Double descent in linear models

Maximilian Kasy

Department of Economics, University of Oxford

Winter 2026

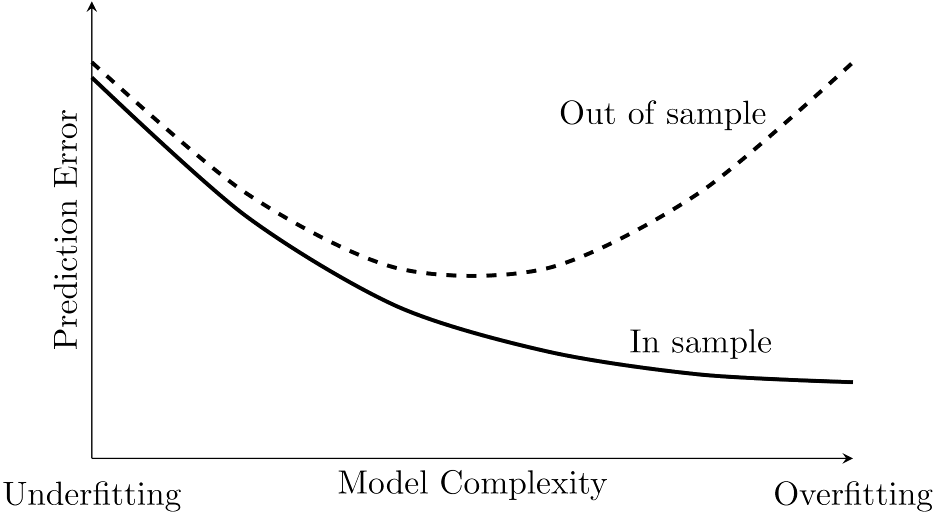
The production function of AI

- How does predictive performance improve as we increase
  - the number of observations n,
  - the model size d?
- $\implies$ Empirical scaling laws (e.g. for LLMs).
- For example (Hoffmann et al., 2022) (*DeepMind*),

$$L(n, d) = \frac{A}{n^a} + \frac{B}{d^b} + L_0.$$

- Prompted "bet of scale" in the industry:
  Keep scaling model size (compute) d, even if data size n is bounded.

Overfitting vs underfitting - the classical picture



Double descent in neural nets

(Belkin et al., 2019)

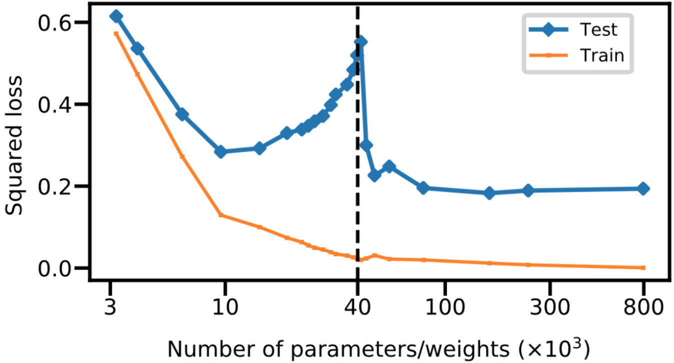


Fig. 3. Double-descent risk curve for a fully connected neural network on MNIST. Shown are training and test risks of a network with a single layer of H hidden units, learned on a subset of MNIST ($n = 4 \cdot 10^3$, $d = 784$, $K = 10$ classes). The number of parameters is $(d + 1) \cdot H + (H + 1) \cdot K$. The interpolation threshold (black dashed line) is observed at $n \cdot K$.

Conundra

1. Is classical learning theory wrong?
2. Is deep learning fundamentally different in some fundamental way?
3. Does double descent mean it pays to keep scaling compute without new data? (A billion dollar question!)

No!

1. With proper regularization and tuning, there is no double descent.
   - e.g. Ridge penalty and tuning using cross validation.
   - e.g. (stochastic) gradient descent with early stopping based on test loss.
2. Double descent only happens when “trained to completion.“
3. Double descent arises equally for linear regression.

(Bach, 2023)

Introduction

The normal linear model

Generalization: Reduced bias

Setup

- Matrix of regressors: $X \in \mathbb{R}^{n \times d}$.
- Vector of outcomes: $y \in \mathbb{R}^n$.
- Vector of coefficients: $\beta \in \mathbb{R}^d$.
- Squared error loss: $R(\beta) = \frac{1}{2}\|X\beta - y\|^2$.

(Following (Bach, 2024), Section 12.2)

OLS regression

- OLS estimator: $\beta^{OLS} = argmin_{\beta} R(\beta)$.
- Full-rank case ($n \geq d$):
$$\beta^{OLS} = (X'X)^{-1}X'y.$$
- Minimum-norm solution for overparameterized case ($n < d$):
$$\beta^{OLS} = X'(XX')^{-1}y.$$

Out of sample prediction error

- Suppose that $x \sim N(0, I)$, and $y|x \sim N(x\beta^*, \sigma^2)$.
- For β non-random:

$$E[\|y - x\beta\|^2] = E[\|y - x\beta^{+x\beta} - x\beta\|^2] = \sigma^2 + \underbrace{\|\beta - \beta^*\|^2}_{\bar{R}(\beta)}.$$

- Decompose the MSE into expected variance and squared bias given X:

$$E[\bar{R}(\beta^{OLS})|X] = E[\|\underbrace{E(\beta^{OLS}|X) - \beta^*}_{\text{Bias}}\|^2] + E[\mathrm{Tr}(\underbrace{Var(\beta^{OLS}|X)}_{\text{Variance}})].$$

- This averages this over the distribution of X.
- Random design, not "fixed design".

Full-rank case

- In the full rank case, $E[\beta^{OLS}] = \beta^*$, and thus
$$E[\bar{R}(\beta^{OLS})|X] = E[\mathrm{Tr}(Var(\beta^{OLS}|X))] = \sigma^2 \cdot E[\mathrm{Tr}((X'X)^{-1})].$$
- The matrix $X'X$ has a Wishart distribution with n degrees of freedom. (`https://en.wikipedia.org/wiki/Wishart_distribution`).
- Since $E[X'X] = n \cdot I$, $E[\mathrm{Tr}((X'X))] = d \cdot n$, and $Tr(E[(X'X)]^{-1}) = \frac{d}{n}$.
- Less obviously, $E[\mathrm{Tr}((X'X)^{-1})] = \frac{d}{n-d-1}$.
- For intuition, note that the inverse is convex, and recall Jensen's inequality.
- It follows that
$$E[\bar{R}(\beta^{OLS})] = \underbrace{\sigma^2 \cdot \frac{d}{n-d-1}}_{\text{Variance}}.$$

The over-parameterised case

- In the over-parameterised case, $\beta^{OLS} = X'(XX')^{-1}y$.
- This estimator is *not* unbiased given X.
- The variance is given by $Var(\beta^{OLS}|X) = \sigma^2 \cdot X'(XX')^{-2}X$, and thus

$$\begin{aligned} E[\mathrm{Tr}(Var(\beta^{OLS}|X))] &= \sigma^2 E[\mathrm{Tr}(X'(XX')^{-2}X)] = \sigma^2 E[\mathrm{Tr}((XX')^{-1})] \\ &= \sigma^2 \frac{n}{d-n-1}. \end{aligned}$$

- Again: Trace of the expectation of an inverse Wishart distribution, but with X and X' switched.

Bias

- As for the bias (given X), we have that $\beta^{OLS} - \beta^* = -P \cdot \beta^*$, where P is the projection $P = I - X'(XX')^{-1}X$.
- By rotational invariance of the distribution of X, $E[\|P\beta^*\|^2]$ depends only on $\|\beta^*\|$, and thus

$$E[\|P\beta^*\|^2] = \|\beta^*\|^2 \cdot \frac{1}{d} \sum_j E[e_j' P^2 e_j] = \|\beta^*\|^2 \cdot \frac{1}{d} E[\mathrm{Tr}(P)].$$

- Since P is a projection matrix on an $d-n$-dimensional subspace, we have $\mathrm{Tr}(P) = d - n$.
- Collecting all our calculations, we get for the overparameterized case that

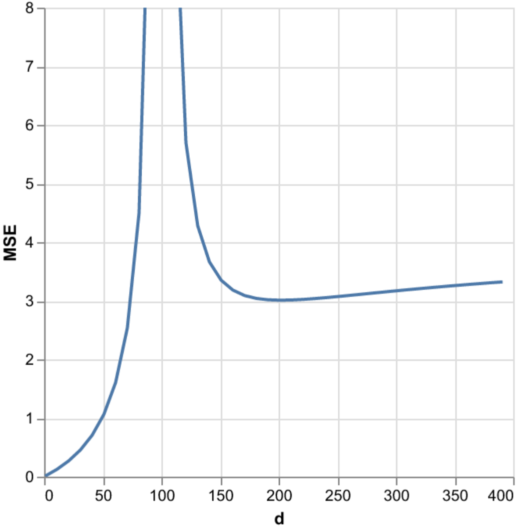
$$E[\bar{R}(\beta^{OLS})] = \underbrace{\sigma^2 \cdot \frac{n}{d-n-1}}_{\text{Variance}} + \underbrace{\|\beta^*\|^2 \cdot \frac{d-n}{d}}_{\text{Bias}}.$$

Putting everything together

$$E[\bar{R}(\beta^{OLS})] = \begin{cases} \sigma^2 \cdot \frac{d}{n-d-1} & d < n \\ \sigma^2 \cdot \frac{n}{d-n-1} + \|\beta^*\|^2 \cdot \frac{d-n}{d} & d > n+1. \end{cases}$$

Illustration

$n = 100$, $\|\beta^*\| = 2$, and $\sigma^2 = 1$:



Introduction

The normal linear model

Generalization: Reduced bias

Generalization: Reduced bias

- Missing from this picture:
  The predictive benefit of adding additional informative regressors.

- Equivalently:
  σ^2 should decrease if we choose larger d.

Subset of regressors

- Suppose that $W \in \mathbb{R}^{n\times m}$ and X is given by the first d columns of W.
- Suppose $w \sim N(0, I_m)$, and $y|w \sim N(w\theta, \tau^2)$.
- Partition $\theta = (\beta, \gamma)$ with dimensions $d, m-d$.
- Then $y|x \sim N(x\beta, \sigma^2)$, where $\sigma^2 = \|\gamma\|^2 + \tau^2$.

Mean square prediction error (MSPE)

- Let $MSPE = \bar{R}(\beta^{OLS}) + \sigma^2$.
- Based on our previous calculations:

$$E[MSPE|\theta] = \begin{cases} (\|\gamma\|^2 + \tau^2) \cdot \left(\frac{d}{n-d-1} + 1\right) & d < n \\ (\|\gamma\|^2 + \tau^2) \cdot \left(\frac{n}{d-n-1} + 1\right) + \|\beta\|^2 \cdot \frac{d-n}{d} & d > n+1. \end{cases}$$

- If we assume that $\theta \sim N(0, \nu^2 \cdot I)$, then

$$E[\|\beta\|^2] = \nu^2 \cdot d, \quad E[\|\gamma\|^2] = \nu^2 \cdot (m-d).$$

Random sketching

- More generally: Power law of coefficients and random sketching.
- Modelling device to mimic increasing model size.
- Cf. (Lin et al., 2025).
- Assumptions:
  - $w \sim N(0, I_m)$, and $y|w \sim N(w\theta, \tau^2)$. (Possibly $m = \infty$.)
  - $\theta_j \approx j^{-\alpha}$. (Power law coefficients.)
  - $x = S \cdot W$ where $S \in \mathbb{R}^{d \times m}$ is a random *sketching matrix*.
- Under these conditions: Geometric decline of $\|\gamma\|^2$ in d.

References

- Bach, F. (2023). High-dimensional analysis of double descent for linear regression with random projections. *arXiv*.

- Bach, F. (2024). *Learning theory from first principles*. MIT press.

- Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, *116*(32), 15849–15854.

- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Las Casas, D. de, Hendricks, L. A., Welbl, J., Clark, A., Hennigan, T., Noland, E., Millican, K., Driessche, G. van den, Damoc, B., Guy, A., Osindero, S., Simonyan, K., Elsen, E., . . . Sifre, L. (2022). *Training compute-optimal large language models*.

- Lin, L., Wu, J., Kakade, S. M., Bartlett, P. L., & Lee, J. D. (2025). Scaling laws in linear regression: Compute, parameters, and data. *arXiv*.

Thank you!