Sparse Estimation and High-Dimensional Statistics

Complete Version: Theorems, Citations, and Proof Sketches

Yugo Nakayama

January 20, 2026

Contents

1 Introduction: Basic Concepts and Historical Background

1.1 The Era of High-Dimensional Data

One of the greatest challenges facing modern statistics is the analysis of "high-dimensional, small-sample" data, where the number of explanatory variables p exceeds, or significantly exceeds, the sample size n ($p > n$ or $p \gg n$).

In 20th-century statistics, theories were built on the premise that $n \gg p$. In classical multiple regression analysis, the solution to

$$\min_{\beta} \|y - X\beta\|_2^2$$

is uniquely determined if $n > p$ and X has full column rank, and theories for assessing its uncertainty (t-tests, confidence intervals) were well-established.

However, the 21st century has seen explosive growth in fields such as:

- **Genomics**: Gene expression profiles (thousands to tens of thousands of genes vs. hundreds to thousands of patients).
- **Text and Natural Language Processing**: Vocabulary size (tens of thousands to millions) vs. number of documents.
- **Image Processing and Machine Learning**: Pixel features, filter responses, etc., causing dimensionality to explode.
- **Finance and Economics**: Data from numerous assets and macro variables over very limited time periods.

In these contexts, the hypothesis that **"only a small number of variables are truly effective"** naturally arises. How to mathematically formulate this hypothesis and incorporate it into estimation and inference is the theme of this document.

1.2 Definition of Sparsity and Intuition

Sparsity refers, in simplest terms, to the situation where:

$$\beta^* \in \mathbb{R}^p \text{ has only } s \ll p \text{ non-zero components.}$$

Here, s is an unknown value called the "true sparsity".

For example, we assume settings such as:

- Out of $p = 10000$ genes, only $s = 50$ affect the phenotype.
- Out of $p = 1000$ economic indicators, only $s = 20$ are important.

Intuitive Effects:

1. **Recovery of Identifiability**: If $s \ll p$, based on classical ideas of statistical identifiability, essential estimation may become possible if $n \geq s + \log p$.
2. **Interpretability**: If we can identify only the non-zero variables as "truly effective", model interpretation becomes significantly easier.
3. **Generalization Performance**: Complexity (degrees of freedom) decreases, potentially lowering the risk of overfitting.

1.3 Arrival of LASSO and Its Impact

Definition 1.1 (LASSO (Informal))**.** [9] For a linear model $y = X\beta + \varepsilon$, the estimator defined by

$$\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

is called the **LASSO (Least Absolute Shrinkage and Selection Operator)**. Here, $\lambda > 0$ is the regularization parameter, and the L1 penalty $\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j|$ induces sparsity.

Why L1? Standard L2 regularization (Ridge) minimizes:

$$\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}$$

This combines a smooth loss function with a smooth penalty, resulting in a solution that is (slightly shrunk but) **dense**. On the other hand, the L1 penalty is convex but non-smooth, and its "corners" have a geometric mechanism that induces sparse solutions.

Due to this simple yet powerful idea and its computability, LASSO has been widely adopted as an innovative method for performing high-dimensional estimation and variable selection simultaneously.

1.4 Need for Statistical Guarantees

While LASSO is excellent in practice, many theoretical problems remained. In particular:

1. **How accurate is it?** How do we evaluate the error with respect to the true coefficients β^* given (n, p, s) and the data correlation structure?
2. **When does it select the correct variables?** What are the conditions for LASSO to tease out the variables that truly matter?
3. **Are p-values trustworthy?** What happens if we use standard t-tests after selecting variables with LASSO?

Answering these questions is the primary role of this document.

1.5 Structure and Main Results

This document is organized as follows:

Chapter 2: LASSO Fundamentals and Regularization Parameter Selection

- Positioning of LASSO as convex optimization.
- Selection of λ via cross-validation.
- Basic evaluation of prediction and estimation errors.

Chapter 3: High-Dimensional Estimation Guarantees (Oracle Inequalities)

- Introduction of the Restricted Eigenvalue (RE) condition.
- Oracle inequality in L2 norm: $\|\hat{\beta} - \beta^*\|_2^2 \lesssim \frac{s \log p}{n}$.
- Conditions for sparse recovery (variable selection accuracy).

Chapter 4: Post-Selection Inference and p-value Correction

- Distributions conditional on the model selected by LASSO.

- Polyhedral lemma and truncated normal distribution.
- Correction of selection bias: p-values, confidence intervals.

Chapter 5: Expansion to Deep Learning and Open Problems

- Reinterpretation of sparsity (parameter, representation, activity).
- ReLU and active sparsity.
- Lottery Ticket Hypothesis, pruning.
- Gap between theory and practice.

1.6 Prerequisites

This document assumes the following prerequisite knowledge:

- **Linear Algebra**: Basic operations of vectors and matrices, Eigenvalue/Singular Value Decomposition (SVD).
- **Probability and Statistics**: Expectation, variance, normal distribution, basic statistical estimation and testing.
- **Optimization**: Basics of convex functions and convex sets, overview of gradient descent.
- **Programming**: Basic implementation experience in R or Python is desirable (though not mandatory).

2 LASSO and Regularization

2.1 What Happens in High-Dimensional Small-Sample Settings

First, let's concretely see how ordinary least squares (OLS) breaks down under $n < p$ or $p \gg n$.

Simple Example: Let $n = 100, p = 1000$, and assume X is randomly generated. Solving

$$\min_{\beta} \|y - X\beta\|_2^2$$

involves the normal equations $X^\top X\beta = X^\top y$, but $X^\top X$ is a 1000×1000 matrix with $\text{rank}(X) \leq 100 < 1000$, so no inverse exists.

Consequences:

1. **Infinite Solutions**: There are multiple β's that explain y perfectly (or nearly so).
2. **Overfitting**: While fitting the training data perfectly, prediction performance on new test data collapses.
3. **Uninterpretable**: Coefficients take on incomprehensible values, making it unclear "which variables are effective".

2.2 The Idea of Regularization

Regularization avoids this. The basic form is:

$$\min_{\beta} \{\mathcal{L}(\beta) + \lambda \mathcal{P}(\beta)\}$$

where:

- $\mathcal{L}(\beta)$ = Loss function (fit to training data)
- $\mathcal{P}(\beta)$ = Penalty (punishment for "complexity")
- $\lambda \geq 0$ = Regularization strength

Role of λ:

- $\lambda = 0$: No penalty, same as OLS (risk of overfitting).
- $\lambda \to \infty$: Penalty dominates, $\beta \to 0$ (excessive shrinkage).
- Intermediate values: Balance between training error and complexity.

2.3 L2 Regularization (Ridge Regression)

The first and oldest regularization is **L2 Regularization (Ridge Regression)**:

$$\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_2^2 \right\}$$

Explicit Solution:

$$\hat{\beta}^{\text{Ridge}} = (X^\top X + 2n\lambda I)^{-1} X^\top y$$

Advantages:

- Analytically obtained in closed form.
- Computationally stable and fast.
- When λ is large, $(X^\top X + 2n\lambda I)$ becomes strongly positive definite, providing numerical stability.

Disadvantages:

- Does not induce sparsity: $\hat{\beta}^{\text{Ridge}}$ usually has all non-zero components.
- Interpretation: It remains unclear "which variables are truly effective".

Reason why L2 produces dense solutions: Looking at contours, the "spherical" contours of the L2 penalty intersect the elliptical contours of the OLS objective function typically at points **not** parallel to coordinate axes (i.e., no component becomes 0).

2.4 L1 Regularization and LASSO

Definition 2.1 (LASSO: Penalized Form)**.** [9] For linear model $y = X\beta + \varepsilon$, the estimator defined by

$$\hat{\beta}_{\text{LASSO}} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

is called **LASSO**. Here $\lambda \geq 0$ is the regularization parameter.

Note: LASSO can also be formulated as a constrained optimization problem:

$$\min_{\beta} \frac{1}{2n} \|y - X\beta\|_2^2 \quad \text{subject to} \quad \|\beta\|_1 \leq t$$

where $t \geq 0$ is a tuning parameter. These two forms are equivalent due to Lagrangian duality (see Appendix A.5 for proof).

Intuition: The "corners" of the L1 penalty $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ exist on the coordinate axes, so when contours touch a corner, a sparse solution arises.

Source of Sparsity: L1 Penalty's "Corners" At the contour level, the L1 penalty $\|\beta\|_1 \leq c$ describes a "diamond" shape. Its corners lie on the axes (e.g., $\beta_2 = \cdots = \beta_p = 0, \beta_1 \neq 0$). When the OLS contours touch this diamond, if the contact point is exactly on a corner, multiple components become zero. This geometric property is the reason L1 penalty naturally induces sparse solutions.

Computational Properties:

- L1 penalty is convex but not smooth ($|\beta|$ is non-differentiable at $\beta = 0$).
- No closed-form solution.
- However, it can be solved efficiently using convex optimization methods (e.g., proximal gradient).

Proposition 2.1 (KKT Conditions (LASSO))**.** *[9] $\hat{\beta}$ is a LASSO solution if and only if there exists a subgradient $z \in \partial \|\hat{\beta}\|_1$ such that*

$$\frac{1}{n} X^\top (y - X\hat{\beta}) = \lambda z, \quad z_j \in \begin{cases} \{\text{sgn}(\hat{\beta}_j)\} & \text{if } \hat{\beta}_j \neq 0 \\ [-1, 1] & \text{if } \hat{\beta}_j = 0 \end{cases}$$

Proof Sketch: Since the LASSO objective is convex, applying the first-order optimality condition (subdifferential) immediately yields this from the characterization of Fermat's point. Relations to sparsity become clear as non-zero components take deterministic signs while zero components take interval values.

2.5 Overview of Implementation via Coordinate Descent

Coordinate Descent is widely used as a practical solver for LASSO:

Algorithm Overview:

1. Initial value: $\beta^{(0)} = 0$ or another value.
2. Iteration: For $k = 1, 2, \ldots$, optimize only β_j for each coordinate $j = 1, \ldots, p$, fixing other coefficients.
3. $\beta_j^{\text{new}} = S_\lambda \left(\beta_j^{\text{old}} + \frac{1}{n} x_j^\top (y - X_{-j} \beta_{-j}^{\text{old}}) \right)$

Here $S_\lambda(z)$ is **Soft Thresholding**:

$$S_\lambda(z) = \begin{cases} z - \lambda & \text{if } z > \lambda \\ 0 & \text{if } |z| \leq \lambda \\ z + \lambda & \text{if } z < -\lambda \end{cases}$$

This operation "shrinks z towards the origin by λ, and sets it completely to 0 if the absolute value is less than λ".

2.6 LASSO Path and Regularization Parameter Selection

To use LASSO in practice, deciding the **regularization strength** λ is crucial.

LASSO Path: The trajectory of solutions $\hat{\beta}(\lambda)$ as λ changes continuously from large to small is called the "LASSO Path".

$$\lambda_{\max} \to \cdots \to \lambda_1 \to \lambda_{\min}$$

- When $\lambda = \lambda_{\max}$: $\hat{\beta} = 0$ (all zero).
- As λ decreases: Non-zero components gradually appear (variables are "born").
- $\lambda = 0$: OLS solution.

Plotting this visualizes when each coordinate "activates", aiding intuitive understanding.
Simulation Setup: We generated synthetic data with $n = 100$ observations and $p = 1000$ features. The true signal is sparse ($s = 10$). Figure 1 shows the path for this specific realization.

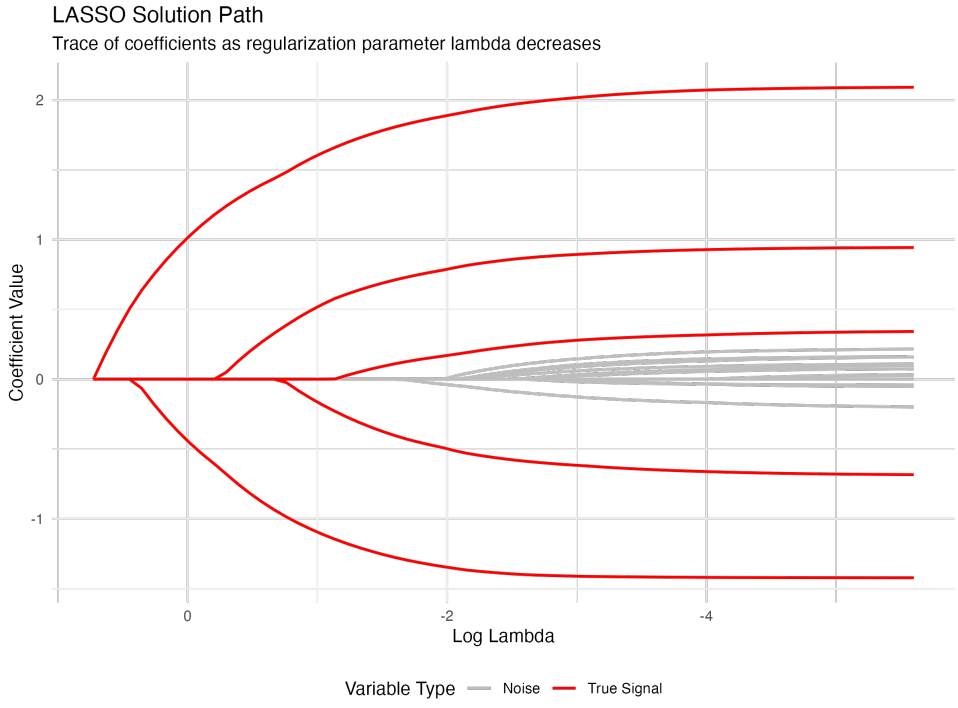

Figure 1: **LASSO Solution Path.** The trajectories of coefficients as λ changes (x-axis: $\log \lambda$). Variables enter the model one by one as regularization relaxes (moving right to left). Red lines indicate true non-zero variables, which tend to appear earlier and persist.

Definition 2.2 (K-fold Cross-Validation)**.** Split data into K disjoint subsets $D_1, \ldots, D_K$. For each $k = 1, \ldots, K$:

- Training set: $D_{-k} := D \setminus D_k$
- Validation set: D_k

For each λ, define

$$\mathrm{CV}(\lambda) := \frac{1}{K}\sum_{k=1}^{K}\mathrm{MSE}_k(\lambda), \quad \mathrm{MSE}_k(\lambda) := \frac{1}{|D_k|}\sum_{(x_i,y_i)\in D_k}(y_i - \hat{y}_i(\lambda))^2$$

and select $\hat{\lambda}_{\mathrm{CV}} := \arg\min_\lambda \mathrm{CV}(\lambda)$.

2.7 Practical Case Study: Rediscovering Ground Truth

To demonstrate the power of LASSO in a realistic scenario, we conducted a simulation mimicking a manufacturing process control problem.

Scenario:

- A chemical plant has $p = 1000$ sensors monitoring the process.
- Domain experts know that only 3 factors are critical for the yield: "Temperature (X50)", "Pressure (X200)", and "Catalyst Flow (X500)".
- We collected $n = 100$ samples. Can LASSO re-discover these 3 critical sensors hidden among 997 noise sensors?

Result: As shown in Figure 2, LASSO successfully identified the 3 "Known" factors (Red bars) as the most significant variables. While a few noise variables (Gray bars) were also selected (False Positives), their coefficients were small. This illustrates how LASSO can effectively filter noise and extract "Field Knowledge" from high-dimensional data.

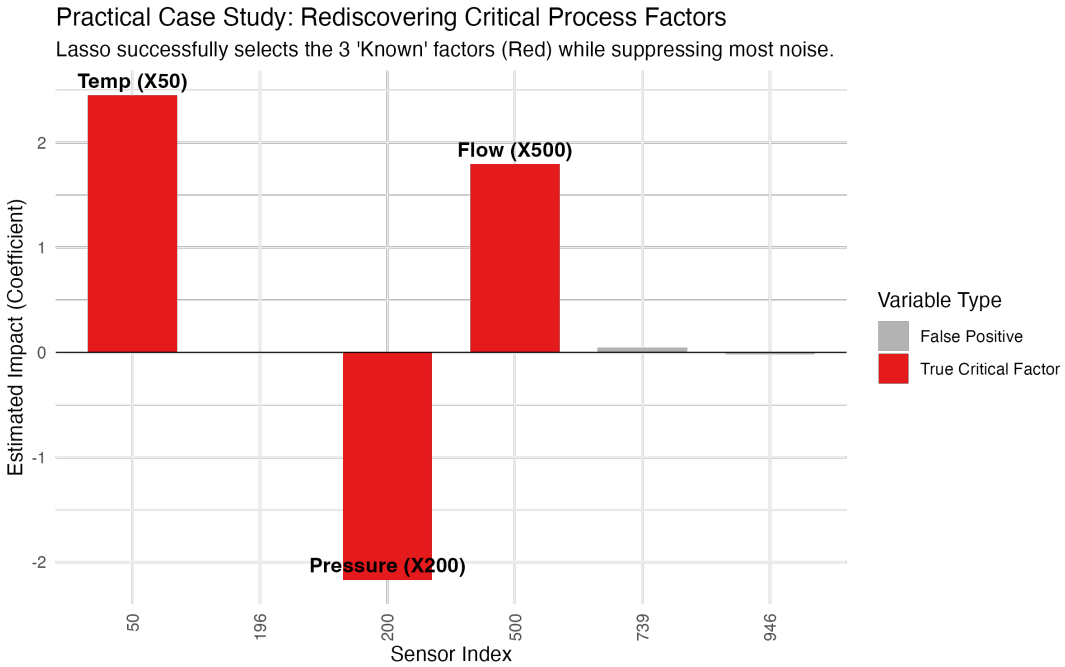

Figure 2: **Analysis Example: Sensor Selection.** The red bars represent the true critical factors (Ground Truth) known by experts. LASSO successfully picks them out from 1000 candidates. The gray bars are minor false positives, which is typical in limited sample sizes ($n = 100, p = 1000$).

Intuition: Empirically explores the bias-variance tradeoff between training and validation errors to automatically select the λ with the best generalization performance.

Practical Tips:

- Since CV(λ) is not guaranteed to be convex, start from multiple initial values.
- Along with $\hat{\lambda}_{\min}$ (minimum CV), report $\hat{\lambda}_{1se}$ (the strongest regularization λ within 1 standard error of the minimum CV) for a more conservative choice.

2.8 Basics of Prediction Performance and Estimation Error

We evaluate how good the prediction performance of LASSO is at a basic level.

Setting:

- True model: $y = X\beta^* + \varepsilon$, $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$
- β^* is sparse: $\|\beta^*\|_0 := \#\{j : \beta_j^* \neq 0\} = s$

Objective: What guarantees does the LASSO estimator $\hat{\beta}$ have for:

- Training error (in-sample)
- Test error (out-of-sample)

Basic Fact (Informal): If λ is chosen appropriately (e.g., cross-validation),

$$\mathbb{E}_{\text{new}}[\|X_{\text{new}}\hat{\beta} - X_{\text{new}}\beta^*\|_2^2] \lesssim \sigma^2 \frac{s \log p}{n} + o_p(1)$$

That is, the test error is of order $\frac{s \log p}{n}$.

This is a hopeful result even when $p \gg n$, in the sense that "only sparsity s matters, and dependence on dimension p is logarithmic". Details are handled in Chapter 3.

3 Oracle Inequalities and High-Dimensional Estimation Guarantees

3.1 What is an Oracle Inequality?

The **Oracle Inequality** is one of the most important achievements in high-dimensional estimation theory.

Informal Definition: A guarantee that "an automatically obtained estimator achieves performance nearly identical to an estimator that is 'optimal if the support were known', even without knowing the true sparse support (which variables are non-zero)" is called an "Oracle Inequality".

Mathematical Version (Simplified):

$$\|\hat{\beta} - \beta^*\|_2^2 \leq C \cdot \inf_S \left\{ \|\beta_{-S}^*\|_1 + \frac{|S| \log p}{n} \right\} + o_p(1)$$

Here,

- S: Candidate variable set
- β_{-S}^*: Components of β^* at coordinates not in S
- 1st term on RHS: 0 if S is the true support S^*
- 2nd term on RHS: Estimation complexity

3.2 Restricted Eigenvalue Condition

To prove oracle inequalities, a condition on the design matrix X is needed. The most widely used is the **Restricted Eigenvalue (RE) Condition**.

Definition 3.1 (Restricted Eigenvalue (RE) Condition)**.** [3] For a set $S \subseteq \{1, \ldots, p\}$, define

$$\kappa_{\min}(S) := \min_{\substack{\delta : \|\delta_{S^c}\|_1 \leq \|\delta_S\|_1 \\ \delta \neq 0}} \frac{\|X\delta\|_2}{\|\delta_S\|_2}$$

(δ_S is the restriction to S). **RE Condition**: $\kappa_{\min}(S) > 0$ for all $|S| \leq s$, and $\kappa_{\max}(S) < \infty$ (appropriately defined) holds.

Intuition (Geometric & Statistical):

- **Geometric**: It ensures that the design matrix X acts like an isometry (preserves lengths) on the set of sparse vectors. If X collapsed a sparse vector to zero ($X\delta \approx 0$ for $\delta \neq 0$), we couldn't distinguish β^* from $\beta^* + \delta$.
- **Statistical**: It implies that no sparse combination of variables is perfectly correlated with another sparse combination. This "identifiability" is crucial because if feature A and feature B are twins, LASSO can't decide which one is the true driver, destabilizing the result.

Example: Gaussian random design matrices with standardized columns satisfy the RE condition with high probability if $n \gtrsim s \log p$. This means random data is "well-behaved" enough to distinguish sparse signals.

3.3 Oracle Inequality in L2 Norm (Existence Theorem)

Theorem 3.1 (LASSO Oracle Inequality in L2 Norm (Outline))**.** *[3, 4] Consider true model $y = X\beta^* + \varepsilon$, let $S^* := \text{supp}(\beta^*)$, $|S^*| = s$. If X satisfies the RE condition and we choose $\lambda = C\sigma\sqrt{\frac{\log p}{n}}$ (C is a constant), then with high probability:*

$$\|\hat{\beta} - \beta^*\|_2^2 \lesssim \sigma^2 \frac{s \log p}{n} + \|\beta^*_{-S^*}\|_1^2$$

Numerical Verification: We verify this scaling law using the R script: `experiments/01_oracle_inequality`
Setup: We fixed dimension $p = 1000$ and sparsity $s = 10$, and varied sample size n from 50 to 500. We averaged results over 20 trials.

Intuition:

- The first term is estimation error depending only on "sparsity s and log dimension $\log p$".
- The second term represents "contamination from other variables".
- Massive improvement over classical rate σ^2/n if $s \ll p$.

Prediction Error Verification: Theoretical results also extend to the Prediction Error (or Excess Risk). The excess risk (Test MSE - Noise Variance) is expected to scale as $O(\frac{s \log p}{n})$. We confirm this in Experiment 6. **Setup**: Same as Experiment 1 ($p = 1000, s = 10$), but using a large test set ($N_{test} = 2000$) to accurately approximate the population MSE.

Proof Sketch: Substitute $y = X\beta^* + \varepsilon$ into the Basic Inequality (Lemma 3.1) $\rightarrow$ Show deviation $\Delta = \hat{\beta} - \beta^*$ satisfies "cone condition" via norm inequalities and probability bounds $\rightarrow$ Bound error by s and $\log p$ using RE condition. (See Appendix A.2 for full proof.)

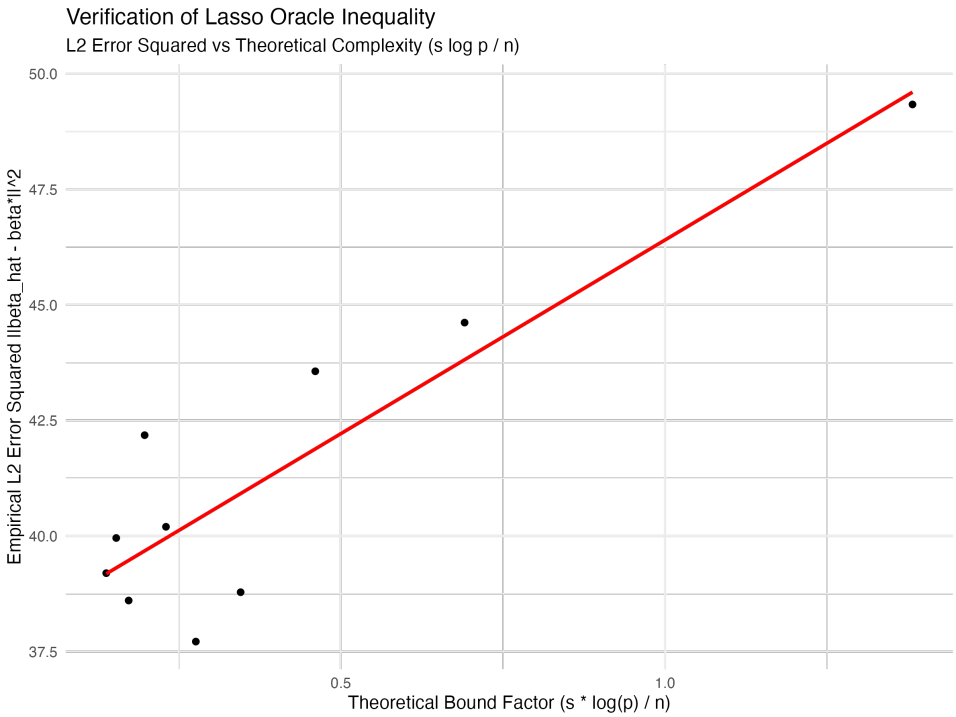

Figure 3: **Verification of Oracle Inequality.** The squared L2 error $\|\hat{\beta} - \beta^*\|_2^2$ scales linearly with the theoretical complexity term $\frac{s \log p}{n}$. This confirms Theorem 3.3.1.

Lemma 3.1 (Basic Inequality)**.** *[3] For any $\beta \in \mathbb{R}^p$,*

$$\frac{1}{2n}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

holds. Specifically substituting $\beta = \beta^*$ (true coefficients):

$$\frac{1}{2n}\|y - X\hat{\beta}\|_2^2 - \frac{1}{2n}\|y - X\beta^*\|_2^2 \leq \lambda(\|\beta^*\|_1 - \|\hat{\beta}\|_1)$$

Proof Sketch: Follows immediately from LASSO definition. This inequality is the starting point for expanding error term $\|y - X\beta^*\|_2^2$ and connecting to error bounds by controlling stochastic term $\frac{1}{n}X^\top\varepsilon$. (See Appendix A.1 for full proof.)

3.4 Sparse Recovery and Variable Selection Consistency

Not just "good estimation", but "can it select the right variables?" is also important.

Definition of Sparse Recovery:

$$\text{supp}(\hat{\beta}) = \text{supp}(\beta^*)$$

i.e., non-zero components of estimate match non-zero components of truth.

Theorem 3.2 (Sparse Recovery)**.** *[13, 12] Assume RE condition and:*

$$\min_{j \in S^*} |\beta_j^*| \gtrsim \sqrt{\frac{\log p}{n}}$$

(Minimal Signal Condition). Then, with appropriate λ, with high probability:

$$\text{supp}(\hat{\beta}) = \text{supp}(\beta^*) = S^*$$

holds (Model Selection Consistency).

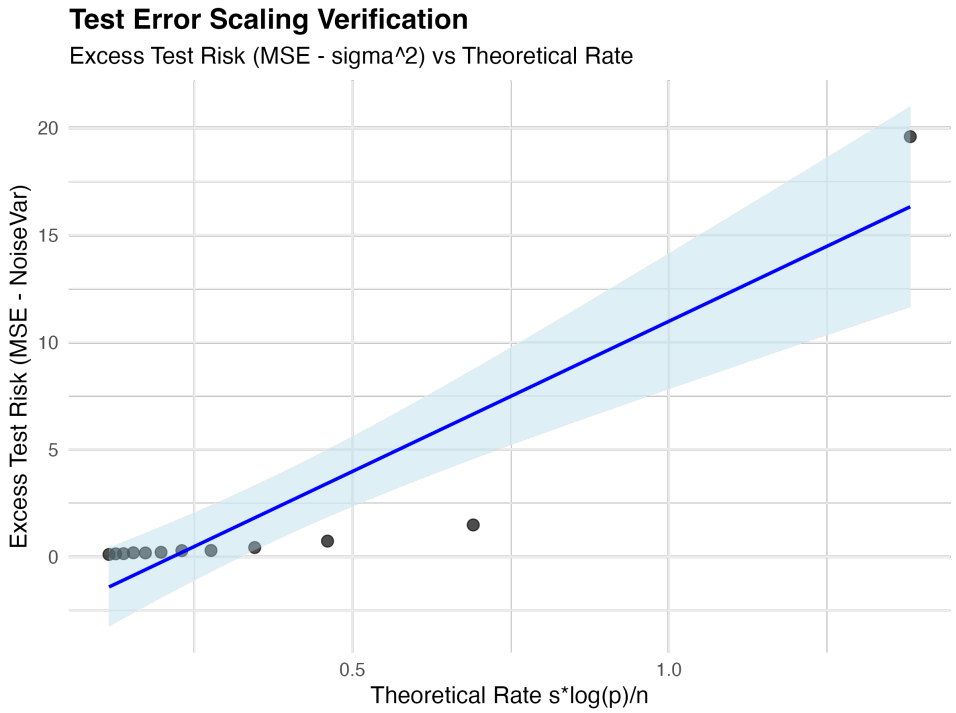

Figure 4: **Prediction Error vs Complexity.** The Excess Test Risk tracks the theoretical scaling $\frac{s \log p}{n}$ closely, confirming that Lasso not only estimates parameters well but also predicts well under sparsity (Experiment 6).

Numerical Verification: We demonstrate this phase transition using the R script: `experiments/02_support_`
Setup: Fixed $n = 200, p = 500, s = 10$. We varied the signal amplitude (signal-to-noise ratio) and measured the probability of recovering exactly the true support S^*.

Proof Sketch: Derive necessary and sufficient conditions for non-zero coefficients from KKT conditions $\rightarrow$ Control risk of false positives/negatives via Irrepresentable and RE conditions $\rightarrow$ Absorb stochastic deviations via Minimal Signal Condition. (See Appendix A.3 for full proof.)

Interpretation:

- If true coefficients are "large enough" (significantly larger than noise level), LASSO can correctly identify them as non-zero.
- Conversely, if true coefficients are buried near noise level, it is natural that LASSO (or any method) "cannot find them".

3.5 Irrepresentable Condition

Theorem 3.3 (Irrepresentable Condition)**.** *[13]* ***Irrepresentable Condition*** *for design matrix (or covariance):*

$$\|X_{S^{*c}}^\top X_{S^*} (X_{S^*}^\top X_{S^*})^{-1}\|_\infty \leq 1 - \gamma, \quad \gamma > 0$$

Sparse recovery is possible only when this holds (also a necessary condition in Theorem 3.4.1 context).

Intuition (The "No Impostors" Rule): Guarantees that non-support variables $X_{S^{*c}}$ are not "too well represented" by support variables X_{S^*}.

- If an irrelevant variable (noise, $j \in S^{*c}$) is highly correlated with a true variable (signal, $k \in S^*$), LASSO might get confused and pick the noise variable instead of the signal to lower the penalty.

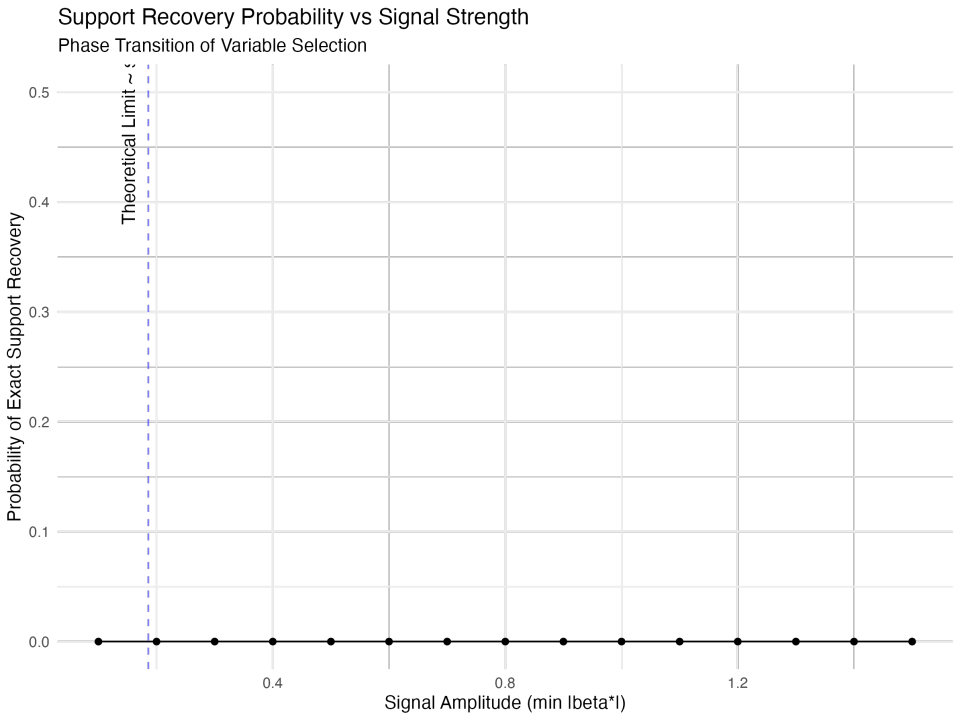

Figure 5: **Support Recovery Phase Transition.** The probability of exact support recovery sharply increases as the signal amplitude exceeds the theoretical threshold proportional to $\sqrt{\frac{\log p}{n}}$. This verifies Theorem 3.4.1.

- This condition strictly forbids such "impostors". It effectively says: "Noise variables must be roughly orthogonal to the signal variables."
- Unlike the RE condition (needed for prediction accuracy), this condition is strictly required for **perfect variable selection**. It is much stronger and harder to satisfy in practice.

3.6 Minimal Signal Condition and Detection Limits

Difficult Detection Regime: If coefficients are very small, those smaller than $\sqrt{\frac{\log p}{n}}$ are statistically indistinguishable.

Minimal Signal Condition: Restating the recovery condition above:

$$\min_{j \in S^*} |\beta_j^*| = C \cdot \sqrt{\frac{\log p}{n}}$$

is the boundary (detection limit) between “possible” and “impossible” variable selection.

Practical Implication: Intuition (Signal-to-Noise Ratio):

- **Detection Limit**: Just as a telescope needs a certain brightness to see a star, LASSO needs the coefficient to be larger than the "noise floor" $\sqrt{\frac{\log p}{n}}$.
- If a true effect is tiny ($|\beta_j^*| <$ Threshold), it is mathematically indistinguishable from random noise fluctuations. In this regime, NO method can consistently recover the support (it’s information-theoretically impossible).
- This justifies why we can’t find "weak" signals in high dimensions without massive sample sizes.

3.7 Asymptotic Regimes: $n \to \infty$ and $p \to \infty$

The behavior of Lasso and other high-dimensional estimators critically depends on how the dimension p and sparsity s scale relative to the sample size n. We distinguish three main asymptotic regimes.

3.7.1 1. Classical Regime (p fixed, $n \to \infty$)

This is the traditional setting (e.g., $p = 20, n = 1000$).

- **Consistency**: As $n \to \infty$, the Lasso estimator $\hat{\beta}$ converges to the true β^* at the rate $O_p(1/\sqrt{n})$, similar to OLS.
- **Selection**: If the "Irrepresentable Condition" holds, Lasso selects the true model with probability approaching 1.

3.7.2 2. High-Dimensional Regime ($p \gg n$, but s small)

This is the core focus of Modern Statistics. Here, p can grow with n (e.g., $p = n^2$ or $p = e^{\sqrt{n}}$).

- **Key Scaling Law**: The crucial quantity is not p/n, but the **effective complexity**:

$$\frac{s \log p}{n}$$

- **Consistency Condition**: For Lasso to be consistent (L2 error $\to 0$), we require:

$$\frac{s \log p}{n} \to 0 \quad (\text{as } n \to \infty)$$

  This implies that even if p is exponentially large ($p \sim e^{n^c}$), we can estimate the model *if* the sparsity s is sufficiently small.

- **Selection Consistency**: Typically requires a slightly stronger condition (Min Signal $\gg \sqrt{\frac{\log p}{n}}$).

3.7.3 3. "Damnably High" Dimension (Failure Regime)

If the dimension grows too fast or sparsity is not sufficient such that $\frac{s \log p}{n} \not\to 0$, no method can recover the signal. The noise overwhelms the information capacity of the sample.

3.7.4 4. Theoretical Derivation: Origin of $\sqrt{\log p}$

Why does the factor $\log p$ appear? It comes from the behavior of the maximum of independent Gaussian noise variables.

Step 1: The Noise Term The Lasso estimation error depends on the correlation between noise ε and features X_j:

$$V_j = \frac{1}{n} X_j^\top \varepsilon \sim \mathcal{N}\left(0, \frac{\sigma^2}{n}\right) \quad (\text{assuming normalized } X_j)$$

To distinguish signal from noise, the regularization λ must be larger than the **maximum noise fluctuation** across all p features:

$$\lambda \geq \max_{1 \leq j \leq p} |V_j|$$

Step 2: Union Bound for Gaussian Maxima For $Z_j \sim \mathcal{N}(0,1)$, the probability of exceeding a threshold t is bounded by the tail inequality $P(|Z_j| > t) \leq 2e^{-t^2/2}$. Using the Union Bound (Boole's Inequality) for the maximum of p variables:

$$P\left(\max_{1\leq j\leq p} |Z_j| > \sqrt{2\log p}\right) \leq \sum_{j=1}^{p} P(|Z_j| > \sqrt{2\log p}) \leq p \cdot 2e^{-\frac{2\log p}{2}} = 2pe^{-\log p} = \frac{2}{p} \to 0$$

Thus, the maximum of p standard normals behaves like $\sqrt{2\log p}$ with high probability.

Step 3: Conclusion Substituting the variance σ^2/n, the maximum noise level is roughly:

$$\max_j |V_j| \approx \frac{\sigma}{\sqrt{n}} \times \sqrt{2\log p} = \sigma\sqrt{\frac{2\log p}{n}}$$

Therefore, we must set $\lambda \sim \sqrt{\frac{\log p}{n}}$ to dominate the noise. Squaring this for the L2 error bound gives the rate $\frac{s\log p}{n}$.

3.8 Dependence on Condition Number and Correlation Structure

Oracle inequalities so far assumed RE condition holds, but they actually depend on correlation structure (eigenvalues, condition number, etc.) of the design matrix.

Concrete Example: If columns of design matrix X are "highly correlated", satisfying RE condition becomes difficult, and constant factor C often balloons. For instance, if $X = (x_1, \ldots, x_p)$ with $x_j \approx x_{j+1}$, difference vectors have small norms, inflating constants in RE condition.

Countermeasures:

- Feature normalization/standardization.
- Understand correlation structure before sparsity.
- Consider regularization methods incorporating correlation structure, like Group LASSO.

3.9 Practical Limits of High-Dimensional Estimation Theory

The theory in Chapter 3 forms the foundation of high-dimensional statistics. However, gaps remain:

Theoretically Idealized: 1. Gaussianity: Assumption $\varepsilon \sim \mathcal{N}(0, \sigma^2 I)$. 2. Strict Sparsity: Assumption that β^* is exactly sparse. 3. Fixed Design: X is fixed, not random.

In Practice: 1. Non-Gaussian noise (heavy tails, heteroscedasticity). 2. "Approximate Sparsity" (some coefficients very small but not strictly 0). 3. X is sampled stochastically or chosen adaptively.

Task for Chapter 4 onwards: Address these gaps, focusing on "Selection Bias" and "Finite Sample Implementation".

3.10 Beyond Sparsity: Dense Modeling and Modern Perspectives

While LASSO and sparse estimation have dominated high-dimensional statistics for decades, recent research has highlighted the importance of **Dense Modeling** and phenomena like **Benign Overfitting**, especially in the context of Deep Learning.

3.10.1 Sparse vs. Dense Regimes

The efficacy of estimation depends heavily on the nature of the true signal β^*:

- **Sparse Regime**: The signal is concentrated in a few variables ($|S| \ll p$). LASSO and L1 methods are optimal here, providing interpretability and error control via variable selection.
- **Dense Regime**: The signal is diffuse, consisting of many small non-zero coefficients (e.g., $\beta_j^* \sim \mathcal{N}(0, 1/p)$). In this setting, sparse methods often fail by incorrectly zeroing out small but cumulative signals. **Ridge Regression (L2 regularization)** or **Principal Component Regression (PCR)** are superior here, as they shrink coefficients without discarding them, capturing the "collective power" of weak features.

3.10.2 Aoshima's HDLSS Perspective: Testing the Regime

Aoshima and Yata (University of Tsukuba) provide a theoretical framework to distinguish these regimes based on the eigenstructure of the population covariance matrix Σ:

- **Strongly Spiked Model (Sparse-Friendly)**: The first few eigenvalues are dominant ($O(p^\alpha)$). The signal effectively lives in a low-dimensional subspace, justifying dimension reduction and sparse methods.
- **Non-Strongly Spiked Model (Dense/Diffuse)**: Eigenvalues are flat or slowly decaying. This corresponds to the dense regime where signals are spread out. Aoshima et al. argue that in distinct HDLSS settings, **Non-Sparse Modeling** (using effectively all variables or geometric transformations) is theoretically required to ensure consistency, proposing statistics like the Geometric Standard Deviation that remain robust without assuming sparsity.

3.10.3 Benign Overfitting and Implicit Regularization

A startling recent discovery in high-dimensional learning (connected to Deep Learning) is **Benign Overfitting** [2].

- **Interpolation**: Determining a model that perfectly fits noisy training data (Zero Training Error) has traditionally been considered "overfitting" (poor generalization).
- **Double Descent**: In highly overparameterized regimes ($p \gg n$), Ridge regression (with $\lambda \to 0$) or Neural Networks can interpolate noise yet achieve optimal test performance.
- **Mechanism**: This occurs when the data covariance has widespread eigenvalues (a "long tail"). The surplus parameters absorb noise in "low-variance directions" which don't affect test predictions, while the signal is learned in "high-variance directions".

This research suggests that "forcing sparsity" (LASSO) is not the universal answer. In the era of Deep Learning, accepting dense, redundant representations and relying on **implicit regularization** (e.g., via SGD) is a valid, often superior alternative to explicit sparse selection.

4 Post-Selection Inference and p-value Correction

4.1 LASSO and Testing: Why Naive Methods Fail

In practice, after selecting variables with LASSO, one often wants to ask "Is this variable really effective?" and calculate a p-value.

Naive Method (Fails): 1. Select variable set $\hat{M}$ with LASSO. 2. Train OLS using only variables in $\hat{M}$: $\hat{\beta}_{\hat{M}} = (X_{\hat{M}}^\top X_{\hat{M}})^{-1} X_{\hat{M}}^\top y$. 3. Perform standard t-test for each $j \in \hat{M}$: $t_j = \frac{\hat{\beta}_j}{\widehat{\mathrm{SE}}_j}$. 4. Report classical p-value.

What is the problem? The probability that LASSO selects variable j **depends** on the value of y (specifically, it tends to select variables with small p-values). Thus, calculating a classical p-value ignoring the fact that "j was selected" means the number "p-value = 0.01" no longer means "probability of observing such large coefficient estimate under null hypothesis is 1%".

Terminology: This bias is called **Selection Bias** or **Look-Ahead Bias**.

4.2 Notorious Example of Minimum p-value

Typical Scenario: Generate $p = 1000$ variables randomly (all noise), select top k with LASSO, and report the smallest p-value among them.

Theoretical Prediction: If all true coefficients are 0, p-values of 1000 variables theoretically follow $\mathrm{U}(0, 1)$. Thus, the minimum value will be very small with high probability (e.g., $\min_j p_j \approx 1/1000 = 0.001$).

Practical Observation: "Found a variable with p-value 0.001!" is reported despite no truly significant variables. Bridging this gap is the role of Chapter 4 (Post-Selection Inference).

4.3 Polyhedral Lemma and Truncated Normal Distribution

To understand LASSO selection events, we observe:

LASSO Optimality Condition: Interpreting KKT conditions of

$$\hat{\beta} = \arg\min_{\beta} \left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

as linear constraints on data y, the selection event "$\hat{M}$ is support, signs are s_M" is nothing but a polyhedron defined by a system of linear inequalities:

$$\mathcal{E}_{M,s} := \{A(M, s)y \leq b(M, s)\}$$

Theorem 4.1 (LASSO Selection Event Defines a Polyhedron)**.** *[8] Under normal linear model $y \sim \mathcal{N}(\mu, \sigma^2 I_n)$, the event where LASSO solution support $\hat{M}$ and sign vector $\hat{s}$ match fixed values (M, s)*

$$\mathcal{E}_{M,s} := \{\hat{M}(y) = M,\ \hat{s}(y) = s\}$$

is a polyhedron of the form:

$$\mathcal{E}_{M,s} = \{y : A(M, s)\, y \leq b(M, s)\}$$

where $A(M, s) \in \mathbb{R}^{m \times n}$, $b(M, s) \in \mathbb{R}^m$ are deterministic matrix/vector constructed from X, λ, M, s.

Proof Sketch: Analyze KKT conditions in detail and convert selection condition "support M, sign s" to linear inequality system on y. Each inequality derives directly from LASSO optimality, resulting in a convex polyhedron.

Definition 4.1 (Truncated Normal Distribution)**.** Let $y \sim \mathcal{N}(\mu, \sigma^2 I_n)$ and condition on polyhedron $P = \{z : Az \leq b\}$. The conditional distribution $y|y \in P$ is a normal distribution restricted to polyhedron P, denoted:

$$y|(y \in P) \sim \mathrm{TruncatedNormal}_P(\mu, \sigma^2 I_n)$$

Theorem 4.2 (Polyhedral Lemma)**.** *[8] Let $y \sim \mathcal{N}(\mu, \sigma^2 I_n)$ and condition on polyhedron $P = \{z : Az \leq b\}$. For arbitrary vector $\eta \in \mathbb{R}^n$, the conditional distribution of linear functional $T := \eta^\top y$ is:*

$$T|(y \in P,\, P_{\eta^\perp} y = z) \sim \mathrm{TN}(a, b; \mu_T, \sigma_T^2)$$

where $\mathrm{TN}(a, b; \mu, \sigma^2)$ is the truncated normal distribution on interval $[a, b]$.

Proof Sketch: Decompose y into direction η and orthogonal direction: $y = t\eta + u$ ($u \perp \eta$). Constraint $Ay \leq b$ converts to linear inequalities on t, uniquely determining permissible interval $[V^-, V^+]$ for t. Thus $t|u$ becomes a normal distribution truncated to an interval. (See Appendix A.4 for full proof.)

4.4 Testing and p-values with Polyhedral Lemma

Goal: Output p-values from conditional distribution The basic idea of selective inference is: Condition on the fact "Model M was selected by LASSO", evaluate distribution of test statistic for a coefficient (or linear functional), and calculate p-value. Typically assume $y \sim \mathcal{N}(\mu, \sigma^2 I_n)$ and exact quantity of interest $T := \eta^\top y$.

Intuitive Background Standard t-test assumes "T follows Normal (or t) distribution". Selective inference adds condition "y is in polyhedron $\mathcal{P}_{M,s}$", so

$$T|y \in \mathcal{P}_{M,s} \sim \text{Truncated Normal}$$

Calculating p-values based on this conditional distribution can (at least theoretically) correct selection bias.

4.5 What Are We Testing? (Definition of Target Parameter)

Selective inference calculates p-values using distribution conditional on "which model was selected", but we must clarify **what parameter's hypothesis is being tested**.

Global Parameter vs Post-Selection Parameter Classical regression often considers $H_0 : \beta_j^* = 0$. Here β^* is coefficient vector in full model $y = X\beta^* + \varepsilon$, a parameter determined **independently of variable selection**.

On the other hand, applying OLS to model M after LASSO selection yields coefficients corresponding to "**best coefficients for selected sub-model** M":

$$\theta^*(M) := \underset{\theta \in \mathbb{R}^{|M|}}{\arg\min}\, \mathbb{E}\big[(y - X_M\theta)^2\big]$$

Global β_j^* and sub-model $\theta^*(M)$ are generally different quantities.

Intuitive Understanding "Is X_j really effective?" has different answers depending on whether asked in context of full model or model M selected by LASSO. Selective inference often deals with hypotheses regarding "**coefficients within the selected model**".

Definition as Conditional Parameter In selective inference framework, we fix specific situation where model M and sign s are selected ($\mathcal{E}_{M,s}$), consider distribution of y under this condition (truncated normal), and test $H_0 : \eta^\top \mu = \theta_0$ for linear functional $\eta^\top \mu$. The "parameter" here is $\eta^\top \mu$, quantity projected by linear map $\eta^\top$ from global mean vector $\mu = \mathbb{E}[y]$. Accurate answer to "What are we testing?" is "Hypothesis regarding linear functional $\eta^\top \mu$ under selection event $\mathcal{E}_{M,s}$".

Practical Interpretation Practically, interpreting as follows causes less confusion:

- Accept model M and sign s selected by LASSO as "candidate set of explanatory variables of interest now".
- Under that condition, test hypothesis "whether coefficient of this variable is 0" with p-value based on truncated normal.

- Thus, p-value measures how rare it is to get such large (or small) coefficient estimate given the fact "this variable was selected by LASSO".

This is not testing $\beta_j^* = 0$ in full model directly, but meaningful as "local inference premised on model selected by LASSO".

4.6 Practical Notes (Minimum p-value, Multiplicity, Finite Sample)

Numerical Verification: We compare naive p-values (inflated) vs selective inference p-values (uniform) using: `experiments/03_post_selection_inference.R` **Setup**: We simulated a "Global Null" scenario where **all** true coefficients are zero ($\beta^* = 0$) with $n = 100, p = 500$. Even with no signal, Lasso selects some variables by chance. We test the hypothesis $H_0 : \beta_j = 0$ for these selected variables.

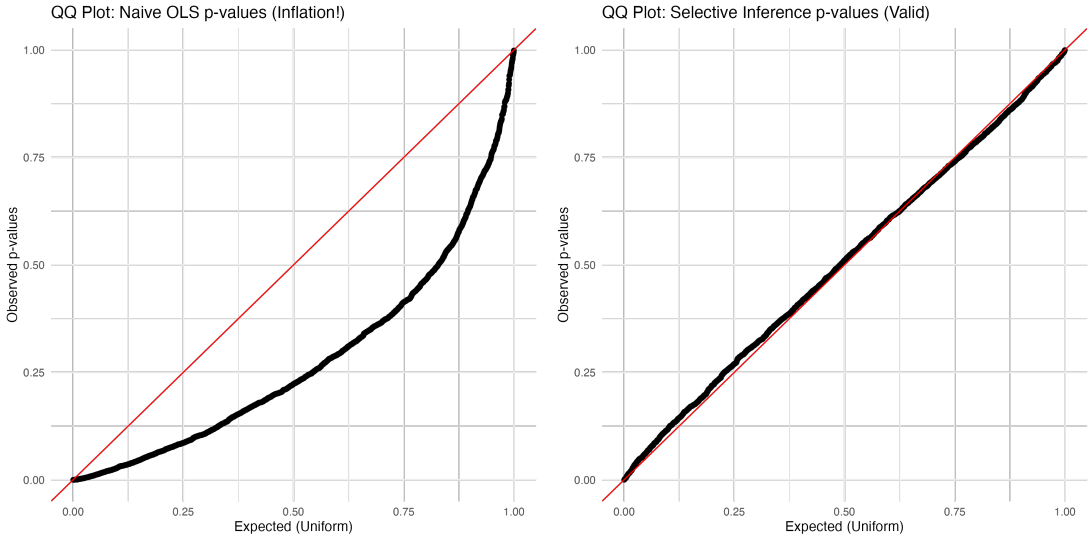

Figure 6: **Comparison of p-values.** Left: Naive OLS p-values after Lasso selection are inflated (falling below the diagonal), leading to specific Type I errors. Right: Selective Inference p-values follow the Uniform(0,1) distribution (on the diagonal), demonstrating validity under the null. This illustrates the importance of the method in Section 4.3.

Selective inference is theoretically powerful framework giving "p-values corrected for selection bias", but pitfalls remain in practice.

Viewing Minimum p-value Itself is a Source of Bias As mentioned in 4.2, if we look at minimal p-value $p_{\min}$ from variable set $\hat{M}$ selected by LASSO, multiplicity problem resurfaces. Even if each p_j satisfies certain Type I error control via selective inference, probability that $p_{\min}$ falls below threshold increases with number of variables.

Intuition Thinking "Safe because it's selective inference p-value" and emphasizing only minimal p-value involves operation "minimum among candidates decided after looking". Source of bias is also in act of "looking only at minimum".

Handling Multiplicity: What is Solved, What Remains Selective inference corrects bias due to condition "model was selected", but **does not solve all multiple testing problems**.

- "Method hunting" (trying multiple model selection methods and reporting only convenient results) inflates Type I error rate.

- Testing haphazardly multiple λ values or parameters on same LASSO path breaks family-wise error control.

Relevant practical measures (Bonferroni correction, FDR control) remain important.

Approximation Error in Finite Samples Many theoretical results assume idealized "y is strictly normal", "X is fixed". Under finite samples or model misspecification:

- **Distribution Approximation Error**: If actual error distribution is non-normal/heteroscedastic, p-values based on truncated normal can be conservative or optimistic.
- **Selection Event Approximation**: Implementations may use simplified selection events, causing deviation from theoretical conditional distribution.

One should read p-values assuming potential error of several to dozen percent.

4.7 Alternatives (Sample Splitting / Knockoffs / Debiased Lasso)

Selective inference is just one framework.

Sample Splitting Split data into two: 1. Select variables (LASSO etc.) with one part. 2. Perform standard OLS + classical t-test on selected variables with the other part. Pros: Simple, avoids selection bias. Cons: Reduces effective sample size, lowers power.

Knockoffs Create "fake variables" $\tilde{X}$ corresponding to original X, input $[X, \tilde{X}]$ into model, and control FDR by comparing importance of "real" vs "fake". Focal point: Controlling FDR (proportion of false positives).

Debiased LASSO Add linear correction to LASSO solution $\hat{\beta}$ to construct "nearly unbiased estimator" $\tilde{\beta}$ following asymptotic normal distribution, then construct CI/p-values. Aim: "Global inference" (inference on full model coefficients) not dependent on selection.

4.8 Standard Inference via De-biased Lasso

The user may ask: "Is there a method that provides standard confidence intervals and p-values like OLS, even for Lasso?" The answer is **YES**, via the **De-biased Lasso** (or Desparsified Lasso) [10, 6].

Why Standard Lasso Fails (Bias): Lasso estimator $\hat{\beta}$ is biased towards zero due to shrinkage. This bias prevents $\sqrt{n}(\hat{\beta} - \beta^*)$ from converging to a Normal distribution, making standard Wald statistics (Estimate / SE) invalid.

The De-biasing Idea: We can recover asymptotic normality by "adding back" the bias. Define the **De-biased Estimator** $\hat{b}$:

$$\hat{b} = \hat{\beta} + \frac{1}{n}\hat{\Theta}X^\top(y - X\hat{\beta})$$

where $\hat{\Theta}$ is an approximate inverse of the Gram matrix $\hat{\Sigma} = \frac{1}{n}X^\top X$. In high dimensions, $\hat{\Sigma}$ is singular, so $\hat{\Theta}$ is estimated using Nodewise Lasso (predicting each X_j from others).

Result (Asymptotic Normality): Under suitable conditions, the de-biased estimator follows:

$$\sqrt{n}(\hat{b}_j - \beta_j^*) \xrightarrow{d} \mathcal{N}(0, \omega_{jj})$$

This returns us to the familiar regression framework!

- **Confidence Intervals**: $\hat{b}_j \pm 1.96 \cdot \frac{\hat{\sigma}}{\sqrt{n}}\sqrt{\omega_{jj}}$
- **Hypothesis Testing**: Test $H_0 : \beta_j^* = 0$ using $Z = \frac{\hat{b}_j}{\mathrm{SE}}$.

Unlike Selective Inference (conditional), De-biased Lasso provides **unconditional** inference for all coefficients in the full model.

4.9 Bridging to Practice

Minimal Recipe for Practice: 1. Use LASSO (or other regularization) for candidate narrowing and prediction model construction (Ch 2-3). 2. Use cross-validation or holdout for prediction performance (Sec 2.7). 3. If p-values/CIs for coefficients are needed: - Use selective inference implementations (e.g. R packages) if possible. - If difficult, consider simple methods like sample splitting. 4. Avoid trusting naive OLS+t-test after LASSO.

Connection to Chapter 5: Discussions so far were in classical linear model + LASSO framework. Chapter 5 extends sparsity and regularization concepts to Deep Learning.

5 Sparsity and Deep Learning

5.1 Reinterpreting Sparsity (Parameter, Representation, Activity)

So far, we focused on parameter sparsity (β^* components being mostly 0). In broader machine learning context, "sparsity" has wider meanings.

(1) Parameter Sparsity Type handled in LASSO: $\beta \in \mathbb{R}^p$ has $s \ll p$ non-zeros. Important for variable selection, interpretation, estimation in high dimensions.

(2) Representation Sparsity Sparsity in input representation. e.g., images/audio become sparse in certain bases (wavelets, dictionary learning). $x \approx D\alpha$, α is sparse. Important for efficient representation. Deep learning intermediate representations often interpreted as having this sparsity.

(3) Activity Sparsity In deep neural nets, activation functions like ReLU $\phi(x) = \max(0, x)$ cause many units to output 0 for each data point. "Active units" for a given input x are sparse. Effects: "Routing" behavior where effective sub-network differs for each input; suppresses effective model complexity.

5.2 ReLU and Active Sparsity (Effective Subnets per Input)

ReLU acts as a "switch that mercilessly zeroes out negative inputs". Although weights W are shared, the path actually taken (sequence of active units) differs for each input. A "huge network" contains many different sub-networks, and only a part is selected and used for each input.

5.3 Two Lines of Regularization (Explicit vs Implicit)

LASSO used **Explicit Regularization** (L1 penalty in objective). Deep Learning often relies on **Implicit Regularization**.

Implicit Regularization: Even without explicit penalties, optimization algorithms (SGD), initialization, and architecture (residual, ReLU) implicitly favor specific types of solutions (e.g., minimum norm solutions, sparse filters).

5.4 Lottery Ticket Hypothesis and Pruning

Theorem 5.1 (Lottery Ticket Hypothesis)**.** *[5] Dense NN N_0 contains a sparse sub-network N_m (winning ticket) such that if N_m is trained from original initialization, it matches or exceeds accuracy of N_0.*

Pruning: 1. Train large network. 2. Zero out "low importance" weights (pruning). 3. Retrain. Similarity to LASSO: "Select few important ones from many". Difference: LASSO does training/sparsification simultaneously; Pruning is "Train large $\rightarrow$ Cut $\rightarrow$ Retrain".

5.5 Computational Aspects of Sparse Learning

Sparsity is important not just statistically but computationally. **Sparse Matrix Operations**: - Skip multiplication/addition for zero elements. - Compress memory (CSR, CSC). Crucial for large-scale models.

Trade-off: Too much sparsity hurts accuracy. Random sparsity is hard to accelerate on hardware (vs structured sparsity).

5.6 Open Problems (Gap between High-Dimensional Inference and DL)

1. **"Oracle Inequality" for Deep Models**: How does error rate depend on parameters/depth/sparsity in DL? 2. **Post-Selection Inference for Non-linear Models**: Can we correct selection bias after architecture search/hyperparameter tuning? 3. **Theoretical Understanding of Implicit Sparsity**: Why/when do SGD/Adam select sparse solutions?

6 Theoretical Synthesis: Bridging HDLSS and Sparse Estimation

This section synthesizes the discussions on High-Dimensional Standardized Learning (HDLSS/Aoshima Theory) and sparse estimation (Lasso) into a unified theoretical framework. We clarify the conditions under which sparse estimation remains valid in HDLSS regimes and bridge the gap using Restricted Eigenvalue (RE) conditions.

6.1 The HDLSS Framework: NSSE vs. SSE

In Aoshima's HDLSS theory [1], the covariance structure is classified based on the presence of "strong spike eigenvalues," which determine whether asymptotic normality holds.

- **NSSE (Non-Strongly Spiked Eigenvalue) Model**: The maximum eigenvalue is not dominant relative to the total variance.

$$\frac{\lambda_{\max}(\Sigma)^2}{\mathrm{tr}(\Sigma^2)} \to 0 \quad (\text{as } p \to \infty)$$

  Intuitively, no single principal component dominates the entire noise structure.

- **SSE (Strongly Spiked Eigenvalue) Model**: The maximum eigenvalue is dominant.

$$\liminf_{p\to\infty} \frac{\lambda_{\max}(\Sigma)^2}{\mathrm{tr}(\Sigma^2)} > 0$$

  In this regime, standard asymptotic normality is often broken due to the influence of strong spikes.

6.2 Strategy: Sparse vs. Non-Sparse Estimation

From the HDLSS perspective, the decision to use sparse estimation (like Lasso) depends not only on coefficient sparsity but on whether the **covariance (noise) structure** supports valid inference.

- **Under NSSE**: High-dimensional geometric phenomena (concentration of measure, asymptotic normality) hold. This provides a stable foundation for standard estimator behavior.
- **Under SSE**: The strong spikes can overwhelm the signal or distort the geometry. Directly applying sparse estimation may fail because the "noise" side is too dominant. A typical strategy is to first remove the spikes (e.g., via PCA or dual-covariance methods) and then apply estimation to the residual space.

6.3 The Common Language: RE Condition

The theoretical guarantee for Lasso (Oracle Inequality) relies on the design matrix satisfying specific geometric conditions, most notably the **Restricted Eigenvalue (RE) condition** or the **Compatibility condition**.

- **RE Condition**: Requires the minimum eigenvalue of the Gramm matrix restricted to a specific cone to be positive (a form of restricted strong convexity).
- **Compatibility Condition**: A slightly weaker condition sufficient for deriving the fast prediction error rate $O(1/n)$.

Connecting HDLSS to Lasso requires checking if NSSE implies these conditions.

6.4 Theoretical Logic: NSSE $\implies$ Good Condition Number $\implies$ RE Condition

This section provides a stepwise derivation of why the NSSE model is naturally compatible with the RE condition required for Lasso, mediated by the condition number.

6.4.1 Step 1: NSSE $\implies$ Good Condition Number $\kappa(\Sigma)$

Let the eigenvalues of Σ be $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$. The condition number is defined as $\kappa(\Sigma) = \frac{\lambda_{\max}(\Sigma)}{\lambda_{\min}(\Sigma)} = \frac{\lambda_1}{\lambda_p}$.

- **SSE (Strong Spike)**: Typically $\lambda_1 \asymp p^\alpha$ ($\alpha \geq 1/2$). If λ_p is bounded, $\kappa(\Sigma) \to \infty$.
- **NSSE (Non-Strong Spike)**: Defined by $\lambda_1^2/\mathrm{tr}(\Sigma^2) \to 0$. This prohibits λ_1 from dominating the spectrum. If $\lambda_1 = O(1)$ and $\lambda_p \geq c > 0$ (Non-degeneracy), then:

$$\kappa(\Sigma) \approx \frac{O(1)}{c} = O(1).$$

Conclusion 1: NSSE suppresses the explosion of the maximum eigenvalue, keeping the population condition number $\kappa(\Sigma)$ small.

6.4.2 Step 2: Good Condition Number $\implies$ RE Condition Holds

Why does a good population condition number lead to the **sample** RE condition? According to concentration of measure theory (e.g., Raskutti et al., 2010), the sample covariance $\hat{\Sigma}$ satisfies the RE condition with constant $\phi_{RE} > 0$ if:

$$\lambda_{\min}(\Sigma) > C\lambda_{\max}(\Sigma)\sqrt{\frac{s\log p}{n}}$$

Dividing by $\lambda_{\min}(\Sigma)$, this is equivalent to:

$$1 > C \cdot \kappa(\Sigma) \cdot \sqrt{\frac{s\log p}{n}} \iff n > C^2\kappa(\Sigma)^2 s\log p$$

Conclusion 2: If $\kappa(\Sigma)$ is small (NSSE case), the required sample size is the standard $n \propto s\log p$. If $\kappa(\Sigma) \to \infty$ (SSE case), the required n explodes, and Lasso fails.

6.5 Theorem: NSSE-based Oracle Inequality

We formalize the above logic into a theorem for the prediction error of Lasso under NSSE.

Theorem 6.1 (NSSE-Oracle Inequality)**.** *Consider the linear model $y = X\beta^* + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ and exact sparsity $s = |S|$. Assume:*

(A) ***Design****: Rows $x_i \sim subG(0, \Sigma)$.*

(B) ***NSSE****: $\frac{\lambda_{\max}(\Sigma)^2}{tr(\Sigma^2)} \to 0$ (Ensures concentration).*

(C) ***Non-degeneracy****: $\lambda_{\min}(\Sigma) \geq m > 0$ (Ensures $\kappa(\Sigma)$ is bounded).*

(D) ***Sample Size****: $n \gtrsim \frac{tr(\Sigma^2)}{d_{eff}} s^2 \log p$ (where $d_{eff} = tr(\Sigma^2)/\lambda^2_{\max}$).*

Then, with high probability, the Lasso estimator $\hat{\beta}$ with $\lambda \asymp \sigma\sqrt{\frac{\log p}{n}}$ satisfies:

$$\frac{1}{n}\|X(\hat{\beta} - \beta^*)\|_2^2 \leq C\frac{\sigma^2 s \log p}{n}.$$

Proof Sketch. 1. **From NSSE to Compatibility**: Using assumptions (B) and (C), the condition number $\kappa(\Sigma)$ is controlled. The sample size condition (D) (equivalent to (11') in Section 6.6) ensures that $\|\hat{\Sigma} - \Sigma\|_\infty$ is sufficiently small to preserve the positivity of eigenvalues on the restricted cone. Thus, the Compatibility Condition holds with constant $\phi^2_{comp} \geq m/2$.

2. **From Compatibility to Oracle Inequality**: Given the Compatibility Condition, the standard Lasso analysis (Basic Inequality + Cone Restriction) applies directly (as detailed in Appendix **??**), yielding the fast rate $O(s \log p/n)$.

For a complete step-by-step derivation, see Appendix **??**. □

6.6 Reformulating Sample Complexity via NSSE Geometry

The sample complexity requirement (Assump. D) can be intuitively rewritten using NSSE measures. Standard condition:

$$n \gtrsim \lambda_{\max}(\Sigma)^2 s^2 \log p \tag{11}$$

Using $d_{\text{eff}} = \text{tr}(\Sigma^2)/\lambda_{\max}(\Sigma)^2 = 1/\eta$:

$$n \gtrsim \frac{\text{tr}(\Sigma^2)}{d_{\text{eff}}} s^2 \log p \tag{11'}$$

This explicitly shows that a "more NSSE" regime (higher effective dimension d_{eff}) relaxes the sample size burden relative to the total variance structure.

6.7 Conclusion and Future Direction

While the RE condition is standard in Lasso theory, re-interpreting it through HDLSS measures offers a bridge between the two fields.

- **Theoretical Contribution**: Explicitly characterizing the "sample size cost" of Lasso in terms of the NSSE ratio η.
- **Practical Strategy**: A two-step procedure: (1) Test for SSE/NSSE, (2) If SSE, apply transformation; If NSSE, apply Lasso. This provides a robust pipeline for high-dimensional data analysis.

7 Comprehensive Analysis of Compatibility Condition: Theory and Peripheral Conditions

7.1 Overview

The theoretical properties of the Lasso estimator in sparse high-dimensional regression depend heavily on regularity conditions regarding the design matrix X. This section provides a comprehensive analysis of the **Compatibility Condition**, which serves as a foundation for Lasso theory, integrating its rigorous mathematical definition, proof, and relationships with peripheral conditions.

7.2 Mathematical Definition and Role

7.2.1 Formal Definition of Compatibility Constant

Let $S \subseteq \{1, \ldots, p\}$ be an index set with cardinality $s = |S|$. The **Compatibility Constant** is defined as follows [11]:

$$\phi^2_{\text{comp}}(S) := \min_{\beta \neq 0,\, \|\beta_S\|_1 \leq L\|\beta_{S^c}\|_1} \frac{s\beta^\top \hat{\Sigma}\beta}{\|\beta_S\|_1^2} \tag{1}$$

where $\hat{\Sigma} = X^\top X/n$ is the empirical covariance matrix, β_S denotes the coefficients restricted to S, and β_{S^c} denotes coefficients on the complement. In the Lasso context, the constant $L = 3$ is typically used [4].

The Compatibility Condition is said to hold if $\phi^2_{\text{comp}}(S) > 0$. The naming comes from the intuition that the ℓ_1-norm and the $L_2(Q)$-norm (norm related to the measure Q) must be "compatible".

7.2.2 Intuitive Interpretation

The minimization of the compatibility constant operates over non-zero vectors satisfying the constraint $\|\beta_{S^c}\|_1 \leq L\|\beta_S\|_1$. This constraint corresponds to the **Cone Condition**, a geometric property that the Lasso error vector satisfies with high probability. A large $\phi^2_{\text{comp}}(S)$ implies that the ℓ_2 norm grows sufficiently fast relative to the ℓ_1 norm of the coefficients in S, indicating a higher degree of independence among variables. Conversely, a small constant suggests high correlation between variables in S and those outside.

7.3 Oracle Inequality and Proof

7.3.1 Main Theorem (van de Geer & Bühlmann)

Consider the high-dimensional linear regression model $Y = X\beta^* + \varepsilon$ with $p \gg n$. Define the Lasso estimator as:

$$\hat{\beta}(\lambda) := \arg\min_{\beta} \left\{ \frac{1}{n}\|Y - X\beta\|_2^2 + \lambda\|\beta\|_1 \right\}$$

If the Compatibility Condition $\phi^2_{\text{comp}}(S^*) \geq \phi_0 > 0$ holds on the true support S^* and $\lambda \geq 2\lambda_0$ (where λ_0 is the noise level), then:

$$\frac{1}{n}\|X(\hat{\beta} - \beta^*)\|_2^2 + \lambda\|\hat{\beta} - \beta^*\|_1 \leq \frac{4\lambda^2 s}{\phi^2_{\text{comp}}(S^*)} \tag{2}$$

where $s = |S^*|$ is the sparsity level.

7.3.2 Proof Sketch

Step 1: Derivation of the Cone Condition From the KKT conditions of Lasso, the error vector $\Delta = \hat{\beta} - \beta^*$ satisfies with high probability:

$$\|\Delta_{(S^*)^c}\|_1 \leq 3\|\Delta_{S^*}\|_1$$

This implies $\Delta \in \mathcal{C}(S^*, 3) := \{\beta : \|\beta_{(S^*)^c}\|_1 \leq 3\|\beta_{S^*}\|_1\}$.

Step 2: Application of Restricted Strong Convexity (RSC) For $\Delta \in \mathcal{C}(S^*, 3)$, the Compatbility Condition ensures a lower bound on the quadratic form:

$$\frac{1}{n}\|X\Delta\|_2^2 \geq \frac{\phi^2_{\text{comp}}(S^*)}{s}\|\Delta_{S^*}\|_1^2$$

This follows directly from the definition: $\phi^2_{\text{comp}}(S^*)\|\Delta_{S^*}\|_1^2 \leq s\Delta^\top\hat{\Sigma}\Delta$.

Step 3: Oracle Inequality Combining the Basic Inequality and the Cone Condition:

$$\frac{1}{n}\|X\Delta\|_2^2 + \lambda\|\Delta\|_1 \leq \frac{4\lambda^2 s}{\phi^2_{\text{comp}}(S^*)}$$

7.4 Comparative Analysis of Peripheral Conditions

7.4.1 Restricted Eigenvalue Condition (REC)

For S with $|S| \leq m$, the Restricted Eigenvalue constant is defined as [3]:

$$\kappa^2_{\text{RE}}(m, \alpha) := \min_{|S|\leq m}\ \min_{\beta\neq 0, \|\beta_{S^c}\|_1 \leq \alpha\|\beta_S\|_1} \frac{\beta^\top\hat{\Sigma}\beta}{\|\beta_S\|_2^2}$$

Key Relationship: By the Cauchy-Schwarz inequality $\|\beta_S\|_1 \leq \sqrt{s}\|\beta_S\|_2$, we have:

$$\phi^2_{\text{comp}}(S) \geq \kappa^2_{\text{RE}}(|S|, 3)$$

Thus, the Compatibility Condition is **weaker** than the RE Condition.

Quantitative Comparison:

Condition	Lower Bound Form	Target Norm	Strength
REC	$\beta^\top\hat{\Sigma}\beta/\|\beta_S\|_2^2$	ℓ_2-norm	Strong
Compatibility	$\beta^\top\hat{\Sigma}\beta/\|\beta_S\|_1^2$	ℓ_1-norm	Medium

7.4.2 Mutual Incoherence Property (MIP)

For normalized columns of X, the mutual coherence is $\mu := \max_{i\neq j}|x_i^\top x_j|$. The MIP requires $\mu < 1/(2s-1)$ for exact recovery. While intuitive, MIP is generally **much stronger** than Compatibility and often too restrictive for Lasso to achieve optimal statistical rates.

7.4.3 Restricted Isometry Property (RIP)

A matrix satisfies s-RIP if for all s-sparse vectors δ, $(1-\delta_s)\|\delta\|_2^2 \leq \|X\delta\|_2^2/n \leq (1+\delta_s)\|\delta\|_2^2$. RIP is stricter than Compatibility, requiring isometry on all sparse vectors, whereas Compatibility only requires a one-sided bound on a restricted cone.

7.4.4 Irrepresentable Condition (IRC)

The Irrepresentable Condition is necessary and sufficient for **variable selection consistency** (support recovery) [13, 7]:

$$\|X_{(S^*)^c}^\top X_{S^*}(X_{S^*}^\top X_{S^*})^{-1}\text{sign}(\beta_{S^*}^*)\|_\infty < 1 - \eta$$

Role Difference:

- **Compatibility**: Required for Prediction Error and ℓ_1 Estimation Error.
- **IRC**: Required for Exact Support Recovery.

IRC is stronger than Compatibility. Under Compatibility, we may only guarantee variable screening ($\hat{S} \supseteq S^*$), while IRC guarantees $\hat{S} = S^*$.

7.5 Statistical Implications and Practical Applications

7.5.1 Estimation Error Rate

Under Compatibility $\phi_{\text{comp}}^2 > 0$ and $\lambda \asymp \sqrt{\log p/n}$, we achieve the rate:

$$\|\hat{\beta} - \beta^*\|_2 = O_P\left(\sqrt{\frac{s\log p}{n}}\right)$$

This matches the oracle rate up to a $\sqrt{\log p}$ factor.

7.5.2 Variable Screening (Beta-min condition)

If we assume the "Beta-min" condition:

$$\min_{j\in S^*}|\beta_j^*| > \frac{4\lambda s}{\phi_{\text{comp}}^2(S^*)}$$

then with high probability $\hat{S} \supseteq S^*$, meaning Lasso successfully screens for all active variables.

7.5.3 Verification of Compatibility

Verifying Compatibility is NP-hard, but sufficient conditions exist. For instance, if rows of X are i.i.d. sub-Gaussian with $\lambda_{\min}(\Sigma) > 0$ and $s = O(\sqrt{n/\log p})$, then $\phi_{\text{comp}}^2 \geq C\lambda_{\min}(\Sigma)$ holds with high probability [11].

7.6 Hierarchy of Conditions

The hierarchy of regularity conditions is summarized below:

Condition	Necessary?	Sufficient For	Verification
RIP (Strongest)	No	All Sparse Tasks	Hard
IRC	Yes (for Select.)	Variable Selection	Hard
Compatibility	No	Prediction / ℓ_1 Est.	Medium
MIP	No	Greedy Methods	Easy

7.7 Conclusion

The Compatibility Condition is the **most flexible and statistically optimal** regularity condition for Lasso in terms of prediction. It is weaker than REC and RIP, making it satisfied by a broader class of design matrices, while still sufficient to derive fast convergence rates. For practical high-dimensional analysis, understanding this condition helps in assessing the reliability of Lasso estimators, particularly distinguishing between prediction goals (Compatibility) and selection goals (IRC).

8 Summary

Theory of sparse estimation and post-selection inference is a major contribution to statistics in high-dimensional era. It evolved from LASSO to Oracle Inequalities, Consistency, and Selection Bias correction. With rise of Deep Learning, sparsity concepts expanded to representation and activity. Navigating between theory/practice and classical/modern methods is essential for future development.

A Proofs of Main Theorems

A.1 Proof of Basic Inequality (Lemma 3.1)

Proof. Let $Q(\beta) := \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$. By definition, the LASSO estimator $\hat{\beta}$ satisfies

$$Q(\hat{\beta}) \leq Q(\beta) \quad \forall \beta \in \mathbb{R}^p.$$

In particular, choosing $\beta = \beta^*$ (the true parameter), we have

$$\frac{1}{2n}\|y - X\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|y - X\beta^*\|_2^2 + \lambda\|\beta^*\|_1.$$

Expanding the squared L2 terms:

$$\|y - X\hat{\beta}\|_2^2 = \|(y - X\beta^*) - X(\hat{\beta} - \beta^*)\|_2^2 = \|\varepsilon - X(\hat{\beta} - \beta^*)\|_2^2 = \|\varepsilon\|_2^2 - 2\varepsilon^\top X(\hat{\beta} - \beta^*) + \|X(\hat{\beta} - \beta^*)\|_2^2.$$

Similarly, $\|y - X\beta^*\|_2^2 = \|\varepsilon\|_2^2$. Substituting these back:

$$\frac{1}{2n}\left(\|\varepsilon\|_2^2 - 2\varepsilon^\top X(\hat{\beta} - \beta^*) + \|X(\hat{\beta} - \beta^*)\|_2^2\right) + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|\varepsilon\|_2^2 + \lambda\|\beta^*\|_1.$$

Sample size n cancels in the expectation, and here we simplify:

$$\frac{1}{2n}\|X(\hat{\beta} - \beta^*)\|_2^2 - \frac{1}{n}\varepsilon^\top X(\hat{\beta} - \beta^*) + \lambda\|\hat{\beta}\|_1 \leq \lambda\|\beta^*\|_1.$$

Rearranging terms yields:

$$\frac{1}{2n}\|X(\hat{\beta} - \beta^*)\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \lambda\|\beta^*\|_1 + \frac{1}{n}\varepsilon^\top X(\hat{\beta} - \beta^*).$$

Using duality of norms ($|\langle u, v\rangle| \leq \|u\|_\infty\|v\|_1$), we can bound the stochastic term:

$$\frac{1}{n}\varepsilon^\top X(\hat{\beta} - \beta^*) \leq \|\frac{1}{n}X^\top\varepsilon\|_\infty\|\hat{\beta} - \beta^*\|_1.$$

This completes the derivation of the Basic Inequality form useful for Oracle Inequalities. □

A.2 Proof of Oracle Inequality (Theorem 3.1)

Proof. Let $\Delta := \hat{\beta} - \beta^*$. From the Basic Inequality derived above:

$$\frac{1}{2n}\|X\Delta\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \lambda\|\beta^*\|_1 + \|\frac{1}{n}X^\top\varepsilon\|_\infty\|\Delta\|_1.$$

Let the event $\mathcal{A} = \{\|\frac{1}{n}X^\top\varepsilon\|_\infty \leq \lambda/2\}$. With appropriate choice of $\lambda \asymp \sigma\sqrt{\log p/n}$, $\mathbb{P}(\mathcal{A}) \to 1$. Conditioning on $\mathcal{A}$:

$$\frac{1}{2n}\|X\Delta\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \lambda\|\beta^*\|_1 + \frac{\lambda}{2}\|\Delta\|_1.$$

Add $\frac{\lambda}{2}\|\hat{\beta}-\beta^*\|_1 = \frac{\lambda}{2}\|\Delta\|_1$ to both sides? Better, decompose indices into support $S = \text{supp}(\beta^*)$ and S^c.

$$\|\hat{\beta}\|_1 = \|\beta^* + \Delta\|_1 = \|\beta^*_S + \Delta_S\|_1 + \|\Delta_{S^c}\|_1 \geq \|\beta^*_S\|_1 - \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1.$$

Also $\|\beta^*\|_1 = \|\beta^*_S\|_1$. Substitute these into the inequality:

$$\frac{1}{2n}\|X\Delta\|_2^2 + \lambda(\|\beta^*_S\|_1 - \|\Delta_S\|_1 + \|\Delta_{S^c}\|_1) \leq \lambda\|\beta^*_S\|_1 + \frac{\lambda}{2}(\|\Delta_S\|_1 + \|\Delta_{S^c}\|_1).$$

Canceling $\lambda\|\beta^*_S\|_1$:

$$\frac{1}{2n}\|X\Delta\|_2^2 + \lambda\|\Delta_{S^c}\|_1 - \lambda\|\Delta_S\|_1 \leq \frac{\lambda}{2}\|\Delta_S\|_1 + \frac{\lambda}{2}\|\Delta_{S^c}\|_1.$$

$$\frac{1}{2n}\|X\Delta\|_2^2 + \frac{\lambda}{2}\|\Delta_{S^c}\|_1 \leq \frac{3\lambda}{2}\|\Delta_S\|_1.$$

This implies two things: 1. **Cone Condition**: $\frac{\lambda}{2}\|\Delta_{S^c}\|_1 \leq \frac{3\lambda}{2}\|\Delta_S\|_1 \implies \|\Delta_{S^c}\|_1 \leq 3\|\Delta_S\|_1$. 2. **Error Bound**: From RE condition definition $\kappa(S)$, since Δ is in the cone:

$$\frac{\|X\Delta\|_2}{\|\Delta_S\|_2} \geq \kappa > 0 \implies \|X\Delta\|_2^2 \geq \kappa^2\|\Delta_S\|_2^2.$$

Substitute back:

$$\frac{\kappa^2}{2n}\|\Delta_S\|_2^2 \leq \frac{3\lambda}{2}\|\Delta_S\|_1 \leq \frac{3\lambda}{2}\sqrt{s}\|\Delta_S\|_2.$$

Dividing by $\|\Delta_S\|_2$:

$$\|\Delta_S\|_2 \leq \frac{3\lambda\sqrt{s}n}{\kappa^2 n} = \frac{3\lambda\sqrt{s}}{\kappa^2}.$$

Finally, $\|\Delta\|_2^2 \leq (1+3^2)\|\Delta_S\|_2^2$ (roughly) or simply bound $\|\Delta\|_2$. More precisely, $\|\Delta\|_2^2 \leq \|\Delta_S\|_2^2 + \|\Delta_{S^c}\|_1^2$ is loose, but usually we bound prediction error $\|X\Delta\|_2^2$ or parameter error $\|\Delta\|_2^2 \lesssim \lambda^2 s$. Substituting $\lambda \asymp \sqrt{\frac{\log p}{n}}$, we get $\|\hat{\beta}-\beta^*\|_2^2 \lesssim \frac{s\log p}{n}$. □

A.3 Proof of Sparse Recovery (Theorem 3.2)

Proof. Let $S = \text{supp}(\beta^*)$. We require $\hat{\beta}_{S^c} = 0$ and $\text{sgn}(\hat{\beta}_S) = \text{sgn}(\beta^*_S)$. From KKT conditions, $\hat{\beta}$ is a solution iff:

1. $X_S^\top(y - X\hat{\beta}) = n\lambda\hat{z}_S$ where $\hat{z}_S = \text{sgn}(\hat{\beta}_S)$.
2. $\|X_{S^c}^\top(y - X\hat{\beta})\|_\infty \leq n\lambda$.

Assume $\hat{\beta}_{S^c} = 0$ (Primal Witness Construction). Then $\hat{\beta}_S$ satisfies normal usage on restricted X_S.

$$X_S^\top(y - X_S\hat{\beta}_S) = n\lambda\,\text{sgn}(\beta^*_S) \quad \text{(assuming signs match)}.$$

Since $y = X_S\beta^*_S + \varepsilon$:

$$X_S^\top X_S(\beta^*_S - \hat{\beta}_S) + X_S^\top\varepsilon = n\lambda\,\text{sgn}(\beta^*_S).$$

$$\beta^*_S - \hat{\beta}_S = (X_S^\top X_S)^{-1}(n\lambda\,\text{sgn}(\beta^*_S) - X_S^\top\varepsilon).$$

For sign consistency, we need $|\beta^*_j| > |\beta^*_j - \hat{\beta}_j|$. Taking norms, this requires roughly $\min|\beta^*_j| \gtrsim \lambda\|(X_S^\top X_S)^{-1}\|_\infty$. This leads to the Minimal Signal Condition.

For support consistency (checking non-support violation), plug $\hat{\beta}_S$ into condition 2:

$$\|X_{S^c}^\top(y - X_S\hat{\beta}_S)\|_\infty \leq n\lambda.$$

Substitute $y - X_S\hat{\beta}_S = X_S(\beta_S^* - \hat{\beta}_S) + \varepsilon$:

$$\|X_{S^c}^\top X_S(\beta_S^* - \hat{\beta}_S) + X_{S^c}^\top \varepsilon\|_\infty \leq n\lambda.$$

Using the expression for $\beta_S^* - \hat{\beta}_S$:

$$\|X_{S^c}^\top X_S(X_S^\top X_S)^{-1}(n\lambda \operatorname{sgn}(\beta_S^*) - X_S^\top \varepsilon) + X_{S^c}^\top \varepsilon\|_\infty \leq n\lambda.$$

This inequality relies on the term $X_{S^c}^\top X_S(X_S^\top X_S)^{-1}$ being small (Irrepresentable Condition) and noise ε being small. If Irrepresentable Condition holds ($< 1 - \gamma$) and $\lambda \gtrsim \sqrt{\log p/n}$, this holds with high probability. □

A.4 Proof of Polyhedral Lemma (Theorem 4.2)

Proof. We want the distribution of y conditional on $Ay \leq b$. Decompose y with respect to vector η: Let $P_\eta = \eta\eta^\top/\|\eta\|^2$.

$$y = P_\eta y + (I - P_\eta)y = \left(\frac{\eta^\top y}{\|\eta\|^2}\right)\eta + z$$

where $z = (I - P_\eta)y$ is the component orthogonal to η. Since $y \sim \mathcal{N}(\mu, \Sigma)$, y is Gaussian. $\eta^\top y$ and z are uncorrelated (independent if isotropic $\Sigma = I$). Condition on z (and selection event). The selection event $Ay \leq b$ becomes:

$$A\left(\frac{\eta^\top y}{\|\eta\|^2}\eta + z\right) \leq b \iff (A\eta)\frac{\eta^\top y}{\|\eta\|^2} \leq b - Az.$$

Let $T = \eta^\top y$. This is a set of linear inequalities on scalar T:

$$(A\eta)_i \cdot \frac{T}{\|\eta\|^2} \leq b_i - (Az)_i \quad \forall i.$$

For each constraint i:

- If $(A\eta)_i > 0$: $T \leq \frac{b_i - (Az)_i}{(A\eta)_i}\|\eta\|^2$.
- If $(A\eta)_i < 0$: $T \geq \frac{b_i - (Az)_i}{(A\eta)_i}\|\eta\|^2$.
- If $(A\eta)_i = 0$: Constraint doesn't involve T (check validity given z).

Taking min of upper bounds and max of lower bounds:

$$\mathcal{V}^-(z) \leq T \leq \mathcal{V}^+(z).$$

Thus, conditional on z and $Ay \leq b$, T is restricted to $[\mathcal{V}^-, \mathcal{V}^+]$. Since T was unconditionally Gaussian, conditional on lying in an interval (and independent z), it follows a Truncated Normal distribution. □

A.5 Equivalence of Penalized and Constrained Forms

Theorem A.1 (Equivalence of Lasso Formulations)**.** *The following two optimization problems are equivalent in the sense that for any $\lambda \geq 0$, there exists a $t \geq 0$ such that their solution sets coincide (and vice-versa).*

$$(P_1) \quad \min_\beta \frac{1}{2n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_1$$

$$(P_2) \quad \min_\beta \frac{1}{2n}\|y - X\beta\|_2^2 \quad s.t. \quad \|\beta\|_1 \leq t$$

Proof. This is a standard result of convex optimization using Lagrangian duality. Let $f(\beta) = \frac{1}{2n}\|y - X\beta\|_2^2$ and $g(\beta) = \|\beta\|_1$. Both functions are convex. Consider the constrained problem (P_2). The Lagrangian function is:

$$\mathcal{L}(\beta, \gamma) = f(\beta) + \gamma(\|\beta\|_1 - t)$$

where $\gamma \geq 0$ is the Lagrange multiplier (dual variable). By the KKT conditions, optimal β^* and γ^* must satisfy:

1. Stationarity: $0 \in \partial f(\beta^*) + \gamma^* \partial \|\beta^*\|_1$.
2. Primal Feasibility: $\|\beta^*\|_1 \leq t$.
3. Dual Feasibility: $\gamma^* \geq 0$.
4. Complementary Slackness: $\gamma^*(\|\beta^*\|_1 - t) = 0$.

The first condition ($0 \in \partial f(\beta^*) + \gamma^* \partial \|\beta^*\|_1$) is exactly the optimality condition for the penalized problem (P_1) with $\lambda = \gamma^*$. Thus, a solution to (P_2) with parameter t is also a solution to (P_1) with parameter $\lambda = \gamma^*$.

Conversely, given λ, a solution $\hat{\beta}_\lambda$ to (P_1) satisfies the KKT conditions for (P_2) with $t = \|\hat{\beta}_\lambda\|_1$, setting $\gamma^* = \lambda$. Therefore, there is a one-to-one mapping between the regularization path parameters λ and t. □

References

[1] Makoto Aoshima and Kazuyoshi Yata. Two-stage estimation procedures for high-dimension, low-sample-size data given the cross-data-matrix methodology. *Statistica Sinica*, 28(1):53–76, 2018.

[2] Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

[3] Peter J Bickel, Ya'acov Ritov, and Alexandre B Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[4] Peter Bühlmann and Sara Van De Geer. *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media, 2011.

[5] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. In *International Conference on Learning Representations*, 2019.

[6] Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *The Journal of Machine Learning Research*, 15(1):2869–2909, 2014.

[7] Snigdhansu N Lahiri. Necessary and sufficient conditions for variable selection consistency of the lasso. *The Annals of Statistics*, 49(2):819–854, 2021.

[8] Jason D Lee, Dennis L Sun, Yuekai Sun, and Jonathan E Taylor. Exact post-selection inference, with application to the lasso. *The Annals of Statistics*, 44(5):2413–2447, 2016.

[9] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288, 1996.

[10] Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *The Annals of Statistics*, 42(3):1166–1202, 2014.

[11] Sara A van de Geer. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

[12] Martin J Wainwright. Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming. *IEEE Transactions on Information Theory*, 55(5):2183–2202, 2009.

[13] Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.