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

Background: Expected Trace of Inverse Wishart distributed matrices

Model and goal. Let $g_1, \dots, g_n \in \mathbb{R}^d$ be i.i.d. $\mathcal{N}(0, \Sigma)$ and define the Wishart matrix

$$W \;=\; \sum_{i=1}^{n} g_i g_i^T \;\sim\; W_d(n, \Sigma),$$

with $n > d+1$ so that W is invertible a.s. and $\mathbb{E}[W^{-1}]$ exists. We will prove, in a short and modular way, that

$$\mathbb{E}[W^{-1}] \;=\; \frac{1}{n-d-1}\,\Sigma^{-1}.$$

We organize the proof into small lemmas and emphasize the distributional input and expectation step.

Step 1: Reduction to identity covariance

Lemma 1 (Linear change of variables)**.** *Let $h_i := \Sigma^{-1/2} g_i \sim \mathcal{N}(0, I_d)$ and $W_0 := \sum_{i=1}^{n} h_i h_i^T \sim W_d(n, I_d)$. Then*

$$W \;=\; \Sigma^{1/2} W_0 \Sigma^{1/2}, \qquad W^{-1} \;=\; \Sigma^{-1/2} W_0^{-1} \Sigma^{-1/2}.$$

Consequently,

$$\mathbb{E}[W^{-1}] \;=\; \Sigma^{-1/2}\,\mathbb{E}[W_0^{-1}]\,\Sigma^{-1/2}.$$

Hence it suffices to prove $\mathbb{E}[W_0^{-1}] = \frac{1}{n-d-1} I_d$ for the identity case. From now on assume $\Sigma = I_d$.

Step 2: Two linear-algebra ingredients

Step 2: Isotropy $\Rightarrow$ diagonal expectation

Lemma 2 (Diagonal form by sign symmetry)**.** *For $\Sigma = I_d$, $\mathbb{E}[W^{-1}]$ is a scalar multiple of the identity:*

$$\mathbb{E}[W^{-1}] \;=\; \alpha\, I_d, \qquad \alpha \;=\; \frac{1}{d}\,\mathbb{E}\big[\,\mathrm{Tr}(W^{-1})\big].$$

Proof. Let $D = \mathrm{diag}(1, \ldots, 1, -1, 1, \ldots, 1)$ flip any fixed coordinate. Since $g_i \overset{d}{=} Dg_i$, we have $W \overset{d}{=} DWD$ and hence $\mathbb{E}[W^{-1}] = \mathbb{E}[DW^{-1}D] = D\,\mathbb{E}[W^{-1}]\,D$. Comparing off-diagonals gives $\mathbb{E}[W^{-1}]_{ij} = 0$ for $i \neq j$. By coordinate permutation symmetry, all diagonal entries are equal. □

Step 3: Trace as a sum of column norms of $G^\dagger$

Let $G \in \mathbb{R}^{d\times n}$ be the data matrix with columns g_i, so $W = GG^T$. Since $n > d$, G has rank d a.s., its Moore–Penrose pseudoinverse is $G^\dagger = G^T(GG^T)^{-1} = G^T W^{-1} \in \mathbb{R}^{n\times d}$, and

$$W^{-1} \;=\; (GG^T)^{-1} \;=\; G^\dagger (G^\dagger)^T, \qquad \mathrm{Tr}(W^{-1}) \;=\; \|G^\dagger\|_F^2.$$

Writing the columns of $G^\dagger$ as $c_1, \ldots, c_d \in \mathbb{R}^n$,

$$\mathrm{Tr}(W^{-1}) \;=\; \sum_{j=1}^{d} \|c_j\|_2^2.$$

Lemma 3 (Reduction to a Hilbert-space minimum-norm problem)**.** *Let $G \in \mathbb{R}^{d\times n}$ have full row rank d, with rows $z_1^T, \ldots, z_d^T$. Fix $j \in \{1, \ldots, d\}$ and set*

$$\mathcal{T}_j := \mathrm{span}\{z_k :\; k \neq j\} \subset \mathbb{R}^n, \qquad \mathcal{S}_j := \mathcal{T}_j^\perp, \qquad q_j := Q_j z_j,$$

where Q_j is the orthogonal projector onto $\mathcal{S}_j$. Consider the convex problem

$$\min \;\|c\|_2 \quad \textit{subject to} \quad z_k^T c = \delta_{kj} \;\; (k = 1, \ldots, d). \tag{1}$$

Then the feasible set of (1) *equals*

$$\mathcal{F}_j \;=\; \{\, c \in \mathcal{S}_j :\; \langle c, q_j\rangle = 1 \,\},$$

and hence (1) *is equivalent to the Hilbert-space problem*

$$\min_{c\in\mathcal{S}_j} \;\|c\|_2 \quad \textit{subject to} \quad \langle c, q_j\rangle = 1, \tag{2}$$

posed in the inner-product space $(\mathcal{S}_j, \langle\cdot,\cdot\rangle)$ inherited from $\mathbb{R}^n$. Moreover, $q_j \neq 0$ and the unique minimizer of (2) *is*

$$c_j^\star \;=\; \frac{q_j}{\|q_j\|_2^2}.$$

Proof. Step 1: Characterize the feasible set. The linear constraints $z_k^T c = \delta_{kj}$ say precisely that c is orthogonal to z_k for $k \neq j$, and has unit inner product with z_j. Thus any feasible c must lie in

$$\mathcal{S}_j = \{v \in \mathbb{R}^n :\; \langle v, z_k\rangle = 0 \text{ for all } k \neq j\}.$$

Write the orthogonal decomposition $z_j = P_{\mathcal{T}_j} z_j + Q_j z_j = P_{\mathcal{T}_j} z_j + q_j$. If $c \in \mathcal{S}_j$, then $\langle c, P_{\mathcal{T}_j} z_j\rangle = 0$, hence the constraint $\langle c, z_j\rangle = 1$ is equivalent to $\langle c, q_j\rangle = 1$. Therefore the feasible set equals

$$\mathcal{F}_j \;=\; \{\, c \in \mathcal{S}_j :\; \langle c, q_j\rangle = 1 \,\}.$$

Step 2: $q_j \neq 0$. Since G has full row rank d, the $d-1$ rows $\{z_k : k \neq j\}$ are linearly independent, so $\dim(\mathcal{S}_j) = n - (d-1) = n - d + 1 \geq 1$. If $q_j = Q_j z_j = 0$, then $z_j \in \mathcal{T}_j$, making the d rows linearly dependent, which contradicts full row rank. Hence $q_j \neq 0$.

Step 3: Reduction to a one-constraint problem in the Hilbert space $\mathcal{S}_j$. By Step 1, the optimization (1) is exactly (2). Since the objective $c \mapsto \frac{1}{2}\|c\|_2^2$ is strictly convex on the Hilbert space $\mathcal{S}_j$ and the constraint is affine and nonempty (because $q_j \neq 0$), the minimizer exists and is unique.

Step 4: Solve (2) *explicitly.* There are two standard ways:

(a) *Cauchy–Schwarz:* For any feasible $c \in \mathcal{S}_j$,

$$1 \;=\; \langle c, q_j \rangle \;\leq\; \|c\|_2 \, \|q_j\|_2,$$

with equality if and only if c is a nonnegative scalar multiple of q_j. Therefore the unique minimizer has the form $c = \theta q_j$ with $\theta > 0$, and the constraint gives $\theta \|q_j\|_2^2 = 1$, hence $c_j^\star = q_j / \|q_j\|_2^2$.

(b) *Lagrange multipliers in* $\mathcal{S}_j$*:* Minimize $f(c) = \frac{1}{2}\|c\|_2^2$ on $\mathcal{S}_j$ subject to $g(c) = \langle c, q_j \rangle - 1 = 0$. The Lagrangian $L(c, \lambda) = \frac{1}{2}\|c\|_2^2 - \lambda(\langle c, q_j \rangle - 1)$ yields the first-order condition $c = \lambda q_j$; imposing $g(c) = 0$ gives $\lambda = 1/\|q_j\|_2^2$ and therefore $c_j^\star = q_j / \|q_j\|_2^2$.

Both routes give the same unique minimizer $c_j^\star$. □

Consequences. Since every solution of $Gc = e_j$ lies in the affine slice $\mathcal{F}_j$, the Moore–Penrose pseudoinverse column c_j is (as discussed in lecture 1), the unique minimum-norm element in $\mathcal{F}_j$. Hence $c_j = c_j^\star = q_j / \|q_j\|_2^2$, which is the claimed formula

$$c_j \;=\; \frac{Q_j z_j}{\|Q_j z_j\|_2^2}, \qquad \|c_j\|_2^2 \;=\; \frac{1}{\|Q_j z_j\|_2^2}.$$

Step 4: Distributional input (tutorial style)

Lemma 4 (Projected Gaussian is chi-square in the projected norm)**.** *Fix j. Condition on the sigma-field generated by $\{z_k : k \neq j\}$. Then Q_j is a deterministic orthogonal projector onto a subspace of dimension*

$$r \;=\; n - (d-1) \;=\; n - d + 1.$$

Since $z_j \sim \mathcal{N}(0, I_n)$ is independent of $\{z_k : k \neq j\}$, the projected vector $Q_j z_j$ is distributed as a standard Gaussian in that r-dimensional subspace. In particular,

$$\|Q_j z_j\|_2^2 \;\sim\; \chi^2_{n-d+1}.$$

Tutorial proof. (i) Dimension count.) Almost surely the $d-1$ rows $\{z_k : k \neq j\}$ are linearly independent (Gaussian rows are in general position), so their span has dimension $d-1$, hence its orthogonal complement has dimension $r = n - (d-1) = n - d + 1$.

(ii) Rotational invariance.) Conditional on $\{z_k : k \neq j\}$, the matrix Q_j is fixed. Because $z_j \sim \mathcal{N}(0, I_n)$ is independent and spherically symmetric, $Q_j z_j$ is Gaussian with mean 0 and covariance $Q_j I_n Q_j = Q_j$. Thus in any orthonormal basis adapted to range(Q_j), the coordinates of $Q_j z_j$ are i.i.d. $\mathcal{N}(0,1)$ on that r-dimensional subspace and 0 elsewhere.

(iii) Norm square.) Therefore $\|Q_j z_j\|_2^2$ is the sum of squares of r independent standard normals, i.e. χ_r^2 with $r = n - d + 1$. □

Proposition 1 (Mean of the reciprocal chi-square)**.** *If $Y \sim \chi^2_\nu$ with $\nu > 2$, then*

$$\mathbb{E}\Big[\frac{1}{Y}\Big] \;=\; \frac{1}{\nu - 2}.$$

Quick integral proof. The density is $f(y) = \dfrac{1}{2^{\nu/2}\Gamma(\nu/2)}\, y^{\nu/2-1}e^{-y/2}$, $y > 0$. Then

$$\mathbb{E}\Big[\frac{1}{Y}\Big] = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^\infty y^{\nu/2-2}e^{-y/2}\, dy = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \cdot 2^{\nu/2-1}\Gamma\Big(\frac{\nu}{2} - 1\Big) = \frac{1}{2} \cdot \frac{\Gamma(\nu/2-1)}{\Gamma(\nu/2)}.$$

Using $\Gamma(x+1) = x\,\Gamma(x)$ with $x = \nu/2 - 1$ gives $\Gamma(\nu/2) = (\nu/2-1)\Gamma(\nu/2-1)$ and hence

$$\mathbb{E}\Big[\frac{1}{Y}\Big] = \frac{1}{2} \cdot \frac{1}{\nu/2 - 1} = \frac{1}{\nu - 2}.$$

□

Step 5: Put it together

Theorem 1 (Expected inverse, identity case)**.** *Assume $\Sigma = I_d$ and $n > d + 1$. Then*

$$\mathbb{E}[W^{-1}] \;=\; \frac{1}{n - d - 1}\, I_d.$$

Proof. By Lemma 2, $\mathbb{E}[W^{-1}] = \alpha I_d$ with $\alpha = \frac{1}{d}\mathbb{E}[\mathrm{Tr}(W^{-1})]$. From Step 4 and Lemma **??**,

$$\mathrm{Tr}(W^{-1}) \;=\; \sum_{j=1}^{d} \|c_j\|_2^2 \;=\; \sum_{j=1}^{d} \frac{1}{\|Q_j z_j\|_2^2}.$$

By Lemma 4, $\|Q_j z_j\|_2^2 \sim \chi^2_{n-d+1}$; hence by Proposition 1 (valid since $n - d + 1 > 2$),

$$\mathbb{E}\big[\|c_j\|_2^2\big] = \mathbb{E}\Big[\frac{1}{\|Q_j z_j\|_2^2}\Big] = \frac{1}{n - d - 1}.$$

This value is the same for each j, so

$$\mathbb{E}[\mathrm{Tr}(W^{-1})] = \sum_{j=1}^{d} \mathbb{E}\big[\|c_j\|_2^2\big] = d \cdot \frac{1}{n - d - 1}, \qquad \alpha = \frac{1}{d}\mathbb{E}[\mathrm{Tr}(W^{-1})] = \frac{1}{n - d - 1}.$$

Thus $\mathbb{E}[W^{-1}] = \frac{1}{n-d-1} I_d$. □

Theorem 2 (Expected inverse, general covariance)**.** *For $W \sim W_d(n, \Sigma)$ with $n > d + 1$,*

$$\mathbb{E}[W^{-1}] \;=\; \frac{1}{n - d - 1}\, \Sigma^{-1}.$$

Proof. Combine Lemma 1 with Theorem 1. □

Remark 1. The identity $\mathrm{Tr}(W^{-1}) = \|G^\dagger\|_F^2 = \sum_{j=1}^d \|c_j\|_2^2$ together with $c_j = (Q_j z_j)/\|Q_j z_j\|_2^2$ is a version of the so-called negative second moment identity.

Lemma 5. *If $Z_1, Z_2, \ldots, Z_m$ are independent standard normal random variables $Z_i \sim \mathcal{N}(0,1)$, then the sum of their squares,*

$$X := Z_1^2 + Z_2^2 + \cdots + Z_m^2,$$

is distributed as a chi-square random variable with m degrees of freedom (denoted χ_m^2). In particular, the probability density function (pdf) of X is

$$f_X(x) = \frac{1}{2^{m/2}\,\Gamma(m/2)}\, x^{\,m/2-1}\, e^{-\,x/2}, \qquad x > 0\,,$$

and the moment-generating function of X is $M_X(t) = \mathbb{E}[e^{tX}] = (1-2t)^{-m/2}$ for $t < \frac{1}{2}$.

Proof. The case $m = 1$ is straightforward: if $Z \sim \mathcal{N}(0,1)$, then $X = Z^2$ has the χ_1^2 distribution. To see this, note that $\Pr(X \le x) = \Pr(-\sqrt{x} \le Z \le \sqrt{x})$. Differentiating the CDF to obtain the pdf, for $x > 0$ we get

$$f_X(x) = \frac{d}{dx}\Pr(|Z| \le \sqrt{x}) = \frac{d}{dx}\big[2\Phi(\sqrt{x}) - 1\big] = \frac{1}{\sqrt{2\pi}}\Big(e^{-\,(\sqrt{x})^2/2}\frac{1}{2\sqrt{x}} + e^{-\,(-\sqrt{x})^2/2}\frac{1}{2\sqrt{x}}\Big) = \frac{1}{\sqrt{2\pi}}\, x^{-1/2} e^{-x/2}\,,$$

since the $N(0,1)$ density is $\frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ and by symmetry the two endpoints give the same contribution. This indeed matches the χ_1^2 density given in the formula (using $\Gamma(1/2) = \sqrt{\pi}$, one can simplify $2^{-1/2}\Gamma(1/2) = \sqrt{\frac{1}{2}}\,\sqrt{\pi} = \sqrt{\frac{\pi}{2}}$, so $1/(2^{1/2}\Gamma(1/2)) = 1/\sqrt{2\pi}$).

For general m, one convenient approach is to use the moment-generating function. Because $Z_1, \ldots, Z_m$ are independent, the MGF of the sum $X = \sum_{i=1}^m Z_i^2$ factors as

$$M_X(t) = \mathbb{E}[e^{tX}] = \prod_{i=1}^m \mathbb{E}[e^{tZ_i^2}] = \Big(\mathbb{E}[e^{tZ_1^2}]\Big)^m.$$

Now for a single standard normal $Z \sim \mathcal{N}(0,1)$, we have

$$\mathbb{E}[e^{tZ^2}] = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{tz^2}\, e^{-z^2/2}\, dz = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} \exp\Big\{-\frac{1}{2}(1-2t)z^2\Big\} dz\,,$$

which converges for $t < \frac{1}{2}$. Evaluating this Gaussian integral gives

$$\mathbb{E}[e^{tZ^2}] = \frac{1}{\sqrt{2\pi}}\sqrt{\frac{2\pi}{1-2t}} = (1-2t)^{-1/2}\,.$$

Therefore $M_X(t) = (1-2t)^{-m/2}$ for $t < 1/2$. But $(1-2t)^{-m/2}$ is exactly the known MGF of the χ_m^2 distribution. (We can confirm this by moment-generating the χ_m^2 density formula: for $X \sim \chi_m^2$ one finds $\mathbb{E}[e^{\,tX}] = (1-2t)^{-m/2}$ as well, using the identity $\int_0^\infty x^{k-1}e^{-ux}\,dx = \frac{\Gamma(k)}{u^k}$ valid for $\Re(u) > 0$.) Since the moment-generating function uniquely characterizes the distribution, we conclude that X indeed has the χ_m^2 distribution with density as given above. □

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 provide 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 both the **under-parameterized regime** (fewer parameters than data points) and the **over-parameterized regime** (more parameters than data). Our derivation will highlight the roles of variance and bias in these regimes, and connect the peak in risk at the *interpolation threshold* (when the number of parameters equals the number of data) to the behavior of the smallest singular values of the data matrix.

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:

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

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 $d \leq n$, and it coincides with the minimum-norm interpolator in that regime.

Proposition 2 (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*

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

Proof. Conditioning on X, the OLS estimator is an unbiased estimator of θ^* (since $\mathbb{E}[Y \mid X] = X\theta^*$). Thus $\mathbb{E}[\hat{\theta}_{\mathrm{OLS}}] = \theta^*$. The excess risk can then be written in terms of the estimator's variance:

$$\mathbb{E}\big[\|\hat{\theta}_{\mathrm{OLS}} - \theta^*\|_2^2\big] = \mathrm{tr}\Big(\,\mathrm{Var}(\hat{\theta}_{\mathrm{OLS}})\Big)\,.$$

Using the formula for OLS, $\hat{\theta}_{\mathrm{OLS}} = (X^\top X)^{-1}X^\top Y = (X^\top X)^{-1}X^\top(X\theta^* + \epsilon)$, we have

$$\hat{\theta}_{\mathrm{OLS}} - \theta^* = (X^\top X)^{-1}X^\top \epsilon\,.$$

Conditioning on X, the covariance is

$$\mathrm{Var}(\hat{\theta}_{\mathrm{OLS}} \mid X) = (X^\top X)^{-1}X^\top \,\mathrm{Var}(\epsilon) X (X^\top X)^{-1} = \sigma^2 (X^\top X)^{-1},$$

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

$$\mathrm{Var}(\hat{\theta}_{\mathrm{OLS}}) = \mathbb{E}_X\big[\,\mathrm{Var}(\hat{\theta}_{\mathrm{OLS}} \mid X)\big] = \sigma^2\,\mathbb{E}\big[(X^\top X)^{-1}\big]\,.$$

In Lecture 2, we showed that for a Wishart matrix $W = X^\top X$ (with n degrees of freedom and covariance I_d), $\mathbb{E}[W^{-1}] = \frac{1}{n-d-1} I_d$ when $n > d+1$. Hence $\mathbb{E}[\mathrm{tr}((X^\top X)^{-1})] = \frac{d}{n-d-1}$. Substituting back gives the result:

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

□

Remark 2. Proposition 2 shows that the test error (excess risk) *increases* as the model size d increases in the under-parameterized regime. In particular, for fixed n, $\frac{d}{n-d-1}$ is an increasing function of d. Intuitively, adding more features/parameters without regularization increases the **variance** of the estimator (since it can fit more noise), while here there is no benefit in bias reduction because the true model θ^* already lies in the parameter space for any d considered. This is the classical picture of **overfitting**: as d grows (approaching n), the model becomes more flexible and training error decreases, but the expected test error grows.

Over-Parameterized Regime ($d > n$)

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 3 (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 2, 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.) 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 3. 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 3 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.