Scaling Laws

Primary references:

"Scaling Laws in Linear Regression: Compute, Parameters, and Data" (Lin et al., 2024)

"Improved Scaling Laws in Linear Regression via Data Reuse" (Lin et al., 2025)

"4+3 Phases of Compute-Optimal Neural Scaling Laws" (Paquette et al., 2024)

"Learning Quadratic Neural Networks in High Dimensions" (Ben Arous et al., 2025)

"Emergence and Scaling Laws in SGD Learning of Shallow Neural Networks" (Ren et al., 2025)

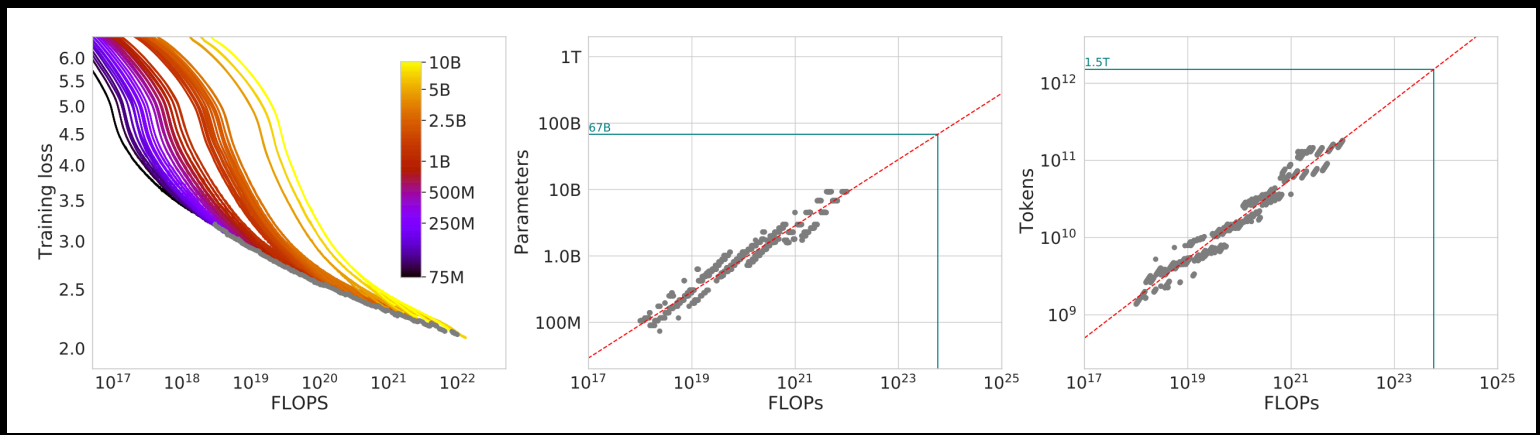
Consider an LLM with d parameters trained on n examples.

Empirically, as d, n grow, a power law trend has been observed in the test loss:

$$R(n,d) = c_1 \cdot n^{-\alpha_n} + c_2 \cdot d^{-\alpha_d}.$$

e.g. "Chinchilla scaling" (Hoffman et al., 2022): $\alpha_n = \alpha_d = 0.5$

One can also ask: If we consider a fixed compute budget $C = nd$ (in FLOPs), what is the optimal choice of (n, d)?



It is also empirically observed that the optimal d, n follow a power law in C.

Question: Is there a theoretically tractable setting in which we can rigorously prove this power law scaling?

- Power Law Random Features (Maloney et al., 2022)

Sketched linear regression with "source and capacity constraints" (Caponnetto and de Vito, 2007)

Input: $x \sim N_d(0, H)$ Design matrix $X \in \mathbb{R}^{n \times d}$

Let $(\lambda_i, v_i)_{i=1}^d$ be the eigenvalues and eigenvectors of H, with $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$.

Capacity condition: $\lambda_i \asymp i^{-a}$ $\forall i \geq 1$ and some $a > 1$

α controls the "effective dimension" of the input data

This is observed in real-world data (e.g. Zipf's Law in natural language)

Target: $\mathbb{E}[y | x, w^*] = \langle x, w^* \rangle$, $\sigma^2 := \mathbb{E}[(y - \langle x, w \rangle)^2]$

For $i \neq j$, $\mathbb{E}[\langle v_i, w^* \rangle \langle v_j, w^* \rangle] = 0$

Source condition: For $i \geq 1$, $\mathbb{E}[\lambda_i \langle v_i, w^* \rangle^2] \asymp i^{-b}$ for some $b > 1$.

The projections of w^* onto the directions of highest variance have power law decay.

Aside: Paquette et al. (2024) assume a simplified special case where $H = D = \text{diag}(j^{-a} : j \in [d])$ and $w^* \in \mathbb{R}^d$ is deterministic with $w_j \asymp j^{-b}$.

Learner: We introduce a notion of model size by assuming the learner only has access to $m \ll d$ random features, i.e., $XS^T \in \mathbb{R}^{n \times m}$, where $S \in \mathbb{R}^{m \times d}$ is a random sketching matrix with $S_{ij} \overset{iid}{\sim} N(0, \frac{1}{m})$, $i \in [d]$, $j \in [m]$.

So the learner is $f_v : \mathbb{R}^d \to \mathbb{R}$ $\quad x \mapsto \langle Sx, v \rangle$ for $v \in \mathbb{R}^m$.

Loss: $\ell(v; x, y) = (y - \langle Sx, v \rangle)^2$

(Lin et al., 2025): Fix dataset $\mathcal{D} = (x_i, y_i)_{i=1}^n$, learn v via multi-pass SGD

$$v_t = v_{t-1} - \gamma_t \nabla \ell(v_{t-1}; x_{i_t}, y_{i_t})$$

$$= v_{t-1} - \gamma_t S x_{i_t} (\langle S x_{i_t}, v_{t-1} \rangle - y_{i_t}), \quad t = 1, \dots, L$$

L: Total number of iterations, $i_t \overset{iid}{\sim} \text{Unif}([n])$

By contrast, GD has updates $\theta_t = \theta_{t-1} - \frac{\gamma_t}{N} \sum_{i=1}^n S x_i (\langle S x_i, \theta_{t-1} \rangle - y_i)$

Risk: $R_m(v_L) = \mathbb{E}_{x,y} \left[(\langle x, S^T v_L \rangle - y)^2 \right]$.

First step: Break down the evolution of the risk over time.

$$\mathbb{E}_{\mathcal{D}, w^*} [R_m(v_L)] = \mathbb{E} \left[(\langle x, S^T v_L \rangle - y)^2 \right]$$

$$= \mathbb{E} \left[(\langle x, S^T v_L - w^* \rangle - \varepsilon)^2 \right] \qquad y = \langle x, w^* \rangle + \varepsilon$$

$$= \sigma^2 + \mathbb{E} \left[\langle x, S^T v_L - w^* \rangle^2 \right]$$

$$= \sigma^2 + \mathbb{E} \left[(\langle x, S^T (v_L - v^*) \rangle + \langle x, S^T v^* - w^* \rangle)^2 \right], \text{ where } v^* = (SHS^T)^{-1} SH w^*$$

$$= \sigma^2 + \mathbb{E} \left[\langle x, S^T v^* - w^* \rangle^2 \right] + \mathbb{E} \left[\langle x, S^T (v_L - v^*) \rangle^2 \right]$$

If you expand, easy to see cross terms have zero mean

$$= \sigma^2 + \underbrace{\mathbb{E} \left[\langle x, S^T v^* - w^* \rangle^2 \right]}_{\text{Error due to sketching}} + \underbrace{\mathbb{E} \left[\langle x, S^T (\theta_L - v^*) \rangle^2 \right]}_{\text{Error due to finite samples}} + \underbrace{\mathbb{E} \left[\langle x, S^T (v_L - \theta_L) \rangle^2 \right]}_{\text{Error due to SGD instead of GD}}$$

$$= \underbrace{\sigma^2}_{\text{Irreducible}} + \underbrace{\mathbb{E} \left[\| S^T v^* - w^* \|_H^2 \right]}_{\text{Approximation}} + \underbrace{\mathbb{E} \left[\| \theta_L - v^* \|_\Sigma^2 \right]}_{\text{Excess}} + \underbrace{\mathbb{E} \left[\| v_L - \theta_L \|_\Sigma^2 \right]}_{\text{Fluctuation}}, \text{ where } \underbrace{\Sigma = SHS^T \in \mathbb{R}^{m \times m}}_{\text{Covariance of sketched features}}$$

Goal: Show power law decay of these sources of error in m, L, n.

One approach: Notice that $\Sigma = SHS^T$ is a random matrix. Use the framework of deterministic equivalents (valid in the regime $m, d \to \infty$ s.t. $\frac{m}{d} = c$) to approximate each error term. (See Paquette et al., 2024.)

Instead, we will discuss the non-asymptotic approach of (Lin et al., 2024; 2025).

Next step: Decompose Excess $\simeq$ Bias + σ^2 Var

$$\text{Excess} = \mathbb{E}\left[\|\theta_L - v^*\|_\Sigma^2\right]$$

$$\text{GD: } \theta_t - v^* = \theta_{t-1} - v^* - \frac{\gamma_t}{N} \sum_{i=1}^{n} S x_i (\langle S x_i, \theta_{t-1} \rangle - y_i)$$

$$= \theta_{t-1} - v^* - \frac{\gamma_t}{N} S X^T (X S^T (\theta_{t-1} - v^*) - \tilde{\varepsilon}) \quad \text{where } \tilde{\varepsilon} = y - X S^T v^*$$

Let $\hat{\Sigma} = \frac{S X^T X S^T}{N}$

$$= (I - \gamma_t \hat{\Sigma})(\theta_{t-1} - v^*) + \frac{\gamma_t}{N} S X^T \tilde{\varepsilon}$$

$$\Rightarrow \theta_L - v^* = \prod_{t=1}^{L} (I - \gamma_t \hat{\Sigma})(\theta_0 - v^*) + V(\hat{\Sigma}) S X^T \tilde{\varepsilon}, \quad \text{where } V(\hat{\Sigma}) = \frac{1}{N} \sum_{t=1}^{L} \gamma_t \prod_{i=t+1}^{L} (I - \gamma_i \hat{\Sigma})$$

Suppose $\theta_0 = 0$. Moreover, one can show $S X^T \perp\!\!\!\perp \tilde{\varepsilon}$ and $\mathbb{E}[\tilde{\varepsilon}_i^2] \simeq \sigma^2$.

$$\Rightarrow \mathbb{E}\left[\|\theta_L - v^*\|_\Sigma^2\right] = \mathbb{E}\left[\left\| \prod_{t=1}^{L} (I - \gamma_t \hat{\Sigma}) v^* \right\|_\Sigma^2\right] + \mathbb{E}\left[\|V(\hat{\Sigma}) S X^T \tilde{\varepsilon}\|_\Sigma^2\right]$$

$$\simeq \text{Bias} + \sigma^2 \text{Var},$$

where $\text{Bias} = \mathbb{E}_{w^*} \mathbb{E}_X \left[\left\| \prod_{t=1}^{L} (I - \gamma_t \hat{\Sigma}) v^* \right\|_\Sigma^2\right]$, $\quad \text{Var} = \mathbb{E}\left[\text{tr}(X S^T V(\hat{\Sigma}) \Sigma V(\hat{\Sigma}) S X^T)\right]$.

Rigorously showing decay rates for each source of error is quite involved, so let us instead focus on intuition.

Recall Capacity condition: $\lambda_i \asymp i^{-a}$ $\forall i \geq 1$ and some $a > 1$

Source condition: For $i \geq 1$, $\mathbb{E}[\lambda_i \langle v_i, w^* \rangle^2] \asymp i^{-b}$ for some $b > 1$.

$\text{Approx} = \mathbb{E}[\|S^T v^* - w^*\|_H^2]$ $\quad S \in \mathbb{R}^{m \times d}$ $\quad v \in \mathbb{R}^m$ $\quad w^* \in \mathbb{R}^d$

Due to sketching, v^* can (at best) only recover the top m components of w (in the appropriate eigenbasis).

Heuristic argument: $\text{Approx} \asymp \sum_{i=m+1}^{\infty} i^{-b} \approx \int_{m+1}^{\infty} x^{-b} dx \asymp m^{1-b}$

Bias is $\mathbb{E}[\|\theta_L - v^*\|_\Sigma^2]$ if there were no label noise.

Consider GD to minimize $\frac{1}{2}\mathbb{E}[\langle Sx, \theta - v^* \rangle^2]$

$$= \frac{1}{2}(\theta - v^*) SHS^T (\theta - v^*)$$

$$= \frac{1}{2} \sum_{i=1}^{m} \tilde{\lambda}_i (\theta_i - v_i^*)^2$$ Assume wlog $(\tilde{\lambda}_i, e_i)_{i=1}^m$ are eigenpairs of SHS^T

Then $|\theta_i - v_i^*|^2 \approx (1 - \gamma \tilde{\lambda}_i)^{2L} \approx e^{-2\gamma \tilde{\lambda}_i L}$ after L iterations of GD

Note: The authors prove the eigenvalues $\tilde{\lambda}_i$ of SHS^T also satisfy power law, i.e., $\tilde{\lambda}_i \asymp i^{-a}$.

$$e^{-2\gamma \tilde{\lambda}_i L} \approx e^{-c} \Leftrightarrow L \approx \frac{c}{2\gamma \tilde{\lambda}_i} \asymp \frac{c i^a}{2\gamma} \Leftrightarrow i \lesssim \left(\frac{2\gamma L}{c}\right)^{\frac{1}{a}}$$

$\Rightarrow$ After L steps, $\asymp (\gamma L)^{\frac{1}{a}}$ components of v^* are learned to high precision.

This leaves an error of $\approx \sum_{i=(\gamma L)^{\frac{1}{a}}+1}^{m} i^{-b} \leq (\gamma L)^{\frac{1-b}{a}}$.

Thm: let $L_{eff} = L/\log L$. Assume $\gamma \lesssim 1/\log n$ and $L_{eff} \lesssim n^2/\gamma$.

$R_m(v_L) = \sigma^2 + \text{Approx} + \text{Bias} + \sigma^2 \text{Var} + \text{Flue}$, with

1) $\text{Approx} \asymp m^{1-b}$

2) $\text{Bias} \lesssim \max\{m^{1-b}, (L_{eff}\gamma)^{(1-b)/a}\}$

$\text{Bias} \gtrsim (L_{eff}\gamma)^{(1-b)/a}$ when $(L_{eff}\gamma)^{1/a} \le m/c$

$\text{Var} \asymp \min\{m, (L_{eff}\gamma)^{1/a}\}/n$.

3) $\mathbb{E}[\text{Fluc}] \lesssim \gamma \log n \cdot \left[(L_{eff}\gamma)^{1/a-1} + \frac{(L_{eff}\gamma)^{1/a}}{n}\right]$

$\mathbb{E}[\text{Fluc}] \gtrsim \gamma (L_{eff}\gamma)^{1/a-1}$ when $L_{eff} \lesssim n/\gamma$ and $(L_{eff}\gamma)^{1/a} \le m/c$ for some $c>0$.

- Emergence and Scaling Laws

Let us move beyond linear regression to study nonlinear targets and feature learning in neural networks.

Analogy: learning many components of target vector $\longrightarrow$ learning many feature vectors

Target: Sum of single-index targets

$$f_*(x) = \sum_{p=1}^{P} a_p \sigma(\langle v_p^*, x \rangle), \quad x \sim N(0, I_d), \qquad \text{No label noise}$$

where $\{v_p^*\}_{p=1}^P$ are orthonormal and $\underbrace{a_p \asymp p^{-\beta} \text{ for all } p \in [P].}_{\text{Source condition}}$

Allow for "extensive width", where $P \to \infty$ as $d \to \infty$.

Ren et al. (2025) assume σ has information exponent > 2 and even.

Ben Arous et al. (2025) assume $\sigma(z) = He_2(z) = z^2 - 1$

Learner: $f(x) = \sum_{k=1}^{m} \|v_k\|^2 \sigma(\langle \bar{v}_k, x \rangle)$, where $\bar{v}_k := \frac{v_k}{\|v_k\|}$.

Algorithm: Online SGD

Loss $\ell(\{v_k\}_{k=1}^m; x) = \frac{1}{2}\left(f_*(x) - f(x; \{v_k\}_{k=1}^m)\right)^2$

$v_k^{(t+1)} = v_k^{(t)} - \eta \nabla_{v_k} \ell(x_t)$, where x_t is a fresh i.i.d. sample at each iteration

Note: When $P=1$ (single-index target), it is known (Ben Arous et al., 2021) that the sample (and thus time) complexity of learning v^* with v is $\eta t \asymp d^{k^*-1}$, where k_* is the information exponent of σ.

$k_* = \mathrm{IE}(\sigma) := \min\{k > 0 : \mathbb{E}_{z\sim N(0,1)}[\sigma(z) He_k(z)] \neq 0\}$,

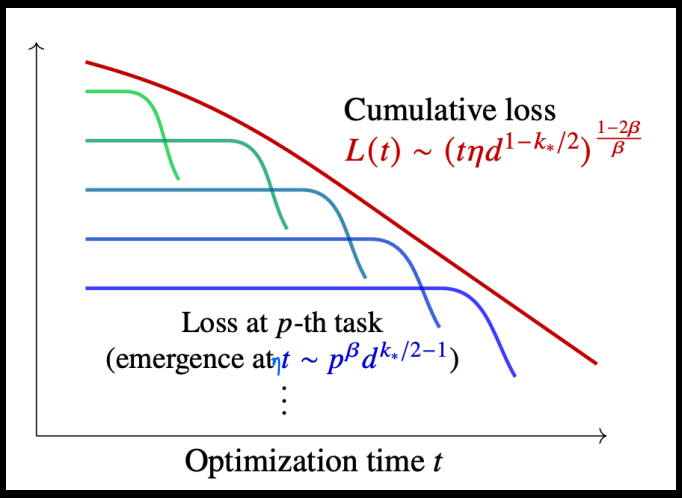
where He_k is the k^{th} (probabilist's) Hermite polynomial.

Thm (Informal): Assuming $\beta > \frac{1}{2}$, $\eta \lesssim d^{-\frac{k^*}{2}} m^{-1} p^{-1}$, we have

(a) Emergence: For $p \le m$, we have $\langle \bar{v}_p, v_p^* \rangle = 1 - o_d(1)$ once $\eta t \asymp p^\beta d^{k^*/2-1}$

Notice if $m = p = 1$ and η is chosen optimally, we return to $t \asymp d^{k^*-1}$.

(b) Scaling Law: The population risk follows the power law $R(t) \asymp \underbrace{(t\eta d^{1-k_*/2})^{\frac{1-2\beta}{\beta}}}_{\text{Excess}} \vee \underbrace{m^{1-2\beta}}_{\text{Approx}}$.



Key idea: Due to small initialization $v_k \sim \mathrm{Unif}(\mathbb{S}^{d-1})$ and square loss, the dynamics of each v_k can be approximately decoupled from the others.

Analysis Steps

1) Population gradient flow: $\dot{v}_k = -\nabla_{v_k} R$

2) Discretization

3) Controlling the fluctuations of SGD (martingale argument)

- Future Directions

  - Incorporating general covariance / capacity constraint into additive target
  - Scaling laws for logistic regression (KC is currently working on this)