High-Dimensional Structure and Sparsity

Statistics · Module 10 — Model Selection Learning

Node 10.5

Position in Knowledge Graph

System: Statistics. Structural Domain: Model Selection and Learning (Module 10). Node 10.5.

When the number of parameters p exceeds the sample size n, classical statistical theory breaks down: the MLE does not exist, consistent estimation in the ambient dimension is impossible, and standard asymptotics (fixed p, $n \to \infty$) are inapplicable. This node develops the theory that replaces the classical paradigm by imposing structural assumptions—primarily sparsity—and establishes oracle inequalities showing that the LASSO (Node 10.4) achieves minimax optimal rates under these assumptions. The connection to compressed sensing provides the information-theoretic foundation.

Formal Definition

Consider the linear model $Y = X\beta_0 + \varepsilon$ where $X \in \mathbb{R}^{n \times p}$, $p \gg n$, and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$.

A vector $\beta_0 \in \mathbb{R}^p$ is **s-sparse** if $\|\beta_0\|_0 = |\{j : \beta_{0j} \neq 0\}| \leq s \ll p$.

The **effective dimension** of the problem is s, not p. The goal is estimation and prediction at rates depending on s and $\log p$, not on p itself.

Intuition

In high dimensions, the parameter space is enormous but the true signal occupies only a low-dimensional subspace. Sparsity says: most coefficients are exactly zero. The statistical challenge is to identify the nonzero coefficients (support recovery) and estimate them accurately, despite having far fewer observations than parameters. The LASSO solves both problems simultaneously under conditions on the design matrix.

The price of high dimensionality is a $\log p$ factor in the estimation rate—the cost of searching through p candidates to find the s active ones.

Structural Role

Dependencies: Builds on penalized likelihood (Node 10.4), cross-validation for tuning (Node 10.3), and asymptotic theory (Node 7.3). **Enables:** This node is a terminal node in Module 10, representing the frontier of modern statistical theory.

High-dimensional statistics replaces the classical assumption $p \ll n$ with the structural assumption $s \ll n$. The restricted eigenvalue condition replaces the non-singularity of $X^\top X$. The oracle inequality characterizes estimation error in terms of intrinsic dimension s rather than ambient dimension p.

Mathematical Development

Why Classical Asymptotics Fail

Classical theory assumes p fixed, $n \to \infty$. The MLE satisfies $\sqrt{n}(\hat{\beta} - \beta_0) \to \mathcal{N}(0, I(\beta_0)^{-1})$. When $p > n$:

1. $X^\top X$ is singular (rank at most n), so $(X^\top X)^{-1}$ does not exist.
2. There are infinitely many β satisfying $X\beta = Y$ exactly (interpolation).
3. The Fisher information matrix is $p \times p$ but has rank $\leq n$.
4. Consistent estimation of β_0 in $\|\cdot\|_2$ is impossible without structural assumptions: the minimax rate over all $\beta_0 \in \mathbb{R}^p$ is $\Theta(p/n)$, which diverges.

Restricted Eigenvalue Condition

The **restricted eigenvalue condition** $\mathrm{RE}(s, c_0)$ holds if there exists $\kappa > 0$ such that

$$\frac{1}{n}\|X\delta\|_2^2 \geq \kappa \|\delta_S\|_2^2$$

for all $\delta \in \mathbb{R}^p$ satisfying $\|\delta_{S^c}\|_1 \leq c_0 \|\delta_S\|_1$, where $S = \mathrm{supp}(\beta_0)$ with $|S| \leq s$.

The cone condition $\|\delta_{S^c}\|_1 \leq c_0\|\delta_S\|_1$ restricts attention to perturbations that are approximately sparse—precisely the directions explored by the LASSO solution path. The RE condition ensures that X is well-conditioned on these directions.

Relation to other conditions. The RE condition is weaker than: - The **restricted isometry property** (RIP): $(1 - \delta_s)\|\delta\|_2^2 \leq \frac{1}{n}\|X\delta\|_2^2 \leq (1 + \delta_s)\|\delta\|_2^2$ for all s-sparse δ. - The **irrepresentability condition**: $\|X_{S^c}^\top X_S (X_S^\top X_S)^{-1}\|_\infty < 1$, required for sign consistency.

All three are satisfied with high probability when X has i.i.d. sub-Gaussian rows and $n \gtrsim s \log p$.

Oracle Inequality for the LASSO

Theorem. Suppose the RE condition holds with constant κ and $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$. Set $\lambda = A\sigma\sqrt{\log p / n}$ for a sufficiently large constant A. Then with probability at least $1 - 2/p$:

$$\frac{1}{n}\|X\hat{\beta}^{\mathrm{LASSO}} - X\beta_0\|_2^2 \leq C \cdot \frac{s\sigma^2 \log p}{n},$$

and

$$\|\hat{\beta}^{\mathrm{LASSO}} - \beta_0\|_1 \leq C' \cdot s\sigma\sqrt{\frac{\log p}{n}},$$

for constants C, C' depending only on κ and A.

Proof sketch.

Step 1 (Basic inequality). By optimality of $\hat{\beta}^{\mathrm{LASSO}}$:

$$\frac{1}{2n}\|Y - X\hat{\beta}\|^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|Y - X\beta_0\|^2 + \lambda\|\beta_0\|_1.$$

Let $\delta = \hat{\beta} - \beta_0$. Expanding:

$$\frac{1}{2n}\|X\delta\|^2 \leq \frac{1}{n}\varepsilon^\top X\delta + \lambda(\|\beta_0\|_1 - \|\hat{\beta}\|_1).$$

Step 2 (Stochastic control). The key event is $\mathcal{E} = \{\max_j |X_j^\top \varepsilon/n| \le \lambda/2\}$. For Gaussian ε:

$$\mathbb{P}\left(\max_{j=1,\dots,p} \left|\frac{1}{n} X_j^\top \varepsilon\right| > t\right) \le 2p \exp\left(-\frac{nt^2}{2\sigma^2}\right).$$

Setting $t = \lambda/2 = (A/2)\sigma\sqrt{\log p/n}$ gives $\mathbb{P}(\mathcal{E}^c) \le 2p^{1-A^2/8} \le 2/p$ for A large enough.

Step 3 (Cone condition). On event $\mathcal{E}$, the triangle inequality gives $\|\delta_{S^c}\|_1 \le 3\|\delta_S\|_1$. This places δ in the cone required by the RE condition.

Step 4 (Combining). $\frac{1}{n}\|X\delta\|^2 \le 2\lambda\|\delta_S\|_1 \le 2\lambda\sqrt{s}\|\delta_S\|_2$. By the RE condition, $\|\delta_S\|_2 \le \frac{1}{\kappa n}\|X\delta\|^2 \cdot \frac{1}{\|\delta_S\|_2}$. Substituting and simplifying gives the oracle rate.

The $\log p$ Factor

The factor $\log p$ arises from controlling the maximum of p sub-Gaussian random variables:

$$\mathbb{E}\left[\max_{j=1,\dots,p} |Z_j|\right] \le C\sqrt{\log p}$$

for Z_j standard Gaussian. This is tight: $\max_j |Z_j| \sim \sqrt{2\log p}$. The union bound over p variables is the information-theoretic price of searching in high dimensions.

Donoho–Tanner Phase Transition

For the noiseless compressed sensing problem ($Y = X\beta_0$, $\varepsilon = 0$), exact recovery of s-sparse β_0 via L^1 minimization exhibits a sharp phase transition.

Define $\rho = s/n$ (sparsity ratio) and $\delta = n/p$ (measurement ratio). The Donoho–Tanner transition curve $\rho^*(\delta)$ satisfies:

$$\begin{aligned} &\text{If } \rho < \rho^*(\delta): \quad \hat{\beta}^{L^1} = \beta_0 \text{ with probability } \to 1. \\ &\text{If } \rho > \rho^*(\delta): \quad \hat{\beta}^{L^1} \neq \beta_0 \text{ with probability } \to 1. \end{aligned}$$

The transition curve is computed from the geometry of random polytopes. For $\delta = 0.5$ (twice as many variables as observations), recovery succeeds when $\rho \lesssim 0.09$, i.e., sparsity is below 9% of the number of measurements.

Restricted Isometry Property (RIP)

A matrix $X \in \mathbb{R}^{n\times p}$ satisfies the **RIP of order** s with constant $\delta_s \in (0,1)$ if

$$(1-\delta_s)\|\beta\|_2^2 \le \frac{1}{n}\|X\beta\|_2^2 \le (1+\delta_s)\|\beta\|_2^2$$

for all s-sparse β. Random matrices with i.i.d. sub-Gaussian entries satisfy RIP with $n = O(s\log(p/s))$ measurements. RIP implies the RE condition and guarantees exact recovery in the noiseless case and stable recovery in the noisy case.

Sign Consistency

Under the irrepresentability condition and a minimum signal strength condition $\min_{j \in S} |\beta_{0j}| \gg \sigma\sqrt{\log p/n}$, the LASSO achieves **sign consistency**:

$$\mathbb{P}\Big(\operatorname{sign}(\hat{\beta}^{\text{LASSO}}) = \operatorname{sign}(\beta_0)\Big) \to 1.$$

Without the irrepresentability condition, the LASSO may include false positives. The adaptive LASSO (with data-dependent weights $w_j = 1/|\hat{\beta}_j^{\text{init}}|^\gamma$ in the penalty) achieves sign consistency under weaker conditions.

Worked Examples

Example 1: Oracle Rate Verification

Setup: $n = 200$, $p = 1000$, $s = 5$, $\sigma = 1$. The oracle rate is $s\sigma^2 \log p/n = 5 \cdot 1 \cdot \log(1000)/200 = 5 \cdot 6.91/200 = 0.173$.

Set $\lambda = 2\sigma\sqrt{\log p/n} = 2\sqrt{6.91/200} = 2 \cdot 0.186 = 0.372$.

In simulation (averaging over 100 replications with $X_{ij} \sim \mathcal{N}(0,1)$ and $\beta_0 = (3, -2, 1.5, -1, 0.8, 0, \ldots, 0)$):

Prediction error: $\frac{1}{n}\|X\hat{\beta}^{\text{LASSO}} - X\beta_0\|^2 \approx 0.15$, consistent with the oracle bound of 0.173 (up to constants).

Support recovery: the LASSO correctly identifies all 5 active variables in 92 of 100 replications. The 8 failures include one false negative (coefficient $\beta_5 = 0.8$ is close to the threshold) and occasional false positives.

Example 2: Phase Transition

Noiseless setting: $p = 500$, $X_{ij} \sim \mathcal{N}(0, 1/n)$. Vary $n \in \{50, 100, 150, 200, 250\}$ and $s \in \{5, 15, 25, 35\}$.

Solve $\min_\beta \|\beta\|_1$ subject to $Y = X\beta$. Record whether $\hat{\beta} = \beta_0$ (exact recovery).

Results: at $n = 100$ ($\delta = 0.2$), recovery succeeds for $s \leq 5$ ($\rho = 0.05$) but fails for $s = 15$ ($\rho = 0.15$). At $n = 250$ ($\delta = 0.5$), recovery succeeds for $s \leq 20$ ($\rho = 0.08$). This matches the Donoho–Tanner prediction: the boundary $\rho^*(\delta)$ is approximately 0.09 at $\delta = 0.5$.

Cross-Links

AI. Sparsity priors underlie feature selection, network pruning, and sparse attention in transformers. The lottery ticket hypothesis posits that overparameterized networks contain sparse subnetworks that train equally well, a neural analogue of sparse recovery.

Economics. High-dimensional methods are standard in treatment effect estimation with many potential confounders (double LASSO), forecasting with large predictor sets, and text-as-data applications where p (vocabulary size) far exceeds n.

Linear Algebra. The RIP is a condition on the spectrum of $X^\top X$ restricted to sparse subspaces. Random matrix theory (Marchenko–Pastur law) describes the bulk eigenvalue distribution of $X^\top X/n$ when $p/n \to \gamma > 0$, explaining why classical OLS fails when $\gamma > 1$.

Quantum Mechanics. Compressed sensing is used in quantum state tomography: a rank-r density matrix in d dimensions has $O(rd)$ free parameters, and nuclear norm minimization recovers it from $O(rd \log^2 d)$ measurements—the matrix analogue of sparse recovery via L^1 minimization.

Structural Compression

Invariant

Under sparsity and the RE condition, the LASSO achieves prediction error $O(s \log p/n)$—the minimax optimal rate for s-sparse regression in $\mathbb{R}^p$.

Mechanism

The L^1 penalty restricts the solution to approximately sparse vectors; the RE condition ensures the design matrix distinguishes sparse directions; the union bound over p variables introduces the $\log p$ factor as the search cost in high dimensions.

Cross-System Echo

The oracle inequality is the statistical analogue of compressed sensing recovery guarantees (RIP). Both reduce high-dimensional recovery to conditions on the measurement matrix. The Donoho–Tanner phase transition is a geometric phenomenon arising from the combinatorics of random polytopes, connecting high-dimensional statistics to convex geometry.

Exercises

1. Show that when $p > n$, the OLS system $X^\top X\beta = X^\top Y$ has infinitely many solutions, and the minimum-norm solution $\hat{\beta}^{\mathrm{MN}} = X^\top (XX^\top)^{-1} Y$ has prediction risk that diverges as $p/n \to \gamma > 1$.
2. Verify the stochastic control step: for $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$ and $\|X_j\|^2 = n$, show that $\mathbb{P}(\max_j |X_j^\top \varepsilon/n| > t) \leq 2p \exp(-nt^2/(2\sigma^2))$ using the union bound and sub-Gaussian tail bounds.
3. Prove the cone condition: if $\lambda \geq 2 \max_j |X_j^\top \varepsilon/n|$ and $\hat{\beta}$ is the LASSO solution, then $\|\hat{\beta}_{S^c} - \beta_{0,S^c}\|_1 \leq 3\|\hat{\beta}_S - \beta_{0,S}\|_1$.
4. For $n = 500$, $p = 5000$, $s = 10$, $\sigma = 2$, compute the theoretical LASSO penalty λ and the oracle prediction error bound.
5. Show that if X satisfies the RIP of order $2s$ with constant $\delta_{2s} < \sqrt{2} - 1$, then the LASSO recovers β_0 exactly in the noiseless case.
6. Explain why the irrepresentability condition is needed for sign consistency but not for prediction consistency. Construct an example where the LASSO achieves the oracle prediction rate but includes a false positive.
7. Argue, structurally, why the $\log p$ factor in the oracle rate is unavoidable (information-theoretically necessary) by considering the number of possible support sets $\binom{p}{s}$ and the bits of information each observation provides about the support.

Statistics · Module 10 — Model Selection Learning · Node 10.5
Concepts: regularization, dimensionality, concentration-of-measure
Prerequisites: 10.4, 10.3, 7.3

hbar.university — own.your.cognition