of Compute-Optimal Neural Scaling Laws" by Paquette et al. [PPXP24].
- **Input:** $\mathbf{x} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{D})$, where $\mathbf{D} := \mathrm{diag}(j^{-2\alpha} : j \in [m])$ for some $\alpha > 0$.
- **Target:** $y = \langle \mathbf{x}, \mathbf{b} \rangle$, where $\mathbf{b} = (b_1, \ldots, b_m) \in \mathbb{R}^m$ is such that $b_j = j^{-\beta}$ for some $\beta \in \mathbb{R}$ satisfying $2\alpha + 2\beta > 1$.

A Tractable Setting

- We introduce the setup in the paper "4+3 Phases of Compute-Optimal Neural Scaling Laws" by Paquette et al. [PPXP24].
- **Input:** $\mathbf{x} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{D})$, where $\mathbf{D} := \mathrm{diag}(j^{-2\alpha} : j \in [m])$ for some $\alpha > 0$.
- **Target:** $y = \langle \mathbf{x}, \mathbf{b} \rangle$, where $\mathbf{b} = (b_1, \ldots, b_m) \in \mathbb{R}^m$ is such that $b_j = j^{-\beta}$ for some $\beta \in \mathbb{R}$ satisfying $2\alpha + 2\beta > 1$.
- **Random features:** We only have access to $\mathbf{W}^T\mathbf{x}$, where $\mathbf{W} \in \mathbb{R}^{m \times d}$ has i.i.d. $\mathcal{N}(0, \frac{1}{d})$ entries, and $d < m$.

A Tractable Setting

- We introduce the setup in the paper "4+3 Phases of Compute-Optimal Neural Scaling Laws" by Paquette et al. [PPXP24].
- **Input:** $\mathbf{x} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{D})$, where $\mathbf{D} := \operatorname{diag}(j^{-2\alpha} : j \in [m])$ for some $\alpha > 0$.
- **Target:** $y = \langle \mathbf{x}, \mathbf{b} \rangle$, where $\mathbf{b} = (b_1, \ldots, b_m) \in \mathbb{R}^m$ is such that $b_j = j^{-\beta}$ for some $\beta \in \mathbb{R}$ satisfying $2\alpha + 2\beta > 1$.
- **Random features:** We only have access to $\mathbf{W}^T \mathbf{x}$, where $\mathbf{W} \in \mathbb{R}^{m \times d}$ has i.i.d. $\mathcal{N}(0, \frac{1}{d})$ entries, and $d < m$.
- **Model:** Fit the target via linear regression with the random features. The risk is then

$$\mathscr{R}_d(\boldsymbol{\theta}) := \mathbb{E}\big[\big(\langle \mathbf{W}^T \mathbf{x}, \boldsymbol{\theta} \rangle - \langle \mathbf{x}, \mathbf{b} \rangle\big)^2 \big| \mathbf{W}\big].$$

The Training Procedure

$$\mathscr{R}_d(\boldsymbol{\theta}) := \mathbb{E}\Big[\big(\langle \mathbf{W}^T\mathbf{x}, \boldsymbol{\theta}\rangle - \langle \mathbf{x}, \mathbf{b}\rangle\big)^2 \Big| \mathbf{W}\Big].$$

- The parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$ is fit by SGD with r iterations. At each iteration t, a fresh batch $\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(B)} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{D})$ is drawn.
    - The labels $y_t^{(i)} = \langle \mathbf{x}_t^{(i)}, \mathbf{b}\rangle$ are assumed noiseless.

The Training Procedure

$$\mathscr{R}_d(\boldsymbol{\theta}) := \mathbb{E}\big[\big(\langle \mathbf{W}^T\mathbf{x}, \boldsymbol{\theta}\rangle - \langle \mathbf{x}, \mathbf{b}\rangle\big)^2 \big| \mathbf{W}\big].$$

- The parameter vector $\boldsymbol{\theta} \in \mathbb{R}^d$ is fit by SGD with r iterations. At each iteration t, a fresh batch $\mathbf{x}_t^{(1)}, \ldots, \mathbf{x}_t^{(B)} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{D})$ is drawn.
  - The labels $y_t^{(i)} = \langle \mathbf{x}_t^{(i)}, \mathbf{b}\rangle$ are assumed noiseless.
- Assuming a step size γ such that $\gamma B < 1$, the SGD updates take the form

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t - \gamma \sum_{i=1}^{B} \mathbf{W}^T\mathbf{x}_t^{(i)}\big(\langle \mathbf{W}^T\mathbf{x}_t^{(i)}, \boldsymbol{\theta}_t\rangle - y_t^{(i)}\big).$$

Risk Dynamics under SGD

- How does the risk $\mathscr{R}$ evolve with d, r?

Risk Dynamics under SGD

- How does the risk $\mathscr{R}$ evolve with d, r?
- [PPXP24] express the risk dynamics as a recurrence that depends on the eigenvalues $\{\lambda_j\}_{j=1}^m$ and eigenvectors $\{\mathbf{u}_j\}_{j=1}^m$ of the random matrix $\hat{\mathbf{K}}_0 = (\mathbf{D}^{1/2}\mathbf{W})(\mathbf{D}^{1/2}\mathbf{W})^T$:

$$\mathscr{R}_d(\boldsymbol{\theta}_r) = \sum_{j=1}^m \left(\mathbf{a}^T \mathbf{u}_j \mathbf{u}_j^T \mathbf{a}\right)\left(1 - 2\gamma B \lambda_j + 2\gamma^2 B^2 \lambda_j^2\right)^r + \sum_{j=1}^d \gamma^2 B \lambda_j^2 \sum_{s=0}^{r-1} \left(1 - 2\gamma B \lambda_j + 2\gamma^2 B^2 \lambda_j^2\right)^{r-1-s} \mathscr{R}(\boldsymbol{\theta}_s),$$

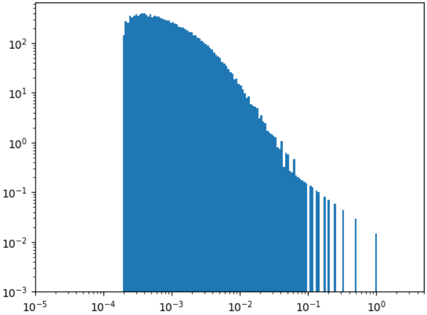
  where $\mathbf{a} = \mathbf{D}^{1/2}\mathbf{b}$.
- $\hat{\mathbf{K}}_0$ has the natural interpretation as a sample covariance matrix for d i.i.d. draws from $\mathcal{N}_m(\mathbf{0}, \mathbf{D})$.

Spectrum of $\hat{\mathbf{K}}_0$

- Recall

$$\mathbf{D} = \mathrm{diag}(j^{-2\alpha} : j \in [m]) \qquad \mathbf{W}_{ij} \sim \mathcal{N}(0, \tfrac{1}{d}).$$

- The spectrum of $\hat{\mathbf{K}}_0 = (\mathbf{D}^{1/2}\mathbf{W})(\mathbf{D}^{1/2}\mathbf{W})^T$ has three components:
  - Point mass at zero: Since $\hat{\mathbf{K}}_0 \in \mathbb{R}^{m\times m}$ has rank at most $d < m$.
  - Bulk: The collection of the smallest nonzero eigenvalues.
  - Spikes: The largest eigenvalues.



Enter Random Matrix Theory

- We can express functions of the spectrum of $\hat{\mathbf{K}}_0$ as contour integrals involving its **resolvent** $(\hat{\mathbf{K}}_0 - z\mathbf{I}_m)^{-1}$ for $z \in \mathbb{C} \setminus \mathbb{R}_0^+$.
- Let Γ be any contour enclosing the eigenvalues of $\hat{\mathbf{K}}_0$. [PPXP24] write

$$\mathscr{R}_d(r) = \mathscr{F}(r) + (\mathscr{K} * \mathscr{R}_d)(r),$$

where

$$\mathscr{F}(r) := -\frac{1}{2\pi i} \oint_\Gamma \mathbf{a}^T (\hat{\mathbf{K}}_0 - z)^{-1} \mathbf{a} \, (1 - 2\gamma B z + 2\gamma^2 B^2 z^2)^r \, dz$$
$$\mathscr{K}(r) := -\frac{\gamma^2 B}{2\pi i} \oint_\Gamma \operatorname{tr} (\hat{\mathbf{K}}_0 - z)^{-1} z^2 (1 - 2\gamma B z + 2\gamma^2 B^2 z^2)^r \, dz.$$

Deterministic Equivalent

- Analysis of the dynamics is rendered tractable by replacing $(\hat{\mathbf{K}}_0 - z)^{-1}$ with a **deterministic equivalent** $\mathbf{R}_0(z)$ satisfying

$$\frac{1}{m}\left| \operatorname{tr}\left(\mathbf{R}_0(z) - (\hat{\mathbf{K}}_0 - z)^{-1} \right) \right| \xrightarrow{\mathbb{P}} 0, \qquad \left| \mathbf{a}^T \left(\mathbf{R}(z) - (\hat{\mathbf{K}}_0 - z)^{-1} \right) \mathbf{a} \right| \xrightarrow{\mathbb{P}} 0$$

as $m, d \to \infty$ such that $m/d \to c > 1$.

Deterministic Equivalent

- Analysis of the dynamics is rendered tractable by replacing $(\hat{\mathbf{K}}_0 - z)^{-1}$ with a **deterministic equivalent** $\mathbf{R}_0(z)$ satisfying

$$\frac{1}{m}\left|\operatorname{tr}\left(\mathbf{R}_0(z) - (\hat{\mathbf{K}}_0 - z)^{-1}\right)\right| \xrightarrow{\mathbb{P}} 0, \qquad \left|\mathbf{a}^T\left(\mathbf{R}(z) - (\hat{\mathbf{K}}_0 - z)^{-1}\right)\mathbf{a}\right| \xrightarrow{\mathbb{P}} 0$$

  as $m, d \to \infty$ such that $m/d \to c > 1$.
- Here,

$$\mathbf{R}_0(z) = \operatorname{diag}((j^{-2\alpha} q(z) - z)^{-1} : j \in [m]),$$

Deterministic Equivalent

- Analysis of the dynamics is rendered tractable by replacing $(\hat{\mathbf{K}}_0 - z)^{-1}$ with a **deterministic equivalent** $\mathbf{R}_0(z)$ satisfying

$$\frac{1}{m}\left|\operatorname{tr}\left(\mathbf{R}_0(z)-(\hat{\mathbf{K}}_0-z)^{-1}\right)\right| \xrightarrow{\mathbb{P}} 0, \qquad \left|\mathbf{a}^T\left(\mathbf{R}(z)-(\hat{\mathbf{K}}_0-z)^{-1}\right)\mathbf{a}\right| \xrightarrow{\mathbb{P}} 0$$

  as $m, d \to \infty$ such that $m/d \to c > 1$.
- Here,

$$\mathbf{R}_0(z) = \operatorname{diag}((j^{-2\alpha}q(z) - z)^{-1} : j \in [m]),$$

  where $q(z)$ is defined implicitly by the equation

$$q(z) := \frac{1}{1 + \frac{1}{d}\sum_{j=1}^{v} \frac{j^{-2\alpha}}{j^{-2\alpha}q(z)-z}}.$$

Deterministic "Approximating" Equation

- We now have a deterministic *convolution-type Volterra equation*

$$\mathcal{R}(r) = \mathcal{F}(r) + \mathcal{K}(r) * \mathcal{R}(r).$$

where

$$\mathcal{F}(r) := -\frac{1}{2\pi i} \oint_{\Gamma} \mathbf{a}^T \mathbf{R}_0(z) \mathbf{a} \, (1 - 2\gamma B z + 2\gamma^2 B^2 z^2)^r \, dz,$$

$$\mathcal{K}(r) := -\frac{\gamma^2 B}{2\pi i} \oint_{\Gamma} \operatorname{tr} \mathbf{R}_0(z) z^2 (1 - 2\gamma B z + 2\gamma^2 B z^2)^r \, dz.$$

A Major Limitation

- Paquette et al. do not show that $\mathcal{R}(r)$, the solution to the deterministic Volterra equation, is in fact "close" to the stochastic dynamics $\mathscr{R}(\boldsymbol{\theta}_r)$.

A Major Limitation

- Paquette et al. do not show that $\mathcal{R}(r)$, the solution to the deterministic Volterra equation, is in fact "close" to the stochastic dynamics $\mathscr{R}(\boldsymbol{\theta}_r)$.

Conjecture [PPXP24]

For $\{\theta_r\}$ the sequence of iterates generated by SGD with $\theta_0 = 0$ and any $\varepsilon > 0$,

$$(1-\varepsilon) \leq \sup_{r \in \mathbb{N}} \left\{ \frac{\mathscr{R}(\theta_r)}{\mathcal{R}(r)} \right\} \leq (1+\varepsilon)$$

with high probability (which tends to 1 as $d \to \infty$).

Decomposing the Volterra Equation

$$\mathcal{R}(r) = \mathcal{F}(r) + \mathcal{K}(r) * \mathcal{R}(r).$$

- With an appropriate choice of contour Γ enclosing the spectrum of $\hat{\mathbf{K}}_0$, [PPXP24] show that, asymptotically,

$$\mathcal{R}(r) \asymp \mathcal{F}_0(r) + \mathcal{F}_{ac}(r) + \mathcal{F}_{pp}(r) + \frac{1}{\gamma B}\mathcal{K}_{pp}(r).$$

Decomposing the Volterra Equation

$$\mathcal{R}(r) = \mathcal{F}(r) + \mathcal{K}(r) * \mathcal{R}(r).$$

- With an appropriate choice of contour Γ enclosing the spectrum of $\hat{\mathbf{K}}_0$, [PPXP24] show that, asymptotically,

$$\mathcal{R}(r) \asymp \mathcal{F}_0(r) + \mathcal{F}_{ac}(r) + \mathcal{F}_{pp}(r) + \frac{1}{\gamma B}\mathcal{K}_{pp}(r).$$

- $\mathcal{F}_0$: Approximation error arising because the input space has a subspace of dimension $(m-d) > 0$ in $\ker(\mathbf{W}^T)$;

Decomposing the Volterra Equation

$$\mathcal{R}(r) = \mathcal{F}(r) + \mathcal{K}(r) * \mathcal{R}(r).$$

- With an appropriate choice of contour Γ enclosing the spectrum of $\hat{\mathbf{K}}_0$, [PPXP24] show that, asymptotically,

$$\mathcal{R}(r) \asymp \mathcal{F}_0(r) + \mathcal{F}_{ac}(r) + \mathcal{F}_{pp}(r) + \frac{1}{\gamma B}\mathcal{K}_{pp}(r).$$

- $\mathcal{F}_0$: Approximation error arising because the input space has a subspace of dimension $(m-d) > 0$ in $\ker(\mathbf{W}^T)$;
- $\mathcal{F}_{ac}$: Error arising from the spectral bulk;

Decomposing the Volterra Equation

$$\mathcal{R}(r) = \mathcal{F}(r) + \mathcal{K}(r) * \mathcal{R}(r).$$

- With an appropriate choice of contour Γ enclosing the spectrum of $\hat{\mathbf{K}}_0$, [PPXP24] show that, asymptotically,

$$\mathcal{R}(r) \asymp \mathcal{F}_0(r) + \mathcal{F}_{ac}(r) + \mathcal{F}_{pp}(r) + \frac{1}{\gamma B}\mathcal{K}_{pp}(r).$$

- $\mathcal{F}_0$: Approximation error arising because the input space has a subspace of dimension $(m-d) > 0$ in $\ker(\mathbf{W}^T)$;
- $\mathcal{F}_{ac}$: Error arising from the spectral bulk;
- $\mathcal{F}_{pp}$: Error arising from the spike eigenvalues;

Decomposing the Volterra Equation

$$\mathcal{R}(r) = \mathcal{F}(r) + \mathcal{K}(r) * \mathcal{R}(r).$$

- With an appropriate choice of contour Γ enclosing the spectrum of $\hat{\mathbf{K}}_0$, [PPXP24] show that, asymptotically,

$$\mathcal{R}(r) \asymp \mathcal{F}_0(r) + \mathcal{F}_{ac}(r) + \mathcal{F}_{pp}(r) + \frac{1}{\gamma B}\mathcal{K}_{pp}(r).$$

- $\mathcal{F}_0$: Approximation error arising because the input space has a subspace of dimension $(m-d) > 0$ in $\ker(\mathbf{W}^T)$;
- $\mathcal{F}_{ac}$: Error arising from the spectral bulk;
- $\mathcal{F}_{pp}$: Error arising from the spike eigenvalues;
- $\mathcal{K}_{pp}$: Error due to SGD noise — under full-batch gradient descent, this term is of strictly lower order.

Asymptotics of $\mathcal{R}$

$$\mathcal{R}(r) \asymp \mathcal{F}_0(r) + \mathcal{F}_{ac}(r) + \mathcal{F}_{pp}(r) + \frac{1}{\gamma B}\mathcal{K}_{pp}(r).$$

- Recall

$$\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{D}) \qquad \mathbf{D} := \mathrm{diag}(j^{-2\alpha} : j \in [m])$$
$$y = \langle \mathbf{x}, \mathbf{b} \rangle \qquad b_j = j^{-\beta} \quad \forall j \in [m].$$

Function	$^*\Gamma(x)$ is the Gamma function
$\mathcal{F}_0(r) \asymp d^{-2\alpha+\max\{0,1-2\beta\}}$	
$\mathcal{F}_{pp}(r) \sim (2\alpha)^{-1} \times \Gamma\big(\frac{\beta}{\alpha} - \frac{1}{2\alpha} + 1\big) \times (2\gamma B \times r)^{-(1+\beta/\alpha)+1/(2\alpha)}$	
$\mathcal{F}_{ac}(r) \leq \begin{cases} C \times \mathcal{F}_0(r), & \text{if } 2\beta > 1,\ 2\alpha < 1 \\ 0, & \text{if } 2\beta < 1 \end{cases}$ for $C > 0$, independent of d If $2\beta > 1, 2\alpha > 1$, $\mathcal{F}_{ac}(r) \sim \big(\sum_{j=1}^{V} j^{-2\beta}\big)(2\alpha)^{-1}\Gamma\big(1 - \frac{1}{2\alpha}\big) \times (2\gamma B \times r)^{-1+1/(2\alpha)} \times d^{-1}$	
$\mathcal{K}_{pp}(r) \sim (2\alpha)^{-1} \times \Gamma\big(2 - \frac{1}{2\alpha}\big) \times (2\gamma B \times r)^{-2+1/(2\alpha)}$	

Image source: [PPXP24]

4(+3) Phases of Compute-Optimal Scaling Laws

- The scaling laws are partitioned into four "phases" in the (α, β) plane depending on the dominant components of $\mathcal{R}$.

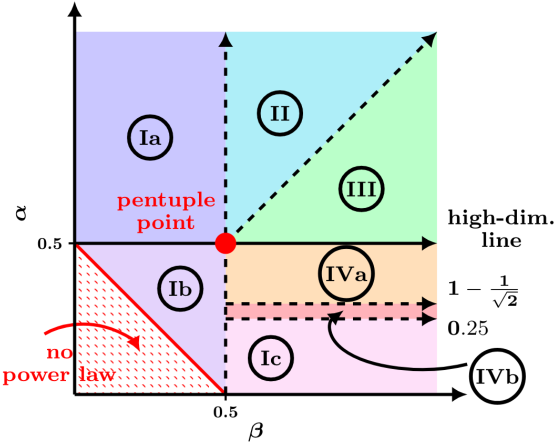


Image source: [PPXP24]

4(+3) Phases of Compute-Optimal Scaling Laws

	Asymptotic $\mathcal{R}(r)$	Trade off	Compute-optimal Curves
Phase I	$\mathcal{F}_{pp}(r) + \mathcal{F}_0(r)$	$\mathcal{F}_{pp} = \mathcal{F}_0$	**Ia** $\mathcal{R}^\star \asymp C^{\left(\frac{1}{2\alpha+1}-1\right)(1+\beta/\alpha-1/(2\alpha))}$ $d^\star \asymp C^{1/(2\alpha+1)}$
			Ib $\mathcal{R}^\star \asymp C^{\frac{1}{2}-\alpha-\beta}$ $d^\star \asymp C^{\frac{1}{2}}$
			Ic $\mathcal{R}^\star \asymp C^{\frac{\alpha(2\alpha+2\beta-1)}{\alpha(2\beta-3)-2\beta+1}}$ $d^\star \asymp C^{\frac{1-2(\alpha+\beta)}{2(\alpha(2\beta-3)-2\beta+1)}}$
Phase II	$\mathcal{F}_{pp}(r) + \mathcal{F}_{ac}(r)$ $+\mathcal{F}_0(r)$	$\mathcal{F}_{pp} = \mathcal{F}_{ac}$	$\mathcal{R}^\star \asymp C^{-\frac{2\alpha+2\beta-1}{2(\alpha+\beta)}}$ $d^\star \asymp C^{(\beta/\alpha)/(1+\beta/\alpha)}$
Phase III	$\mathcal{F}_{ac}(r) + \mathcal{F}_0(r)$ $+\frac{1}{\gamma B}\mathcal{K}_{pp}(r)$	$\frac{1}{\gamma B}\mathcal{K}_{pp} = \mathcal{F}_{ac}$	$\mathcal{R}^\star \asymp C^{(1-4\alpha)/(4\alpha)}$ $d^\star \asymp C^{1/2}$
Phase IV	$\mathcal{F}_{pp}(r) + \mathcal{F}_0(r)$ $+\frac{1}{\gamma B}\mathcal{K}_{pp}(r)$	**IVa** $\frac{1}{\gamma B}\mathcal{K}_{pp} = \mathcal{F}_0$	$\mathcal{R}^\star \asymp C^{-\alpha}$ $d^\star \asymp C^{1/2}$
		IVb $\frac{1}{\gamma B}\mathcal{K}_{pp} = \mathcal{F}_{pp}$	$\mathcal{R}^\star \asymp C^{\frac{(1-2\alpha)(2\alpha+2\beta-1)}{(2(2\alpha\beta+\alpha-2\beta))}}$ $d^\star \asymp C^{(\alpha-\beta)/(2\alpha\beta+\alpha-2\beta)}$

Limitations of [PPXP24]

- We highlight three important limitations of the paper:

Limitations of [PPXP24]

- We highlight three important limitations of the paper:
  1. The deterministic Volterra equation is not rigorously proven to approximate the true stochastic risk dynamics;

Limitations of [PPXP24]

- We highlight three important limitations of the paper:
  (1) The deterministic Volterra equation is not rigorously proven to approximate the true stochastic risk dynamics;
  (2) Both the model and target are linear;

Limitations of [PPXP24]

- We highlight three important limitations of the paper:
  (1) The deterministic Volterra equation is not rigorously proven to approximate the true stochastic risk dynamics;
  (2) Both the model and target are linear;
  (3) The feature map $\mathbf{W}^T$ is fixed and random instead of being learnable by SGD.

Limitations of [PPXP24]

- We highlight three important limitations of the paper:
  (1) The deterministic Volterra equation is not rigorously proven to approximate the true stochastic risk dynamics;
  (2) Both the model and target are linear;
  (3) The feature map $\mathbf{W}^T$ is fixed and random instead of being learnable by SGD.
- Our proposed extension outlines a potential step towards addressing the third limitation.

Feature Learning Step as a Rank-1 Update

Theorem ([BES+22] Theorem 3, Informal)

Consider a linear random features regression problem with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, $y = \mu\langle \mathbf{x}, \tilde{\mathbf{b}}\rangle$ *such that* $||\tilde{\mathbf{b}}||_2 = 1$ *and* $\mu > 0$.
Let $\mathbf{W}_1$ *denote the result of a single full-batch gradient descent step on a training set of size* n *with* $\gamma = \Theta(1)$. *Then* $\mathbf{W}_1\mathbf{W}_1^T$ *has a spike eigenvalue* λ_1 *and associated eigenvector* $\mathbf{v}_1$ *such that, in the limit* $n, d, m \to \infty$ *at a proportional rate,*

$$\lambda_1 \to \Theta(\mu) \qquad 1 - |\langle \mathbf{v}_1, \tilde{\mathbf{b}}\rangle|^2 \to \Theta(\mu^{-2}).$$

Feature Learning Step as a Rank-1 Update

Theorem ([BES+22] Theorem 3, Informal)

Consider a linear random features regression problem with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_m)$, $y = \mu\langle \mathbf{x}, \tilde{\mathbf{b}} \rangle$ *such that* $||\tilde{\mathbf{b}}||_2 = 1$ *and* $\mu > 0$.
Let $\mathbf{W}_1$ *denote the result of a single full-batch gradient descent step on a training set of size* n *with* $\gamma = \Theta(1)$. *Then* $\mathbf{W}_1\mathbf{W}_1^T$ *has a spike eigenvalue* λ_1 *and associated eigenvector* $\mathbf{v}_1$ *such that, in the limit* $n, d, m \to \infty$ *at a proportional rate,*

$$\lambda_1 \to \Theta(\mu) \qquad 1 - |\langle \mathbf{v}_1, \tilde{\mathbf{b}} \rangle|^2 \to \Theta(\mu^{-2}).$$

- Back in our setting with $\mathbf{x} \sim \mathcal{N}_m(\mathbf{0}, \mathbf{D})$ and $y = \langle \mathbf{x}, \mathbf{b} \rangle$, this motivates us to consider a spiked random features model where

$$\tilde{\mathbf{W}} = \tau \mathbf{b} \mathbf{1}^T + \mathbf{Z}, \qquad \mathbf{Z}_{ij} \sim \mathcal{N}(0, \tfrac{1}{d}), \tau = \Theta(\tfrac{1}{\sqrt{d}}).$$

Deterministic Equivalent for Spiked Power-Law Model

- To use the ideas from [PPXP24], we require a new deterministic equivalent $\mathbf{R}(z)$ for $(\mathbf{D}^{1/2}\tilde{\mathbf{W}}\tilde{\mathbf{W}}^T\mathbf{D}^{1/2} - z)^{-1}$.

Deterministic Equivalent for Spiked Power-Law Model

- To use the ideas from [PPXP24], we require a new deterministic equivalent $\mathbf{R}(z)$ for $(\mathbf{D}^{1/2}\tilde{\mathbf{W}}\tilde{\mathbf{W}}^T\mathbf{D}^{1/2} - z)^{-1}$.
- We derive, using [HLN07]:

$$\mathbf{R}(z) = -\frac{1}{z}\left(\mathbf{I}_m + \tfrac{s}{d}\mathbf{D}\right)^{-1} - \frac{\frac{\tau^2 d}{z}(1+\frac{t}{d})^{-1}\left(\mathbf{I}_m + \frac{s}{d}\mathbf{D}\right)^{-1}\mathbf{a}\mathbf{a}^T\left(\mathbf{I}_m + \frac{s}{d}\mathbf{D}\right)^{-1}}{1 - \frac{\tau^2 d}{z}(1+\frac{t}{d})^{-1}\mathbf{a}^T(\mathbf{I}_m + \frac{s}{d}\mathbf{D})^{-1}\mathbf{a}},$$

where

$$s + \frac{d}{z(1+\frac{t}{d})} + \frac{\frac{\tau^2 d}{z}(1+\frac{t}{d})^{-2}\sum_{j=1}^m \frac{j^{-2\alpha-2\beta}}{1+\frac{s}{d}j^{-2\alpha}}}{1 - \frac{\tau^2 d}{z}(1+\frac{t}{d})^{-1}\sum_{j=1}^m \frac{j^{-2\alpha-2\beta}}{1+\frac{s}{d}j^{-2\alpha}}} = 0$$

$$t + \frac{1}{z}\sum_{j=1}^m \frac{j^{-2\alpha}}{1+\frac{s}{d}j^{-2\alpha}} + \frac{\frac{\tau^2 d}{z}(1+\frac{t}{d})^{-1}\sum_{j=1}^m \frac{j^{-4\alpha-2\beta}}{(1+\frac{s}{d}j^{-2\alpha})^2}}{1 - \frac{\tau^2 d}{z}(1+\frac{t}{d})^{-1}\sum_{j=1}^m \frac{j^{-2\alpha-2\beta}}{1+\frac{s}{d}j^{-2\alpha}}} = 0.$$

Deterministic Equivalent for Spiked Power-Law Model

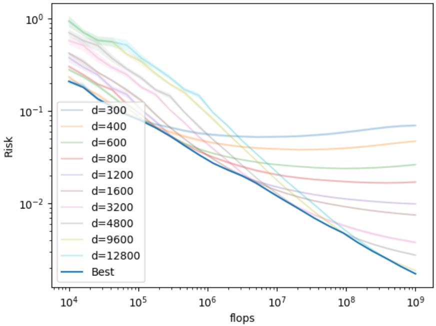
- The major drawback of the above deterministic equivalent is that it is not an explicit function of $q(z)$.
- As a result, the technical estimates from [PPXP24] that lead to bounds on $\mathcal{F}$ and $\mathcal{K}$ do not immediately carry over.

Deterministic Equivalent for Spiked Power-Law Model

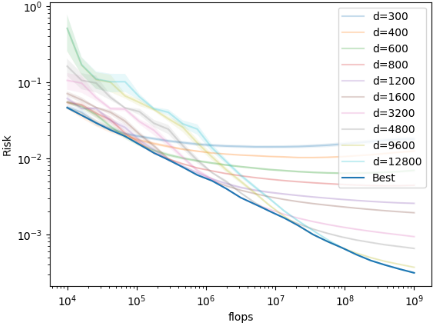
- The major drawback of the above deterministic equivalent is that it is not an explicit function of $q(z)$.
- As a result, the technical estimates from [PPXP24] that lead to bounds on $\mathcal{F}$ and $\mathcal{K}$ do not immediately carry over.
- **Whether a more tractable deterministic equivalent can be found remains an open problem.**

Simulation Study

- The case $(\alpha, \beta) = (0.5, 0.7)$
  - On boundary of Phases III and IVa
  - Theory (for $\tau = 0$): $\mathcal{R}^* \asymp C^{-1/2}$



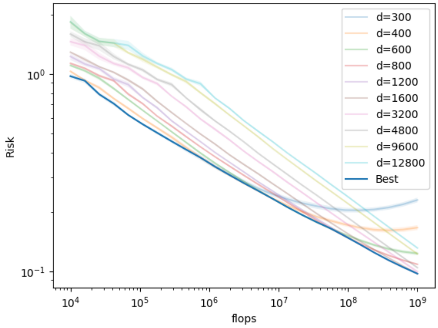
(a) $\tau = 0$
$\hat{\mathscr{R}}(C) = 9.862 \cdot C^{-0.4175}$



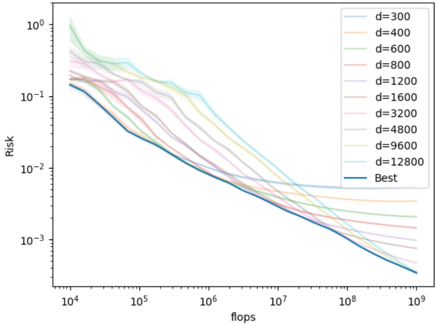
(b) $\tau = 1/\sqrt{d}$
$\hat{\mathscr{R}}(C) = 2.756 \cdot C^{-0.4503}$

Simulation Study

- The case $(\alpha, \beta) = (0.6, 0.2)$
  - Phase Ia
  - Theory (for $\tau = 0$): $\mathcal{R}^* \asymp C^{-0.2727}$



(a) $\tau = 0$
$\hat{\mathscr{R}}(C) = 1.761 \cdot C^{-0.2002}$



(b) $\tau = 1/\sqrt{d}$
$\hat{\mathscr{R}}(C) = 2.361 \cdot C^{-0.5058}$

References I

[AAA+23] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[ABA+23] Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[BES+22] Jimmy Ba, Murat A Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, and Greg Yang. High-dimensional asymptotics of feature learning: How one gradient step improves the representation. *Advances in Neural Information Processing Systems*, 35:37932–37946, 2022.

References II

[HBM+22] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pages 30016–30030, 2022.

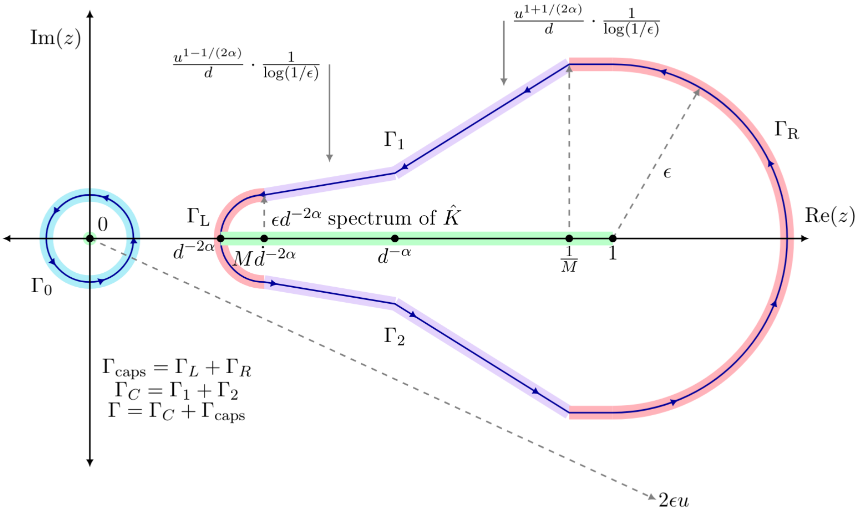
[HLN07] Walid Hachem, Philippe Loubaton, and J Najim. Deterministic equivalents for certain functionals of large random matrices. *Annals of Applied Probability*, 17(1):875–930, 2007.

References III

[KMH+20] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[PPXP24] Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4+ 3 phases of compute-optimal neural scaling laws. *arXiv preprint arXiv:2405.15074*, 2024.

The Choice of Contour Γ



Along Γ_0

Proposition (Behaviour of $q(z)$ near zero, [PPXP24] Proposition 7.3)

The function $q(z)$ is analytic in a neighbourhood of $z = 0$ of radius $c(\alpha)d^{-2\alpha}$ for some $c(\alpha) > 0$. Furthermore, q is negative on $(0, cd^{-2\alpha})$, vanishes at 0, and has $|m'(0) + \kappa(m/d)d^{2\alpha}| \leq Cd^{2\alpha-1}$ for all d sufficiently large, where $\kappa(m/d)$ solves

$$\int_0^{m/d} \frac{\kappa}{\kappa + x^{2\alpha}}\, dx = 1. \tag{1}$$

Proposition (Contribution along Γ_0, [PPXP24] Proposition 8.1)

The function $\mathcal{F}_0(r)$ is constant and

$$\left| \mathcal{F}_0(0) - \sum_{j=1}^{m} \frac{j^{-2\alpha-2\beta}}{1 + j^{-2\alpha} d^{2\alpha} \kappa(m/d)} \right| \lesssim Cd^{-2\alpha+(2\beta-1)_+-1}.$$

Along Γ_C

Proposition (Contribution along Γ_C, [PPXP24] Proposition 8.3)

There exists $M(u_0, u_1) > 0$ and $C(r)$ so that if $\gamma Br \in [M, d^{2\alpha}/M]$, then

$$\frac{1}{C(r)}\big(\mathcal{F}_{pp}(r) + \mathcal{F}_{ac}(r)\big) \leq \mathcal{F}_C(r) \leq C(r)\big(\mathcal{F}_{pp}(r) + \mathcal{F}_{ac}(r)\big),$$

where

$$\mathcal{F}_{pp}(r) := \frac{1}{2\alpha}\int_0^1 u^{(2\beta-1)/2\alpha} \exp(-2\gamma Bru)\, du$$

and

$$\mathcal{F}_{ac}(r) := \frac{c_\beta}{2\alpha}\int_{d^{-2\alpha}}^1 u^{-1/2\alpha} d^{-1} \exp(-2\gamma Bru)\, du,$$

with $c_\beta = \sum_{j=1}^{\infty} j^{-2\beta}$ if $\beta > 1/2$ and $c_\beta = 0$ otherwise.

Along Γ_C

- Proving the previous proposition requires several technical estimates of $q(z)$ along Γ_C.
- Interpretation: $\mathcal{F}_{pp}$ captures the contribution of the spike eigenvalues, while $\mathcal{F}_{ac}$ captures the spectral bulk.
  - $\mathcal{F}_{pp}$ describes the dynamics of learning the high-variance directions.
  - If β is sufficiently large (i.e., the task is "sufficiently easy"), then the lesser eigenvalues do not contribute (i.e., $\mathcal{F}_{ac}$ does not matter).

Along Γ_C

- The dominant contribution to the kernel function $\mathcal{K}$ also arises along Γ_C.

Proposition (Approximation of $\mathcal{K}$, [PPXP24] Proposition 9.1)

Suppose $\alpha > 1/4$. Then there exists a positive function $C(r)$ such that

$$\frac{1}{C(r)}\mathcal{K}_{pp}(r) \leq \mathcal{K}(r) \leq C(r)\mathcal{K}_{pp}(r),$$

where

$$\mathcal{K}_{pp}(r) := \frac{\gamma^2 B}{2\alpha}\int_0^1 u^{1-\frac{1}{2\alpha}} \exp(-2\gamma Bru)\, du.$$

$C(r)$ is bounded independent of d if $\gamma Br < d^{2\alpha}M$ for some $M > 0$.

- Interpretation: This captures the effect of SGD noise. Indeed, if we instead had $B = n \propto r$ (full-batch GD), then due to the assumption $\gamma B < 1$, this would be dominated by $\mathcal{F}$.

Along Γ_L and Γ_R

Proposition (Contribution along Γ_{caps}, [PPXP24] Proposition 8.2)

Define

$$\mathcal{F}_{caps} := \int_{\Gamma_L \cup \Gamma_R} \mathbf{a}^T \mathbf{R}_0(z)\mathbf{a}(1 - 2\gamma B z + 2\gamma^2 B^2 z^2)^r \, dz.$$

There exist positive functions $f(r)$, $g(r)$ and a constant K satisfying $f(r) \leq K \exp(-c\gamma B r d^{-2\alpha})$ and $g(r) \leq K \exp(-c\gamma B r)$ so that

$$|\mathcal{F}_{caps}(r)| \leq K \cdot f(r) d^{-2\alpha + (1-2\beta)_+} + K \cdot g(r).$$

- Hence, so long as $\gamma B r \gtrsim d^{\varepsilon}$ for some $\varepsilon > 0$, then $\mathcal{F}_{\text{caps}}$ is of at most the same order as $\mathcal{F}_0$.

Making the Volterra Equation Explicit

Theorem (Approximation Solution for $\mathcal{R}$, [PPXP24] Theorem 2.1)

Suppose that γ, B are such that $\gamma < 1/2$ and $1/4(1 - \sqrt{1 - \frac{4}{B}}) > \gamma$. Moreover assume $2\alpha + 2\beta > 1$ and $\alpha > 1/4$.[a] There exist $M > 0$ and $A = A(\alpha, \beta, M)$ such that if $\gamma B r > M$, then

$$\mathcal{F}(r) + (\mathcal{K} * \mathcal{F})(r) \leq \mathcal{R}(r) \leq \mathcal{F}(r) + A(\mathcal{K} * \mathcal{F})(r).$$

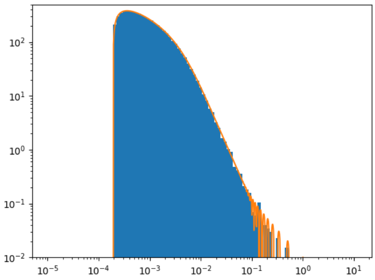
Moreover, the convolution can be bounded as

$$(\mathcal{K} * \mathcal{F})(r) \asymp \mathcal{F}(r) + \frac{1}{\gamma B}\mathcal{K}(r).$$

[a] The authors conjecture that this result also holds for $\alpha \leq 1/4$.

Deterministic Approximation to Spectrum of $\hat{\mathbf{K}}_0$

- Stieltjes inversion: The "spectral density" of $\hat{\mathbf{K}}_0$ at each $\lambda \geq 0$ can be approximated via $\frac{1}{\pi} \lim_{\eta \to 0^+} \frac{1}{d} \operatorname{tr} \mathbf{R}_0(\lambda + i\eta)$.



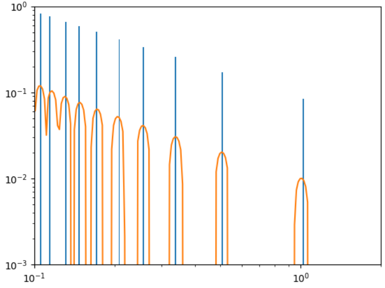


Figure: Simulation for $m = 2000$, $d = 1000$, $\alpha = 0.5$, $\beta = 0.7$, $\tau = 0$.

Change to Spectrum under Rank-1 Perturbation

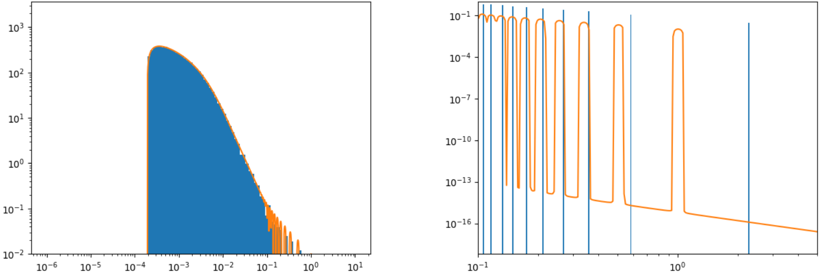


Figure: Simulation for $\alpha = 0.5$, $\beta = 0.7$, $m = 2000$, $d = 1000$, $\tau = 1/\sqrt{d}$.