STAT 33612 Autumn 2025 Scribe: Lin Fang
Nov 4, 2025 Lecturer: Frederic Koehler
These notes have not received the scrutiny of publication. They could be missing important references, etc.

Linear Regression and Gordon's Theorem

Summary Notes (from Nov 4 Lectures)

Lecture Nov 4: Linear Regression and Gordon's Theorem

Linear Regression Model

We consider the linear regression model

$$Y = Xw^\star + \xi,$$

where

- $Y \in \mathbb{R}^n$ is the response vector,
- $X \in \mathbb{R}^{n \times p}$ is the design matrix,
- $w^\star \in \mathbb{R}^p$ is the unknown parameter vector,
- $\xi \sim \mathcal{N}(0, \delta^2 I_n)$ is Gaussian noise.

Ordinary Least Squares (OLS)

The ordinary least–squares estimator solves

$$\hat{w} = \arg\min_{w \in \mathbb{R}^p} \|Y - Xw\|_2^2.$$

When $X^\top X$ is invertible, the solution has the closed form

$$\hat{w}_{\text{OLS}} = (X^\top X)^{-1} X^\top Y.$$

Random Design Assumption

Assume the rows of X are i.i.d. Gaussian:

$$X = \begin{pmatrix} X_1^\top \\ \vdots \\ X_n^\top \end{pmatrix}, \qquad X_i \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, I_p).$$

We are interested in the estimation error

$$\|\hat{w} - w^\star\|_2^2.$$

"Easy" Case: Fixed p, $n \to \infty$

Classical asymptotic statistics gives

$$\frac{1}{n} X^\top X \xrightarrow{\text{a.s.}} I_p \quad \text{as } n \to \infty.$$

In this regime,

$$\|\hat{w} - w^\star\|_2^2 \to 0,$$

and more precisely the error is of order

$$\|\hat{w} - w^\star\|_2^2 \asymp \delta^2 \frac{p}{n}.$$

High-Dimensional Regime

The more interesting case is when both p and n grow:

$$p, n \to \infty, \qquad \frac{p}{n} \to \gamma \in (0, 1).$$

In this high-dimensional limit, one can show (under the Gaussian design)

$$\|\hat{w} - w^\star\|_2^2 \approx \frac{\delta^2 \gamma}{1 - \gamma}.$$

As $\gamma \uparrow 1$, the factor $\frac{\gamma}{1-\gamma}$ diverges, so any fixed noise level $\delta^2 > 0$ leads to exploding estimation error.

Why Recovery Fails When $\gamma > 1$

Consider the noiseless case $\delta = 0$:

$$Y = X w^*.$$

When $\gamma = p/n > 1$, the linear system above is underdetermined. For typical Gaussian X we have

$$\dim(\ker X) = p - \operatorname{rank}(X) \approx (\gamma - 1)n,$$

Since $\gamma = p/n$, we can rewrite

$$p = n\gamma = n + (\gamma - 1)n,$$

so there are n directions determined by the rows of X and roughly $(\gamma - 1)n$ additional directions lying in $\ker(X)$. so all solutions lie in a large affine space

$$\{w : Xw = Y\} = w_0 + \ker(X).$$

Put a Gaussian prior $w^* \sim \mathcal{N}(0, \tau^2 I_p)$. Given (X, Y) with $\delta = 0$, the posterior distribution $p(w \mid X, Y)$ is simply this Gaussian prior restricted to $w_0 + \ker(X)$. In the directions inside $\ker X$ the data provide no information, so the posterior variance in those directions remains of order τ^2. Thus the Bayes risk goes to infinity as the prior variance $\tau^2 \to \infty$.

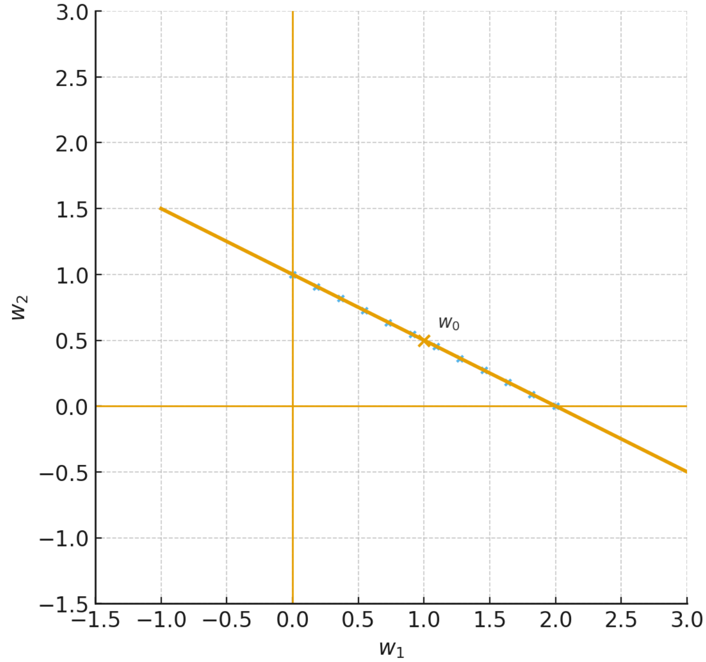

Figure 1: The line represents all solutions of $Xw = Y$, and the posterior mass (dots) is spread along this line

Expected Minimal Training Error

We also study the minimal empirical loss:

$$\min_w \frac{1}{n}\|Y - Xw\|_2^2.$$

Theorem. Under the model $Y = Xw^\star + \xi$ with $\xi \sim \mathcal{N}(0, \delta^2 I_n)$ and Gaussian design as above,

$$\mathbb{E}\left[\min_w \frac{1}{n}\|Y - Xw\|_2^2\right] = \delta^2\Big(1 - \frac{p}{n}\Big).$$

In particular, when $p = n$ the expected minimal training error is 0. This corresponds to the interpolation regime where there exists an estimator $\hat{w}$ such that $Y = X\hat{w}$ exactly.

Idea of the proof (geometric). Let $\hat{w}_{\text{OLS}}$ be an OLS solution. Then we can decompose

$$Y = X\hat{w}_{\text{OLS}} + r,$$

where $X\hat{w}_{\text{OLS}}$ is the orthogonal projection of Y onto the subspace $\text{span}(X)$, and r is the residual in the orthogonal complement $\text{span}(X)^\perp$. For a Gaussian design, this residual has distribution

$$r \sim \mathcal{N}(0, \delta^2 P_{\text{span}(X)^\perp}),$$

so $\mathbb{E}\|r\|_2^2 = \delta^2(n - p)$. Dividing by n gives the formula above.

Preliminaries: Gaussian Comparison

We recall a basic Gaussian comparison principle. Let $X = (X_a)_{a\in I}$ and $Y = (Y_a)_{a\in I}$ be two centered Gaussian processes. If their increments satisfy

$$\text{Var}(X_a - X_b) \le \text{Var}(Y_a - Y_b), \qquad \forall a, b \in I,$$

then

$$\mathbb{E}\max_{a\in I} X_a \le \mathbb{E}\max_{a\in I} Y_a.$$

This kind of comparison will be strengthened by Slepian and Gordon's theorems.

Slepian's Theorem

Let $X = (X_a)_{a\in I}$ and $Y = (Y_a)_{a\in I}$ be two mean-zero Gaussian processes such that

1. $\mathbb{E}X_a^2 = \mathbb{E}Y_a^2$ for all $a \in I$;
2. $\mathrm{Var}(X_a - X_b) \le \mathrm{Var}(Y_a - Y_b)$ for all $a, b \in I$.

Then for all real z,

$$\mathbb{P}\Big(\max_{a\in I} X_a \ge z\Big) \le \mathbb{P}\Big(\max_{a\in I} Y_a \ge z\Big).$$

Integrating this inequality over z yields

$$\mathbb{E}\max_{a\in I} X_a \le \mathbb{E}\max_{a\in I} Y_a.$$

Gordon's Theorem (Generalized Slepian)

Let $X = (X_{ij})_{i\in I,\, j\in J}$ and $Y = (Y_{ij})_{i\in I,\, j\in J}$ be two mean-zero Gaussian processes. Assume:

1. For all (i,j), $\mathbb{E}X_{ij}^2 = \mathbb{E}Y_{ij}^2$;
2. For all $i \in I$ and $j, k \in J$,
$$\mathrm{Var}(X_{ij} - X_{ik}) \le \mathrm{Var}(Y_{ij} - Y_{ik});$$
3. For all $i \ne e$ and $j, k \in J$,
$$\mathrm{Var}(X_{ij} - X_{ek}) \ge \mathrm{Var}(Y_{ij} - Y_{ek}).$$

Then, for any array of thresholds $(\lambda_{ij})_{i\in I,\, j\in J}$,

$$\mathbb{P}\Big(\forall i \in I,\ \exists j \in J \text{ s.t. } X_{ij} \ge \lambda_{ij}\Big) \le \mathbb{P}\Big(\forall i \in I,\ \exists j \in J \text{ s.t. } Y_{ij} \ge \lambda_{ij}\Big).$$

If $I = \{1\}$, the statement reduces to Slepian's theorem.

Gaussian Min–Max Corollary

Let $A \in \mathbb{R}^{m\times n}$ be a random matrix with i.i.d. entries $A_{ij} \sim \mathcal{N}(0,1)$. Let

$$\tilde{g} \sim \mathcal{N}(0,1), \quad g \sim \mathcal{N}(0, I_m), \quad h \sim \mathcal{N}(0, I_n),$$

all independent. Consider index sets $X \subset \mathbb{R}^n$, $Y \subset \mathbb{R}^m$, and a deterministic function $\psi : X \times Y \to \mathbb{R}$.

Define two Gaussian processes indexed by $(x, y) \in X \times Y$:

$$B_{x,y} = \langle y, Ax\rangle + \tilde{g}\,\|x\|\,\|y\|,$$
$$D_{x,y} = \|x\|\langle g, y\rangle + \|y\|\langle h, x\rangle.$$

Gordon's theorem implies a comparison between the probabilities of min–max events:

$$\mathbb{P}\Big(\min_{x\in X}\max_{y\in Y}\{D_{x,y} + \psi(x,y)\} \ge c\Big) \ \le\ \mathbb{P}\Big(\min_{x\in X}\max_{y\in Y}\{B_{x,y} + \psi(x,y)\} \ge c\Big)$$

for any $c \in \mathbb{R}$. Often the (auxiliary) process $D_{x,y}$ is simpler to analyze, and this inequality allows us to control properties of the (primary) optimization involving $B_{x,y}$.

Lecture Nov 6: From Gordon to the Min–Max Inequality

Checking Gordon's Conditions

Sketch why the specific choices of $B_{x,y}$ and $D_{x,y}$ satisfy Gordon's assumptions.

Matching variances. One can compute

$$\mathbb{E}\langle y, Ax\rangle^2 = \|x\|^2\|y\|^2, \qquad \mathbb{E}(\tilde{g}\,\|x\|\,\|y\|)^2 = \|x\|^2\|y\|^2,$$

and these terms are independent, hence

$$\mathbb{E}B_{x,y}^2 = 2\|x\|^2\|y\|^2.$$

A similar computation for $D_{x,y}$ using independence of g and h shows

$$\mathbb{E}D_{x,y}^2 = 2\|x\|^2\|y\|^2.$$

Therefore, $\mathbb{E}B_{x,y}^2 = \mathbb{E}D_{x,y}^2$.

Covariance comparison (idea). For (x, y) and (x', y'), one computes

$$\mathbb{E}[B_{x,y}B_{x',y'}] = \langle x, x'\rangle\langle y, y'\rangle + \|x\|\|x'\|\|y\|\|y'\|,$$

and

$$\mathbb{E}[D_{x,y}D_{x',y'}] = \|x\|\|x'\|\langle y, y'\rangle + \|y\|\|y'\|\langle x, x'\rangle.$$

Subtracting gives

$$\mathbb{E}[D_{x,y}D_{x',y'} - B_{x,y}B_{x',y'}] = -\big(\|x\|\|x'\| - \langle x, x'\rangle\big)\big(\|y\|\|y'\| - \langle y, y'\rangle\big).$$

By choosing (x, y) and (x', y') to correspond to the different index pairs in Gordon's theorem ("same row, different column" vs. "different row"), this sign structure yields the required inequalities on the variances of increments.

From a Min–Max to a Gordon Event

Consider the random quantity

$$\Phi_B = \min_{x\in X}\max_{y\in Y}\{B_{x,y} + \psi(x, y)\}.$$

The event $\{\Phi_B \geq c\}$ can be rewritten as

$$\{\forall x \in X,\ \exists y \in Y \text{ such that } B_{x,y} \geq c - \psi(x, y)\}.$$

This has exactly the form of the event in Gordon's theorem with thresholds $\lambda_{x,y} = c - \psi(x, y)$. Applying the comparison inequality to $B_{x,y}$ and $D_{x,y}$ gives the Gaussian min–max inequality stated above.

These Gaussian comparison tools, together with the random matrix structure of X, underlie precise high-dimensional characterizations of the error of OLS and related estimators in linear regression.