Scaling Laws in Linear Regression: Compute, Parameters, and Data

Lin et al.

Gregory Levy

Machine Learning and Economics Reading and Discussion Group

27/01/2025

Motivation: Why Don't Big Models Overfit?

- Empirical observation: test error improves polynomially with model size (M) and data size (N):

$$\mathcal{R}(M,N) \approx \mathcal{R}^* + \frac{c_1}{M^{a_1}} + \frac{c_2}{N^{a_2}}$$

  for irreducible risk $\mathcal{R}^* > 0$, constants $a_1, a_2, c_1, c_2 > 0$ independent of M, N

- But from statistical theory:

$$MSE = \text{Bias}^2 + \text{Variance}$$

  - Bias decreases as $M \uparrow$ (better approximation)
  - Variance increases as $M \uparrow$ (overfitting, memorising noise)

Theory Versus Practice

- **Conflict:** theory predicts "U-shaped" curve:
  - Model improves with size initially
  - Eventually model gets too big for data ⇒ performance crashes
  - But in modern neural networks, we never see the crash

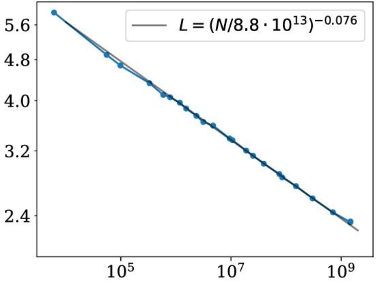

Figure: Scaling law for language models: test loss vs. number of parameters (Kaplan et al., 2020).

Lin et al.: Simplify for Precise Analysis

- Big neural networks are complicated (e.g., transformers)
- **Solution:** focus on simple linear regression case:
    - Input: data x live in infinite-dimensional space ("true", complex world)
    - Model: can only use $M < \infty$ covariates $\Rightarrow$ compress infinite world into M features via Gaussian sketching[1]
    - Training: one-pass SGD

[1] $\tilde{x} = Sx$, where S is $M \times \infty$ random Gaussian matrix

Error Decomposition

- **Decomposition:** risk (error) broken down into:
  - **Approximation Error:** model is too small to represent complexity of data (decreasing in M)
  - **Bias Error:** insufficient data to converge to best solution (decreasing in N)
  - **Variance Error:** "memorising" noise in the specific training samples (usually increasing in M)

Disappearing Variance

- In one-pass SGD, variance term is higher-order ⇒ effectively vanishes
- Why?
  - Implicit regularisation!
  - SGD prefers “minimum norm” solution even in the absence of explicit penalty term in loss (e.g., Ridge/LASSO)[2]
  - So, model can avoid memorising noise even when capacity M is sufficient to do so

[2] Zou et al. (2021).

Disappearing Variance

- Effective production function:

$$\mathcal{R}(M,N) = \mathcal{R}^* + \underbrace{\Theta\left(\frac{1}{M^{a-1}}\right) + \tilde{\Theta}\left(\frac{1}{(N\gamma)^{(a-1)/a}}\right)}_{\text{leading order given by the sum of Approx and Bias}}$$

$$\mathrm{Var} = \underbrace{\tilde{\Theta}\left(\frac{\min\{M, (N\gamma)^{1/a}\}}{N}\right)}_{\text{higher order, thus unobservable}}$$

- Economic interpretation:
  - Returns to M never negative $\Rightarrow$ only limited by approximation (need bigger M), bias (need bigger N), or compute ($C \approx MN$)

Allocating Resources

- Optimisation problem:
  - Budget of compute is C
  - Compute cost $C \approx M \times N$
  - How to choose M, N to minimise risk?
- Solution:
  - Optimal ratio: $M \propto C^{\frac{1}{b+1}}, N \propto C^{\frac{b}{b+1}}$[3]
- Comparison with **Chinchilla**[4]**:**
  - Famous paper suggesting that N and M should scale equally
  - This paper: optimal ratio depends on structure of data through spectral decay b

[3] $b > 1$ controls the decay of the optimal model parameter $\mathbf{w}^*$ (= signal, since $y = \mathbf{x}^T\mathbf{w}^* + \epsilon$), and so is a measure of the difficulty of the task where larger $b \rightarrow$ simpler.

[4] Hoffman et al. (2022).

Takeaways

- **Key points:**
  - Variance term negligible $\Rightarrow$ want to uniformly increase parameters/data
  - However, “optimal” AI production function depends on the structure of the data:
    - Harder problems require larger N (input of intermediate goods?) relative to M (capital?)

Key Limitations

1. Linear Model:
   - Focus on linear setting is tractable, shows scaling applies even in simple settings
   - But most models (NNs) are non-linear: features $\neq$ weights
   - Feature learning ("grokking") is arguably where a lot of interesting stuff happens
2. One-pass SGD:
   - Seeing data once brings theoretical neatness
   - But real models train for multiple epochs
   - Authors admit: multi-pass SGD may cause variance to return, but their theory can't handle extra complexity
3. Data assumptions:
   - Assume data have Gaussian distribution, power-law spectrum
   - Real data could be meaningfully different (heavy-tailed, structured) $\Rightarrow$ exponents in scaling law may not hold

Future Weeks

- Tuesday 10th February (2:30pm): double descent in linear models, presented by Max Kasy
- Tuesday 24th February (2:30pm): empirical scaling laws in LLMs, presented by Thomas Foster
- Tuesday 10th March (2:30pm): scaling and the means of prediction, presented by Aarushi Kalra