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

Background: Wishart Matrices and Rotational Invariance

A *Wishart matrix* $W_N \sim \mathcal{W}_N(P, I_N)$ is defined as $W_N = XX^T$, where X is an $N \times P$ data matrix with independent standard Gaussian entries. In other words, $W_N = \sum_{j=1}^{P} x_j x_j^T$, where each $x_j \in \mathbb{R}^N$ is an independent N-dimensional random vector drawn from $\mathcal{N}(0, I_N)$. The parameter P (with $P \geq N$) is the number of degrees of freedom (samples), and I_N is the $N \times N$ identity (the covariance matrix of each x_j). The Wishart matrix W_N is symmetric positive definite (invertible with probability 1 when $P \geq N$). Importantly, W_N is *rotationally invariant*: since each x_j has an isotropic Gaussian distribution, for any fixed orthonormal $U \in \mathbb{R}^{N \times N}$ we have $Ux_j \stackrel{d}{=} x_j$. It follows that $UW_N U^T \stackrel{d}{=} W_N$. In particular, $\mathbb{E}[W_N] = PI_N$ and, by symmetry, $\mathbb{E}[W_N^{-1}]$ must be a scalar multiple of the identity matrix I_N.

Theorem 1 (Trace of the Inverse Wishart)**.** *Let $W_N = XX^T$ be an $N \times N$ Wishart matrix formed from an $N \times P$ data matrix X with i.i.d. standard normal entries. Assume $P > N + 1$, and define the aspect ratio $\gamma = P/N$. Then*

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

In particular, as $N \to \infty$ with $P/N \to \gamma > 1$, this quantity converges to $1/(\gamma - 1)$.

Proof. First, by the rotational symmetry of the Wishart distribution noted above, we know $\mathbb{E}[W_N^{-1}] = cI_N$ for some constant c. Equivalently, all the diagonal entries of $\mathbb{E}[W_N^{-1}]$ are equal (to c) and the off-diagonal entries are zero. To determine c, it suffices to compute the expected trace $\mathbb{E}[\mathrm{Tr}(W_N^{-1})]$, since $\mathrm{Tr}(W_N^{-1})$ is the sum of the N diagonal entries of W_N^{-1}. In fact, $\mathbb{E}[\mathrm{Tr}(W_N^{-1})] = Nc$, so we seek to prove

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

To derive this trace identity, we apply Stein's lemma (integration by parts for Gaussian expectations). Let $x_1, x_2, \ldots, x_P \in \mathbb{R}^N$ denote the P independent column vectors of X, so that $W_N = \sum_{j=1}^{P} x_j x_j^T$. For each fixed $i \in \{1, 2, \ldots, P\}$, consider the scalar quantity $Q_i = x_i^T W_N^{-1} x_i = x_i^T (XX^T)^{-1} x_i$, which is the ith quadratic form (Mahalanobis distance) of x_i with respect to W_N^{-1}. We will compute $\mathbb{E}[Q_i]$ by conditioning on all other samples $\{x_j : j \neq i\}$ and applying integration by parts in the Gaussian variable x_i.

Because $x_i \sim \mathcal{N}(0, I_N)$ (independent of the other columns), Stein's lemma gives $\mathbb{E}[\, x_i^T G(x_i)\,] = \mathbb{E}[\, \nabla_{x_i} \cdot G(x_i)\,]$ for any differentiable vector-valued function $G : \mathbb{R}^N \to \mathbb{R}^N$. We apply this to $G(x_i) =$

$W_N^{-1}x_i$, which depends on x_i both explicitly and through W_N^{-1}. Using the identity $\partial(W_N^{-1}) = -\,W_N^{-1}(\partial W_N)\,W_N^{-1}$ (for infinitesimal perturbations) together with $\partial(x_i x_i^T) = (\partial x_i)x_i^T + x_i(\partial x_i)^T$, one finds that differentiating $W_N^{-1}x_i$ with respect to x_i yields

$$\nabla_{x_i} \cdot (W_N^{-1}x_i) = \mathrm{Tr}(W_N^{-1}) - 2\,x_i^T W_N^{-1} x_i \left(\nabla_{x_i} \cdot (W_N^{-1}x_i)\right).$$

Solving for the divergence, we get

$$\nabla_{x_i} \cdot (W_N^{-1}x_i) = \frac{\mathrm{Tr}(W_N^{-1})}{1+2Q_i}.$$

Now taking expectations on both sides, Stein's formula implies

$$\mathbb{E}[Q_i] = \mathbb{E}\left[\frac{\mathrm{Tr}(W_N^{-1})}{1+2Q_i}\right].$$

Because all P columns are i.i.d., $\mathbb{E}[Q_i]$ does not depend on i. We can therefore sum the above identity over $i = 1$ to P. Using the facts that $\sum_{i=1}^P Q_i = \mathrm{Tr}(W_N^{-1}W_N) = N$ for every realization, and $\sum_{i=1}^P x_i^T W_N^{-2} x_i = \mathrm{Tr}(W_N^{-2}W_N) = \mathrm{Tr}(W_N^{-1})$ for every realization, the sum gives

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

Here the first term comes from summing $\mathrm{Tr}(W_N^{-1})/(1+2Q_i)$ over i and exchanging expectation and summation, the second term comes from summing the $-2Q_i\,\mathrm{Tr}(W_N^{-1})$ contributions, and the third term comes from summing the $-2Q_i\,(\nabla \cdot (W_N^{-1}x_i))$ contributions (which effectively adds up to $-\mathbb{E}[\mathrm{Tr}(W_N^{-1})]$ as noted above). Simplifying the right-hand side yields

$$N = (P - N - 1)\,\mathbb{E}[\mathrm{Tr}(W_N^{-1})]$$

, so $\mathbb{E}[\mathrm{Tr}(W_N^{-1})] = \frac{N}{P-N-1}$ as claimed. Dividing both sides by N yields the stated result for the normalized trace.[1] □

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.

[1] This integration-by-parts derivation is inspired by *jlewk*'s answer to the MathOverflow question *"How to derive the mean of inverse Wishart distribution?"* (posted Feb. 4, 2022).

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 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*

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

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 1 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 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.) 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 $\hat{\theta}_{\text{MN}}$ 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.

Double Descent Curve and Singular Values

Combining the results from Propositions 1 and 2, we can sketch the behavior of the expected excess risk as a function of the model dimension d (for fixed n and noise level σ^2). For $d < n$, the excess risk grows approximately like $\sigma^2 \frac{d}{n-d}$, blowing up to $+\infty$ as $d \to n^-$ (since the estimator becomes arbitrarily unstable as one approaches the interpolation threshold). At $d = n$, the training data are exactly interpolated and the classical least-squares solution is not unique; indeed, in the presence of any noise, the test error formally diverges. For $d > n$, the excess risk is finite again, and initially very large just above n (due to the variance term). As d increases beyond n, the risk falls (the second descent) as the variance term $\sigma^2 \frac{n}{d-n-1}$ decreases. Eventually, however, the bias term $\frac{d-n}{d}\|\theta^*\|_2^2$ becomes significant for very large d, preventing the risk from going to zero. In fact, as $d \to \infty$, the variance term vanishes but the bias term approaches $\|\theta^*\|_2^2$, so the excess risk approaches $\|\theta^*\|_2^2$. The overall shape of $\mathbb{E}[R(\hat{\theta})]$ versus d therefore features a divergence at $d \approx n$ and a second decrease for $d > n$, which is the hallmark of **double descent** (test error decreasing twice as model size increases).

It is insightful to consider the role of the design matrix singular values in this phenomenon. The matrix X has rank $\min(n, d)$. In the under-parameterized regime ($d < n$), all d singular values of X are typically well-behaved (bounded away from zero with high probability), but as d grows large relative to n, the smallest singular value of X approaches 0. At $d = n$, X becomes square and singular (with probability 1), meaning the least-squares problem is ill-conditioned and the OLS estimator $\hat{\theta}_{\text{OLS}}$ amplifies the noise dramatically (leading to the spike in risk). In the over-parameterized regime ($d > n$), X has n nonzero singular values; as d increases further, these n singular values are influenced by having more random columns, but the smallest of the nonzero singular values tends to be better bounded away from zero once d is significantly larger than n. In fact, adding many extra random features tends to "fill out" the space and can improve the conditioning of $XX^\top$ (the eigenvalues of $XX^\top$ concentrate around their mean when d is large). This explains why the variance term shrinks with growing d in the over-parameterized regime. On the other hand, adding additional random features also means that a larger portion of the true parameter θ^* lies in the nullspace of X, which increases the bias.

Remark 3. The double descent behavior demonstrates that the conventional wisdom of "more parameters = more overfitting = higher test error" is not universally true. However, it is important to note that double descent can be mitigated by regularization. For example, if we use ridge regression (or any form of ℓ_2-regularization) and choose the regularization parameter optimally as a function of model size d, the risk curve becomes monotonic (avoiding the large peak at $d \approx$

n). Regularization ensures that the smallest singular values of the effective design matrix are bounded away from zero, thus controlling the variance. In practice, the strong performance of highly over-parameterized models (such as deep neural networks) is often attributed to implicit forms of regularization that avoid the worst effects of the interpolation threshold.