FAST LOCAL SEARCH FOR SPARSE LINEAR REGRESSION WITH TIGHT APPROXIMATION BOUNDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Finding the best fit with at most k features in sparse linear regression is computationally challenging, while existing local search methods often rely on strong curvature conditions or repeatedly scan all features when searching for an exchange. We propose *Subsampled Ridge-Exchange Local Search* (SRE-LS), which uses ridge-based support exchanges and balanced candidate sampling while maintaining a support of size k. We prove that SRE-LS achieves an $O(1/\rho)$ approximation with probability at least $1-\delta$ in $O((nd+nk^2)\log(1/\delta))$ time, where ρ is a response orthogonal regularity parameter. We also establish a matching $\Omega(1/\rho)$ worst-case inapproximability bound for randomized polynomial-time algorithms, unless $\mathrm{NP} \subseteq \mathrm{BPP}$, showing that the dependence on ρ is tight. Experiments on synthetic and real-world datasets show that SRE-LS achieves regression loss comparable to existing local-search methods while requiring less running time.

1 INTRODUCTION

Sparse linear regression is a fundamental problem in statistics and machine learning, with applications in variable selection, interpretable modeling, and high-dimensional estimation Tibshirani (1996); Zou & Hastie (2005); Hastie et al. (2009). In modern high-dimensional settings, the number of features can substantially exceed the number of observations, which often poses challenges for statistical estimation and variable selection (Fan & Lv, 2010; Bühlmann & Van De Geer, 2011; Wainwright, 2019). Sparse linear regression addresses this issue by restricting the predictor to a small subset of features while maintaining a good fit to the response vector. In this paper, we consider the following sparse linear regression problem.

Sparse Linear Regression Problem. Given a matrix $A \in \mathbb{R}^{n\times d}$, a response vector $b \in \mathbb{R}^n$, and an integer sparsity parameter k, the goal is to find $x \in \mathbb{R}^d$ satisfying $\|x\|_0 \le k$ that minimizes $\|Ax-b\|_2$. Let $f^\star = \mathrm{OPT}_k(A,b) \coloneqq \min_{\|x\|_0\le k} \|Ax-b\|_2$ denote the optimal residual. We focus on the regime $k \le \min\{n,d\}$ and assume throughout that $\|A_j\|_2 \le 1$ for every $j \in [d]$ and $0 < \|b\|_2 \le 1$, where A_j denotes the jth column of A.

Common approaches to sparse regression include regularization, greedy and thresholding algorithms, and combinatorial optimization methods Pati et al. (1993); Tibshirani (1996); Fan & Li (2001); Candes & Tao (2007); Blumensath & Davies (2009); Bertsimas et al. (2016). Regularization and greedy methods are computationally attractive, but their guarantees typically concern statistical recovery, regularized objectives, or convergence to a stationary solution, rather than approximation of the optimal residual under the same sparsity budget. Combinatorial methods enforce sparsity directly, but can be expensive to solve exactly. In this work, we focus on approximating the optimal k-sparse residual in polynomial time under the same sparsity constraint.

From an optimization perspective, obtaining approximation guarantees for arbitrary design matrices is challenging. The exact ℓ_0-constrained problem is NP-hard (Natarajan, 1995). For fixed n and ϵ, geometric and data-structural approaches obtain $(1+\epsilon)$-approximations for sparse regression and its variants, but retain a d^{k-1} dependence in query or running time (Har-Peled et al., 2018; Cardinal & Ooms, 2022). Under standard complexity assumptions, no polynomial-time algorithm can guarantee a finite multiplicative residual approximation over all instances, even with a constant-factor relaxation of the sparsity constraint (Price et al., 2022). In particular, when $\mathrm{OPT}_k(A,b) = 0$, any finite multiplicative guarantee requires an exact fit. Conditional hardness also persists for

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

certain additive-error guarantees with relaxed sparsity, even when a k-sparse exact fit exists (Foster et al., 2015). Motivated by these barriers, we focus on polynomial-time multiplicative residual approximation for instances with $\mathrm{OPT}_k(A, b) > 0$, without relaxing the sparsity constraint.

A closely related line of work studies sparse regression through single-column exchanges while maintaining exactly k selected columns. Fujii (2020) proposed a non-oblivious local search method whose curvature-based guarantee yields a squared-residual bound with an additive term depending on $\|b\|_2^2$. The guarantee assumes restricted strong concavity, requiring a positive restricted minimum eigenvalue over supports of size at most $2k$, and its approximation factor depends on this eigenvalue. The method scores exchanges using gradient and coefficient information, requiring a full-gradient computation at each iteration and $O(ndT + nk^2)$ time over T iterations. Thiery & Ward (2022) developed distorted local search under two-sided weak submodularity for general matroid constraints. Its residual guarantee contains an additive term, and exchange-score estimation requires repeated subset-regression evaluations with polynomial dependence on k. A complementary approach is adaptively regularized hard thresholding (ARHT) (Axiotis & Sviridenko, 2021), which achieves arbitrarily small additive error in squared residual under restricted strong convexity and smoothness. Its guarantee allows $O(k\widetilde{\kappa})$ selected columns, where $\widetilde{\kappa}$ is a restricted condition number. A detailed comparison is given in Table 1. These limitations motivate efficient exact-k local search with a multiplicative residual approximation without assuming restricted strong convexity.

A common difficulty in designing local-search algorithms for sparse regression is that the gain from adding a column is coupled with the loss from removing a selected column. A large insertion gain alone does not imply a large decrease after refitting, since one selected column must also be removed. A natural approach is to leverage ridge regularization, which provides the curvature needed to control such exchanges without assuming restricted strong convexity of the original objective. However, the resulting exchange analysis controls a regularized objective, which may remain large even for a sparse solution with small residual because of the coefficient penalty. Thus, reducing the regularized objective does not directly imply a multiplicative residual approximation. A separate computational challenge is to identify useful exchanges without evaluating all d candidate columns, since computing all exchange scores incurs $O(nd)$ time in each iteration.

1.1 Our Contributions

In this paper, we propose *Subsampled Ridge-Exchange Local Search* (SRE-LS), a randomized local-search algorithm for sparse regression that maintains exactly k selected columns. SRE-LS uses ridge-based exchange scores to control complete support replacements without assuming restricted strong convexity of the original objective. A key part of the analysis is to convert the decrease in the regularized objective into a multiplicative bound for the original residual through an exchange bound that captures the interaction between the response and sparse column spaces. This interaction is measured by response-orthogonal regularity ρ. For normalized instances with $\rho > 0$, SRE-LS achieves an $O(1/\rho)$ residual approximation with constant success probability. To implement these exchanges efficiently, SRE-LS constructs balanced candidate groups independently of the exchange scores and samples one group in each iteration. This limits candidate evaluation to $O(d/k)$ columns per iteration and $O(d)$ over k iterations, giving $O(nd + nk^2)$ running time with constant probability. We further establish an $\Omega(1/\rho)$ approximation lower bound for polynomial time algorithms, showing that the dependence on response-orthogonal regularity is tight up to constant factors. Specifically, our main contributions are summarized below. (See Table 1)

- For normalized instances with response-orthogonal regularity $\rho > 0$, SRE-LS maintains exactly k selected columns and achieves an $O(1/\rho)$ residual approximation with probability at least $1 - \delta$ in $O((nd + nk^2)\log(1/\delta))$ time.
- We prove an $\Omega(1/\rho)$ approximation lower bound for randomized polynomial-time algorithms on normalized rational instances with $\mathrm{OPT}_k > 0$, unless $\mathrm{NP} \subseteq \mathrm{BPP}$. Thus, the $O(1/\rho)$ dependence is tight up to constant factors.
- Experiments on synthetic and real datasets show that SRE-LS achieves regression loss comparable to the evaluated local search methods while reducing the running time, particularly on larger datasets.

Remark 1.1. For a column-normalized design A and nonzero response b, every instance with $f^\star > 0$ satisfies $\rho \in (0, 1]$. Thus, the guarantee does not require ρ to be bounded below by a fixed con-

Table 1: Comparison of local search approximation guarantees for sparse regression with $f^\star = \mathrm{OPT}_k(A,b) := \min_{\|x\|_0 \le k} \|Ax-b\|_2 > 0$. Here, $A \neq 0$, $\|A_j\|_2 \le 1$ for all $j \in [d]$, $0 < \|b\|_2 \le 1$, $\varepsilon, \eta, \delta \in (0,1)$, and $1 < k \le \min\{n,d\}$. The explained-gain guarantees of Fujii and Thiery–Ward are converted into residual bounds, where $\mathcal{E}_T$ and $\mathcal{E}_\eta$ are the approximation factors defined in the table notes.

Method	**Columns**	**Residual bound**	**Running time**	**Conditions**
Fujii (2020)[2]	k	$\sqrt{\mathcal{E}_T \cdot (f^\star)^2 + (1-\mathcal{E}_T)\|b\|_2^2}$	$O(ndT + nk^2)$	RSC/RSS[1], $\beta_{2k}^- > 0$
Axiotis & Sviridenko (2021)[3]	$O(k\widetilde{\kappa})$	$\sqrt{(f^\star)^2 + \varepsilon}$	$O(ndk \cdot \widetilde{\kappa}\Lambda_\varepsilon)$	RSC/RSS[1]
Thiery & Ward (2022)[4]	k	$\sqrt{\mathcal{E}_\eta \cdot (f^\star)^2 + (1-\mathcal{E}_\eta)\|b\|_2^2}$	$\widetilde{O}\left(\frac{ndk^6}{\eta^3}\right)$[5]	Weak submodularity
SRE-LS[6]	k	$O\left(\frac{1}{\rho}\right) \cdot f^\star$	$O\left((nd+nk^2)\log\frac{1}{\delta}\right)$	–
Lower bound	k	$\Omega\left(\frac{1}{\rho}\right) \cdot f^\star$	Randomized polynomial time	NP $\not\subseteq$ BPP

[1] *Restricted strong convexity and smoothness (RSC/RSS).* For $F(x) = \frac{1}{2}\|Ax-b\|_2^2$, the lower and upper r-sparse curvature constants β_r^- and β_r^+ are the tight constants in $\beta_r^-\|v\|_2^2 \le \|Av\|_2^2 \le \beta_r^+\|v\|_2^2$ for all r-sparse v. The condition $\beta_{2k}^- > 0$ is precisely $(2k)$-RSC, which means that every set of at most $2k$ columns of A is linearly independent.

[2] For Fujii (2020), $\mathcal{E}_T = \theta^2(1-e^{-T/(k\theta)})$, where $\theta = \beta_{2k}^-/\beta_2^+$ and T is the number of local-search iterations.

[3] For Axiotis & Sviridenko (2021, Theorem 10), the guarantee assumes RSC/RSS over the sparse directions considered by their algorithm, and $\widetilde{\kappa}$ denotes the corresponding restricted condition number. Let $\Lambda_\varepsilon = L_\varepsilon^2 \log(2 + dL_\varepsilon)$ with $L_\varepsilon = \log(2 + \|b\|_2^2/\varepsilon)$.

[4] For Thiery & Ward (2022, Theorem 11), $\mathcal{E}_\eta = [\gamma^2(1-e^{-\phi})/\phi - O(\eta)]_+$, where γ and β are the lower and upper submodularity ratios, and $[t]_+ = \max\{t,0\}$. The parameter η controls the accuracy.

[5] We use $\widetilde{O}$ notation to hide polylogarithmic factors in d, k, and $\frac{1}{\eta}$.

[6] SRE-LS *guarantees.* The response-orthogonal regularity $\rho \in (0,1]$ is fixed by the input (A,b,k) for the column-normalized design (Definition 3.4).

stant. The parameter ρ measures the minimum singular value obtained after projecting linearly independent sets of at most k columns onto $b^\perp$, and allows the original design to contain repeated or linearly dependent columns. Standard restricted isometry and coherence conditions further imply explicit lower bounds on ρ when the response is sufficiently separated from the corresponding column spans (Candes et al., 2006; Bickel et al., 2009). Notably, SRE-LS does not compute ρ and its guarantee applies to every instance with $\rho > 0$.

1.2 Other Related Work

Sparse regression has been extensively studied through convex and nonconvex regularization. The LASSO and related sparse estimators have been analyzed for variable selection, estimation, and inference (Tibshirani, 1996; Zou, 2006; Candes & Tao, 2007; Takahashi & Kabashima, 2018), while nonconvex alternatives include SCAD and MCP (Fan & Li, 2001; Zhang, 2010; Wang et al., 2013; Loh & Wainwright, 2015). Exact sparsity has also been studied using mixed-integer optimization, conic relaxations, and branch-and-bound methods (Bertsimas et al., 2016; Dong et al., 2015; Hazimeh et al., 2022). Related algorithmic approaches include sparse ridge approximation (Xie & Deng, 2020) and greedy or iterative methods such as OMP and OMPR, forward stagewise regression, iterative hard thresholding, and CoSaMP (Pati et al., 1993; Hastie et al., 2007; Blumensath & Davies, 2009; Needell & Tropp, 2009; Jain et al., 2011). These lines of work focus on statistical recovery, regularized objectives, or practical exact optimization, and do not provide polynomial-time multiplicative approximation of the residual relative to $f^\star$ under the same sparsity constraint.

Existing approximation algorithms for general fixed designs have an exponential dependence on k. For fixed n and ε, Har-Peled et al. (2018) obtain $(1+\varepsilon)$-approximate queries in $\widetilde{O}(d^{k-1})$ time after $\widetilde{O}(d^k)$ preprocessing, while Cardinal & Ooms (2022) obtain a randomized $(1+\varepsilon)$-approximation

without preprocessing in $O(d^{k-1}\log^{n-k+2} d)$ time. Under the condition $f^\star = 0$, Price et al. (2022) obtain an additive error solution in $(d\|b\|_2/\varepsilon)^{O(k)}\,\mathrm{poly}(n,k)$ time. Thus, these results do not provide a multiplicative approximation for $f^\star > 0$ in time polynomial in n, d, and k. Other approaches address complementary aspects of sparse regression without resolving the support-selection bottleneck. Oblivious sketching reduces the measurement dimension (Mai et al., 2023), while feature adaptation and distributed methods target different statistical or computational settings (Kelner et al., 2023; Ji et al., 2023; Maros & Scutari, 2022).

Sparse regression is NP-hard in general (Natarajan, 1995). On exact-fit instances, Price et al. (2022) rule out multiplicative approximation even with a constant factor sparsity violation, while Foster et al. (2015) prove hardness under simultaneous residual and sparsity relaxations, assuming that NP admits no randomized quasipolynomial-time algorithms. Under the minimum-weight k-clique conjecture, Gupte & Vaikuntanathan (2021) give a $d^{k-o(1)}$ time lower bound for exact threshold decision when $k \geq 4$. Our lower bound in Appendix B gives a matching worst-case $\Omega(1/\rho)$ hardness result when $f^\star > 0$, unless $\mathrm{NP} \subseteq \mathrm{BPP}$.

2 Preliminaries

For a positive integer n, let $[n] = \{1, \ldots, n\}$. For $x \in \mathbb{R}^d$ and $1 \leq p < \infty$, define $\|x\|_p = \left(\sum_{i=1}^d |x_i|^p\right)^{1/p}$. The support of x is $\mathrm{supp}(x) = \{i \in [d] : x_i \neq 0\}$, and $\|x\|_0 = |\,\mathrm{supp}(x)|$. We call x k-sparse if $\|x\|_0 \leq k$. For $i \in [d]$, let $e^i \in \mathbb{R}^d$ denote the i-th standard basis vector. The identity matrix of order m is denoted by I_m. For a matrix $A \in \mathbb{R}^{n\times d}$, we write $A = [a_1, \ldots, a_d]$, where a_j is its j-th column, and let A_{ij} denote its (i,j)-th entry. Its Frobenius and spectral norms are $\|A\|_F = \left(\sum_{i=1}^n \sum_{j=1}^d A_{ij}^2\right)^{1/2}$, and $\|A\|_2 = \max_{\|x\|_2=1} \|Ax\|_2$, respectively. We write $A^\top$ for the transpose and $A^\dagger$ for the Moore–Penrose pseudoinverse. For a matrix M with at least one column, define $\sigma_{\min}(M) = \min_{\|x\|_2=1} \|Mx\|_2$. Thus, $\sigma_{\min}(M) = 0$ whenever M is not full column rank. When M has positive rank, $\sigma^+_{\min}(M)$ is its smallest positive singular value. We use the notation $\tilde{O}(f(n))$ denotes $O(f(n)\cdot \mathrm{poly}(\log f(n)))$. For $I \subseteq [n]$ and $S \subseteq [d]$, let $A_{I,S}$ denote the submatrix with rows in I and columns in S. We denote $A_{[n],S}$ as A_S, and let x_S denote the restriction of x to S.

3 Local Search for Sparse Linear Regression

In this section, we develop *Subsampled Ridge-Exchange Local Search* (SRE-LS), a randomized local-search algorithm that maintains exactly k selected columns. Let $U = [u_1, \ldots, u_d]$ be obtained by normalizing the nonzero columns of A, leaving zero columns unchanged. For a design matrix M and a support $S \subseteq [d]$, write $f(M, b, S) = \min_{w\in\mathbb{R}^{|S|}} \|M_S w - b\|_2$. The column normalization preserves the least-squares residual on every support, and hence $|u_j|_2 \leq 1$ and $f(U, b, S) = f(A, b, S)$ for all $S \subseteq [d]$ with $|S| \leq k$. Proofs of auxiliary results are deferred to the appendix.

3.1 Algorithm Overview

A support exchange in sparse regression couples the gain from inserting a new column with the loss from removing a selected one. Thus, a large insertion gain alone does not imply a useful exact-k update after refitting. SRE-LS handles this coupling through the ridge support energy

$$\mathcal{E}(S) = \min_{w\in\mathbb{R}^{|S|}} \|U_S w - b\|_2^2 + \|w\|_2^2.$$

The ridge term makes every restricted system positive definite and allows the insertion gain and deletion loss to be controlled within the same quadratic objective. At a current support S, the algorithm assigns each sampled candidate an exchange score for the full support replacement, combining its insertion gain with the loss from removing a selected column with a small ridge coefficient. A positive score guarantees a decrease after the complete support replacement and refitting, while the support size remains exactly k.

To quantify the decrease available from these exchanges, we compare the current support with an arbitrary k-sparse vector y fixed before the candidate groups are constructed. Let T be a size-k

Algorithm 1 Subsampled Ridge-Exchange Local Search (SRE-LS)

Input Design matrix $A \in \mathbb{R}^{n \times d}$, response vector $b \in \mathbb{R}^n$, and $1 \le k \ll \min\{n, d\}$.
Output A support $S \subseteq [d]$ with $|S| = k$, and the corresponding least squares coefficients $\hat{w}$.

Normalize the nonzero columns of A to obtain U, and choose any fixed support $S \subseteq [d]$ with $|S| = k$.
Set $m \leftarrow \lceil d/(4k) \rceil$ and $d' \leftarrow 4km$.
Add dummy labels $d+1, \dots, d'$ so that there are exactly d' labels.
Draw three independent uniform permutations of $[d']$ and ignore the dummy labels when forming the candidate groups.
Compute $R_S = (U_S^\top U_S + I_{|S|})^{-1}$, $x_S = R_S U_S^\top b$, and $r_S = b - U_S x_S$.
Form candidate groups $\mathcal{B}_1, \dots, \mathcal{B}_{4k} \subseteq [d]$ according to equation 1.
for $t = 0, \dots, k-1$ **do**
  $q \in \arg\min_{j \in S} (x_S)_j^2$.
  Draw $s_t \sim \mathrm{Unif}([4k])$ independently.
  $\mathcal{C}_t \leftarrow \mathcal{B}_{s_t} \setminus S$.
  if $\mathcal{C}_t \neq \varnothing$ **then**
    For each $h \in \mathcal{C}_t$, compute $g_h = -2\langle u_h, r_S \rangle$, $\chi_S(q,h) = \lambda_{\max}\left(U_{\{q,h\}}^\top U_{\{q,h\}} + I_2\right)$, and $\widehat{\Gamma}_S(q,h) = \frac{g_h^2}{4\chi_S(q,h)} - \chi_S(q,h)(x_S)_q^2$.
    $h_t \in \arg\max_{h \in \mathcal{C}_t} \widehat{\Gamma}_S(q,h)$.
    if $\widehat{\Gamma}_S(q,h_t) > 0$ **then**
      $\mathcal{A} \leftarrow S \cup \{h_t\}$.
      Obtain $R_{\mathcal{A}}$ by block insertion.
      Compute $x_{\mathcal{A}}$.
      $j^+ \in \arg\min_{j \in \mathcal{A}} (x_{\mathcal{A}})_j^2 / (R_{\mathcal{A}})_{jj}$.
      Update $S \leftarrow \mathcal{A} \setminus \{j^+\}$.
      Obtain R_S by block deletion and update x_S and r_S.
    end if
  end if
end for
$\widehat{w} \in \arg\min_{w \in \mathbb{R}^k} \|A_S w - b\|_2$.
return $(S, \widehat{w})$.

set containing its support. We sum the exchange scores over the missing columns in $T \setminus S$. The resulting deletion penalties can be charged to the ridge coefficient mass on $S \setminus T$, while the quadratic comparison with y contains the same coefficient quantity through $\|y - x\|_2^2$, where x denotes the current ridge coefficients. This yields a lower bound on the total positive exchange score whenever $\mathcal{E}(S)$ is sufficiently above $\|Uy - b\|_2^2 + O(\|y\|_2^2)$. For a large energy gap, the quadratic term gives a stronger lower bound on the total positive exchange score. Thus, the ridge exchange scores can be combined to compare the current support with an arbitrary sparse vector without requiring the columns in $S \cup T$ to be linearly independent.

Evaluating these scores over all candidate columns would cost $O(nd)$ in each iteration. To reduce the candidate evaluations, SRE-LS constructs $O(k)$ balanced candidate groups independently of the exchange scores and samples one group in each iteration. For the fixed comparison support T, with constant probability its relevant columns are separated across these groups. Conditional on this event, uniform group sampling preserves an $\Omega(1/k)$ fraction of their total positive exchange score in expectation, even though the current support depends on the previously sampled groups. Combining this sampling property with the aggregate exchange bound over k iterations yields

$$\mathbb{E}\left[f(U, b, S_k)^2 \mid \mathscr{B}\right] \le \|Uy - b\|_2^2 + O(\|y\|_2^2),$$

with only $O(d)$ candidate scores are evaluated over all k iterations.

The remaining coefficient term prevents this comparison from immediately giving a multiplicative residual guarantee. For an optimal sparse prediction, we choose a representation supported on linearly independent columns and project it onto $b^\perp$. Since the response vanishes under this projection, the projected prediction has norm at most $f^\star$. Response-orthogonal regularity ρ lower bounds the projected singular values, yielding an optimal sparse vector $x^\star$ with $\rho\|x^\star\|_2 \le f^\star$. Substituting $x^\star$ into the preceding comparison gives an $O(1/\rho)$ residual approximation. Since only $O(d)$ candidate scores are evaluated over k iterations, SRE-LS runs in $O(nd + nk^2)$ time with constant probability.

3.2 SUPPORT REPLACEMENTS

In the following, we first construct a score that guarantees a decrease after a complete support replacement. Directly analyzing the least-squares objective is difficult caused from the insertion gain and deletion loss are coupled. Next, we evaluate replacements through the ridge support energy.

Definition 3.1. For a support $S \subseteq [d]$, the *ridge support energy* is the minimum regularized loss on the selected columns

$$\mathcal{E}(S) = \min_{w \in \mathbb{R}^{|S|}} \|U_S w - b\|_2^2 + \|w\|_2^2.$$

For nonempty S, the ridge energy has a unique minimizer even when the selected columns are dependent. Let $R_S = (U_S^\top U_S + I_{|S|})^{-1}$. The ridge term makes the system positive definite even when the selected columns are dependent, giving the unique minimizer

$$x_S = R_S U_S^\top b.$$

Let $r_S = b - U_S x_S$ denote the corresponding residual. Since $f(U, b, S)$ minimizes the residual on the same support, $f(U, b, S)^2 \le \|r_S\|_2^2 \le \mathcal{E}(S)$.

A replacement must remove one selected column together with the insertion. To limit the deletion term, choose $q \in \arg\min_{j \in S} (x_S)_j^2$ and pair each candidate $h \notin S$ with q when computing its exchange score. For each candidate, define the residual correlation $g_h = -2\langle u_h, r_S \rangle$. The joint quadratic change on q, h is controlled by the pair curvature

$$\chi_S(q, h) = \lambda_{\max}\left(U_{\{q,h\}}^\top U_{\{q,h\}} + I_2\right).$$

The insertion gain and deletion loss are combined through the exchange score

$$\widehat{\Gamma}_S(q, h) = \frac{g_h^2}{4\chi_S(q, h)} - \chi_S(q, h)(x_S)_q^2.$$

The first term measures the gain from adding h, while the second bounds the loss from removing q. The following lemma shows that this score guarantees a decrease after the full support replacement.

Lemma 3.2. *For a size-k support S, $q \in S$, and $h \notin S$, the score lower-bounds the ridge-energy decrease from replacing q by h*

$$\mathcal{E}(S \setminus \{q\} \cup \{h\}) \le \mathcal{E}(S) - \widehat{\Gamma}_S(q, h).$$

On the support $\mathcal{A} = S \cup \{h\}$, removing any $j \in \mathcal{A}$ has the exact deletion cost

$$\mathcal{E}(\mathcal{A} \setminus \{j\}) - \mathcal{E}(\mathcal{A}) = \frac{(x_{\mathcal{A}})_j^2}{(R_{\mathcal{A}})_{jj}}.$$

The first bound follows from a quadratic upper bound, while the deletion formula follows from the block-inverse identity. After inserting h and refitting on $\mathcal{A}$, the algorithm removes the column with the smallest deletion cost in theorem 3.2. Since $q \in \mathcal{A}$ is a feasible choice, the actual deletion incurs no larger increase than removing q. Hence, every candidate with $\widehat{\Gamma}_S(q, h) > 0$ decreases the ridge energy while restoring a size-k support.

For the comparison analysis, we use a simpler lower bound on the exchange score. The column normalization implies $1 \le \chi_S(q, h) \le 3$, and therefore

$$\Gamma_S(q, h) = \frac{g_h^2}{12} - 3(x_S)_q^2 \le \widehat{\Gamma}_S(q, h).$$

The algorithm selects candidates using $\widehat{\Gamma}_S$, while Γ_S is used only in the analysis. The gradient–coefficient form is related to the exchange criterion of Fujii (2020). Here, the curvature depends on the selected pair, and the deletion is chosen after insertion and ridge refitting.

Balanced candidate groups. Evaluating the exchange scores over all d columns in every iteration would be expensive. To restrict each iteration to $O(d/k)$ candidates, let $m = \lceil d/(4k) \rceil$ and $d' = 4km$, and introduce dummy labels $d+1, \ldots, d'$. Draw three independent uniform permutations π_1, π_2, π_3 of $[d']$. For each permutation, divide the labels into $4k$ consecutive blocks of size m, and define

$$\mathcal{B}_s = \bigcup_{\ell=1}^{3} \{j \in [d] \mid (s-1)m < \pi_\ell(j) \leq sm\}, \tag{1}$$

where $s \in [4k]$, each real column belongs to at most three groups, and every group contains at most $3m = O(d/k)$ columns.

3.3 Analysis of Sampled Exchanges and Residual Approximation

We now show that the balanced candidate groups retain enough of the exchange score associated with a fixed sparse comparison. Fix a k-sparse vector y and a size-k set T containing its support before constructing the groups $\mathscr{B}$. With constant probability, the columns in T can be assigned to distinct groups containing them, and we denote this event by M_T. Conditional on this event, the best candidate in a uniformly sampled group captures an $\Omega(1/k)$ fraction of the total positive exchange score over $T \setminus S_t$ in expectation. The matching and sampling argument is given in Appendix A.2.

The main comparison bound relates the total positive exchange score to the ridge-energy gap from y. Let $q_t \in \arg\min_{j \in S_t} (x_{S_t})_j^2$. For nonzero y, define $D_t = \big(\mathcal{E}(S_t) - \|Uy - b\|_2^2 - 9\|y\|_2^2\big)_+$. Then, the simplified exchange scores satisfy

$$\sum_{h \in T \setminus S_t} [\Gamma_{S_t}(q_t, h)]_+ \geq 3D_t + \frac{(D_t - 9\|y\|_2^2)_+^2}{12\|y\|_2^2}.$$

The linear term contracts the energy gap, while the quadratic term controls the regime in which the initial ridge energy is much larger than the comparison level. Combining this bound with the group-sampling estimate and Lemma 3.2 over k rounds gives the following residual comparison.

Lemma 3.3. *For every fixed k-sparse vector y and size-k set T containing its support, conditioning on any grouping $\mathscr{B} \in \mathsf{M}_T$ gives $\mathbb{E}\left[f(U, b, S_k)^2 \mid \mathscr{B}\right] \leq \|Uy - b\|_2^2 + 48\|y\|_2^2$, where the expectation is over the sampled groups in the k rounds. Consequently, with constant probability over the construction and selection of candidate groups,*

$$f(U, b, S_k) \leq 2\sqrt{\|Uy - b\|_2^2 + 48\|y\|_2^2}.$$

The comparison in Lemma 3.3 still contains the coefficient term $\|y\|_2^2$. For an optimal sparse fit, this term cannot be bounded directly by the residual, since nearly dependent columns may represent the same prediction using large coefficients. To address this issue, we introduce *response-orthogonal regularity*.

Definition 3.4. Let $b \neq 0$ and let $P_b^\perp = I_n - bb^\top/\|b\|_2^2$ be the orthogonal projection onto $b^\perp$. For a nonzero design $M \in \mathbb{R}^{n \times d}$, let $\mathcal{I}_k(M) = \{S \subseteq [d] : 1 \leq |S| \leq k,\ \mathrm{rank}(M_S) = |S|\}$ denote its independent column sets of size at most k. The *response-orthogonal regularity* is

$$\bar{\rho}_k(M, b) = \min_{S \in \mathcal{I}_k(M)} \sigma_{\min}(P_b^\perp M_S).$$

For $M = 0$, set $\bar{\rho}_k(0, b) = 0$.

This quantity relates the coefficient norm of a sparse fit to its component orthogonal to the response.

Lemma 3.5. *Let $\rho = \bar{\rho}_k(U, b)$ and $U \neq 0$. Then $\rho > 0$ if and only if $f^\star > 0$. When $\rho > 0$, a minimum-norm optimal k-sparse solution $x^\star$ satisfies $\rho\|x^\star\|_2 \leq f^\star$.*

The minimum in definition 3.4 excludes dependent supports, so dependent columns do not force ρ to vanish.Substituting Lemma 3.5 into Lemma 3.3 gives the main approximation guarantee.

Theorem 3.6. *Let $\rho = \bar{\rho}_k(U, b) > 0$, $\epsilon, \delta \in (0, 1)$. Algorithm 1 returns a support $\widehat{S}$ with $|\widehat{S}| = k$ such that, with probability at least $1 - \delta$,*

$$f(A, b, \widehat{S}) \leq \frac{14}{\rho} \mathrm{OPT}_k(A, b).$$

Then, the SRE-LS *algorithm runs in $O((nd + nk^2) \log \frac{1}{\delta})$ time.*

3.4 HARDNESS OF APPROXIMATION

We next show that the $1/\rho$ dependence in the approximation guarantee is unavoidable in the worst case, even for normalized instances with positive optimal residual.

Theorem 3.7. *Unless* NP $\subseteq$ BPP*, no randomized algorithm polynomial in the rational input length achieves a uniform $\alpha(\rho)$ residual approximation with probability at least $2/3$ while selecting at most k columns, where $\alpha(t) = o(1/t)$ as t approaches zero. This holds even for unit column norms, $\|b\|_2 = 1$, and $\rho > 0$, with $k \geq 2$ part of the input.*

The reduction starts from a gap version of Max k-Cover and appends an orthogonal block whose scale a controls the response-orthogonal regularity. The resulting instances satisfy $\rho = \Theta(a)$ and have positive optimal residual. The YES instances have $f^\star = O(a)$, while the NO instances have $f^\star = \Omega(1)$. Therefore, a uniform $o(1/\rho)$ approximation would distinguish the two cases for sufficiently small fixed $a > 0$, where the detailed analysis are given in Appendix B.

4 NUMERICAL EXPERIMENTS

We evaluate the proposed SRE-LS on one synthetic and four real-world datasets against existing local search methods for sparse regression. All experiments are conducted on a server with two AMD EPYC 7742 CPUs and 1.5 TB of RAM. All methods are implemented in MATLAB R2024b.

Datasets. We first consider a synthetic common-factor design with $n = 10{,}000$ observations and $d = 5{,}000$ columns. Let $g_0, g_1, \ldots, g_d \in \mathbb{R}^n$ and $Z \in \mathbb{R}^{n\times d}$ have independent standard Gaussian entries. For a correlation parameter $\rho \in (0, 1)$, we construct $A_0(:, j) = \sqrt{\rho}\, g_0 + \sqrt{1-\rho}\, g_j$, normalize the columns of A_0, add the perturbation $(\sigma/\sqrt{n})Z$ with $\sigma = 0.1$, and normalize the resulting columns again. We use $k = 20$ and select the support of $x^\star$ uniformly at random. Its nonzero entries contain an equal number of $+1$ and -1 coefficients. The response is generated as $b = Ax^\star + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2 I_n)$, and $\sigma_\epsilon^2 = \frac{\|Ax^\star\|_2^2}{n\,\mathrm{SNR}}$ with $\mathrm{SNR} = 20$. For real-world experiments, we use the feature matrices from MNIST ($n = 10{,}000, d = 784$), RCV1 ($n = 804{,}414, d = 47{,}236$), AmazonCat-13K ($n = 1{,}186{,}239, d = 203{,}882$), and Webspam ($n = 350{,}000, d = 680{,}715$), where the Webspam matrix is obtained after removing all-zero columns. For each dataset and sparsity level k, we select a support uniformly from the nonzero columns and construct a binary k-sparse vector $x^\star \in \{0, 1\}^d$. The response is generated as $b = Ax^\star + \nu \frac{\|Ax^\star\|_2}{\|e\|_2} e$, where $e \sim \mathcal{N}(0, I_n)$ and $\nu = 1/\sqrt{20}$. Here $x^\star$ is the planted vector used to generate the response and is not assumed to be an optimal k-sparse solution after noise is added. All four real-world datasets are publicly available and widely used in large-scale learning benchmarks.

Algorithms. We compare SRE-LS with four baselines: FujiiLS (Fujii, 2020), OMPR (Jain et al., 2011), Thiery–Ward (Thiery & Ward, 2022), and ARHT (Axiotis & Sviridenko, 2021). FujiiLS, Thiery–Ward, and ARHT are included in the theoretical comparison in Table 1. We additionally include OMPR as a closely related support-replacement baseline with RIP-based recovery guarantees.

Methodology. For the synthetic experiment, we vary the common correlation parameter ρ while fixing the sparsity level k. For the real-world datasets, we vary k over dataset-dependent ranges to evaluate scalability. We report the regression loss $f(A, b, S) = \min_{\mathrm{supp}(x)\subseteq S} \|Ax - b\|_2$, and the running time. All methods operate on column-normalized designs and use the same randomly generated size-k initial support. The returned support is refit by least squares on the original design before evaluation. SRE-LS uses $\lambda = 0.05$ and performs $10k$ sampled rounds. For ARHT, whose theoretical guarantee allows a larger output sparsity, we set the output support size to k for the same-sparsity experimental comparison. Randomized methods are run 10 times, while deterministic methods are run once.

Results for Sparse Regression. As shown in Figure 1, on the synthetic dataset, SRE-LS achieves regression loss close to OMPR, Thiery–Ward, and ARHT across the tested values of ρ, with substantially lower running time than Thiery–Ward and ARHT. On MNIST and RCV1 with $k = 20$, its regression loss is within 0.7% of the best value. SRE-LS is $17.6\times$ and $820\times$ faster than Thiery–Ward and ARHT on MNIST, and $25.8\times$ and $416\times$ faster on RCV1, respectively. On the larger

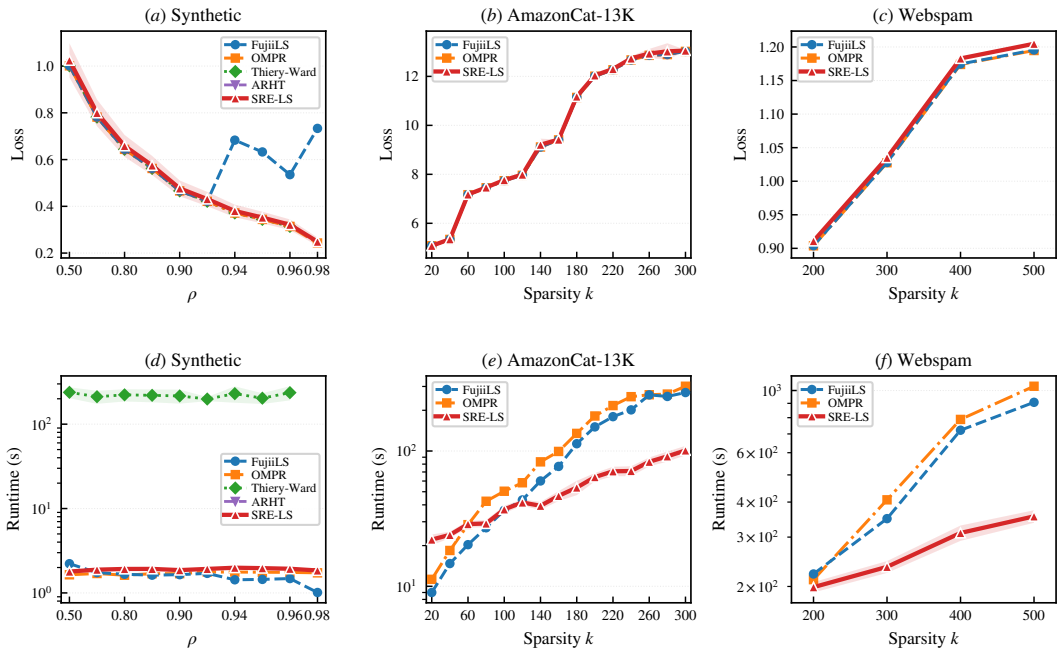

Figure 1: Comparison of regression loss and running time for local-search methods on a synthetic benchmark and two real-world datasets. (I) Panels (a) and (d) show the synthetic results with $n = 10{,}000$, $d = 5{,}000$, $k = 20$, $\sigma = 0.1$, and $\mathrm{SNR} = 20$, with varying synthetic parameter ρ. (II) Panels (b) and (e) report results on AmazonCat-13K ($n = 1{,}186{,}239, d = 203{,}882$), while panels (c) and (f) report results on Webspam ($n = 350{,}000, d = 680{,}715$), with varying k and $\mathrm{SNR} = 20$.

Table 2: Comparison results of five local search methods on MNIST ($n = 10{,}000, d = 784$) and RCV1 ($n = 804{,}414, d = 47{,}236$) with $k = 20$ and $\mathrm{SNR} = 20$.

Dataset	Metric	FujiiLS	OMPR	Thiery–Ward	ARHT	**SRE-LS**
MNIST	Loss	14776.2	**14758.8**	15544.7 ± 177.0	14769.8 ± 4.8	14854.8 ± 99.7
	Time (s)	0.39	0.37	5.47 ± 0.40	254.1 ± 93.0	$\mathbf{0.31 \pm 0.06}$
RCV1	Loss	**4.227**	**4.227**	5.265 ± 0.000	4.229 ± 0.000	4.229 ± 0.000
	Time (s)	**5.46**	6.07	279.3 ± 55.6	4511.5 ± 120.2	10.84 ± 0.53

AmazonCat-13K and Webspam datasets, SRE-LS maintains regression loss comparable to FujiiLS and OMPR and becomes faster as k increases. At $k = 500$ on Webspam, it is $2.6\times$ faster than FujiiLS and $2.9\times$ faster than OMPR. Additional results in Section C show a similar trend on RCV1 and AmazonCat-13K. The candidate-subsampling ablation reduces candidate evaluations by 54–$133\times$ and running time by 34–$54\times$ for $k = 40$–100, with less than 0.4% change in regression loss. Overall, SRE-LS achieves regression loss comparable to the evaluated methods while providing substantial running-time improvements, particularly on larger instances.

5 Conclusion

We presented SRE-LS, which uses ridge-based support exchanges with balanced subsampling for sparse linear regression. The algorithm maintains exactly k selected columns and performs $O(d)$ candidate evaluations over k rounds. For instances with response-orthogonal regularity $\rho > 0$, it achieves an $O(1/\rho)$ approximation to the residual with probability at least $1 - \delta$ in $O((nd + nk^2)\log\frac{1}{\delta})$ time. We also establish a matching worst-case $\Omega(1/\rho)$ inapproximability bound for randomized algorithms with polynomial time. An interesting direction is to extend this local search strategy to structured sparse models or constrained regression problems while preserving efficient candidate evaluation and exact structural constraints.

AI USE STATEMENT

In this paper, generative AI tools were used to assist with language editing, improve the clarity of the manuscript, and diagnostic code. The authors reviewed and revised all AI-assisted material and take responsibility for the final content of the manuscript, including the text produced with the assistance of generative AI. We take full responsibility for the final content and have carefully reviewed and revised all AI-assisted material.

ETHICS STATEMENT

This work studies approximation algorithms for sparse regression and does not involve human subjects or the collection of personal data. The experiments use standard publicly available benchmark datasets and synthetic data. We do not identify specific ethical concerns associated with the proposed methodology beyond those arising from the downstream use of sparse regression models in application specific settings.

REPRODUCIBILITY STATEMENT

We provide complete proofs of all theoretical results in the appendix, including the assumptions and auxiliary results used in the analysis. The algorithmic procedures and implementation details are described in the main text and appendix. The experimental section specifies the datasets, data construction and parameter settings, evaluation metrics, and computational environment, with additional results and ablations provided in appendix.

REFERENCES

Kyriakos Axiotis and Maxim Sviridenko. Sparse convex optimization via adaptively regularized hard thresholding. *Journal of Machine Learning Research*, 22(122):1–47, 2021.

Dimitris Bertsimas, Angela King, and Rahul Mazumder. Best subset selection via a modern optimization lens. *The Annals of Statistics*, 44:813–852, 2016.

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.

Thomas Blumensath and Mike E Davies. Iterative hard thresholding for compressed sensing. *Applied and Computational Harmonic Analysis*, 27(3):265–274, 2009.

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

Emmanuel Candes and Terence Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

Emmanuel J Candes, Justin K Romberg, and Terence Tao. Stable signal recovery from incomplete and inaccurate measurements. *Communications on Pure and Applied Mathematics: A Journal Issued by the Courant Institute of Mathematical Sciences*, 59(8):1207–1223, 2006.

Jean Cardinal and Aurélien Ooms. Algorithms for approximate sparse regression and nearest induced hulls. *Journal of Computational Geometry*, 13(1):377–398, 2022.

Ali Civril. A note on the hardness of sparse approximation. *Information Processing Letters*, 113 (14-16):543–545, 2013.

Hongbo Dong, Kun Chen, and Jeff Linderoth. Regularization vs. relaxation: A conic optimization perspective of statistical variable selection. *arXiv preprint arXiv:1510.06083*, 2015.

Jianqing Fan and Runze Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

Jianqing Fan and Jinchi Lv. A selective overview of variable selection in high dimensional feature space. *Statistica Sinica*, 20(1):101–148, 2010.

Uriel Feige. A threshold of $\ln n$ for approximating set cover. *Journal of the ACM*, 45(4):634–652, 1998.

Dean Foster, Howard Karloff, and Justin Thaler. Variable selection is hard. In *Proceedings of the 28th Conference on Learning Theory*, pp. 696–709, 2015.

Kaito Fujii. Approximation guarantees of local search algorithms via localizability of set functions. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 3327–3336, 2020.

Aparna Gupte and Vinod Vaikuntanathan. The fine-grained hardness of sparse linear regression. *arXiv preprint arXiv:2106.03131*, 2021.

Sariel Har-Peled, Piotr Indyk, and Sepideh Mahabadi. Approximate sparse linear regression. In *Proceedings of the 45th International Colloquium on Automata, Languages, and Programming*, pp. 77:1–77:14, 2018.

Trevor Hastie, Jonathan Taylor, Robert Tibshirani, and Guenther Walther. Forward stagewise regression and the monotone lasso. *Electronic Journal of Statistics*, 1:1–29, 2007.

Trevor Hastie, Robert Tibshirani, and Jerome H Friedman. *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer, 2009.

Hussein Hazimeh, Rahul Mazumder, and Ali Saab. Sparse regression at scale: Branch-and-bound rooted in first-order optimization. *Mathematical Programming*, 196(1-2):347–388, 2022.

Prateek Jain, Ambuj Tewari, and Inderjit Dhillon. Orthogonal matching pursuit with replacement. In *Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 1215–1223, 2011.

Yao Ji, Gesualdo Scutari, Ying Sun, and Harsha Honnappa. Distributed sparse regression via penalization. *Journal of Machine Learning Research*, 24(272):1–62, 2023.

Jonathan Kelner, Frederic Koehler, Raghu Meka, and Dhruv Rohatgi. Feature adaptation for sparse linear regression. In *Proceedings of the 37th Annual Conference on Neural Information Processing Systems*, pp. 29951–29995, 2023.

Po-Ling Loh and Martin J Wainwright. Regularized m-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *The Journal of Machine Learning Research*, 16(1):559–616, 2015.

Tung Mai, Alexander Munteanu, Cameron Musco, Anup Rao, Chris Schwiegelshohn, and David Woodruff. Optimal sketching bounds for sparse linear regression. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, pp. 11288–11316, 2023.

Marie Maros and Gesualdo Scutari. Acceleration in distributed sparse regression. In *Proceedings of the 36th Annual Conference on Neural Information Processing Systems*, pp. 36832–36844, 2022.

Balas Kausik Natarajan. Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, 24(2):227–234, 1995.

Deanna Needell and Joel A Tropp. Cosamp: Iterative signal recovery from incomplete and inaccurate samples. *Applied and Computational Harmonic Analysis*, 26(3):301–321, 2009.

Yagyensh Chandra Pati, Ramin Rezaiifar, and Perinkulam Sambamurthy Krishnaprasad. Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar conference on Signals, Systems and Computers*, pp. 40–44, 1993.

Eric Price, Sandeep Silwal, and Samson Zhou. Hardness and algorithms for robust and sparse optimization. In *Proceedings of the 39th International Conference on Machine Learning*, pp. 17926–17944, 2022.

Takashi Takahashi and Yoshiyuki Kabashima. A statistical mechanics approach to de-biasing and uncertainty estimation in lasso for random measurements. *Journal of Statistical Mechanics: Theory and Experiment*, 2018(7):073405, 2018.

Theophile Thiery and Justin Ward. Two-sided weak submodularity for matroid constrained optimization and regression. In *Proceedings of the 35th Conference on Learning Theory*, pp. 3605–3634, 2022.

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

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Lan Wang, Yongdai Kim, and Runze Li. Calibrating non-convex penalized regression in ultra-high dimension. *Annals of Statistics*, 41(5):2505–2536, 2013.

Weijun Xie and Xinwei Deng. Scalable algorithms for the sparse ridge regression. *SIAM Journal on Optimization*, 30(4):3359–3386, 2020.

Cun-Hui Zhang. Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, 38(2):894–942, 2010.

Hui Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429, 2006.

Hui Zou and Trevor Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320, 2005.

A Omitted Proofs for SRE-LS

We prove the local-search results in Section 3, using the notation introduced there.

A.1 Support replacements and incremental updates

To analyze a support replacement, we use the ridge-regularized objective

$$\Phi(z) = \|Uz - b\|_2^2 + \|z\|_2^2.$$

For a support S, let x be the zero extension of its ridge minimizer. The ridge term makes the restricted problem strictly convex even when the selected columns are linearly dependent. Hence, $\Phi(x) = \mathcal{E}(S)$ and $\nabla_S \Phi(x) = 0$. The ridge energy also bounds the squared residual

$$f(U, b, S)^2 \le \mathcal{E}(S) \le |b|_2^2.$$

The first inequality follows by dropping the ridge penalty, and the second by evaluating the objective at the zero vector.

Proof of Lemma 3.2. We first prove that the proposed exchange score provides a guaranteed decrease of the ridge energy. Fix a size-k support S, a selected column $q \in S$, and an unselected column $h \notin S$. Let $\chi = \chi_S(q, h)$. The exchange direction

$$v = -x_q e_q - \frac{g_h}{2\chi} e_h$$

removes the contribution of q and adds a step along h. Since the current point is optimal on S, the directional derivative in the selected coordinates vanishes, while the derivative in coordinate h equals g_h. Hence, the quadratic expansion of Φ gives

$$\begin{aligned}\Phi(x + v) &= \Phi(x) + \nabla\Phi(x)^\top v + v^\top (U^\top U + I) v \\ &\le \mathcal{E}(S) - \frac{g_h^2}{2\chi} + \chi \left(x_q^2 + \frac{g_h^2}{4\chi^2} \right) \\ &= \mathcal{E}(S) - \widehat{\Gamma}_S(q, h).\end{aligned}$$

The resulting vector is supported on $S \setminus \{q\} \cup \{h\}$, so minimizing over the replacement support establishes equation 3.2.

The pair curvature is bounded since the columns are normalized. In particular, the largest eigenvalue of the two-column ridge Hessian is bounded by its trace, which gives $1 \le \chi \le 3$. The conservative score used in the comparison analysis follows from the fact that $g_h^2/(4z) - z x_q^2$ decreases as z increases. Therefore, we have

$$\widehat{\Gamma}_S(q, h) \ge \Gamma_S(q, h). \tag{2}$$

We next prove the exact deletion identity used after insertion. Let $\mathcal{A} = S \cup \{h\}$ and let $j \in \mathcal{A}$ be the column removed after refitting on the enlarged support. By reordering the coordinates, we have

$$U_{\mathcal{A}}^\top U_{\mathcal{A}} + I = \begin{bmatrix} K_{\mathcal{C}} & p \\ p^\top & a \end{bmatrix},$$

where $\mathcal{C} = \mathcal{A} \setminus \{j\}$. The Schur complement $\omega = a - p^\top K_{\mathcal{C}}^{-1} p$ is positive because the ridge term makes the matrix positive definite. The block inverse formula gives $(R_{\mathcal{A}})_{jj} = \omega^{-1}$, and the corresponding ridge coefficient satisfies $(x_{\mathcal{A}})_j = \frac{z_j - p^\top K_{\mathcal{C}}^{-1} z_{\mathcal{C}}}{\omega}$, where $z_{\mathcal{A}} = U_{\mathcal{A}}^\top b$. Finally, the optimal ridge value can be written as

$$\mathcal{E}(\mathcal{A}) = \|b\|_2^2 - z_{\mathcal{A}}^\top R_{\mathcal{A}} z_{\mathcal{A}}.$$

Substituting the block inverse expression shows that the increase caused by deleting j is exactly

$$\mathcal{E}(\mathcal{C}) - \mathcal{E}(\mathcal{A}) = \frac{(z_j - p^\top K_{\mathcal{C}}^{-1} z_{\mathcal{C}})^2}{\omega} = \frac{(x_{\mathcal{A}})_j^2}{(R_{\mathcal{A}})_{jj}},$$

which proves equation 3.2. □

The coordinate q used in the exchange score need not be the column removed after insertion. Suppose the selected candidate is h_t and $\widehat{\Gamma}_{S_t}(q, h_t) > 0$. Since q remains a feasible deletion from $\mathcal{A} = S_t \cup h_t$, the minimizing deletion rule and Lemma 3.2 give

$$\mathcal{E}(S_{t+1}) \le \mathcal{E}(\mathcal{A} \setminus \{q\}) \le \mathcal{E}(S_t) - \widehat{\Gamma}_{S_t}(q, h_t).$$

Since h_t maximizes the implemented score over $\mathcal{C}_t$ and $\widehat{\Gamma}_S(q, h) \ge \Gamma_S(q, h)$, we get

$$\mathcal{E}(S_t) - \mathcal{E}(S_{t+1}) \ge \max_{h \in \mathcal{C}_t} [\Gamma_{S_t}(q, h)]_+,$$

where $[a]_+ = \max\{a, 0\}$ and an empty maximum is zero. If $\mathcal{C}_t$ is empty or no sampled candidate has a positive implemented score, all sampled conservative scores are nonpositive and both sides are zero. Thus, the ridge energy is nonincreasing in every round, and the support always contains exactly k indices.

The exchange computation requires only the ridge inverse $R_S = (U_S^\top U_S + I)^{-1}$, the coefficient vector x_S, and the residual r_S. Solving a new ridge regression after every accepted exchange would be unnecessarily expensive. We maintain these quantities using block inverse updates.

When a column $h \notin S$ is inserted, let $\mathcal{A} = S \cup \{h\}$ and define $p = U_S^\top u_h$, $w = R_S p$, and

$$a = 1 + \|u_h\|_2^2 - p^\top w.$$

This scalar is positive because

$$a = 1 + u_h^\top (I - U_S R_S U_S^\top) u_h,$$

and $I - U_S R_S U_S^\top = (I + U_S U_S^\top)^{-1}$ is positive definite. The block inverse formula gives

$$R_{\mathcal{A}} = \begin{bmatrix} R_S + w w^\top / a & -w/a \\ -w^\top / a & 1/a \end{bmatrix}. \tag{3}$$

After deleting $j \in \mathcal{A}$, let $\mathcal{C} = \mathcal{A} \setminus \{j\}$. The ridge inverse on the retained support is

$$R_{\mathcal{C}} = (R_{\mathcal{A}})_{\mathcal{C},\mathcal{C}} - \frac{(R_{\mathcal{A}})_{\mathcal{C},j} (R_{\mathcal{A}})_{j,\mathcal{C}}}{(R_{\mathcal{A}})_{jj}}. \tag{4}$$

The updated coefficients and residual follow from equation 3.2. These updates remain valid when the selected columns are linearly dependent because the ridge matrices are positive definite.

A.2 Expected exchange score under group sampling

We next bound the largest positive exchange score available in a uniformly sampled candidate group. Since the current support and scores depend on earlier rounds, we condition on the candidate groups and the search history. If the columns of a fixed comparison support can be assigned to distinct groups containing them, the conditional expected maximum is at least $1/(4k)$ of the total positive score over the missing comparison columns. The following lemma establishes this assignment with constant probability and gives the sampling bound for arbitrary nonnegative weights.

Lemma A.1. *Fix a size-k set T before constructing $\mathscr{B}$. With probability at least $199/200$ over the construction of $\mathscr{B}$, there is an injection $\mu : T \to [4k]$ such that $h \in \mathcal{B}_{\mu(h)}$ for every $h \in T$. Denote this event by M_T. On any fixed realization $\mathscr{B} \in \mathsf{M}_T$, let $H_t = T \setminus S_t$ and let $w_t(h) \ge 0$, $h \in H_t$, be determined by $\mathcal{F}_t$ and $\mathscr{B}$. Then, it holds that*

$$\mathbb{E}\left[\max_{h \in H_t \cap \mathcal{B}_{s_t}} w_t(h) \,\middle|\, \mathcal{F}_t, \mathscr{B} \right] \ge \frac{1}{4k} \sum_{h \in H_t} w_t(h),$$

where an empty maximum is zero.

Proof. Let $h \in \mathcal{B}_s$ as a bipartite graph between the columns in T and the $4k$ groups. The claim is immediate for $k = 1$. For $k \ge 2$, failure of M_T implies by Hall's theorem that some $H \subseteq T$ with

$|H| = s \geq 2$ has neighbors contained in a set $C \subseteq [4k]$ of size $s-1$. For fixed H and C, the probability that all columns in H fall into the blocks indexed by C in one permutation is

$$\frac{\binom{m(s-1)}{s}}{\binom{d'}{s}} \leq \left(\frac{s-1}{4k}\right)^s.$$

Since the three permutations are independent, a union bound gives

$$\Pr(\mathsf{M}_T^c) \leq \sum_{s=2}^{k} \binom{k}{s}\binom{4k}{s-1}\left(\frac{s-1}{4k}\right)^{3s}.$$

Using $\binom{N}{r} \leq (eN/r)^r$ and $s \leq k$, each summand is at most

$$\binom{k}{s}\binom{4k}{s-1}\left(\frac{s-1}{4k}\right)^{3s} \leq \frac{e^{2s-1}(s-1)^{2s+1}}{4^{2s+1}s^s k^{s+1}} \leq \frac{e^{2s-1}}{4^{2s+1}}\left(1-\frac{1}{s}\right)^{2s+1} \leq \frac{e^{-3}}{4}\left(\frac{e^2}{16}\right)^s.$$

Therefore, we obtain

$$\Pr(\mathsf{M}_T^c) \leq \frac{e^{-3}}{4}\frac{(e^2/16)^2}{1-e^2/16} < \frac{1}{200}.$$

Now, we fix $\mathscr{B} \in \mathsf{M}_T$ and an injection $\mu : T \to [4k]$ with $h \in \mathcal{B}_{\mu(h)}$. Condition on $\mathcal{F}_t$ and $\mathscr{B}$, so H_t and the weights $w_t(h)$ are fixed. For each group, let

$$W_s = \max_{h \in H_t \cap \mathcal{B}_s} w_t(h),$$

with $W_s = 0$ when the intersection is empty. Since the groups $\mu(h)$ are distinct and contain their assigned columns, we have

$$\sum_{s=1}^{4k} W_s \geq \sum_{h \in H_t} W_{\mu(h)} \geq \sum_{h \in H_t} w_t(h).$$

The index s_t is uniform on $[4k]$ and independent of $\mathcal{F}_t$, yielding

$$\mathbb{E}\left[\max_{h \in H_t \cap \mathcal{B}_{s_t}} w_t(h) \,\middle|\, \mathcal{F}_t, \mathscr{B}\right] = \frac{1}{4k}\sum_{s=1}^{4k} W_s \geq \frac{1}{4k}\sum_{h \in H_t} w_t(h).$$

Conditioning on $\mathcal{F}_t$ allows both S_t and the weights to depend on all previously sampled groups. □

A.3 Exchange Scores for a Sparse Comparison

Fix a sparse comparison vector y and a size-k set T containing its support. We next lower-bound the total positive exchange score over the missing columns $T \setminus S$. Ridge optimality converts an energy gap into large residual correlations with these columns, while the smallest selected ridge coefficient controls the total deletion penalty.

Lemma A.2. *Fix y with $\|y\|_0 \leq k$ and a size-k set $T \supseteq \operatorname{supp}(y)$. For a size-$k$ support S, let x be the zero-extended ridge minimizer and choose $q \in \arg\min_{j \in S} x_j^2$. Let $R = \|y\|_2$ and*

$$D = \left(\mathcal{E}(S) - \|Uy - b\|_2^2 - 9R^2\right)_+.$$

For $R > 0$, the missing columns satisfy

$$\sum_{h \in T \setminus S} [\Gamma_S(q,h)]_+ \geq 3D + \frac{(D - 9R^2)_+^2}{12R^2}.$$

If $R = 0$, then $D = 0$ and the bound is interpreted as zero.

Proof. The claim is immediate when $R = 0$ or $D = 0$. Suppose $R > 0$ and $D > 0$. Let $H = T \setminus S$ and $Q = S \setminus T$. Since $|S| = |T| = k$, we have $|H| = |Q|$. Define $u = \|x_Q\|_2^2$ and $v = \|y_H\|_2^2$. We

have $0 < v \le R^2$. Indeed, if $v = 0$, then y is supported on S, and evaluating the ridge objective at y gives $\mathcal{E}(S) \le \|Uy - b\|_2^2 + R^2$, contradicting $D > 0$.

Let $\Phi(z) = \|Uz - b\|_2^2 + \|z\|_2^2$ and $g = \nabla\Phi(x)$. Since x minimizes the ridge objective on S, $g_S = 0$. The exact quadratic expansion at x gives

$$\Phi(y) = \mathcal{E}(S) + g_H^\top y_H + \|U(y - x)\|_2^2 + \|y - x\|_2^2 \ge \mathcal{E}(S) + g_H^\top y_H + u + v.$$

Using $\Phi(y) = \|Uy - b\|_2^2 + R^2$ and $\mathcal{E}(S) = \|Uy - b\|_2^2 + 9R^2 + D$, we obtain

$$-g_H^\top y_H \ge D + 8R^2 + u + v \ge D + u + 9v.$$

Therefore, Cauchy–Schwarz yields

$$\|g_H\|_2^2 \ge \frac{(D + u + 9v)^2}{v}.$$

The choice of q controls the deletion penalties. Since $x_q^2 \le x_j^2$ for every $j \in Q$,

$$|H| x_q^2 = |Q| x_q^2 \le u.$$

On H, the components of g are the insertion gradients used in $\Gamma_S(q, h)$. Hence, we obtain

$$\sum_{h \in H} \Gamma_S(q, h) = \frac{\|g_H\|_2^2}{12} - 3|H| x_q^2 \ge \frac{(D + u + 9v)^2}{12v} - 3u.$$

Completing the square gives

$$\frac{(D + u + 9v)^2}{12v} - 3u = 3D + \frac{(D + u - 9v)^2}{12v} \ge 3D + \frac{(D - 9R^2)_+^2}{12R^2},$$

where the last inequality uses $u \ge 0$ and $v \le R^2$. Finally, $\sum_{h \in H} [\Gamma_S(q, h)]_+ \ge \sum_{h \in H} \Gamma_S(q, h)$, which proves the claim. □

A.4 Residual Bound for a Fixed Sparse Comparison

We now combine the matching bound with the exchange-score lower bound to compare the output after k rounds with a fixed sparse vector y. The quadratic term for large energy gaps removes the dependence on the initial support and reduces the final excess to $O(\|y\|_2^2)$.

Proof of Lemma 3.3. Fix y and a size-k set $T \supseteq \mathrm{supp}(y)$ before constructing the groups. Condition on any realization $\mathscr{B} \in \mathsf{M}_T$. Let $J_t = \mathcal{E}(S_t)$, $R = \|y\|_2$, and

$$C = \|Uy - b\|_2^2 + 9R^2.$$

If $R = 0$, then $y = 0$ and $J_t \le \|b\|_2^2 = C$ for every t, since the zero vector is feasible for the ridge objective. Hence the claimed moment bound follows immediately. Suppose $R > 0$. For $s \ge 0$, define

$$\psi_R(s) = 3s + \frac{(s - 9R^2)_+^2}{12R^2}.$$

Before sampling the group in round t, we assign each $h \in T \setminus S_t$ the weight $w_t(h) = [\Gamma_{S_t}(q_t, h)]_+$. Since $\widehat{\Gamma}_{S_t}(q_t, h) \ge \Gamma_{S_t}(q_t, h)$ and Lemma 3.2 guarantees the corresponding ridge-energy decrease, the actual decrease is at least the largest such weight in the sampled group. Applying Theorems A.1 and A.2 gives

$$\mathbb{E}[J_t - J_{t+1} \mid \mathcal{F}_t, \mathscr{B}] \ge \frac{1}{4k} \psi_R((J_t - C)_+). \tag{5}$$

Let $\mu_t = \mathbb{E}[J_t \mid \mathscr{B}]$ and $a_t = (\mu_t - C)_+$. The function $z \mapsto \psi_R((z - C)_+)$ is convex and nondecreasing, so Jensen's inequality applied to eq. (5) gives

$$a_{t+1} \le \left[a_t - \frac{\psi_R(a_t)}{4k} \right]_+. \tag{6}$$

In particular, a_t is nonincreasing.

We next show that $a_k \le 39R^2$. This is immediate if $a_k \le 9R^2$. Otherwise, $a_t > 0$ for every $t \le k$, and eq. (6) gives

$$a_t - a_{t+1} \ge \frac{\psi_R(a_t)}{4k}.$$

Since ψ_R is nondecreasing, we have

$$\int_{a_{t+1}}^{a_t} \frac{du}{\psi_R(u)} \ge \frac{a_t - a_{t+1}}{\psi_R(a_t)} \ge \frac{1}{4k}.$$

Summing over the k rounds yields

$$\frac{1}{4} \le \int_{a_k}^{a_0} \frac{du}{\psi_R(u)} \le \int_{a_k}^{\infty} \frac{du}{\psi_R(u)}.$$

Since $a_k > 9R^2$, we have $\psi_R(u) = (u + 9R^2)^2/(12R^2)$ for every $u \ge a_k$. Therefore, we get

$$\frac{1}{4} \le \frac{12R^2}{a_k + 9R^2},$$

which gives $a_k \le 39R^2$.

Consequently, we have

$$\mathbb{E}[f(U, b, S_k)^2 \mid \mathscr{B}] \le \mu_k \le C + 39R^2 = \|Uy - b\|_2^2 + 48\|y\|_2^2. \tag{7}$$

For every fixed $\mathscr{B} \in \mathsf{M}_T$, Markov's inequality then gives

$$f(U, b, S_k) \le 2\sqrt{\|Uy - b\|_2^2 + 48\|y\|_2^2}$$

with conditional probability at least $3/4$. If the quantity inside the square root is zero, the conclusion holds by eq. (7).

Finally, Theorem A.1 gives $\Pr(\mathsf{M}_T) \ge 199/200$. Since the conditional success bound holds for every realization in M_T, we conclude that

$$\Pr\left\{ f(U, b, S_k) \le 2\sqrt{\|Uy - b\|_2^2 + 48\|y\|_2^2} \right\} \ge \frac{199}{200} \cdot \frac{3}{4} = \frac{597}{800}.$$

□

A.5 Residual approximation for Local Search

The following result extends the coefficient bound to supports with linearly dependent columns.

Lemma A.3. *Let $M \in \mathbb{R}^{n \times d}$ and $b \in \mathbb{R}^n$ be nonzero, let $1 \le k \le d$, and write $\rho = \bar{\rho}_k(M, b)$. Then $\rho > 0$ if and only if $\mathrm{OPT}_k(M, b) > 0$. When $\rho > 0$, for every $S \subseteq [d]$ with $|S| \le k$ and every $v \in \mathrm{col}(M_S)$, it holds that*

$$\rho \|M_S^\dagger v\|_2 \le \|P_b^\perp v\|_2.$$

In particular, $\rho \|M_S^\dagger b\|_2 \le f(M, b, S)$. Hence a minimum-norm optimal k-sparse solution $x^\star$ satisfies $\rho \|x^\star\|_2 \le \mathrm{OPT}_k(M, b)$.

Proof. Since $M \ne 0$ and $k \ge 1$, $\mathcal{I}_k(M)$ is nonempty. Suppose $\rho = 0$. Then, some $T \in \mathcal{I}_k(M)$ admits a nonzero vector z with $P_b^\perp M_T z = 0$. Since M_T has full column rank, $M_T z \ne 0$. Hence, $M_T z = cb$ for some $c \ne 0$, and $b = M_T(z/c)$, which leads to $\mathrm{OPT}_k(M, b) = 0$.

If $\mathrm{OPT}_k(M, b) = 0$, then b lies in the span of at most k columns of M. Choosing an independent subset T spanning b gives a nonzero vector z with $M_T z = b$. Thus, we have $P_b^\perp M_T z = 0$, so $\sigma_{\min}(P_b^\perp M_T) = 0$ and $\rho = 0$. This proves the equivalence.

Now, we suppose $\rho > 0$ and fix $v \in \mathrm{col}(M_S)$. The claim is immediate for $v = 0$. Otherwise, choose an independent subset $T \subseteq S$ spanning $\mathrm{col}(M_S)$, and write $v = M_T w$. Extending w by zero to S gives a representation of v on S. Since $M_S^\dagger v$ is the minimum-norm form, we have

$$\|M_S^\dagger v\|_2 \le \|w\|_2.$$

By the definition of ρ on the independent support T, it holds that

$$\rho\|M_S^\dagger v\|_2 \le \rho\|w\|_2 \le \|P_b^\perp M_T w\|_2 = \|P_b^\perp v\|_2.$$

For the least squares bound, let $v = M_S M_S^\dagger b$, the orthogonal projection of b onto $\mathrm{col}(M_S)$. Then $M_S^\dagger v = M_S^\dagger b$, and

$$\|P_b^\perp v\|_2 = \|P_b^\perp (v-b)\|_2 \le \|v-b\|_2 = f(M,b,S).$$

Applying the preceding bound gives $\rho\|M_S^\dagger b\|_2 \le f(M,b,S)$. Finally, applying this inequality to the support of the optimal sparse solution gives $\rho\|x^\star\|_2 \le \mathrm{OPT}_k(M,b)$. □

Proof of Theorem 3.6. Let $f^\star = \mathrm{OPT}_k(U,b) = \mathrm{OPT}_k(A,b)$. Choose a optimal k-sparse vector $x^\star$ by fixed tie-breaking, and fix a size-k set $T^\star$ containing its support. These choices depend only on the input and are made before constructing the candidate groups. By Lemma A.3, we get

$$\rho\|x^\star\|_2 \le f^\star.$$

Moreover, every nonzero column of U has unit norm and forms an admissible singleton in the definition of ρ, so $0 < \rho \le 1$.

Applying Lemma 3.3 with $y = x^\star$ gives, with probability at least $p_0 = 597/800$,

$$\begin{aligned} f(U,b,S_k) &\le 2\sqrt{(f^\star)^2 + 48\|x^\star\|_2^2} \\ &\le 2\sqrt{1+\frac{48}{\rho^2}}\, f^\star \le \frac{14}{\rho} f^\star. \end{aligned}$$

The column normalization preserves every selected column span, yielding $f(A,b,S_k) = f(U,b,S_k)$. The accepted exchanges preserve the support size, while skipped rounds leave it unchanged, so $|S_k| = k$.

We next bound the running time of Algorithm 1. Specifically, normalizing the columns and constructing the candidate groups take $O(nd)$ time. Since every sampled group contains at most $3\lceil d/(4k)\rceil$ columns, the total number of inspected candidates over the k rounds is at most $3k\left\lceil \frac{d}{4k}\right\rceil = O(d)$. Evaluating a candidate takes $O(n)$ time, giving $O(nd)$ total candidate-evaluation time. The initial ridge state can be computed in $O(nk^2 + k^3)$ time. Then, each round of local search requires $O(nk + k^2)$ time, and the total maintenance cost over k rounds is $O(nk^2 + k^3)$. The final least squares fit on the selected support has the same order. Thus, with constant probability, the SRE-LS requires $O(nd + nk^2)$ time. By repeating the algorithm independently $O(\log(1/\delta))$ times and returning the support with the smallest residual raises the success probability to $1-\delta$ while preserving the approximation guarantee. The total running time is $O\left((nd + nk^2)\log\frac{1}{\delta}\right)$. □

B Hardness of Approximation

We prove Theorem 3.7 by reducing a gap version of Max k-Cover to sparse regression. The construction appends an orthogonal block that makes the optimal residual positive and controls the response-orthogonal regularity. We first present a real-valued construction to illustrate the reduction, and then give a rational realization with exact unit column norms. In the following, polynomial time is measured in the binary input length.

B.1 Max k-Cover Gap

We use the Max k-Cover gap from Civril (2013, Theorem 2.3), following Feige (1998). For some fixed rational $\gamma \in (0,1)$, it is NP-hard to distinguish a YES instance, in which k pairwise disjoint sets cover $[m]$, from a NO instance, in which every collection of at most k sets covers at most $(1-\gamma)m$ elements. We remove empty sets and duplicate a nonempty set as needed so that a duplicate pair exists and $d > k$. These modifications preserve both sides of the gap, and $k \ge 2$ remains part of the input.

For each set $\mathcal{C}_j$, let $s_j = |\mathcal{C}_j|$ and define $A_{e,j} = \frac{\mathbf{1}\{e \in \mathcal{C}_j\}}{\sqrt{s_j}}$. Let $b_0 = \frac{\mathbf{1}_m}{\sqrt{m}}$. Then, every column of A and the response b_0 have unit norm. Thus, every column of A and the response b_0 have unit norm. In a YES instance, let $x^\star_j = \sqrt{s_j/m}$ on the k sets of a disjoint cover and zero elsewhere. Since these sets partition $[m]$, we have $\|x^\star\|_2 = 1$ and $Ax^\star = b_0$. In a NO instance, the support of any k-sparse vector selects at most k sets and leaves at least γm coordinates uncovered. The predictor vanishes on these coordinates, giving

$$\|Ax - b_0\|_2 \geq \sqrt{\gamma}$$

for every k-sparse vector x.

B.2 Construction at a Fixed Regularity Scale

We first give a real-valued construction that explains how the regularity scale is introduced. For $0 < a < 1$, extend the response to $b_a = (b_0^\top, 0_d^\top)^\top$ and define

$$U_a = \begin{bmatrix} \sqrt{1-a^2}\, A \\ a I_d \end{bmatrix}.$$

The columns of U_a and the response b_a have unit norm. For every vector z, we have

$$\|P^\perp_{b_a} U_a z\|_2^2 = (1-a^2)\|P^\perp_{b_0} A z\|_2^2 + a^2 \|z\|_2^2.$$

Hence every column subset is independent and $\bar{\rho}_k(U_a, b_a) \geq a$. For a duplicate pair i, j, the vector $(e_i - e_j)/\sqrt{2}$ has zero upper image and projected norm a, giving

$$\bar{\rho}_k(U_a, b_a) = a.$$

In a YES instance, let $y = \sqrt{1-a^2}\, x^\star$. Since $Ax^\star = b_0$ and $\|x^\star\|_2 = 1$, its squared residual is $a^4 + a^2(1-a^2) = a^2$. In a NO instance, at least γm upper coordinates remain uncovered, so every k-sparse predictor has residual at least $\sqrt{\gamma}$. The lower diagonal block also rules out an exact fit. Thus the construction has positive optimum in both cases and separates the two residual scales by a factor of order $1/a$.

The square roots in this construction need not be rational. We next realize the same gap using rational entries and exact unit column norms. Replicate each universe element m times. The new universe has size $N = m^2$, each replicated set $\mathcal{C}'_j$ has size $t_j = m s_j$, and the covering promise is unchanged. The upper response $\mathbf{1}_N/m$ is rational and has unit norm.

Fix a rational $a \in (0, 1/4]$. For each column, choose a positive dyadic rational r_j satisfying

$$\frac{1-2a}{t_j(1+2a)} \leq r_j^2 \leq \frac{1-a}{t_j(1+a)}. \tag{8}$$

Duplicate sets use the same value of r_j. The interval for r_j^2 has width $\Theta(a/t_j)$, while both endpoints are $\Theta(1/t_j)$. Hence, the corresponding interval for r_j has width $\Omega(a/\sqrt{t_j})$, and a suitable dyadic rational with $O(\log t_j + \log(1/a))$ bits can be found in polynomial time.

Define $\zeta_j = (1 - t_j r_j^2)/(1 + t_j r_j^2)$ and $\alpha_j = 2r_j/(1+t_j r_j^2)$. Then, $a \leq \zeta_j \leq 2a$ and $t_j \alpha_j^2 + \zeta_j^2 = 1$. Let $Z_j = \alpha_j \mathbf{1}_{\mathcal{C}'_j}$ and define

$$V_a = \begin{bmatrix} Z \\ \operatorname{diag}(\zeta_1, \ldots, \zeta_d) \end{bmatrix}.$$

Let $v = ((\mathbf{1}_N/m)^\top, 0_d^\top)^\top$ be the padded response. The design has rational unit columns and full column rank, and $\|v\|_2 = 1$. For every x, the diagonal block gives

$$\|P^\perp_v V_a x\|_2^2 = \|P^\perp_{\mathbf{1}_N} Z x\|_2^2 + \sum_j \zeta_j^2 x_j^2 \geq a^2 \|x\|_2^2.$$

Hence $\bar{\rho}_k(V_a, v) \geq a$. For the duplicate pair, the normalized coefficient difference has zero upper image and lower-image norm at most $2a$. Therefore, we have

$$a \leq \bar{\rho}_k(V_a, v) \leq 2a.$$

In a YES instance, let $S^\star$ be a disjoint cover and set $y_j = 1/(m\alpha_j)$ for $j \in S^\star$. This fits the upper response exactly. Since the selected replicated sets partition the universe,

$$\|y\|_2^2 = \sum_{j \in S^\star} \frac{t_j}{N(1-\zeta_j^2)} \le \frac{1}{1-4a^2}.$$

Only the lower block contributes to the residual, which gives

$$0 < \mathrm{OPT}_k(V_a, v) \le \frac{2a}{\sqrt{1-4a^2}} \le 3a.$$

The strict inequality follows because the positive diagonal block precludes an exact fit.

In a NO instance, every choice of at most k columns leaves at least γN upper coordinates uncovered. These coordinates contribute $1/m^2$ each to the squared residual, giving

$$\mathrm{OPT}_k(V_a, v) \ge \sqrt{\gamma}.$$

Consequently, both promise cases satisfy $a \le \bar{\rho}_k(V_a, v) \le 2a$, while

$$\mathrm{OPT}_k(V_a, v) \in \begin{cases} (0, 3a], & \text{in the YES case,} \\ [\sqrt{\gamma}, \infty), & \text{in the NO case.} \end{cases}$$

The construction has $m^2 + d$ rows and polynomial encoding length. The dyadic values r_j use $O(\log t_j + \log(1/a))$ bits, and the resulting rational values α_j and ζ_j also have polynomial bit length. Thus, the construction can be produced in polynomial time without evaluating the regularity parameter.

B.3 APPROXIMATION LOWER BOUND

The rational construction above gives hardness on instances whose response-orthogonal regularity lies in a fixed interval $[a, 2a]$. We first establish this fixed-scale lower bound and then derive Theorem 3.7.

Proof of Theorem 3.7. We first establish a fixed-scale claim. There exist absolute constants $c, a_0 > 0$ such that, for every fixed rational $0 < a \le a_0$, no randomized algorithm polynomial in the rational input length can achieve a c/a residual approximation on all rational instances with unit column norms, $\|b\|_2 = 1$, and $a \le \bar{\rho}_k(U, b) \le 2a$, unless $\mathrm{NP} \subseteq \mathrm{BPP}$. The hard instances have full column rank and positive optimum.

Let $c > 0$ be a rational constant satisfying $3c < \sqrt{\gamma}/2$, and let $a_0 > 0$ be rational with $a_0 \le \min 1/4, c/2$. For any fixed rational $a \in (0, a_0]$, the construction above maps a Max k-Cover instance to (V_a, v, k). The construction satisfies

$$a \le \bar{\rho}_k(V_a, v) \le 2a.$$

In a YES instance, $\mathrm{OPT}_k(V_a, v) \le 3a$. Hence, with probability at least $2/3$, a c/a approximation would return a support whose residual is at most

$$\frac{c}{a}\,\mathrm{OPT}_k(V_a, v) \le 3c < \frac{\sqrt{\gamma}}{2}.$$

In a NO instance, every support of size at most k has residual at least $\sqrt{\gamma}$.

The two cases can be distinguished in polynomial time from the returned support. For a nonempty support S, let $G_S = (V_a)_S^\top (V_a)_S$ and $z_S = (V_a)_S^\top v$. Since V_a has full column rank, G_S is rational and positive definite, and the fitted squared residual is

$$f(V_a, v, S)^2 = 1 - z_S^\top G_S^{-1} z_S.$$

This quantity can be evaluated in polynomial bit complexity. The empty support has squared residual one. The reduction accepts when the squared residual is less than $\gamma/4$. A YES instance is accepted with probability at least $2/3$, whereas a NO instance is always rejected. The construction and verification are polynomial in the covering input length, contradicting $\mathrm{NP} \not\subseteq \mathrm{BPP}$. This proves the fixed-scale claim.

Table 3: Results on RCV1 ($n = 804{,}414, d = 47{,}236$) for varying k with $\mathrm{SNR} = 20$. Thiery–Ward and ARHT are omitted since their evaluations exceed one hour.

k	Method	Loss ↓	Time (s) ↓
60	FujiiLS	**5.745**	14.45
	OMPR	**5.745**	15.58
	SRE-LS	5.747 ± 0.002	23.81
80	FujiiLS	**6.892**	19.80
	OMPR	**6.892**	20.75
	SRE-LS	6.896 ± 0.001	25.36
100	FujiiLS	**8.382**	31.05
	OMPR	**8.382**	33.59
	SRE-LS	8.388 ± 0.003	**23.56**
120	FujiiLS	**10.40**	41.63
	OMPR	**10.40**	43.63
	SRE-LS	10.44 ± 0.09	**28.39**

It remains to derive the stated uniform lower bound. Suppose a randomized polynomial-time algorithm guarantees a uniform approximation factor $\alpha(\rho) = o(1/\rho)$. Since $\alpha(r) = o(1/r)$, there exists a fixed rational $a \in (0, a_0]$ such that $r\alpha(r) \le c$ for every $r \in [a, 2a]$. Then, we conclude that

$$\alpha(r) \le \frac{c}{r} \le \frac{c}{a}$$

throughout this interval, contradicting the fixed-scale claim. □

C Additional Experimental Results

In this section, we provide additional results on RCV1, AmazonCat-13K, and Webspam, together with an ablation of candidate subsampling. We follow the same experimental settings as in Section 4 and report regression loss and running time across different sparsity levels k. As shown in Tables 3 to 5, SRE-LS maintains regression loss close to FujiiLS and OMPR while requiring less running time on the larger settings. On RCV1, SRE-LS is faster than both methods at $k = 100$ and 120. The same behavior is observed on AmazonCat-13K and Webspam, where the runtime difference increases with k. In particular, at $k = 500$ on Webspam, SRE-LS is $2.6\times$ faster than FujiiLS and $2.9\times$ faster than OMPR.

We further evaluate candidate subsampling in Table 6. The sampled variant evaluates one balanced candidate group in each round, while the full variant evaluates all unselected columns. For $k = 40$–100, subsampling reduces candidate evaluations by 53–133$\times$ and running time by 34–54$\times$, with less than 0.4% change in regression loss. This shows that the balanced sampling scheme generally reduces computation while preserving the regression performance of the full exchange procedure.

Table 4: Results on AmazonCat-13K ($n = 1{,}186{,}239, d = 203{,}882$) across different values of k with $\mathrm{SNR} = 20$.

k	Method	Loss ↓	Time (s) ↓
100	FujiiLS	**7.740**	**36.46**
	OMPR	**7.740**	50.43
	SRE-LS	7.766 ± 0.049	37.05
120	FujiiLS	**7.960**	43.48
	OMPR	**7.960**	58.26
	SRE-LS	7.989 ± 0.039	**41.63**
140	FujiiLS	**9.116**	60.07
	OMPR	**9.116**	83.14
	SRE-LS	9.210 ± 0.246	**39.38**
160	FujiiLS	**9.406**	76.93
	OMPR	**9.406**	99.28
	SRE-LS	9.420 ± 0.001	**46.60**
180	FujiiLS	11.14	113.2
	OMPR	11.14	135.0
	SRE-LS	$\mathbf{11.18 \pm 0.08}$	**53.69**

Table 5: Results on Webspam ($n = 350{,}000, d = 680{,}715$) across different k with $\mathrm{SNR} = 20$.

k	Method	Loss ↓	Time (s) ↓
200	FujiiLS	**0.904**	221.6
	OMPR	**0.904**	211.5
	SRE-LS	0.911 ± 0.001	**198.8**
300	FujiiLS	**1.028**	349.1
	OMPR	**1.028**	407.4
	SRE-LS	1.035 ± 0.001	**235.0**
400	FujiiLS	**1.175**	721.7
	OMPR	**1.174**	788.6
	SRE-LS	1.183 ± 0.001	**310.4**
500	FujiiLS	**1.196**	907.6
	OMPR	**1.195**	1035.7
	SRE-LS	1.205 ± 0.001	**355.7**

Table 6: Ablation of candidate subsampling from SRE-LS on MNIST ($n = 10{,}000, d = 784$). The sampled variant evaluates one balanced candidate group in each round, while the full variant evaluates all unselected columns. Both variants otherwise use the same exchange rule, $\lambda = 0.05$, $T = 10k$, and initial support. The sampled variant is averaged over 10 runs.

	Sampled candidates			Full candidates		
k	Loss	Time (s)	Candidate evaluation	Loss	Time (s)	Candidate evaluation
40	32301.51 ± 15.87	**0.53**	**5,543**	**32280.53**	28.48	2.98×10^5
60	59721.59 ± 41.94	**0.96**	**5,418**	**59709.42**	39.51	4.34×10^5
80	$\mathbf{90259.94 \pm 54.23}$	**1.24**	**5,275**	90456.03	49.83	5.63×10^5
100	$\mathbf{107926.78 \pm 30.46}$	**1.72**	**5,128**	108327.14	59.10	6.84×10^5