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

Background: Expected Trace of Inverse Wishart distributed matrices

Theorem 1 (Sherman–Morrison Theorem)**.** *If A is an invertible $n \times n$ matrix and u, v are column vectors in $\mathbb{R}^n$ such that $1 + v^T A^{-1} u \neq 0$, then $A + uv^T$ is invertible and its inverse is given by*

$$(A + uv^T)^{-1} = A^{-1} - \frac{A^{-1} u\, v^T A^{-1}}{1 + v^T A^{-1} u}\,.$$

Proof. We first "guess" the formula as follows[1]. Note that it is sufficient to prove the Sherman-Morrison Formula for $A = I$. Indeed suppose that we proved it for $A = I$ then

$$\begin{aligned}(A + \mathbf{u}\mathbf{v}^T)^{-1} = (A(I + (A^{-1}\mathbf{u})\mathbf{v}^T))^{-1} &= \left(I - \frac{(A^{-1}\mathbf{u})\mathbf{v}^T}{1 + \mathbf{v}^T(A^{-1}\mathbf{u})}\right) A^{-1} \\ &= A^{-1} - \frac{A^{-1}\mathbf{u}\mathbf{v}^T A^{-1}}{1 + \mathbf{v}^T A^{-1}\mathbf{u}}.\end{aligned}$$

Here is a derivation of the Sherman-Morrison Formula for the case $A = I$ when $\|u\| < 1$ and $\|v\| < 1$.

$$\begin{aligned}(I + \mathbf{u}\mathbf{v}^T)^{-1} &= \sum_{k=0}^{\infty} (-1)^k (\mathbf{u}\mathbf{v}^T)^k = I + \sum_{k=1}^{\infty} (-1)^k (\mathbf{u}\mathbf{v}^T)^k \\ &= I + \mathbf{u}\left(\sum_{k=1}^{\infty} (-1)^k (\mathbf{v}^T\mathbf{u})^{k-1}\right)\mathbf{v}^T = I - \mathbf{u} \times \frac{1}{1 + \mathbf{v}^T\mathbf{u}} \times \mathbf{v}^T \\ &= I - \frac{\mathbf{u}\mathbf{v}^T}{1 + \mathbf{v}^T\mathbf{u}}\end{aligned}$$

Now we rigorously verify that the proposed inverse indeed satisfies $(A+uv^T)\left(A^{-1} - \frac{A^{-1}u\,v^T A^{-1}}{1+v^T A^{-1}u}\right) = I$ (and likewise for the reversed order, since matrix inverses are unique). Let

$$B := A^{-1} - \frac{A^{-1} u\, v^T A^{-1}}{1 + v^T A^{-1} u}\,.$$

We compute the product $(A + uv^T)B$ step by step:

[1] `https://math.stackexchange.com/questions/1705671/proof-of-the-sherman-morrison-formula`

Distribute B:

$$(A + uv^T)B = AB + (uv^T)B\,.$$

Simplify AB: By definition of B,

$$AB = AA^{-1} - A\Big(\frac{A^{-1}u\,v^TA^{-1}}{1+c}\Big) = I - \frac{AA^{-1}u\,v^TA^{-1}}{1+c} = I - \frac{u\,v^TA^{-1}}{1+c}\,,$$

where $c := v^TA^{-1}u$.

Simplify $(uv^T)B$: First note that $v^TA^{-1}u = c$ is a scalar. Then

$$(uv^T)B = u\Big(v^TA^{-1} - \frac{v^TA^{-1}u\,v^TA^{-1}}{1+c}\Big) = u\,v^TA^{-1} - \frac{c}{1+c}\,u\,v^TA^{-1}\,.$$

Simplifying the scalar coefficient:

$$u\,v^TA^{-1} - \frac{c}{1+c}\,u\,v^TA^{-1} = \frac{1+c-c}{1+c}u\,v^TA^{-1} = \frac{1}{1+c}u\,v^TA^{-1}\,.$$

Combine the terms: Now add the results of the previous steps:

$$(A + uv^T)B = \Big(I - \frac{u\,v^TA^{-1}}{1+c}\Big) + \frac{1}{1+c}u\,v^TA^{-1} = I + \Big(-\frac{1}{1+c} + \frac{1}{1+c}\Big)u\,v^TA^{-1} = I\,,$$

since the two terms involving uv^TA^{-1} cancel out exactly.

By a nearly identical computation for $B(A + uv^T)$ (or by the symmetry of the argument, as $1 + v^TA^{-1}u$ is scalar), one finds $B(A + uv^T) = I$ as well. Hence $B = (A + uv^T)^{-1}$, proving the Sherman–Morrison formula. □

Theorem 2. *Let $W = XX^T$ be an $N \times N$ Wishart random matrix with P degrees of freedom. Equivalently, $X \in \mathbb{R}^{N\times P}$ has independent columns $x_1, \ldots, x_P \sim \mathcal{N}(0, I_N)$. Assume $P > N + 1$ so that W is almost surely invertible and $\mathbb{E}[W^{-1}]$ exists. Then*

$$\mathbb{E}[\mathrm{Tr}(W^{-1})] \;=\; \frac{N}{P - N - 1}\,.$$

Proof. For each $j \in \{1, \ldots, P\}$, let Q_j denote the orthogonal projection onto the subspace orthogonal to all columns except x_j. In other words, Q_j projects onto the subspace orthogonal to $\mathrm{span}\{x_i : i \neq j\}$ (the column space spanned by all x_i with $i \neq j$). Equivalently, $P_j := I - Q_j$ is the projection onto the subspace $\mathrm{span}\{x_i : i \neq j\}$ itself, so P_jx_j is the component of x_j in the span of the other columns, and $Q_jx_j = x_j - P_jx_j$ is the component of x_j orthogonal to all the other x_i.

Lemma 1. *For each $j \in \{1, \ldots, P\}$, the vector $W^{-1}x_j$ lies in the direction of Q_jx_j. In fact,*

$$W^{-1}x_j = \frac{Q_jx_j}{\|Q_jx_j\|_2^2}\,.$$

Proof. Intuitively, since W^{-1} is the inverse of the Gram matrix W, the vector $W^{-1}x_j$ should "zero out" all correlations with columns x_i for $i \neq j$, leaving only a component in the unique direction independent of those columns. We can make this rigorous in two ways:
(A) Using orthogonality conditions: Because W is invertible (assume the x_i are linearly independent so that W is full-rank), we have $W^{-1}W = I$. If we assemble the column vectors into a matrix

$X := [x_1 \quad x_2 \quad \cdots \quad x_P]$ (so $W = XX^T$), then invertibility of W also implies X is square and invertible (hence P equals the dimension of each x_i). One can verify the identity:

$$X^T W^{-1} X = I_P,$$

which can be seen by multiplying X^T on the left of $W^{-1}W = I$ and X on the right (or by noting $W^{-1} = (XX^T)^{-1} = X^{-T}X^{-1}$ when X is invertible). The (i,j)-entry of $X^T W^{-1} X = I$ is exactly $x_i^T W^{-1} x_j$. Therefore:

- For $i \neq j$: $x_i^T (W^{-1} x_j) = 0$. This means $W^{-1}x_j$ is orthogonal to every other column x_i $(i \neq j)$.
- For $i = j$: $x_j^T (W^{-1} x_j) = 1$. This is a normalization condition ensuring $W^{-1}x_j$ has inner product 1 with x_j itself.

These two conditions uniquely characterize $W^{-1}x_j$ as the vector in the one-dimensional subspace $\text{span}\{Q_j x_j\}$ that has unit inner product with x_j. In particular, $W^{-1}x_j$ must be proportional to $Q_j x_j$ (since $Q_j x_j$ is, by definition, orthogonal to all other x_i).
(B) Using block matrix inversion / Sherman–Morrison: We can also derive this result by viewing W as a rank-1 update of the matrix containing all columns except x_j. Let $W_{(j)} := \sum_{i \neq j} x_i x_i^T$ be the Gram matrix of all columns except x_j. Then $W = W_{(j)} + x_j x_j^T$. By the Sherman–Morrison rank-1 update formula (proven later in Theorem 1), the inverse of W can be written as:

$$W^{-1} = W_{(j)}^{-1} - \frac{W_{(j)}^{-1} x_j x_j^T W_{(j)}^{-1}}{1 + x_j^T W_{(j)}^{-1} x_j}. \qquad (**)$$

Now, apply this inverse to x_j:

$$W^{-1}x_j = W_{(j)}^{-1}x_j - \frac{W_{(j)}^{-1}x_j \left(x_j^T W_{(j)}^{-1} x_j\right)}{1 + x_j^T W_{(j)}^{-1} x_j}.$$

Let $\alpha := x_j^T W_{(j)}^{-1} x_j$ (a scalar). The second term on the right is $\frac{\alpha}{1+\alpha} W_{(j)}^{-1} x_j$. Thus,

$$W^{-1}x_j = \Big(1 - \frac{\alpha}{1+\alpha}\Big) W_{(j)}^{-1} x_j = \frac{1}{1+\alpha} W_{(j)}^{-1} x_j.$$

This shows $W^{-1}x_j$ is collinear with $W_{(j)}^{-1}x_j$. But note that $W_{(j)}$ acts on the subspace spanned by $\{x_i : i \neq j\}$, and $W_{(j)}^{-1}x_j$ essentially produces the unique vector in $\text{span}\{x_i : i \neq j\}$ that $W_{(j)}$ would map to the projection $P_j x_j$. In fact, $W_{(j)}(W_{(j)}^{-1}x_j) = P_j x_j$ (since $W_{(j)}$ cannot "see" the component of x_j orthogonal to the span of the other columns). This implies $W_{(j)}^{-1}x_j$ lies in the subspace spanned by the other columns (the range of $W_{(j)}^{-1}$) and satisfies $W_{(j)}(W_{(j)}^{-1}x_j) = P_j x_j$. Now consider $Q_j x_j = x_j - P_j x_j$, the portion of x_j in the nullspace of $W_{(j)}$. Because $W_{(j)}^{-1}x_j$ has no component in that nullspace (it lies entirely in the column space of the other x_i), the vector $W^{-1}x_j = \frac{1}{1+\alpha} W_{(j)}^{-1} x_j$ is orthogonal to $Q_j x_j$. In other words, all of $W^{-1}x_j$ lies in the subspace orthogonal to $\text{span}\{x_i : i \neq j\}$. The only nonzero component of $W^{-1}x_j$ is along the $Q_j x_j$ direction. (This conclusion is consistent with what we found in (A).)

Regardless of the approach, we have established that $W^{-1}x_j$ is parallel to $Q_j x_j$. Therefore, we can write

$$W^{-1}x_j = c_j\,(Q_j x_j)$$

for some scalar c_j depending on j. To determine c_j, we use the normalization condition $x_j^T(W^{-1}x_j) = 1$ (from approach (A) or by multiplying equation $(**)$ on the right by x_j and noting $x_j^T W_{(j)}^{-1} x_j = \alpha$):

$$1 = x_j^T(W^{-1}x_j) = x_j^T(c_j\,Q_j x_j) = c_j\,x_j^T(Q_j x_j)\,.$$

Since Q_j is an orthogonal projection, $x_j^T(Q_j x_j) = (Q_j x_j)^T(Q_j x_j) = \|Q_j x_j\|_2^2$ is just the squared norm of the orthogonal component of x_j. Thus,

$$c_j\,\|Q_j x_j\|_2^2 = 1 \quad\Longrightarrow\quad c_j = \frac{1}{\|Q_j x_j\|_2^2}\,.$$

So we obtain the explicit formula for the inverse acting on a column x_j:

$$W^{-1}x_j = \frac{Q_j x_j}{\|Q_j x_j\|_2^2}\,.$$

In words, $W^{-1}x_j$ is the unique vector proportional to $Q_j x_j$ having inner product 1 with x_j. (This matches our expectations: $W^{-1}x_j$ is pointing along the part of x_j that is independent of the other columns, scaled so that it "picks out" x_j from the mix when multiplied by W.) □

Now, take Euclidean norms on both sides of the formula above. Using the fact that scaling a vector scales its norm by the absolute value of the scalar, we get:

$$\|W^{-1}x_j\|_2 = \frac{\|Q_j x_j\|_2}{\|Q_j x_j\|_2^2} = \frac{1}{\|Q_j x_j\|_2}\,.$$

Squaring both sides yields the neat relationship:

$$\|W^{-1}x_j\|_2^2 = \frac{1}{\|Q_j x_j\|_2^2}\,. \tag{*}$$

Finally, we sum equation (*) over $j = 1$ to P and use the cyclic property of the trace (noting that $\|W^{-1}x_j\|_2^2 = x_j^T(W^{-1})^T W^{-1}x_j = x_j^T W^{-1}W^{-1}x_j$, and $\sum_j x_j x_j^T = W$):

$$\sum_{j=1}^{P}\|W^{-1}x_j\|_2^2 = \sum_{j=1}^{P}\frac{1}{\|Q_j x_j\|_2^2}\,.$$

But notice that $\sum_{j=1}^{P}\|W^{-1}x_j\|_2^2$ is exactly the trace of W^{-1} when W^{-1} is multiplied by W on either side. More explicitly, using the identity $\mathrm{Tr}(AB) = \mathrm{Tr}(BA)$ and letting $E_{jj} = x_j x_j^T$ (the matrix with x_j on the jth column and zeros elsewhere), we have:

$$\begin{aligned}\mathrm{Tr}(W^{-1}) = \mathrm{Tr}(W^{-1}W) &= \mathrm{Tr}\Big(W^{-1}\sum_{j=1}^{P}x_j x_j^T\Big)\\ &= \mathrm{Tr}\Big(\sum_{j=1}^{P}W^{-1}x_j x_j^T\Big) = \sum_{j=1}^{P}\mathrm{Tr}\big(W^{-1}x_j x_j^T\big)\,.\end{aligned}$$

Since $W^{-1}x_j x_j^T$ is a rank-1 matrix with the single nonzero column $W^{-1}x_j$ and corresponding row x_j^T, its trace is $x_j^T(W^{-1}x_j) = 1$. Therefore each term in the sum is 1, and $\mathrm{Tr}(W^{-1}) = P$. On the other hand, summing (*) gives $\sum_j \|W^{-1}x_j\|_2^2 = \sum_j \frac{1}{\|Q_j x_j\|_2^2}$. We conclude:

$$\mathrm{Tr}(W^{-1}) = \sum_{j=1}^{P} \frac{1}{\|Q_j x_j\|_2^2}\,.$$

Alternatively, apply Sherman-Morrison to W^{-1}: in our context, set $A = W_{(j)} = \sum_{i\neq j} x_i x_i^T$ and $u = v = x_j$. The formula then reads:

$$(W_{(j)} + x_j x_j^T)^{-1} = W_{(j)}^{-1} - \frac{W_{(j)}^{-1} x_j\, x_j^T W_{(j)}^{-1}}{1 + x_j^T W_{(j)}^{-1} x_j}\,.$$

This is exactly formula (**) used above. It is worth noting that in this scenario $W_{(j)}$ is singular as a matrix on the full space (since x_j lies in its nullspace), but one can interpret $W_{(j)}^{-1}$ as a well-defined inverse on the $(P-1)$-dimensional subspace $\mathrm{span}\{x_i : i \neq j\}$. The Sherman–Morrison result still applies as a limit or by working in a basis adapted to the decomposition of x_j into $P_j x_j$ and $Q_j x_j$. The conclusion, as we saw, is that all of the effect of inverting W relative to inverting $W_{(j)}$ is concentrated in the $Q_j x_j$ direction, yielding the formula $W^{-1}x_j = \frac{Q_j x_j}{\|Q_j x_j\|_2^2}$ rigorously.

Summing over $j = 1$ to P, and using the cyclic property of trace, we obtain

$$\mathrm{Tr}(W^{-1}) \;=\; \sum_{j=1}^{P} \|W^{-1}x_j\|_2^2 \;=\; \sum_{j=1}^{P} \frac{1}{\|Q_j x_j\|_2^2}\,.$$

Next, by symmetry and independence of the columns, each random variable $\|Q_j x_j\|_2^2$ has the same chi-square distribution. In fact, one can show that

$$\|Q_j x_j\|_2^2 \;\sim\; \chi^2_{P-N+1}\,,$$

a chi-square with $P - N + 1$ degrees of freedom. (Intuitively, this holds because $Q_j x_j$ is the component of a standard Gaussian $x_j \in \mathbb{R}^N$ that lies outside the span of $P-1$ independent Gaussian vectors x_i $(i \neq j)$, so the dimension of the subspace for $Q_j x_j$ is $N-(P-1) = P-N+1$.) Consequently, $\frac{1}{\|Q_j x_j\|_2^2}$ has the distribution of the reciprocal of a χ^2_{P-N+1} variable. It is well-known that if $Y \sim \chi^2_\nu$ with $\nu > 2$, then $\mathbb{E}[Y^{-1}] = \frac{1}{\nu-2}$. Applying this to $Y = \|Q_j x_j\|_2^2$ (with $\nu = P-N+1$) gives

$$\mathbb{E}\left[\frac{1}{\|Q_j x_j\|_2^2}\right] \;=\; \frac{1}{(P-N+1)-2} \;=\; \frac{1}{P-N-1}\,.$$

Now, an important symmetry of the Wishart distribution (with identity covariance) is that it is isotropic on $\mathbb{R}^N$. This implies $\mathbb{E}[W^{-1}]$ is a scalar multiple of the identity matrix I_N (and in particular, all off-diagonal entries of $\mathbb{E}[W^{-1}]$ are zero and all diagonal entries are equal). Let $c = \mathbb{E}[(W^{-1})_{ii}]$ for any fixed i. Since $\mathrm{Tr}(W^{-1}) = \sum_{i=1}^{N}(W^{-1})_{ii}$, taking expectation gives $\mathbb{E}[\mathrm{Tr}(W^{-1})] = N\,c$. On the other hand, from our previous calculation we also have

$$\mathbb{E}[\mathrm{Tr}(W^{-1})] \;=\; \sum_{j=1}^{P} \mathbb{E}\Big[\frac{1}{\|Q_j x_j\|_2^2}\Big] \;=\; P \cdot \frac{1}{P-N-1}\,.$$

Equating these two expressions for $\mathbb{E}[\mathrm{Tr}(W^{-1})]$ and solving for c gives $c = \frac{1}{P-N-1}$. Therefore $\mathbb{E}[W^{-1}] = \frac{1}{P-N-1} I_N$, and multiplying by N yields the stated formula for $\mathbb{E}[\mathrm{Tr}(W^{-1})]$. □

Lemma 2. *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^2_m). 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^2_1 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^2_1 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^2_m distribution. (We can confirm this by moment-generating the χ^2_m density formula: for $X \sim \chi^2_m$ 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^2_m distribution with density as given above. □

Proposition 1. *Let U be a chi-square distributed random variable with ν degrees of freedom ($U \sim \chi^2_\nu$). If $\nu > 2$, then U has a finite mean for its reciprocal, given by*

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

(If $\nu \le 2$, then $\mathbb{E}[1/U]$ diverges to $+\infty$.)

Proof. Since $U \sim \chi^2_\nu$, its probability density is $f_U(u) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)}\, u^{\nu/2-1} e^{-u/2}$ for $u > 0$. Then for $\nu > 2$ we can compute

$$\mathbb{E}\Big[\frac{1}{U}\Big] = \int_0^\infty \frac{1}{u}\, f_U(u)\, du = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \int_0^\infty u^{\nu/2-2} e^{-u/2}\, du.$$

Make the change of variable $t = u/2$, so that $u = 2t$ and $du = 2\,dt$. The integral becomes

$$\int_0^\infty u^{\nu/2-2} e^{-u/2}\, du = \int_0^\infty (2t)^{\nu/2-2} e^{-t}\, 2\, dt = 2^{\nu/2-1} \int_0^\infty t^{\nu/2-2} e^{-t}\, dt = 2^{\nu/2-1}\, \Gamma\Big(\frac{\nu}{2} - 1\Big)\,,$$

which is valid provided $\nu/2 - 2 > -1$ (i.e. $\nu > 2$) so that the integral converges. Therefore

$$\mathbb{E}[1/U] = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} \cdot 2^{\nu/2-1}\, \Gamma\Big(\frac{\nu}{2} - 1\Big) = \frac{\Gamma(\frac{\nu}{2} - 1)}{\Gamma(\frac{\nu}{2})} \cdot \frac{2^{\nu/2-1}}{2^{\nu/2}} = \frac{1}{2}\, \frac{\Gamma(\frac{\nu}{2} - 1)}{\Gamma(\frac{\nu}{2})} = \frac{\Gamma(\frac{\nu}{2} - 1)}{\Gamma(\frac{\nu}{2})}\,.$$

Using the identity $\Gamma(x+1) = x\,\Gamma(x)$ with $x = \frac{\nu}{2} - 1$, we have $\Gamma(\frac{\nu}{2}) = (\frac{\nu}{2} - 1)\,\Gamma(\frac{\nu}{2} - 1)$. Hence

$$\frac{\Gamma(\frac{\nu}{2} - 1)}{\Gamma(\frac{\nu}{2})} = \frac{1}{\frac{\nu}{2} - 1} = \frac{2}{\nu - 2}\,.$$

Finally, note that $\frac{2}{\nu-2}$ simplifies to $\frac{1}{\nu-2}$ (since $\nu - 2$ in the denominator has the same dimension as ν which is an integer; indeed $\frac{2}{\nu-2} = \frac{1}{(\nu-2)/2} = \frac{1}{\frac{\nu}{2}-1}$ as above, and recall $\nu/2 - 1$ was an integer for even ν, but even if ν is not integer the algebra still yields $1/(\nu-2)$ in simplest terms). Thus we conclude $\mathbb{E}[1/U] = 1/(\nu - 2)$, as claimed.

For completeness, we note that when $\nu \le 2$, the integral $\int_0^\infty u^{\nu/2-2} e^{-u/2}\, du$ diverges at the lower limit $u \to 0$, which reflects the fact that the distribution of $1/U$ has a heavy tail at infinity (the density behaves like $u^{\nu/2-2}$ near 0, which is not integrable if $\nu/2 - 2 \le -1$). In those cases $\mathbb{E}[1/U]$ is infinite. □

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:

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

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 $d \leq n$, 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, $\hat{\theta}_{\text{OLS}} = (X^\top X)^{-1} X^\top Y = (X^\top X)^{-1} X^\top (X\theta^* + \epsilon)$, we have

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

Conditioning on X, the covariance is

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

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

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

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 $\mathbb{E}[\text{tr}((X^\top X)^{-1})] = \frac{d}{n-d-1}$. Substituting back gives the result:

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

□

Remark 1. 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*:

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

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

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

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

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

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 2. 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.