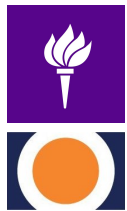
Learning shallow neural networks in high dimensions: SGD dynamics and scaling laws

Denny Wu
dennywu@nyu.edu

Center for Data Science, New York University
Center for Computational Mathematics, Flatiron Institute

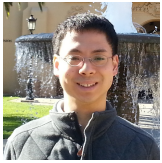


Introduction

- [RN**W**L25] *Emergence and scaling laws for SGD learning of shallow neural networks.*
- [BEV**W**25] *Learning quadratic neural networks in high dimensions: SGD dynamics and scaling laws.*


Gerard Ben Arous


Murat A. Erdogdu


Jason D. Lee

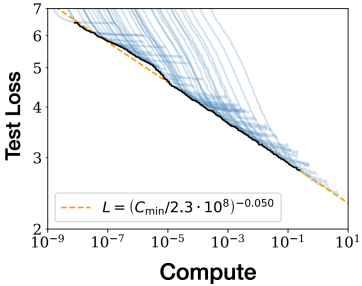

Eshaan Nichani


Yunwei Ren


N. Mert Vural

Neural Scaling Laws & Emergence

Neural Scaling Laws [Hestness et al. 17] [Kaplan et al. 20] [Hoffmann et al. 22].
Increasing compute and data leads to predictable *power-law decay* in the loss.

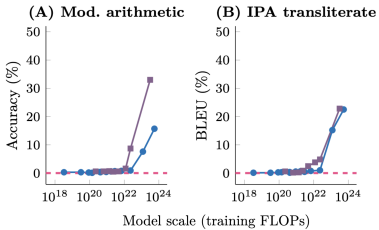


Functional form: $\mathcal{L} \propto D^{-\alpha} + N^{-\beta} + C$.

- D – number of data points.
- N – number of trainable parameters.

Emergent Capabilities [Wei et al. 22] [Ganguli et al. 22] [Schaeffer et al. 23].
Learning of individual tasks (skills) exhibits *sharp transition* with scale.

☹ Unpredictable scaling in skill acquisition.

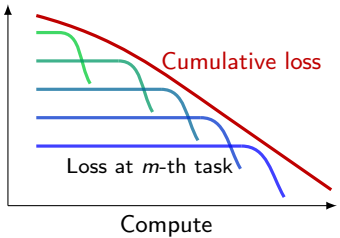


Neural Scaling Laws & Emergence

Question: How do we reconcile the *emergent behavior* in skill acquisition and the *smooth power-law decay* in the cumulative loss?

Hypothesis: Additive Model [Michaud et al. 24] [Nam et al. 24]

- Cumulative objective can be decomposed into a large number of distinct "skills", the learning of each exhibits abrupt phase transitions.
- Juxtaposition of numerous emergent learning curves at different timescales results in a predictable power-law rate in the cumulative loss.



Our goal: theoretical justification of the *additive model* hypothesis in the context of SGD learning of shallow neural network.

Emergence in Gradient-based Feature Learning

Gaussian single-index model: $f_*(\boldsymbol{x}) = \sigma_*(\langle \boldsymbol{x}, \boldsymbol{\theta} \rangle), \ \boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d)$.

- Requires learning the direction $\boldsymbol{\theta} \in \mathbb{R}^d$ and link function $\sigma_* : \mathbb{R} \to \mathbb{R}$.
  - Learning algorithm should adapt to *low-dimensional structure*.

Hermite expansion: $\sigma_*(z) = \sum_{j=0}^{\infty} \alpha_j^* \mathrm{He}_j(z), \ \alpha_j^* = \mathbb{E}[\sigma_*(z)\mathrm{He}_j(z)]$.

Definition: information exponent [Ben Arous et al. 2021]

The information exponent of σ_* is defined as $k = \mathrm{IE}(\sigma_*) = \min\{k \in \mathbb{N}_+ : \alpha_k^* \neq 0\}$.

Intuition: the amount of information in the gradient at *random initialization*.

Theorem ([Ben Arous et al. 21], [Bietti et al. 22], [Damian et al. 23]...)

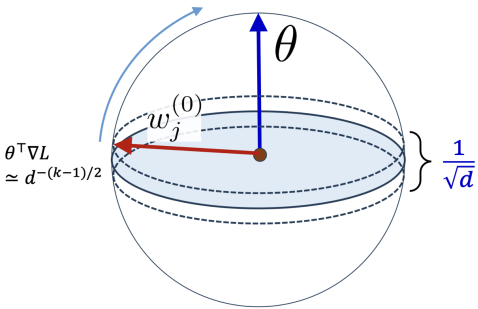
A two-layer neural network can learn single-index target f_* with information exponent k using $n \simeq T \gtrsim d^{\Theta(k)}$ *samples and SGD steps.*

Emergence in Gradient-based Feature Learning

Definition: information exponent [Ben Arous et al. 2021]

The information exponent of σ_* is defined as $k = \mathrm{IE}(\sigma_*) = \min\{k \in \mathbb{N}_+ : \alpha_k^* \neq 0\}$.

Phenomenon: most training examples in online SGD are used to escape from the high-entropy *equator* ($d^{-1/2}$ overlap) around random initialization.



Emergent learning curve

- **Search Phase.** Online SGD exhibits extensive loss plateau up to $T \asymp d^{k-1}$ steps.
- **Descent Phase.** Loss sharply decreases in $T = \tilde{\Theta}(1)$.

Scaling Laws for Shallow Neural Networks?

Target Function: Width-M_* two-layer neural network

$$f_*(\boldsymbol{x}) = \sum_{m=1}^{M_*} a_m \cdot \sigma(\langle \boldsymbol{x}, \boldsymbol{\theta}_m \rangle), \quad \sum a_m^2 = 1, \ \{\boldsymbol{\theta}_m\}_{m=1}^{M_*} \text{ orthonormal}.$$

- **Extensive width.** $M_* \asymp d^\alpha$ for $\alpha > 0$.
  - Large number of "tasks" $\Rightarrow$ *infinite-dimensional* effective dynamics.
- **Large condition number.** $a_{\text{max}}/a_{\text{min}} \sim \text{Poly}(M_*)$.
  - Covers **power-law decay** in second-layer $a_m \asymp m^{-\beta}, \beta \in [0, \infty)$.

Goal: characterize the optimization and sample complexity of SGD training:

- *sharp recovery time (emergence)* for individual single-index tasks $\boldsymbol{\theta}_m$.
- *power-law scaling* in the cumulative mean squared error (MSE) objective.

Prior Results: Well-Conditioned Regime

Theorem ([OSSW24] Sample Complexity of SGD Training)

Assume $k > 2$, *layer-wise (online) SGD training of two-layer neural network with* N *neurons achieves* ε *population loss using*

$$n = \tilde{O}_d\big(M_* d^{k-1} \vee M d \varepsilon^{-2}\big), \qquad N = \tilde{O}_d\big(M_*^{\kappa+1/2} \varepsilon^{-1}\big),$$

where $\kappa = |a_{\max}/a_{\min}| \geq 1$ *is the condition number.*

Comparison against prior results (assume degree-p link σ)

- **Kernel ridge regression** requires $n \gtrsim d^p$ samples.
  - KRR does not adapt to low-dimensional structures.
- **GD-based training for multi-index model** requires $n \gtrsim \big(d^{\Theta(k)}\big) \vee M_*^p$ samples.
  - Does not account for the *additive* (ridge-separable) structure of f_*
    $\Rightarrow$ sample complexity worsen as M_* becomes large.

Prior Results: Well-Conditioned Regime

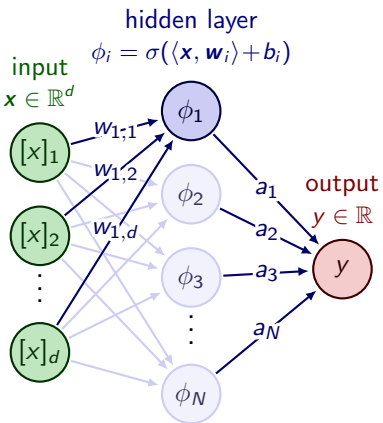
Width-N Two-layer NN: $f_{\mathrm{NN}}(\boldsymbol{x}) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_i \rangle + b_i)$.

❒ **Correlation loss analysis.**

- Under small initialization, MSE loss is approximated by correlation (i.e., interaction between neurons ignored).
- Correlation loss yields alignment with teacher neurons.

❒ **Layer-wise training.**

- After $\{\boldsymbol{\theta}_m\}_{m=1}^{M_*}$ recovered, "fine-tune" second-layer to account for varying a_m.
- Convex problem with closed-form solution (under ℓ_2 penalty).

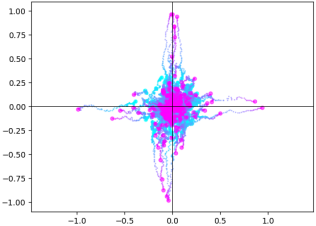


Prior Results: Well-Conditioned Regime

❒ How does *high information exponent* $k > 2$ simplify the analysis?

Correlation loss dynamics: each neuron is attracted by the sum of M_* tasks,

$$\boldsymbol{w}^{(t+1)} \approx \boldsymbol{w}^{(t+1)} + \eta \sum_{m=1}^{M_*} a_m \langle \boldsymbol{w}^{(t)}, \boldsymbol{\theta}_m \rangle^{k-1} \boldsymbol{\theta}_m + Z^t.$$



$\Rightarrow$ For *k sufficiently large*, the sum is approximated by $\max_m \{a_m \langle \boldsymbol{w}^{(t)}, \boldsymbol{\theta}_m \rangle^{k-1}\}$.

☺ **Decoupled dynamics:** each student neuron converges to $\boldsymbol{\theta}_m$ with *largest initial overlap*.

❒ Why can't we handle *large condition number* $\kappa = a_{\max}/a_{\min} \gg 1$?

- At initialization $\langle \boldsymbol{w}^{(0)}, \boldsymbol{\theta}_m \rangle^2 \asymp d^{-1}$ with high probability.

☹ All student neurons converge to *large* a_m *directions*, unless $N \sim \exp(M_*)$.

Decoupling via MSE Loss + Simultaneous Training

- **Student model:** *2-homogeneous* two-layer neural network (matching σ)

$$f(\boldsymbol{x}) = \sum_{i=1}^{N} \|\boldsymbol{w}_i\|_2^2 \cdot \sigma(\langle \boldsymbol{x}, \boldsymbol{w}_i \rangle / \|\boldsymbol{w}_i\| \rangle).$$

- **Online SGD training:** at time t, compute MSE gradient update

$$\boldsymbol{w}_i(t+1) = \boldsymbol{w}_i(t) - \eta \nabla_{\boldsymbol{w}_i} (f_*(\boldsymbol{x}) - f(\boldsymbol{x}))^2, \quad \boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{I}_d).$$

Single-stage training! No layer-wise learning & reinitialization, etc.

How does this resolve exponential dependence on condition number κ?

- **Tangent & radial dynamics.** Starting from small initialization, SGD first "rotates" $\boldsymbol{w}_i(t)$ to align with one of the target neurons $\boldsymbol{\theta}_m$, after which $\|\boldsymbol{w}_i(t)\| \to a_m$ rapidly.
- **"Automatic deflation".** After $\boldsymbol{\theta}_m$ is recovered, it gets "deleted" from the MSE loss, so that subsequent student neurons can converge to other target directions.

Main Theorem: Complexity of SGD Learning

Theorem ([RNWL25] Complexity of SGD learning)

Assume $k > 2$ and $M_* \ll d^{0.1}$, $\bar{m} < M_*$. If we train a student network with $N = \tilde{\Theta}(\bar{m})$ neurons using online SGD with $\eta \asymp \frac{a_{\bar{m}}}{d^{k/2}\mathrm{poly}(M_*)}$, then w.h.p.,

- *If $m < \bar{m}$, alignment with $\boldsymbol{\theta}_m$ emerges at* $T_m \asymp a_m^{-1} \cdot \eta^{-1} d^{k/2-1}$.
- *All directions up to $\bar{m}$ are learned at* $n \asymp T \asymp a_{\bar{m}}^{-2} d^{k-1} \mathrm{poly}(M_*)$.

Corollary: ignoring logarithmic factors, achieving small population loss (i.e., learning *all* tasks) requires $N \asymp M_*$ neurons and $n \asymp T \asymp a_{\min}^{-2} d^{k-1} \mathrm{poly}(M_*)$.

- Prior works required sample size or compute *exponentially* large in the condition number: $n, N \asymp \exp\left(\frac{a_{\max}}{a_{\min}}\right)$ [Li et al. 22] [Oko et al. 24].
- Our learning procedure does not involve reinitialization [Ge et al. 21] or Stiefel constraint [Ben Arous et al. 24].

Neural Scaling Laws for MSE Loss

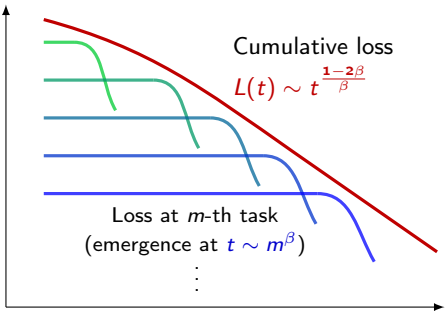
Corollary ([RNWL25] Emergence & Scaling Laws)

Assume $a_m \asymp m^{-\beta}$ *for* $\beta > 1/2$ *and fixed learning rate* η*, then we have*

1. ***Emergence:*** *the m-th teacher neuron is learned at time* $t \cdot \eta \sim d^{k/2-1} m^{\beta}$.
2. ***Scaling law:*** *population MSE decays* $\mathcal{L}(t) \sim N^{1-2\beta} \vee \left(t \cdot \eta d^{1-k/2}\right)^{\frac{1-2\beta}{\beta}}$.

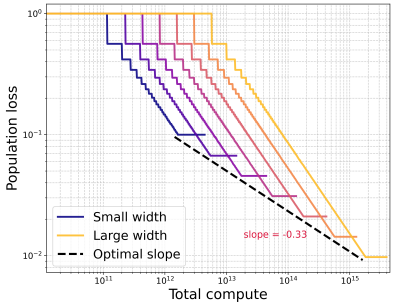
Intuition: Assume *decoupled* learning of different components,

- Direction $\boldsymbol{\theta}_m$ learned at step $T_m \propto m^{\beta} \eta^{-1} d^{k/2-1}$.
- $\mathcal{L}(t) \approx \sum_{m=1}^{M_*} a_m^2 \mathbb{I}\{t < T_m\} \Rightarrow \mathcal{L}(T_m) \approx \int_m^{\infty} s^{-2\beta} \, \mathrm{d}s \sim m^{1-2\beta}$.
- $N < M_*$ student neurons reaches $\sum_{s>N} a_s^2 \asymp N^{1-2\beta}$ error.

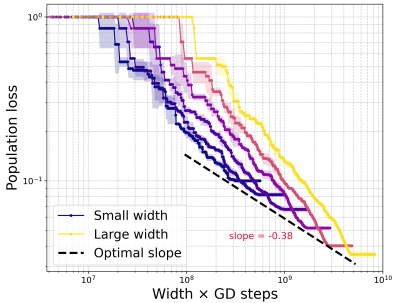


Neural Scaling Laws for MSE Loss

- $N^{1-2\beta}$ – approximation barrier, determined by the student network width.
- $\left(t\eta d^{1-k/2}\right)^{\frac{1-2\beta}{\beta}}$ – optimization error, determined by number of SGD steps.



(a) Theoretical scaling law ($\beta = 0.8$).



(b) Empirical scaling law ($\beta = 0.8$).

Discretizing Properly

Note: the continuous-time (or constant-η) rate can be misleading!

- For a fixed step size, online SGD is unstable for tasks with sufficiently small a_m.
  $\Rightarrow \eta$ should decay with T_m to resolve smaller signal directions.

Adaptive learning rate for m-th task. Consider learning the top-m neurons.

- "Optimal" learning rate: $\eta \asymp \frac{a_m}{d^{k/2}\mathrm{poly}(M_*)}$, $a_m \asymp m^{-\beta}$.
- Direction $\boldsymbol{\theta}_m$ now learned at $n = T_m \sim m^\beta \eta^{-1} d^{k/2-1} = m^{2\beta} d^{k-1}\mathrm{poly}(M_*)$.

Sample complexity: under this learning rate, for the first m neurons

$$\mathcal{L}(n) \sim \left(\frac{n}{d^{k-1}\mathrm{poly}(M_*)}\right)^{\frac{1-2\beta}{2\beta}}$$

☺ This matches the optimal rate for weak ℓ_p ball [Johnstone 17].

Proof Sketch

- **Decoupled gradient dynamics.**

  Let $\boldsymbol{w}_{\iota(p)}$ denote the neuron that eventually converged to direction $\boldsymbol{\theta}_p$, and $m_{p,q}(t) = \langle \boldsymbol{w}_p(t)/\|\boldsymbol{w}_p(t)\|, \boldsymbol{\theta}_q \rangle$ measures the overlap at time t.

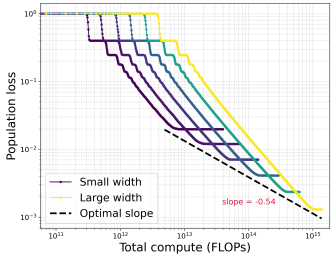
  - **Claim 1 - decoupling**. When all $m^2_{\iota(p),q}$ $(q \neq p)$ are small, the learning of different directions can be approximately decoupled.
  - **Claim 2 - sharp transitions**. Since the norm of $\boldsymbol{w}_p$ grows rapidly after alignment is achieved, all $m^2_{\iota(p),q}$ $(q \neq p)$ remain small when the q-th direction is recovered by $\boldsymbol{w}_{\iota(q)}$ – after which $\boldsymbol{\theta}_q$ no longer affects the learning dynamics ("automatic deflation").

- **From gradient flow to online SGD.**
  - Martingale decomposition in [Ben Arous et al. 21] + a refined stochastic induction argument from [Ren & Lee 24].
  - "Unstable" discretization that couples the online SGD dynamics for learning the top-p teacher neurons.

SGD Learning of Quadratic Neural Networks

Question: how do we handle the *lower information exponent* $k = 2$ setting?

Empirical observation: the same emergence & scaling law phenomenon appears in SGD training of quadratic neural networks.



☹ Learning of different teacher directions $\boldsymbol{\theta}_m$ *no longer decoupled*.

☺ For *quadratic* nonlinearity $\sigma(z) = z^2 - 1$, the population dynamics admits *closed-form* description (see e.g., [Martin et al. 2023]).

Claim. Dynamics of the overlap Gram matrix $\boldsymbol{G}(t) = \boldsymbol{\Theta}^\top \boldsymbol{U}(t)\boldsymbol{U}(t)^\top \boldsymbol{\Theta} \in \mathbb{R}^{M_* \times M_*}$, where $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{Q}^{1/2}$ is the *polar decomposition*, is given by a Matrix Riccati ODE,

$$\partial_t \boldsymbol{G}(t) = \frac{1}{\|\boldsymbol{A}\|_F}(\boldsymbol{A}\boldsymbol{G}(t) + \boldsymbol{G}(t)\boldsymbol{A} - 2\boldsymbol{G}(t)\boldsymbol{A}\boldsymbol{G}(t)), \text{ where } \boldsymbol{A}_{j,j} = a_j.$$

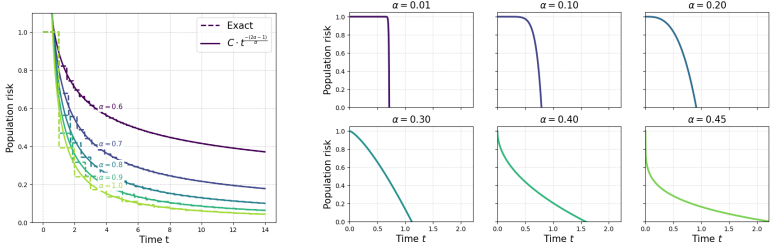
Risk Scaling of Gradient Flow

Theorem ([BEVW25] Time Complexity of Gradient Flow)

Assume $\sigma(z) = z^2 - 1$, $M_* \asymp d^\alpha$, $\alpha \in [0, 1)$, *and* $a_m \asymp m^{-\beta}$. *Then as* $d \to \infty$,

- $\beta > \frac{1}{2}$: *Let* $N = \Theta_d(1)$. $\quad \mathcal{L}(t \cdot \log d) \sim t^{\frac{1-2\beta}{\beta}} \vee N^{1-2\beta}$.
- $\beta < \frac{1}{2}$: *Let* $\frac{N}{M_*} \to \varphi \in (0, \infty)$. $\mathcal{L}(t \cdot M_* \log d) \sim \left[\left(1 - t^{\frac{1-2\beta}{\beta}}\right) \vee \left(1 - \varphi^{1-2\beta}\right)\right]_+$.

Remark: when $\beta > 1/2$ the second-layer coefficients $\{a_m\}_{m=1}^{M_*}$ are square-summable.



(a) Light-tailed ($\alpha > 1/2$). **(b)** Heavy-tailed ($\alpha < 1/2$).

Discretization and Sample Complexity

Algorithm: Stage-wise training with online SGD

Phase I: Online SGD on Stiefel manifold (feature learning)

for $t = 0$ ***to*** T_1 **do**

$\widetilde{\boldsymbol{W}}_t = \boldsymbol{W}_{t-1} - \eta \nabla_{\mathrm{St}} L(\boldsymbol{W}_{t-1}; (\boldsymbol{x}_t, y_t))$; // Online SGD step

$\boldsymbol{W}_t = \widetilde{\boldsymbol{W}}_t \big(\widetilde{\boldsymbol{W}}_t^\top \widetilde{\boldsymbol{W}}_t \big)^{-1/2}$; // Polar retraction

end

Phase II: Closed-form for radial component (fine-tuning)

$\boldsymbol{W}_t^{\mathrm{final}} = \boldsymbol{W}_t \boldsymbol{\Omega}_*, \ \boldsymbol{\Omega}_* = \underset{\boldsymbol{\Omega} \in \mathbb{R}^{N \times N}}{\operatorname{argmin}} \sum_{i=1}^{n'} L\big(\boldsymbol{W}_t \boldsymbol{\Omega}; (\boldsymbol{x}_{t+i}, y_{t+i})\big)$; // closed-form

- **Phase I:** Online SGD to recover the M_*-dimensional subspace spanned by f_*.
  - Discretizes the *directional* component of gradient flow; dominates the statistical and computational complexity.
  - **Challenge:** need to control Martingale terms in *operator norm* since $M_* \gg 1$.
- **Phase II:** "Fine-tuning" on M_*-dimensional subspace (artifact of Stiefel SGD ☹).
  - Closed-form solution; required sample size scales as $n' = \tilde{\Theta}(N^2) \ll Nd$.

Discretization and Sample Complexity

Theorem ([BEVW25] Sample Complexity of Online SGD)

To track the gradient flow risk curve, it suffices to set the SGD step size as follows.

- $\beta > \frac{1}{2}$: *set* $\eta^{-1} \asymp d\,\mathrm{polylog}(d)$*; hence achieving* $o_d(1)$ *population MSE requires*

$$n \simeq T \simeq d\,\mathrm{polylog}(d).$$

- $\beta < \frac{1}{2}$: *set* $\eta^{-1} \asymp dM_*^{\beta}\,\mathrm{polylog}\big(1 + \frac{d}{M_*}\big)$*; this yields*

$$n \simeq T \simeq dM_*^{1+\beta}\,\mathrm{polylog}\Big(1 + \frac{d}{M_*}\Big).$$

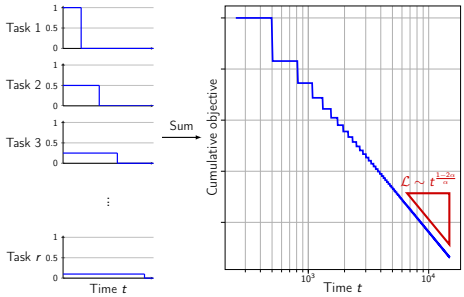
- $\beta > \frac{1}{2}$: $n = \tilde{\Theta}(d)$ – *information theoretically optimal* up to polylog factors.
- $\beta = 0$: $n = \tilde{\Theta}(dM_*)$ – optimal complexity for estimating rank-M_* subspace.
  - $M_* = 1$: *phase retrieval.* Log factors in the time and sample complexity.
  - $\alpha \to 1$: *proportional regime.* Polylogarithmic factors start to diminish.
- $\beta \in \big(0, \frac{1}{2}\big)$: SGD rate likely not sharp in terms of M_* dependence.

Conclusion

SGD learning of two-layer NN

(with high information exponent)

- Recovery of individual neurons exhibits *sharp transition (emergence)*.
- Cumulative MSE loss has predictable *power-law scaling*.



Future Directions

- Anisotropic input: $\boldsymbol{x} \sim \mathcal{N}(0, \boldsymbol{\Sigma})$, where $\boldsymbol{\Sigma} = \sum_{i=1}^{d} \lambda_i \boldsymbol{v}_i \boldsymbol{v}_i^\top$, $\lambda_i \asymp i^{-\gamma}$.
  - Two-parameter scaling law? (source & capacity conditions)
- $k = 1$ (e.g., ReLU) or heterogeneous information exponents?
- Beyond additive structure $\Rightarrow$ scaling laws for *compositional generalization*?

References

Thank you! Happy to take questions :)

- Kaplan et al., 2020. *Scaling laws for neural language models*.
- Ben Arous et al., 2021. *Stochastic gradient descent on non-convex losses from high-dimensional inference*.
- Rong et al., 2021. *Deflation process in over-parametrized tensor decomposition*.
- Hoffmann et al., 2022. *Training compute-optimal large language models*.
- Wei et al., 2022. *Emergent abilities of large language models*.
- Damien et al., 2022. *Neural networks can learn representations with gradient descent*.
- Martin et al., 2023. *On the impact of overparameterization on the training of a shallow neural network in high dimensions*.
- Oko et al., 2024. *Learning sum of diverse features: computational hardness and efficient gradient-based training for ridge combinations*.
- Michaud et al., 2024. *The quantization model of neural scaling*.
- Nam et al., 2024. *An exactly solvable model for emergence and scaling laws*.
- Paquette et al., 2024. *4+3 phases of compute-optimal neural scaling laws*.