Differentially Private High-dimensional Variable Selection via Integer Programming

NeurIPS 2025

Petros Prastakos, MIT Operations Research Center
Kayhan Behdin, LinkedIn
Rahul Mazumder, MIT Sloan School of Management

NEURAL INFORMATION PROCESSING SYSTEMS

Sparse variable selection

Formulation:

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \boldsymbol{x}_i^T \boldsymbol{\beta}) \ \text{ s.t. } \ \|\boldsymbol{\beta}\|_0 \le s,\ \|\boldsymbol{\beta}\|_2^2 \le r^2 \tag{1}$$

- ℓ is convex function of $\boldsymbol{\beta}$
- When $\ell(y_i, \boldsymbol{x}_i^T \boldsymbol{\beta}) = (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2$ the problem is commonly called Best Subset Selection (BSS) [Miller, 2002]
- An important methodological problem, which has been shown to have favorable statistical properties over its convex relaxation under certain settings [Hazimeh and Mazumder, 2020, Guo et al., 2020].
- Can be computationally challenging
- Recent work uses Mixed Integer Programming (MIP) to solve large BSS instances [Bertsimas and Van Parys, 2020, Hazimeh et al., 2022]

How to privatize sparse variable selection?

- (ε, δ)-Differentially Private (DP) Algorithm $\mathcal{A}$:

$$\mathbb{P}(\mathcal{A}(\mathcal{D}) \in K) \leq e^{\varepsilon}\mathbb{P}(\mathcal{A}(\mathcal{D}') \in K) + \delta$$

  for any measurable event $K \subset \text{range}(\mathcal{A})$ and for any pair of neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ [Dwork et al., 2014].
- **Current Algorithms for DP sparse linear regression:**
  - convex relaxations, private Lasso [Thakurta and Smith, 2013, Kifer et al., 2012]
  - Markov chain mixing [Roy and Tewari, 2023]
- **Goal:** Designing scalable, pure DP ($\delta = 0$) estimators for ℓ_0-sparse variable selection (i.e., optimal location of nonzeros in feature vector $\boldsymbol{\beta}$), using integer programming techniques for selection and sampling.

First attempt – exponential mechanism

- The exponential mechanism $\mathcal{A}_E(\cdot)$ that follows

$$\mathbb{P}(\mathcal{A}_E(\mathcal{D}) = o) \propto \exp\left(-\frac{\varepsilon \mathcal{R}(o, \mathcal{D})}{2\Delta}\right), \quad \forall o \in \mathcal{O}$$

  ensures $(\varepsilon, 0)$-DP [McSherry and Talwar, 2007].
- Outcome set $\mathcal{O} = \{S \subseteq [p] : |S| = s\}$ (all subsets of size s from $\{1, ..., p\}$)
- The objective for each subset S is

$$\mathcal{R}(S, \mathcal{D}) = \min_{\boldsymbol{\beta} \in \mathbb{R}^{|S|}} \sum_{i=1}^{n} \ell(y_i, (\boldsymbol{x}_i)_S^T \boldsymbol{\beta}) \ \text{ s.t. } \ \|\boldsymbol{\beta}\|_2^2 \leq r^2$$

- The global sensitivity is

$$\Delta = \max_{S \in \mathcal{O}} \max_{\mathcal{D}, \mathcal{D}' \text{ are neighbors}} \mathcal{R}(S, \mathcal{D}) - \mathcal{R}(S, \mathcal{D}').$$

Challenges

- **Issue:** The exponential mechanism may require evaluating all $\binom{p}{s}$ subsets $\rightarrow$ infeasible for large p.
- **Our Question:** *Is it necessary to have access to $\mathcal{R}(S, \mathcal{D})$ for all $S \in \mathcal{O}$ in the exponential mechanism?*

Our contributions

1. **Sampling via integer programming:** Design **Top-R** and **Mistakes** mechanisms leveraging MIP to approximate full exponential mechanism.
2. **Pure-DP guarantees:** Achieve pure-DP for general convex loss functions without exhaustive enumeration.
3. **Support recovery guarantees:** Provide theoretical support recovery guarantees under standard high-dimensional assumptions in the case of BSS that match with the SOTA
4. **Numerical experiments:** Demonstrate strong empirical performance (support recovery) for ℓ_0-sparse regression and classification with up to $p = 10^4$ features

Top-R: order supports by objective

Best support $\hat{S}_1$ (lowest objective)

2nd best support $\hat{S}_2$

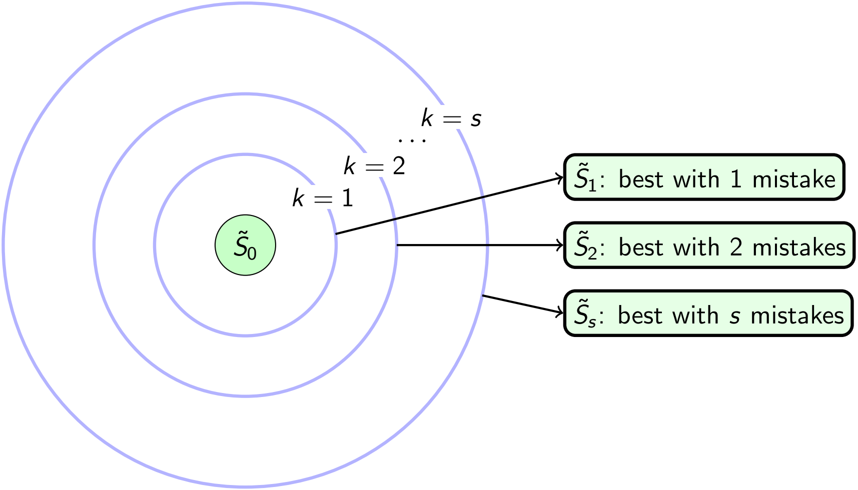
$\vdots$

R-th best support $\hat{S}_R$

All other $\binom{p}{s} - R$ supports sampled *uniformly*

- Approximate exponential mechanism using p_0 on $[R+1]$:
  $k \leq R: \; p_0(k) \propto \exp\big(-\varepsilon\,\mathcal{R}(\hat{S}_k, \mathcal{D})/(2\Delta)\big)$;
  $R{+}1: \; p_0(R{+}1) \propto \big(\binom{p}{s} - R\big)\exp\big(-\varepsilon\,\mathcal{R}(\hat{S}_R, \mathcal{D})/(2\Delta)\big)$.
- Draw $a \sim p_0$; if $a \leq R$ return $\hat{S}_a$, else sample uniformly from the remaining supports.

Mistakes: bucket by number of mistakes from the best



All supports that make k mistakes relative to the best support $\tilde{S}_0$ form a bucket P_k. We assign every $S \in P_k$ the objective of the *best-in-bucket* $\tilde{S}_k$, then sample via the exponential mechanism.

Optimization algorithm

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \boldsymbol{x}_i^T \boldsymbol{\beta}) + \frac{\lambda}{2n}\|\boldsymbol{\beta}\|_2^2 \ \text{ s.t. } \ \|\boldsymbol{\beta}\|_0 \le s,\ \|\boldsymbol{\beta}\|_2^2 \le r^2$$

Our outer approximation formulation:

$$\min_{\boldsymbol{z}} \quad c(\boldsymbol{z}) \ \text{ s.t. } \ \boldsymbol{z} \in \{0,1\}^p,\ \sum_{i=1}^{p} z_i \le s$$

where

$$c(\boldsymbol{z}) = \min_{\|\boldsymbol{\beta}\|_2^2 \le r^2} \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, \boldsymbol{x}_i^T \boldsymbol{\beta}) + \frac{\lambda}{2n}\sum_{i=1}^{p} \frac{\beta_i^2}{z_i}$$

Key idea: pick points $\widehat{\boldsymbol{z}} \in (0,1]^p$ and approximate c from below by adding the cutting planes

$$\eta \ \ge \ c(\widehat{\boldsymbol{z}}) + \nabla c(\widehat{\boldsymbol{z}})^\top (\boldsymbol{z} - \widehat{\boldsymbol{z}}).$$

Each new cut tightens the piecewise-linear lower bound.

Support recovery theoretical guarantees

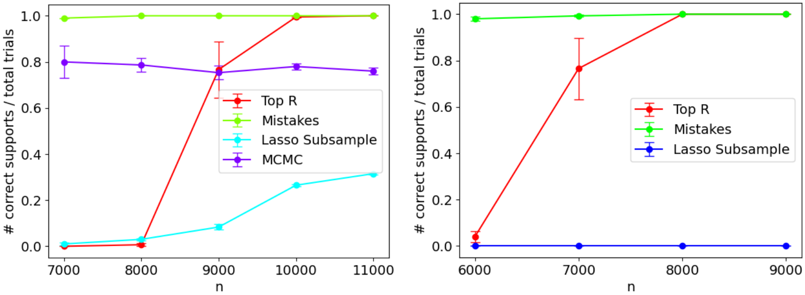
Suppose that data is generated using

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}^* + \boldsymbol{\epsilon}$$

where $\{\epsilon_i\}_{i\in[n]}$ are i.i.d. zero-mean sub-Gaussian random variables and $\boldsymbol{\beta}^*$ is the unknown feature vector with $\|\boldsymbol{\beta}^*\|_0 = s$.
Under standard boundedness and regularity conditions, we have that, in the case of least squares,

- if $\beta_{\min} := \min_{j\in\{i:\beta_i^*\neq 0\}} |\beta_j^*| \gtrsim \sqrt{\max\{1, s^2/\epsilon\}(\log p)/n}$, **Top-R** recovers the true support with high probability, which matches the non-private minimax-optimal $\sqrt{(\log p)/n}$ threshold by Guo et al. [2020] in the low-privacy regime.
- if $\beta_{\min} \gtrsim \sqrt{\max\{1, 1/\epsilon\}(s\log p)/n}$, **Mistakes** recovers the true support with high probability, which matches with the state-of-the-art condition of Roy and Tewari [2023] in the high-privacy regime.

Numerical experiments – support recovery



Least squares Hinge loss

Figure: Proportion of correct supports over number of trials for varying n, with $p = 10{,}000$, $s = 5$, SNR=5, $\rho = 0.1$, and $\epsilon = 1$. The objective function is least squares (left panel) and hinge loss (right panel).

Experiments (cont.) – prediction accuracy and utility loss

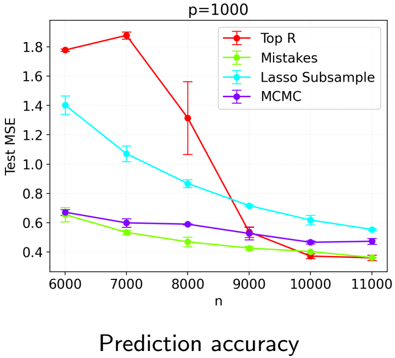


Figure: Numerical experiments for prediction accuracy (left panel) and utility loss (right panel) for varying n, with $p = 1{,}000$, $s = 5$, SNR=5, $\rho = 0.1$, and $\epsilon = 2$.

References

Dimitris Bertsimas and Bart Van Parys. Sparse high-dimensional regression: Exact scalable algorithms and phase transitions. 2020.

Cynthia Dwork, Aaron Roth, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4):211–407, 2014.

Yongyi Guo, Ziwei Zhu, and Jianqing Fan. Best subset selection is robust against design dependence. *arXiv preprint arXiv:2007.01478*, 2020.

Hussein Hazimeh and Rahul Mazumder. Fast best subset selection: Coordinate descent and local combinatorial optimization algorithms. *Operations Research*, 68(5):1517–1537, 2020.

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.

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *Conference on Learning Theory*, pages 25–1. JMLR Workshop and Conference Proceedings, 2012.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, pages 94–103. IEEE, 2007.

Alan Miller. *Subset selection in regression*. CRC Press, 2002.

Saptarshi Roy and Ambuj Tewari. On the computational complexity of private high-dimensional model selection via the exponential mechanism. *arXiv preprint arXiv:2310.07852*, 2023.

Abhradeep Guha Thakurta and Adam Smith. Differentially private feature selection via stability arguments, and the robustness of the lasso. In *Conference on Learning Theory*, pages 819–850. PMLR, 2013.