Introduction to Theory of Deep Learning
Lecture 3: Mickey Mouse Proof for Double Descent, Part 1

Background: Expected Trace of Random Projection Matrix

Theorem 1. *Let $X \in \mathbb{R}^{n\times d}$ have i.i.d. $\mathcal{N}(0,1)$ entries. If $d > n+1$, then*

$$\mathbb{E}\Big[X^\top (XX^\top)^{-1} X\Big] \;=\; \frac{n}{d} I_d.$$

Proof. Define

$$M := X^\top (XX^\top)^{-1} X \;\in\; \mathbb{R}^{d\times d}.$$

We wish to compute $\mathbb{E}[M]$.

—

Step 1: Rotational invariance. Let $U \in O(d)$ be any orthogonal matrix. Since X has i.i.d. standard Gaussian entries, XU has the same distribution as X. Then

$$M(XU) = (XU)^\top \big((XU)(XU)^\top\big)^{-1}(XU) = U^\top X^\top (XX^\top)^{-1} XU = U^\top M(X) U.$$

Taking expectations and using distributional equality,

$$\mathbb{E}[M] = \mathbb{E}[M(XU)] = U^\top \mathbb{E}[M] U.$$

Thus, $\mathbb{E}[M]$ commutes with all orthogonal matrices U. By Lemma 1, this implies

$$\mathbb{E}[M] = \alpha I_d$$

for some scalar α.

—

Step 2: Trace computation. To determine α, take traces:

$$\mathrm{Tr}(M) = \mathrm{Tr}\big(X^\top (XX^\top)^{-1} X\big).$$

Using cyclicity of trace,

$$\mathrm{Tr}(M) = \mathrm{Tr}\big((XX^\top)^{-1} XX^\top\big) = \mathrm{Tr}(I_n) = n.$$

Therefore,

$$\mathbb{E}[\mathrm{Tr}(M)] = n.$$

On the other hand,

$$\mathrm{Tr}(\mathbb{E}[M]) = \mathrm{Tr}(\alpha I_d) = \alpha d.$$

Equating gives $\alpha d = n$, so

$$\alpha = \frac{n}{d}.$$

—

Step 3: Conclusion. Thus,

$$\mathbb{E}[M] = \frac{n}{d} I_d.$$

□

Lemma 1 (Schur's Lemma for Real Orthogonal Representations)**.** *Suppose $A \in \mathbb{R}^{d\times d}$ satisfies $UA = AU$ for all $U \in O(d)$. Then $A = \alpha I_d$ for some $\alpha \in \mathbb{R}$.*

Proof. First, since A commutes with all diagonal sign-flip matrices (elements of $O(d)$ of the form $D = \mathrm{diag}(\pm 1, \ldots, \pm 1)$), we deduce that all off-diagonal entries of A must vanish. Indeed, if D flips only the i-th coordinate, then $(DA)_{ij} = -A_{ij}$ while $(AD)_{ij} = A_{ij}$, forcing $A_{ij} = 0$ for $i \neq j$. Thus A is diagonal.

Now, let $R \in O(d)$ be a permutation matrix swapping coordinates i and j. Commutation $AR = RA$ forces the i-th and j-th diagonal entries of A to be equal. By varying i, j, we see all diagonal entries are equal. Hence $A = \alpha I_d$. □

Geometric Intuition

The matrix

$$M = X^\top (XX^\top)^{-1} X$$

is the orthogonal projection in $\mathbb{R}^d$ onto the row space of X. Indeed, for any $v \in \mathbb{R}^d$, we have $Mv = X^\top (XX^\top)^{-1}(Xv)$, which is the least-norm solution w to $Xw = Xv$. Thus Mv is exactly the projection of v onto $\mathrm{Row}(X)$.

Therefore M is an idempotent projection matrix with rank n. Its trace is n, the dimension Rotational invariance ensures that the row space is uniformly distributed among all N-dimensional subspaces of $\mathbb{R}^d$. The expected projection operator onto a random n-dimensional subspace is isotropic, hence equal to $(n/d)I_d$. This provides a geometric explanation of the result.

Mickey Mouse Proof for Double Descent

In modern machine learning, it has been observed that increasing model complexity (e.g. number of parameters) can sometimes *improve* generalization even after reaching a point where the model exactly fits the training data. This phenomenon, known as **double descent**, contradicts the classical U-shaped risk curve from basic learning theory. In this lecture, we start working on a rigorous analysis of double descent in the simple setting of linear regression. We will quantify how the **test risk** (expected error on new data) behaves as a function of model dimension, in the the **under-parameterized regime** (fewer parameters than data points) and the **over-parameterized regime** (more parameters than data).

Assumption 1 (Linear Gaussian Model)**.** *We consider a linear regression model with n training examples. Each data point consists of a feature vector $x_i \in \mathbb{R}^d$ and a scalar response $y_i \in \mathbb{R}$, for $i = 1, \ldots, n$. We assume:*

- *The features x_i are drawn i.i.d. from a d-dimensional Gaussian distribution with mean zero and covariance I_d (the $d \times d$ identity).*
- *The responses follow $y_i = x_i^\top \theta^* + \epsilon_i$, where $\theta^* \in \mathbb{R}^d$ is the (unknown) true parameter vector and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is independent noise.*

We denote by $X \in \mathbb{R}^{n \times d}$ the design matrix whose i-th row is $x_i^\top$, and by $Y \in \mathbb{R}^n$ the vector of responses. The loss function is the squared error, and we measure performance via the ***expected risk*** *$R(\theta) = \mathbb{E}_{(x,y)}[(y - x^\top \theta)^2]$. In this well-specified linear model, the minimal risk (achieved by the Bayes-optimal predictor $f(x) = x^\top \theta^*$) is $R^* = \sigma^2$, and the* ***excess risk*** *of an estimator $\hat{\theta}$ is*

$$R(\hat{\theta}) - R^* = \mathbb{E}_{(x,y)}\big[\, (\hat{\theta} - \theta^*)^\top x x^\top (\hat{\theta} - \theta^*) \,\big] = \mathbb{E}\big[\|\hat{\theta} - \theta^*\|_2^2\big]\,,$$

where the expectation is over both a fresh test example and the training data (we used $\mathbb{E}[xx^\top] = I_d$).

Under Assumption 1, our goal is to analyze the excess risk $\mathbb{E}\big[\|\hat{\theta} - \theta^*\|_2^2\big]$ for the empirical risk minimizer in two regimes: *(a)* when $d < n$ (under-parameterization), and *(b)* when $d > n$ (over-parameterization). We will see that in case (a) the risk increases as the model size d increases (a classical regime of **overfitting** when d is large relative to n), whereas in case (b) increasing d (further over-parameterizing the model) can actually *decrease* the risk again. This non-monotonic behavior as a function of d is the double descent phenomenon.

Under-Parameterized Regime ($n > d$)

When the number of samples exceeds the number of parameters, the empirical least-squares solution is unique and given by the ordinary least squares (OLS) estimator:

$$\hat{\theta}_{\mathrm{OLS}} = \arg\min_{\theta \in \mathbb{R}^d} \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top \theta)^2 = (X^\top X)^{-1} X^\top Y\,,$$

provided $X^\top X$ is invertible. (For $n > d$, $X^\top X$ is invertible almost surely under the Gaussian assumption.) The OLS solution interpolates the training data with zero training error when $d \le n$, and it coincides with the minimum-norm interpolator in that regime.

Proposition 1 (Excess Risk in the Under-Parameterized Case)**.** *Assume $n > d + 1$ (so that $X^\top X$ is invertible and the expectation below is finite). Then the expected excess risk of the OLS estimator is*

$$\mathbb{E}\big[\|\hat{\theta}_{\mathrm{OLS}} - \theta^*\|_2^2\big] = \sigma^2 \frac{d}{n - d - 1}\,.$$

When the number of parameters exceeds the number of samples, the least-squares problem has infinitely many minimizers that achieve zero training error (the training data can be perfectly interpolated). In practice, gradient descent or other implicit algorithms bias the solution toward the minimum ℓ_2-norm solution. We will analyze the *minimum-norm interpolating estimator*:

$$\hat{\theta}_{\mathrm{MN}} = \arg\min\{\|\theta\|_2 : X\theta = Y\}\,.$$

It can be shown that $\hat{\theta}_{\mathrm{MN}} = X^\top (XX^\top)^{-1} Y$ (this is the Moore-Penrose pseudoinverse solution). Equivalently, $\hat{\theta}_{\mathrm{MN}}$ can be written as

$$\hat{\theta}_{\mathrm{MN}} = X^\top (XX^\top)^{-1} X\, \theta^* + X^\top (XX^\top)^{-1} \epsilon\,,$$

since $Y = X\theta^* + \epsilon$. Define the matrix

$$P := X^\top (XX^\top)^{-1} X\,,$$

which is the orthogonal projection onto the column space of $X^\top$ (a subspace of $\mathbb{R}^d$ of dimension n). Notice that P is a $d \times d$ symmetric idempotent matrix ($P^2 = P$) of rank n. Using this notation, we can express the error as

$$\begin{aligned}\hat{\theta}_{\mathrm{MN}} - \theta^* &= P\,\theta^* - \theta^* + X^\top (XX^\top)^{-1}\epsilon \\ &= -\,(I - P)\,\theta^* + X^\top (XX^\top)^{-1}\epsilon\,.\end{aligned}$$

This decomposition separates the estimation error into two parts: a **bias term** $-(I-P)\theta^*$ stemming from the fact that in the over-parameterized regime the estimator cannot recover any component of θ^* lying in the nullspace of X, and a **variance term** $X^\top (XX^\top)^{-1}\epsilon$ due to noise amplification.

Proposition 2 (Excess Risk in the Over-Parameterized Case)**.** *Assume $d > n + 1$. Then the expected excess risk of the minimum-norm interpolator $\hat{\theta}_{\mathrm{MN}}$ is*

$$\mathbb{E}\big[\|\hat{\theta}_{\mathrm{MN}} - \theta^*\|_2^2\big] = \sigma^2 \frac{n}{d-n-1} + \frac{d-n}{d}\,\|\theta^*\|_2^2\,.$$

Proof. Using the decomposition above, let $a := -(I - P)\theta^*$ and $b := X^\top (XX^\top)^{-1}\epsilon$. We have $\hat{\theta}_{\mathrm{MN}} - \theta^* = a + b$. By construction, $\mathbb{E}[b \mid X] = 0$ (since $\mathbb{E}[\epsilon] = 0$ and ϵ is independent of X) and a is deterministic given X. Therefore the cross-term has zero mean:

$$\mathbb{E}[a^\top b] = \mathbb{E}_X\big[\, a^\top \mathbb{E}(b \mid X)\,\big] = 0\,.$$

It follows that

$$\mathbb{E}\big[\|\hat{\theta}_{\mathrm{MN}} - \theta^*\|_2^2\big] = \mathbb{E}\big[\|a\|_2^2\big] + \mathbb{E}\big[\|b\|_2^2\big]\,,$$

i.e. the bias and variance contributions add.

For the variance term: condition on X and compute $\|b\|_2^2 = \epsilon^\top (XX^\top)^{-1}\epsilon$. Taking expectation over ϵ (with X fixed) yields

$$\mathbb{E}\big[\|b\|_2^2 \mid X\big] = \sigma^2\ \mathrm{tr}\Big((XX^\top)^{-1}\Big),$$

since $\mathrm{Var}(\epsilon) = \sigma^2 I_n$. Thus

$$\mathbb{E}\big[\|b\|_2^2\big] = \sigma^2\,\mathbb{E}_X\Big[\,\mathrm{tr}\big((XX^\top)^{-1}\big)\Big]\,.$$

Under our Gaussian model, $XX^\top$ is an $n \times n$ Wishart matrix with d degrees of freedom. By an analogous result to the one used in Proposition 1, we have $\mathbb{E}[(XX^\top)^{-1}] = \frac{1}{d-n-1} I_n$ for $d > n + 1$. Therefore $\mathbb{E}[\mathrm{tr}((XX^\top)^{-1})] = \frac{n}{d-n-1}$. This gives

$$\mathbb{E}\big[\|b\|_2^2\big] = \sigma^2 \frac{n}{d-n-1}\,.$$

For the bias term: note that $a = -(I - P)\theta^*$ is a d-vector. Since P is the projection onto an n-dimensional random subspace of $\mathbb{R}^d$, by symmetry we have

$$\mathbb{E}[P] = \frac{n}{d} I_d\,.$$

(Indeed, for any fixed unit vector $u \in \mathbb{R}^d$, $u^\top P u$ is the squared length of the projection of u onto the n-dimensional subspace $\mathrm{Col}(X^\top)$, which in expectation is n/d by rotational invariance. For details, refer to background material.) Thus $\mathbb{E}[I - P] = I_d - \frac{n}{d} I_d = \frac{d-n}{d} I_d$. It follows that

$$\mathbb{E}\big[\|a\|_2^2\big] = \mathbb{E}\big[\theta^{*T}(I-P)\theta^*\big] = \theta^{*T}\,\mathbb{E}[I-P]\,\theta^* = \frac{d-n}{d}\|\theta^*\|_2^2\,.$$

Combining the two parts, we obtain

$$\mathbb{E}\big[\|\hat{\theta}_{\mathrm{MN}} - \theta^*\|_2^2\big] = \sigma^2 \frac{n}{d-n-1} + \frac{d-n}{d}\,\|\theta^*\|_2^2\,,$$

as claimed. □

Remark 1. The excess risk in the over-parameterized regime consists of a variance term (the first term, decreasing in d) and a bias term (the second term, increasing in d). Just above the interpolation threshold (d slightly larger than n), the variance term is very large (due to the factor $\frac{1}{d-n-1}$) while the bias term is small, so the overall risk is high. As d grows further, the variance term shrinks (since adding more parameters beyond the n data points dilutes the effect of noise), but the bias term grows (since $\hat{\theta}_{\mathrm{MN}}$ cannot recover components of θ^* in directions with no data). This trade-off means that the risk in Proposition 2 will typically decrease for a range of $d > n$ and then eventually increase towards the limit $\|\theta^*\|_2^2$ as $d \to \infty$. In other words, as a function of model size d, the over-parameterized risk exhibits a **U-shape**: it drops after $d = n$ (a "second descent"), achieves a minimum at some larger d, and then rises toward an asymptote.