Sharp feature-learning transitions and Bayes-optimal neural scaling laws in extensive-width networks

Minh-Toan Nguyen Jean Barbier

NeurIPS 2026

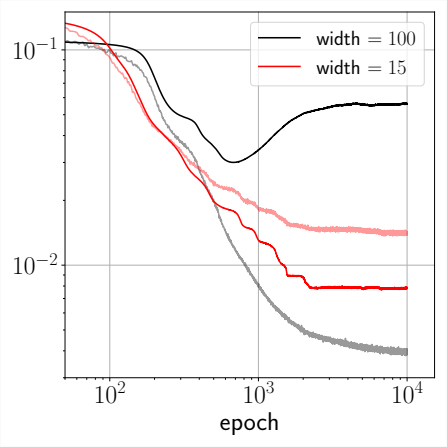
International Centre for Theoretical Physics, Trieste

Neural scaling laws

- Test loss of large networks follows *power laws* in data n, model size, compute (Hestness et al. 2017; Kaplan et al. 2020; Hoffmann et al. 2022)
- A popular explanation: the target has a *hierarchy of features*
  - features have decaying importance $v_1 \geq v_2 \geq \dots$
  - a learner acquires them roughly in order (Hutter 2021; Michaud et al. 2023 "quantization model"; Ren et al. 2025; Defilippis et al. 2025, 2026)
- In a teacher–student / distillation setting, this raises a design question:

At a fixed data budget n, how wide should the student be?

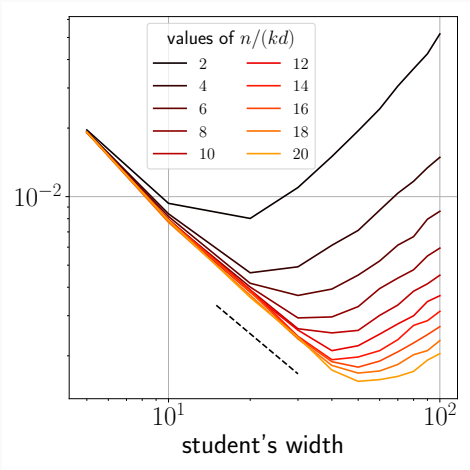
An experiment: Adam students on a power-law teacher



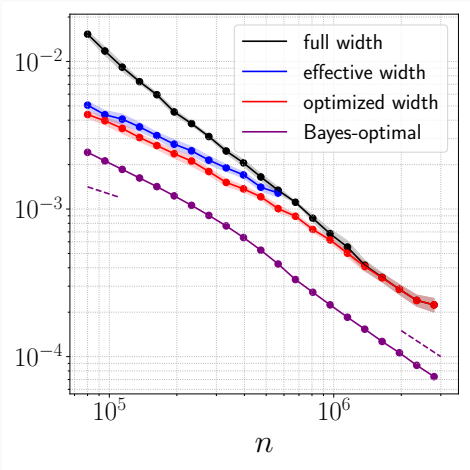
- Teacher: one hidden layer, $k = 100$, $d = 200$, $\sigma(x) = \tanh(2x)$
- Readout $v_i \propto i^{-\beta}$, $\beta = 1$
- Student: same architecture, width k_s, trained with ADAM
- Fixed sample rate $n/(kd) = 2$

Observation
A *smaller* student (solid, test error) beats the full-width student $k_s = k$.

Two empirical regularities



(i) Error vs. width k_s: a *critical width*; below it, slope $1 - 2\beta$.



(ii) Error vs. n near optimal width: slopes $\frac{1}{2\beta} - 1$, then -1.

Questions

Lift all computational constraints: take the *Bayes-optimal* viewpoint.

Q1
Given a dataset, *which* features of the teacher are statistically recoverable?

Q2
How does the recoverable set of features determine the scaling law of the Bayes-optimal error $\varepsilon^{\mathrm{BO}}$ in n?

Short answers

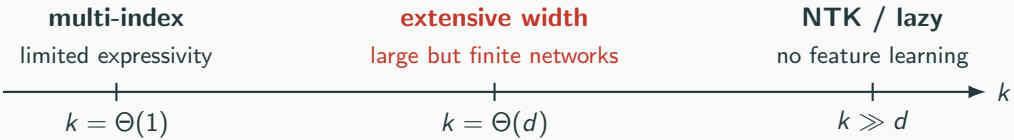
Q1 Feature i is learned iff an effective SNR crosses a threshold λ_σ; features are learned one by one, each with a discontinuous jump.

Q2 With k_c = number of learned features (*effective width*),

$$\varepsilon^{\mathrm{BO}} = \Theta\Big(\frac{k_c\, d}{n}\Big) \qquad \text{in both scaling regimes.}$$

Tool: a heuristic leave-one-out (cavity) argument reducing the network to k coupled single-feature problems, checked by Hamiltonian Monte Carlo.

Why extensive width $k = \Theta(d)$?



- Minimal setting for non-trivial feature learning beyond $n = \Theta(d)$
  - at $n = \Theta(d)$, the BO problem reduces to linear regression (Cui et al. 2023; Camilli et al. 2025)
- Only recently tractable: Maillard et al. 2024 (quadratic σ), Barbier et al. 2025 (replicas, dense readouts)
- This work: structured readouts via a leave-one-out (cavity) argument, no replicas

Setting

Teacher and data

Inputs $\mathbf{x}_\mu \sim \mathcal{N}(0, \mathbf{I}_d)$, noise $z_\mu \sim \mathcal{N}(0,1)$, $\mu = 1, \dots, n$:

$$y_\mu^{\text{out}} = \mathbf{v}^\top \sigma(\mathbf{W}^0 \mathbf{x}_\mu) + \sqrt{\Delta}\, z_\mu = \sum_{i=1}^{k} v_i\, \sigma(\mathbf{w}_i^0 \cdot \mathbf{x}_\mu) + \sqrt{\Delta}\, z_\mu .$$

- row $\mathbf{w}_i^0 \in \mathbb{R}^d$ of $\mathbf{W}^0 \in \mathbb{R}^{k \times d}$ = a *feature*; $\mathbf{W}^0$ uniform on the Stiefel manifold (orthonormal rows)
- readout $\mathbf{v}$ known, $\|\mathbf{v}\| = 1$, $v_1 \geq v_2 \geq \cdots \geq v_k \geq 0$: v_i^2 is the strength of feature i

Readouts considered.

dense	$\sqrt{k}\,\mathbf{v}$ has a limiting density
power-law	$v_i \propto i^{-\beta}$ (focus: $\beta > 1/2$)
exponential	$v_i \propto e^{-ci}$
ultra-sparse	all but $\Theta(1)$ entries vanish

Assumptions and learners

Regime: $d \gg 1, \qquad k = \Theta(d), \qquad n \gg d.$

Activation: $\sigma(z) = \sum_\ell \frac{\mu_\ell}{\ell!} \mathrm{He}_\ell(z)$ with

$$\mu_2 = \mathbb{E}_z[\sigma(z)(z^2 - 1)] = 0 \qquad \text{(e.g. any odd } \sigma\text{)}.$$

The linear part $\mu_0 + \mu_1 z$ can be removed at no cost $\Rightarrow$ *information exponent* ≥ 3.

Two learners.

- *Bayesian student*: knows $\sigma, \Delta, \mathbf{v}$, the prior; predicts with the posterior mean

  $$\hat{y}(\mathbf{x}) = \mathbb{E}[\mathbf{v}^\top \sigma(\mathbf{W}\mathbf{x}) \mid \mathcal{D}] =: \langle \mathbf{v}^\top \sigma(\mathbf{W}\mathbf{x}) \rangle,$$

  whose MSE is the *Bayes-optimal error* $\varepsilon^{\mathrm{BO}}$.
- ADAM *student*: width $k_s \leq k$, weights $(\mathbf{W}, \mathbf{v})$ trained; rows renormalized each epoch.

Building block: one feature

Two functions of the activation

For $\sigma \in L^2(\mathcal{N}(0,1))$ define the kernel

$$g_\sigma(q) := \mathbb{E}[\sigma(z_1)\sigma(z_2)], \quad \mathrm{Corr}(z_1, z_2) = q \qquad \Longrightarrow \qquad g_\sigma(q) = \sum_{\ell \geq 0} \frac{\mu_\ell^2}{\ell!} q^\ell,$$

and the overlap function

$$m_\sigma(\lambda) := \arg \max_{q \in [0,1)} \Big\{ \lambda\, g_\sigma(q) + q + \ln(1-q) \Big\}, \qquad \lambda_\sigma := \sup\{\lambda \geq 0 : m_\sigma(\lambda) = 0\}.$$

- $\lambda g_\sigma(q)$: gain from fitting the data; $\quad q + \ln(1-q) \leq 0$: entropic cost
- m_σ is non-decreasing; λ_σ is the critical SNR

Result A: the high-sample-rate GLM

Single-feature model
Infer $\mathbf{w}^0 \sim \mathcal{U}(\mathbb{S}^{d-1})$ from $y_\mu = \sigma(\mathbf{w}^0 \cdot \mathbf{x}_\mu) + \sqrt{\Delta}\, z_\mu,\ \mu \in [n]$, with $n \gg d \gg 1$, $\Delta \gtrsim 1$, and

$$\lambda := \frac{n}{d\,\Delta} \qquad \text{(samples per parameter / noise)}.$$

Result A
For a posterior sample $\mathbf{w}^1$, the overlap concentrates:

$$\mathbb{E}\left\langle \left(\mathbf{w}^0 \cdot \mathbf{w}^1 - m_\sigma(\lambda)\right)^2 \right\rangle = o_d(1).$$

- Equivalently: the high-sample-rate, high-noise limit ($n/d \to \infty$, $\Delta \to \infty$, ratio fixed) of the GLM formula of Barbier et al. (2019)

Where the potential comes from (heuristic)

Assume $\mathbf{w}^0 \cdot \mathbf{w}^1 \simeq q$. Focus on one coordinate w_j^0 and linearize σ around the rest:

$$y_\mu \simeq \sigma(A_{\mu j}) + \sigma'(A_{\mu j})\, w_j^0 x_{\mu j} + \sqrt{\Delta}\, z_\mu, \qquad A_{\mu j} = \textstyle\sum_{l \neq j} w_l^0 x_{\mu l}.$$

This is a Gaussian channel for w_j^0; its SNR is

$$\begin{aligned} \frac{1}{\Delta} \sum_{\mu=1}^{n} \langle \sigma'(A_{\mu j}) \rangle^2 x_{\mu j}^2 &\simeq \frac{n}{\Delta}\, \mathbb{E} \langle \sigma'(\mathbf{w} \cdot \mathbf{x}) \rangle^2 \\ &= \lambda d\, g_{\sigma'}(q), \qquad \text{since } \mathbf{w} \cdot \mathbf{x} \overset{d}{\simeq} \sqrt{q}\, \xi_0 + \sqrt{1-q}\, \xi_1 \\ &= \lambda d\, g'_\sigma(q), \qquad \text{since } g_{\sigma'} = g'_\sigma. \end{aligned}$$

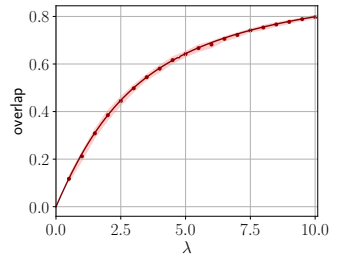
A Gaussian channel with SNR $d \cdot s$ has overlap $q/(1-q) = s$. Hence

$$\frac{q}{1-q} = \lambda\, g'_\sigma(q) \quad \Longleftrightarrow \quad \frac{d}{dq}\Big[\lambda g_\sigma(q) + q + \ln(1-q)\Big] = 0.$$

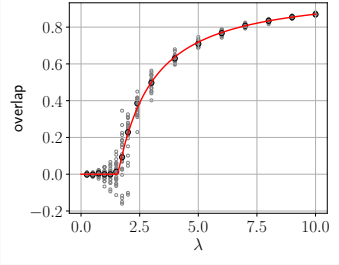
Why the transition is discontinuous

Near $q = 0$: $q + \ln(1-q) = -\frac{q^2}{2} + O(q^3)$, and $\lambda g_\sigma(q) = \lambda \sum_\ell \frac{\mu_\ell^2}{\ell!} q^\ell$.

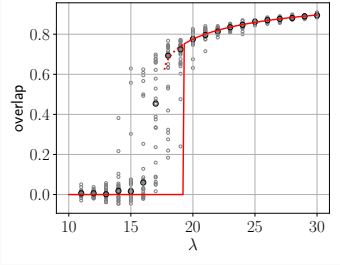
first nonzero μ_ℓ	potential near 0	transition
$\ell = 1$	$\lambda\mu_1^2 q - q^2/2$	none: $q > 0$ for all $\lambda > 0$
$\ell = 2$	$(\lambda\mu_2^2 - 1)\, q^2/2$	continuous, at $\lambda = 1/\mu_2^2$
$\ell \geq 3$	$-q^2/2 + O(\lambda q^3)$	discontinuous: $q = 0$ always a local max



ReLU



$|x|$



$\tanh(2x) - cx$

Main result: many features

Leave-one-out: isolate feature i

$$y_\mu^{\text{out}} = v_i\,\sigma(\mathbf{w}_i^0\cdot\mathbf{x}_\mu) + R_{\mu i}(\mathbf{W}^0) + \sqrt{\Delta}\,z_\mu, \qquad R_{\mu i}(\mathbf{W}) := \sum_{j\neq i} v_j\,\sigma(\mathbf{w}_j\cdot\mathbf{x}_\mu).$$

For a posterior sample $\mathbf{W}$, split

$$R_{\mu i} = \underbrace{\langle R_{\mu i}\rangle}_{\text{function of } \mathcal{D}\text{: subtract}} + \underbrace{\xi_{\mu i}}_{\text{posterior fluctuation}}.$$

- Variance of the fluctuation, with q_j the overlap of feature j:

$$\langle \xi_{\mu i}^2\rangle \simeq \sum_{j\neq i} v_j^2\big(g_\sigma(1) - g_\sigma(q_j)\big) =: \varepsilon_{-i} \simeq \varepsilon.$$

- **Cavity assumption:** $(\xi_{\mu i})_\mu$ are asymptotically independent Gaussians (CLT over $k-1$ weakly coupled terms).

Each feature sees a GLM with inflated noise

After subtracting $\langle R_{\mu i} \rangle$:

$$\text{observation}_\mu = v_i\, \sigma(\mathbf{w}_i^0 \cdot \mathbf{x}_\mu) + \sqrt{\Delta + \varepsilon}\ \text{noise}_\mu.$$

Divide by v_i and apply Result A with noise $(\Delta + \varepsilon)/v_i^2$:

$$\lambda = \frac{n}{d} \cdot \frac{v_i^2}{\Delta + \varepsilon} =: \mathrm{snr}_i \qquad \Longrightarrow \qquad q_i \simeq m_\sigma(\mathrm{snr}_i).$$

- snr_i depends on: feature strength v_i^2, sample rate n/d, effective noise $\Delta + \varepsilon$
- All features are coupled only through the scalar ε

Assumptions used (posterior concentration), for $\mathbf{W}^1$ a posterior sample:

$$\mathbb{E}\left\langle (\mathbf{w}_i^0 \cdot \mathbf{w}_j^1)^2 \right\rangle = O(d^{-1})\ (i \neq j), \qquad \mathbb{E}\left\langle (\mathbf{w}_i^0 \cdot \mathbf{w}_i^1 - q_i)^2 \right\rangle = O(d^{-1}).$$

First one proved for dense readouts.

Result 1: fixed-point equations

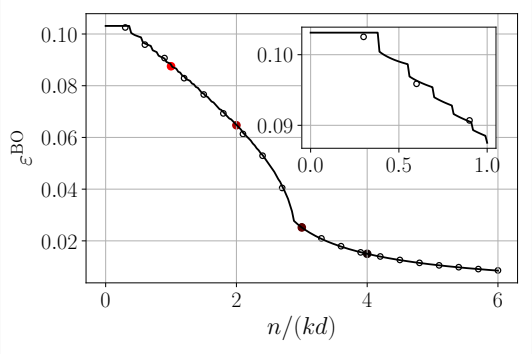
Result 1 (overlaps and Bayes-optimal error)
The overlaps $(q_i)_{i=1}^k$ solve

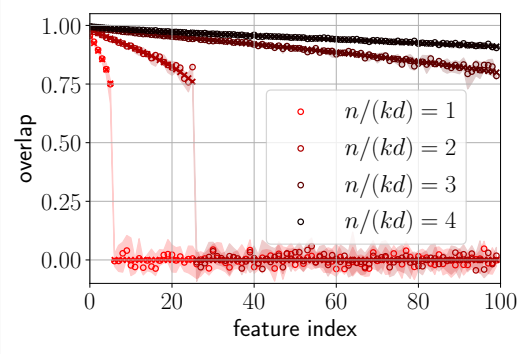
$$q_i = m_\sigma\Big(\frac{n v_i^2/d}{\Delta+\varepsilon}\Big), \qquad \varepsilon = \sum_{i=1}^{k} v_i^2\big(g_\sigma(1) - g_\sigma(q_i)\big),$$

and $\varepsilon^{\mathrm{BO}} \sim \varepsilon$.

- Feature i is *learnable* iff $q_i > 0$ iff $\mathrm{snr}_i > \lambda_\sigma$
- *Effective width*: $k_c := |\{i : q_i > 0\}| \le k$
- Numerically: a single scalar equation in ε; unique solution in all our experiments
- Valid for arbitrary readouts; dense case recovers Barbier et al. (2025) (replicas)

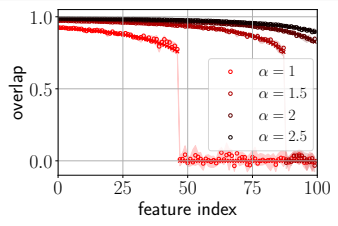
A staircase of transitions



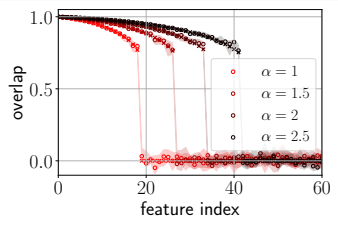


As n grows, $\mathrm{snr}_1 > \mathrm{snr}_2 > \dots$ cross λ_σ one at a time: each crossing is a jump in q_i and a drop in $\varepsilon^{\mathrm{BO}}$.
Lines/crosses: theory. Circles: HMC ($d = 200$, $k = 100$, $\beta = 0.3$).

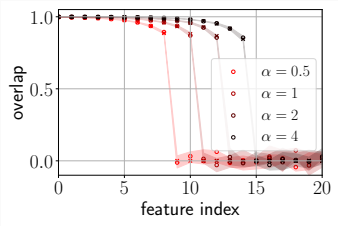
Other readouts



dense



power-law, $\beta = 0.7$



$v_j \propto e^{-j/5}$

- Crosses: theory; circles: HMC, already accurate at $d = 200$
- Caveat: HMC is initialized at the teacher for the first k_c rows, so it checks overlaps and error, not k_c itself

Result 2: decomposition of the error

Split the sum at k_c ($q_i = 0$ for $i > k_c$) and write $v_i^2 = (\Delta + \varepsilon)\frac{d}{n}\mathrm{snr}_i$:

$$\begin{aligned}\varepsilon^{\mathrm{BO}} &\sim \sum_{i \le k_c} v_i^2 \big(g_\sigma(1) - g_\sigma(q_i)\big) + g_\sigma(1) \sum_{i > k_c} v_i^2 \\ &= (\Delta + \varepsilon)\frac{d}{n} \sum_{i \le k_c} \underbrace{\mathrm{snr}_i \big(g_\sigma(1) - g_\sigma(m_\sigma(\mathrm{snr}_i))\big)}_{\in [c_1, c_2] \text{ since } \mathrm{snr}_i > \lambda_\sigma} + g_\sigma(1) \sum_{i > k_c} v_i^2 .\end{aligned}$$

Result 2

$$\varepsilon^{\mathrm{BO}} = \Theta\Big(\frac{k_c\, d}{n}\Big) + \Theta\Big(\sum_{j > k_c} v_j^2\Big).$$

- each learned feature costs $\Theta(d/n) =$ (its d parameters)$/n$
- each unlearned feature j costs its full strength $\Theta(v_j^2)$

Scaling laws

Effective width for power-law readouts

Take $v_i^2 = i^{-2\beta}/z_k(\beta)$ with $\beta > 1/2$, so $z_k(\beta) = \sum_{j \le k} j^{-2\beta} = \Theta(1)$. Also $\Delta + \varepsilon = \Theta(1)$.

$$\begin{aligned} \text{feature } i \text{ learnable} &\iff \frac{n}{d} \cdot \frac{v_i^2}{\Delta + \varepsilon} > \lambda_\sigma \\ &\iff i^{-2\beta} \gtrsim \frac{d}{n} \\ &\iff i \lesssim \left(\frac{n}{d}\right)^{1/(2\beta)}. \end{aligned}$$

Hence

$$k_c = \min\left\{k,\ \Theta\big((n/d)^{1/(2\beta)}\big)\right\}, \qquad \text{crossover at } n^\star = \Theta(k^{2\beta} d)$$

($n^\star$: the weakest feature k becomes learnable).

Result 3: two regimes, one law

Feature-learning regime $n \ll k^{2\beta} d$: $k_c = \Theta((n/d)^{1/(2\beta)})$.

$$\frac{k_c d}{n} = \Theta\Big((n/d)^{\frac{1}{2\beta}-1}\Big),$$

$$\sum_{j>k_c} v_j^2 = \Theta\big(k_c^{1-2\beta}\big), \qquad \text{since } \textstyle\sum_{j>K} j^{-2\beta} \asymp K^{1-2\beta}$$

$$= \Theta\Big((n/d)^{\frac{1}{2\beta}-1}\Big) \qquad \text{same order as the first term.}$$

Refinement regime $n \gg k^{2\beta} d$: $k_c = k$, no unlearned features.

Result 3 (Bayes-optimal scaling law, $\beta > 1/2$)

$$\varepsilon^{\mathrm{BO}} = \Theta\Big(\frac{k_c\, d}{n}\Big) = \begin{cases} \Theta_{k,d}\big(n^{\frac{1}{2\beta}-1}\big) & n \ll k^{2\beta} d \quad \text{(feature learning)} \\ \Theta_{k,d}\big(n^{-1}\big) & n \gg k^{2\beta} d \quad \text{(refinement)} \end{cases}$$

Interpretation

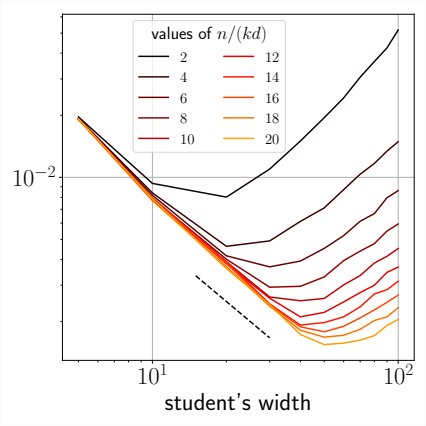
$$\varepsilon^{\mathrm{BO}} = \Theta\Big(\Big(\frac{n}{k_c d}\Big)^{-1}\Big): \quad \text{inverse of the } \textit{effective sampling rate}.$$

- The teacher behaves as a network with k_c features, each needing $\Theta(d)$ samples
- Feature learning: k_c grows with n and partially offsets the $1/n$ decay $\Rightarrow$ slower exponent $\frac{1}{2\beta} - 1 > -1$
- Refinement: $k_c = k$ frozen $\Rightarrow$ parametric rate n^{-1}
- A first-principles version of the *quantization model* (Michaud et al. 2023): discrete units learned in order, capacity-limited

Compressibility. When $k_c < k$, the data are statistically equivalent (same $\varepsilon^{\mathrm{BO}}$) to those of the width-$k_c$ teacher with extra noise

$$y_\mu = \mathbf{v}_{1:k_c}^\top \sigma(\mathbf{W}^0_{1:k_c}\mathbf{x}_\mu) + \sqrt{\Delta + \Delta'}\, z_\mu, \qquad \Delta' = g_\sigma(1) \textstyle\sum_{j>k_c} v_j^2.$$

Back to Adam: scaling in model size



Width k_s below the critical width, feature-learning regime:

- learns $k_s \leq k_c$ features (cannot beat Bayes)
- as in Result 2: error $\Theta(\frac{k_s d}{n}) + \Theta(\sum_{j>k_s} v_j^2)$

The second term dominates:

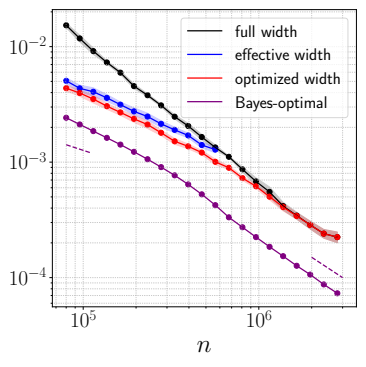
$$\sum_{j>k_s} v_j^2 \geq \sum_{j>k_c} v_j^2 \gtrsim \frac{k_c d}{n} \geq \frac{k_s d}{n}.$$

Therefore

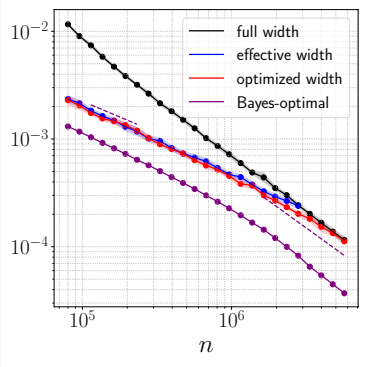
$$\begin{aligned} \text{error}(k_s) &= \Theta\Big(\sum_{j>k_s} j^{-2\beta}\Big) \\ &= \Theta\Big(\int_{k_s}^{k} x^{-2\beta}\, dx\Big) \\ &= \Theta\big(k_s^{1-2\beta}\big). \end{aligned}$$

Above k_c: fits non-learnable directions, error grows.

Back to Adam: scaling in data



$\beta = 1$



$\beta = 1.2$

- Width $k_s = k_c$ from theory: ADAM matches the BO exponents
- Conjecture: ADAM learns feature i iff $\mathrm{snr}_i > \lambda_\sigma^{\mathrm{algo}} \geq \lambda_\sigma$ (computational-to-statistical gap, as for AMP; for $\sigma = \mathrm{He}_3$ the GLM is tensor PCA). A constant-factor gap leaves exponents unchanged.
- Small deviation at large n: $\approx 5\%$ weakest features not learned by ADAM

Proportional regime $k \sim \gamma d$, $n \sim \alpha d^2$, $d \to \infty$

Same equations give the rate in d at fixed (α, γ):

readout	$\varepsilon^{\mathrm{BO}}$	k_c
dense	$\Theta(1)$	$\Theta(d)$
power-law, $\beta < 1/2$	$\Theta(1)$	$\Theta(d)$
power-law, $\beta = 1/2$	$\Theta\left(\frac{\ln \ln d}{\ln d}\right)$	$\Theta\left(\frac{d}{\ln d}\right)$
power-law, $\beta > 1/2$	$\Theta\left(d^{\frac{1}{2\beta}-1}\right)$	$\Theta\left(d^{\frac{1}{2\beta}}\right)$
exponential	$\Theta\left(\frac{\ln d}{d}\right)$	$\Theta(\ln d)$
ultra-sparse	$\Theta\left(\frac{1}{d}\right)$	$\Theta(1)$

- Sparser readout $\Rightarrow$ smaller effective width $\Rightarrow$ smaller error
- Exact constants and the limiting overlap profile $q_{\lfloor x k_c \rfloor} \to m_\sigma(\lambda_\sigma x^{-2\beta})$ in the paper

Two hidden layers

Two hidden layers of extensive width

$$y_\mu = \mathbf{v}^\top \sigma\big(\mathbf{W}^{(2)}\sigma(\mathbf{W}^{(1)}\mathbf{x}_\mu)\big) + \sqrt{\Delta}\, z_\mu, \qquad \mathbf{W}^{(1)} \in \mathbb{R}^{k_1 \times d},\ \mathbf{W}^{(2)} \in \mathbb{R}^{k_2 \times k_1},$$

$k_1, k_2 = \Theta(d)$. Overlaps: q_0 (first layer, all rows equal by symmetry), q_i (row i of the second layer).

Potential, adapted from the replica-symmetric one of Barbier et al. (2025), with $\alpha_1 = \frac{n}{k_1 d}$:

$$F(q_0, \mathbf{q}) = -\alpha_1 \ln(\Delta + \varepsilon) + q_0 + \ln(1 - q_0) + \frac{1}{d}\sum_{i=1}^{k_2} \big(q_i + \ln(1 - q_i)\big),$$

$$\varepsilon = \sum_{i=1}^{k_2} v_i^2 \big(1 - g_\sigma(G q_i)\big), \qquad G = g_\sigma(q_0).$$

- At fixed q_0: a one-layer problem with input dimension k_1 and the noisy activation σ_G, so $q_i = m_{\sigma_G}(\mathrm{snr}_i)$, $\mathrm{snr}_i = \frac{(n/k_1)v_i^2}{\Delta + \varepsilon}$
- $q_0 = 0 \iff \mathbf{q} = 0$: neither layer is learned without the other

Result 4: two-layer scaling laws

Result 4

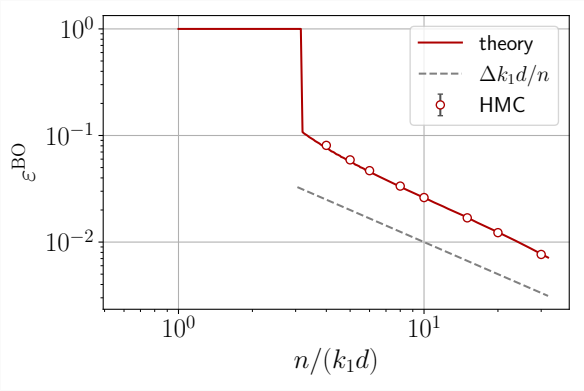
1. **First transition.** Nothing is learned for $n < \alpha_{1c} k_1 d$, *whatever the readout*; then q_0 and the top q_i jump together.
2. **Error.** For every readout and $n \gg k_1 d$: $\varepsilon^{\mathrm{BO}} = \Theta(k_1 d/n)$.
3. **Effective width.** Power-law, $\beta > 1/2$: $k_c = \min\{k_2, \Theta((n/k_1)^{1/(2\beta)})\}$.

Reason for 2: the error splits as

$$\varepsilon^{\mathrm{BO}} = \underbrace{\Theta\Big(\frac{k_1 d}{n}\Big)}_{\text{first layer}} + \underbrace{\Theta\Big(\frac{k_c k_1}{n}\Big)}_{\text{learned rows}} + \underbrace{O\Big(\frac{(k_2 - k_c)k_1}{n}\Big)}_{\text{unlearned rows}} = \Theta\Big(\frac{k_1 d}{n}\Big),$$

since $k_c \le k_2 = \Theta(d)$. The first layer is learned as a whole: the readout only changes the constant, not the exponent.

Two layers: HMC check



- $d = k_1 = k_2 = 100$, $\beta = 1$, $\Delta = 0.1$
- Solid: theory from F; circles: HMC; dashed: $\Delta k_1 d/n$
- One big jump at $n = \Theta(k_1 d)$ removes most of the error
- Later jumps (one per row) have size $\Theta(k_1/n)$, a fraction $\Theta(1/d)$ of $\varepsilon^{\mathrm{BO}}$: invisible

Conclusion

Relation to prior work

- **Defilippis et al. (2025, 2026)**: multi-index $k = \Theta(1)$, activations with $\mu_2 \neq 0$; same BO data exponent $\frac{1}{2\beta} - 1$, error decomposition and k_c anticipated there
  $\rightarrow$ suggests universality across width regimes and activation classes
- **Ren et al. (2025)**: same sequential learning, dynamically under SGD, $k \ll d$
- **Barbier et al. (2025)**: replicas, dense readouts, multi-layer; recovered here
- **Michaud et al. (2023)**: quantization model, postulated $\rightarrow$ derived here

New here: sharp individual overlaps at $k = \Theta(d)$, general readouts, $\mu_2 = 0$; k_c as a prescription for the student width.

Limitations and open problems

- Heuristic: Gaussian cavity fields and overlap concentration are assumed, not proved
- HMC initialized partly at the teacher: checks overlaps and error, not k_c
- $\mu_2 = 0$ needed for discontinuous transitions; scaling laws expected to survive $\mu_2 \neq 0$
- Isotropic Gaussian inputs
- Open: prove the leave-one-out reduction; the algorithmic threshold $\lambda_\sigma^{\mathrm{algo}}$ for ADAM; compute-optimal scaling

Summary

1. Each feature of an extensive-width network behaves like a single-feature GLM with noise $\Delta + \varepsilon^{\mathrm{BO}}$:
$$q_i = m_\sigma\Big(\frac{n v_i^2/d}{\Delta+\varepsilon}\Big), \qquad \varepsilon = \sum_i v_i^2\big(g_\sigma(1) - g_\sigma(q_i)\big).$$
2. Features are learned one by one, each through a sharp jump at $\mathrm{snr}_i = \lambda_\sigma$.
3. The *effective width* k_c unifies both scaling regimes:
$$\varepsilon^{\mathrm{BO}} = \Theta\Big(\frac{k_c\, d}{n}\Big) \sim n^{\frac{1}{2\beta}-1} \text{ then } n^{-1}.$$
4. An ADAM student of width $\approx k_c$ attains these exponents; wider students overfit.
5. Two layers: nothing learned below $n = \Theta(k_1 d)$, then $\varepsilon^{\mathrm{BO}} = \Theta(k_1 d/n)$ for every readout.

Code: `https://github.com/Minh-Toan/scaling-laws`

Thank you!

Backup: uniqueness and the potential

The fixed-point equations can have several solutions (e.g. $\sigma = \mathrm{He}_3/\sqrt{6}$, dense, $n/(kd) = 1.07$, $\Delta = 0.02$). Select the maximizer of

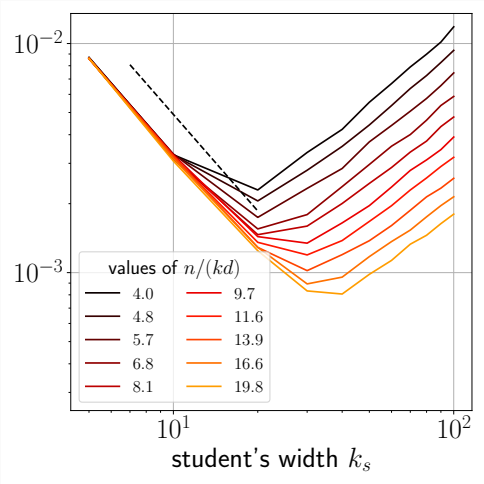
$$F(\mathbf{q}) = -\frac{n}{d}\ln(\Delta+\varepsilon) + \sum_{i=1}^{k}\big[q_i + \ln(1-q_i)\big], \qquad \varepsilon = \sum_i v_i^2\big(g_\sigma(1) - g_\sigma(q_i)\big).$$

Maximizing in q_i alone, with $a = \Delta + \varepsilon_{-i}$ and $\ln(a+x) \simeq \ln a + x/a$:

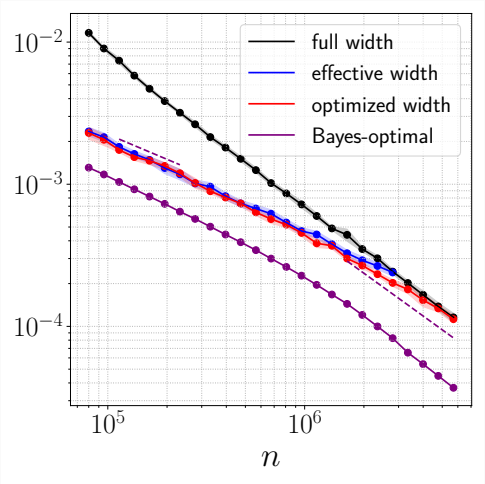
$$\begin{aligned} F &\simeq -\frac{n}{d}\cdot\frac{v_i^2\big(g_\sigma(1) - g_\sigma(q_i)\big)}{\Delta+\varepsilon_{-i}} + q_i + \ln(1-q_i) + \text{const} \\ &= \frac{nv_i^2/d}{\Delta+\varepsilon_{-i}}\, g_\sigma(q_i) + q_i + \ln(1-q_i) + \text{const}, \end{aligned}$$

whose maximizer is $m_\sigma(\mathrm{snr}_i)$: the fixed-point equations again. Stationary points are informative about AMP and the computational gap.

Backup: Adam at $\beta = 1.2$



slope $1 - 2\beta = -1.4$



slopes $-7/12$ and -1

Teacher $\tanh(2x)$, $k = 100$, $d = 200$, $\Delta = 0.01$; 3 batches, 10^4 epochs, lr 0.003; 9 runs.