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

## Background: Expected Trace of Inverse Wishart distributed matrices

Expected Inverse Wishart: A Tutorial Proof via Sherman–Morrison (with $n > d + 1$)

**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

**Lemma 2** (Sherman–Morrison, invertible base)**.** *Let $A \in \mathbb{R}^{m \times m}$ be invertible and $u, v \in \mathbb{R}^m$. If $1 + v^T A^{-1} u \neq 0$, then $A + uv^T$ is invertible and*

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

*Proof.* Set $c := v^T A^{-1} u$ and define $B := A^{-1} - \dfrac{A^{-1} u\, v^T A^{-1}}{1+c}$. Then

$$(A + uv^T)B = AB + (uv^T)B = \Big(I - \tfrac{uv^T A^{-1}}{1+c}\Big) + \Big(uv^T A^{-1} - \tfrac{c}{1+c} uv^T A^{-1}\Big) = I.$$

Similarly $B(A + uv^T) = I$. Hence $B = (A + uv^T)^{-1}$. □

**Lemma 3** (Leave-one-out base is invertible a.s.)**.** *Fix $\ell \in \{1, \ldots, n\}$ and let $W_{(-\ell)} := \sum_{i \neq \ell} g_i g_i^T$. If $n > d + 1$, then $W_{(-\ell)}$ is invertible almost surely.*

*Proof.* Let $G \in \mathbb{R}^{d\times n}$ have columns $g_1, \ldots, g_n$ and $G_{(-\ell)} \in \mathbb{R}^{d\times(n-1)}$ be $G$ with the $\ell$-th column removed. Since $n - 1 \geq d$, a random Gaussian $G_{(-\ell)}$ has full row rank $d$ a.s. (the event of rank $< d$ is a zero set of polynomial equations in the entries). Thus $W_{(-\ell)} = G_{(-\ell)} G_{(-\ell)}^T$ is positive definite and invertible a.s. □

*Remark* 1 (Simple Sherman–Morrison update, no subspace decomposition)*.* By Lemmas 2 and 3, for each $\ell$ we can write

$$W \;=\; W_{(-\ell)} + g_\ell g_\ell^T, \qquad W^{-1} \;=\; W_{(-\ell)}^{-1} - \frac{W_{(-\ell)}^{-1} g_\ell g_\ell^T W_{(-\ell)}^{-1}}{1 + g_\ell^T W_{(-\ell)}^{-1} g_\ell}.$$

Pedagogically, when $n > d + 1$ the span of $\{g_i : i \neq \ell\}$ is already $\mathbb{R}^d$ a.s., so no column-space orthogonal complements are needed to justify this update.

## Step 3: Isotropy ⇒ diagonal expectation

**Lemma 4** (Diagonal form by sign symmetry)**.** *For $\Sigma = I_d$, $\mathbb{E}[W^{-1}]$ is a scalar multiple of the identity:*

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

*Proof.* Let $D = \operatorname{diag}(1, \ldots, 1, -1, 1, \ldots, 1)$ 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 4: 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

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

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

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

**Lemma 5** (Geometric formula for $c_j$)**.** *Let $z_1, \ldots, z_d \in \mathbb{R}^n$ denote the rows of $G$. Then $GG^\dagger = I_d$ implies $z_k \cdot c_j = \delta_{kj}$. If $Q_j$ is the orthogonal projector onto $\big(\mathrm{span}\{z_k : k \neq j\}\big)^\perp$, then*

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

*Proof.* The relations $z_k \cdot c_j = \delta_{kj}$ say exactly that $c_j$ is orthogonal to all $z_k$ $(k \neq j)$ and has unit inner product with $z_j$. Thus $c_j$ lies on the ray spanned by $Q_j z_j$, say $c_j = \theta_j Q_j z_j$. The constraint $1 = c_j^T z_j = \theta_j \|Q_j z_j\|_2^2$ gives $\theta_j = \|Q_j z_j\|_2^{-2}$ and the claims follow. $\square$

## Step 5: Distributional input (tutorial style)

**Lemma 6** (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. $\chi^2_r$ with $r = n - d + 1$. $\square$

**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

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

$\square$

## Step 6: 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 4, $\mathbb{E}[W^{-1}] = \alpha I_d$ with $\alpha = \frac{1}{d}\mathbb{E}[\mathrm{Tr}(W^{-1})]$. From Step 4 and Lemma 5,

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

By Lemma 6, $\|Q_j z_j\|_2^2 \sim \chi^2_{n-d+1}$; 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* 2. 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.