

Sparse High-Dimensional Regression: Exact Scalable Algorithms and Phase Transitions

Dimitris Bertsimas & Bart Van Parys (2020)

November 30, 2025

Background

Problem (Best Subset Selection).

Given input data $X = (x_1, \dots, x_n)^\top \in \mathbb{R}^{n \times p}$ and response $Y = (y_1, \dots, y_n)^\top \in \mathbb{R}^n$, we seek a **k -sparse** linear predictor:

$$\min_{w \in \mathbb{R}^p} \frac{1}{2\gamma} \|w\|_2^2 + \frac{1}{2} \|Y - Xw\|_2^2 \quad \text{s.t.} \quad \|w\|_0 \leq k.$$

Why sparsity?

- ▶ In high dimensions ($p \gg n$), restricting to few variables improves **interpretability** and guards against **overfitting**. [3]
- ▶ Many scientific domains require **explicit variable selection** (e.g., genetics, networks, text), so the goal is not just prediction but identifying a small set of truly relevant features. [3]

Challenges

Computational barrier.

- ▶ Best subset selection solves the *right* ℓ_0 problem, but is **NP-hard** and traditionally scales poorly. [4, 1]

Convex surrogates are scalable but imperfect.

- ▶ ℓ_1 relaxation (Lasso) is efficient, yet may yield **biased estimates** and **unstable supports**. [3, 5]

Gap. Can we obtain **exact** sparse regression *at scale*?

This paper: a new optimization view + scalable algorithm for high-dimensional best subset selection. [1]

Main Idea & Contributions

Main idea.

Reformulate best subset selection into a **convex integer optimization** problem in binary variables, and solve it via a tailored **cutting-plane / outer-approximation** algorithm.

Contributions.

- ▶ **New reformulation** eliminates big- M constants and yields a pure binary convex program.
- ▶ **Scalable algorithm** with fast updates and warm starts, enabling problems with n, p up to 10^5 .
- ▶ **Empirical insights**: reveals a **statistical & computational phase transition** where exact subset selection becomes easy beyond a sample-size threshold.

Talk roadmap.

Reformulation \rightarrow Algorithm \rightarrow Phase transition theory \rightarrow Experiments.

Reformulation I: From ℓ_0 to Binary Selection

Start from the ℓ_0 problem.

We want at most k nonzero coefficients in w :

$$\min_{w \in \mathbb{R}^p} \frac{1}{2\gamma} \|w\|_2^2 + \frac{1}{2} \|Y - Xw\|_2^2 \quad \text{s.t.} \quad \|w\|_0 \leq k. \quad (1)$$

Introduce binary selectors.

Let $s \in \{0, 1\}^p$ indicate which variables are used:

$$S_k^p := \left\{ s \in \{0, 1\}^p : \mathbf{1}^\top s \leq k \right\}.$$

Then subset selection becomes: choose $s \in S_k^p$ and fit w only on active features.

Reformulation II-a: Kernel (Gram) from a Selected Subset

For each column $X_{:j} \in \mathbb{R}^n$, define a rank-1 Gram matrix

$$K_j := X_{:j} X_{:j}^\top \in \mathbb{R}^{n \times n}.$$

Reformulation II-a: Kernel (Gram) from a Selected Subset

For each column $X_{:j} \in \mathbb{R}^n$, define a rank-1 Gram matrix

$$K_j := X_{:j} X_{:j}^\top \in \mathbb{R}^{n \times n}.$$

Given $s \in S_k^p$, the selected design matrix is

$$X_s = [X_{:j_1}, \dots, X_{:j_k}] \quad \text{where } s_{j_t} = 1.$$

Reformulation II-a: Kernel (Gram) from a Selected Subset

For each column $X_{:j} \in \mathbb{R}^n$, define a rank-1 Gram matrix

$$K_j := X_{:j} X_{:j}^\top \in \mathbb{R}^{n \times n}.$$

Given $s \in S_k^p$, the selected design matrix is

$$X_s = [X_{:j_1}, \dots, X_{:j_k}] \quad \text{where } s_{j_t} = 1.$$

Gram expansion = sum of column outer products.

$$X_s X_s^\top = \sum_{t=1}^k X_{:j_t} X_{:j_t}^\top = \sum_{j=1}^p s_j X_{:j} X_{:j}^\top.$$

Reformulation II-a: Kernel (Gram) from a Selected Subset

For each column $X_{:j} \in \mathbb{R}^n$, define a rank-1 Gram matrix

$$K_j := X_{:j} X_{:j}^\top \in \mathbb{R}^{n \times n}.$$

Given $s \in S_k^p$, the selected design matrix is

$$X_s = [X_{:j_1}, \dots, X_{:j_k}] \quad \text{where } s_{j_t} = 1.$$

Gram expansion = sum of column outer products.

$$X_s X_s^\top = \sum_{t=1}^k X_{:j_t} X_{:j_t}^\top = \sum_{j=1}^p s_j X_{:j} X_{:j}^\top.$$

Define

$$K_s := \sum_{j=1}^p s_j K_j \quad \Rightarrow \quad \boxed{K_s = X_s X_s^\top}.$$

Intuition: the subset Gram matrix is the sum of Gram contributions from each selected feature.

Reformulation II-b: Eliminating w via a Kernel View

Inner ridge fit for a fixed subset.

For $s \in S_k^p$, solve ridge regression on X_s :

$$w_s^* = \arg \min_{w_s} \frac{1}{2} \|Y - X_s w_s\|_2^2 + \frac{1}{2\gamma} \|w_s\|_2^2. \quad (7)$$

Optimal value depends only on s .

Using the closed-form ridge solution (Woodbury identity) and $K_s = X_s X_s^\top$, the optimal objective becomes

$$c(s) = \frac{1}{2} Y^\top (I_n + \gamma K_s)^{-1} Y. \quad (10)$$

Resulting pure binary optimization.

$$\min_{s \in S_k^p} c(s) \quad \text{with } c(s) \text{ convex in } s. \quad (\text{CIO})$$

So best subset selection becomes a **convex integer optimization** problem in binary variables, with no big- M constants.

Algorithm Intuition

We have a convex-integer problem.

After reformulation:

$$\min_{s \in S_k^p} c(s), \quad s \in \{0, 1\}^p,$$

where $c(s) = \frac{1}{2}Y^\top(I_n + \gamma K_s)^{-1}Y$ is **convex and smooth in s** .

Key idea.

Convexity implies a global linear lower bound at any point $s^{(t)}$:

$$c(s) \geq c(s^{(t)}) + \nabla c(s^{(t)})^\top (s - s^{(t)}).$$

Each bound is a **cutting plane**.

Outer approximation.

Collect many cuts to form a tight lower envelope, then search over binary s using a master MIO that gets tighter each iteration [2].

Takeaway: exploit convex geometry to guide combinatorial search.

Algorithm Overview

Algorithm: Cutting-Plane / Outer Approximation for Best Subset Selection

Input: Data (X, Y) , sparsity level k , ridge parameter γ , tolerance ε .

Output: Globally optimal subset s^* (and coefficients w^*).

1. **Initialize.** Obtain an initial subset $s^{(0)}$ (e.g., from SOCP/greedy warm start), set cut set $\mathcal{C} \leftarrow \emptyset$.

2. **Evaluate objective and gradient.** For current $s^{(t)}$, compute

$$c(s^{(t)}) = \frac{1}{2} Y^\top (I_n + \gamma K_{s^{(t)}})^{-1} Y, \quad \nabla c(s^{(t)}).$$

3. **Add a cutting plane (linear lower bound).** Append the cut

$$\eta \geq c(s^{(t)}) + \nabla c(s^{(t)})^\top (s - s^{(t)}),$$

and update $\mathcal{C} \leftarrow \mathcal{C} \cup \{\text{new cut}\}$.

4. **Solve the master MIO.**

$$\min_{s \in S_k^p, \eta} \eta \quad \text{s.t. all cuts in } \mathcal{C}.$$

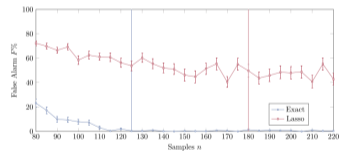
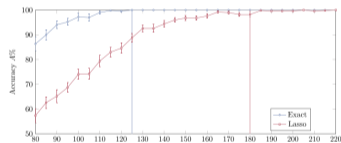
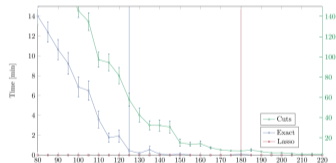
Let the solution be $(s^{(t+1)}, \eta^{(t+1)})$.

5. **Check convergence.** If $\eta^{(t+1)} \geq c(s^{(t+1)}) - \varepsilon$, stop and output $s^* = s^{(t+1)}$.

Empirical Teaser: Phase Transition

What happens as sample size n increases?

Exact best subset selection exhibits a sharp **phase transition**: from poor recovery to near-perfect recovery, and (surprisingly) from hard to easy computation.



Theory Setup for Phase Transition

Data model.

Assume a sparse linear model with true support S^* :

$$Y = Xw^* + \varepsilon, \quad |S^*| = k, \quad \varepsilon \sim \mathcal{N}(0, \sigma^2 I_n).$$

Two regimes.

- ▶ **Undersampled regime** ($n < n_t$): many subsets fit similarly well \Rightarrow hard recovery and heavy computation.
- ▶ **Oversampled regime** ($n > n_t$): true subset separates clearly \Rightarrow accurate recovery and easy computation.

Goal of theory.

Characterize the threshold n_t and show exact subset selection succeeds *earlier* than ℓ_1 surrogates.

Main Theoretical Message

Theorem. Statistical phase transition

Assume the design is uncorrelated ($\rho = 0$), set the ridge parameter $\gamma = 1/n$, and suppose $p - k > k$. Then there exist numerical constants $c_8, c_9 > 0$ (independent of n, k, p, σ^2) such that for all $\theta \geq 1$,

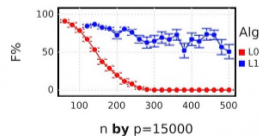
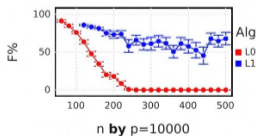
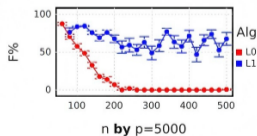
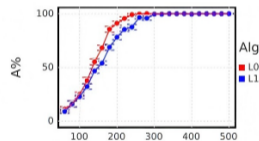
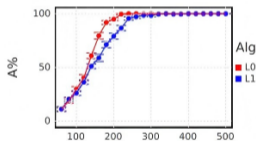
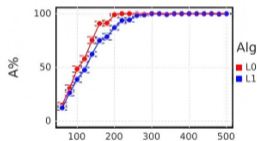
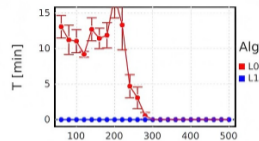
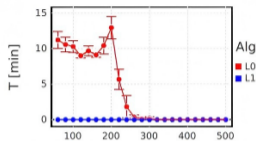
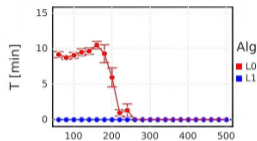
$$n \geq \theta n_1 \implies \mathbb{P}[s^\star = s_1^\star = s^{\text{true}}] \geq 1 - c_9 \exp(-\theta c_8).$$

Corollary. Earlier success than Lasso

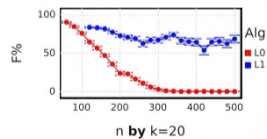
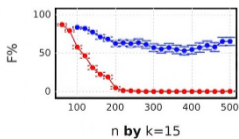
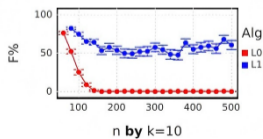
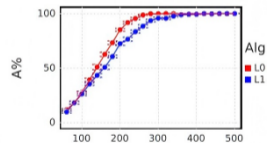
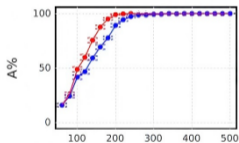
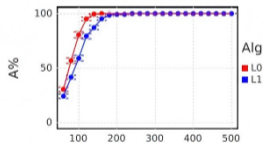
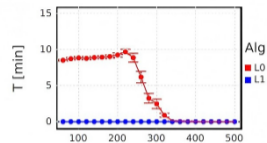
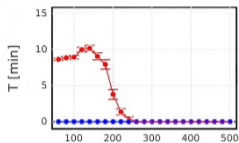
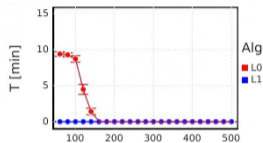
Exact ℓ_0 subset selection achieves full support recovery once n exceeds its threshold (around n_0 / n_t), and this occurs *strictly earlier* than the Lasso accuracy threshold n_1 :

$$n_t^{(\ell_0)} < n_1^{(\ell_1)}.$$

Phase Transition vs. Dimension p



Phase Transition vs. Sparsity k



References I

- [1] Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44(2):813–852, 2016.
- [2] Marco A. Duran and Ignacio E. Grossmann. An outer-approximation algorithm for a class of mixed-integer nonlinear programs. *Mathematical Programming*, 36(3):307–339, 1986.
- [3] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical Learning with Sparsity: The Lasso and Generalizations*. CRC Press, 2015.
- [4] B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.
- [5] Robert Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 58(1):267–288, 1996.