ON CORESET FOR LASSO REGRESSION PROBLEM WITH SENSITIVITY SAMPLING

Anonymous authors
Paper under double-blind review

ABSTRACT

In this paper, we study coreset construction for LASSO regression, where a coreset is a small, weighted subset of the data that approximates the original problem with provable guarantees. For unregularized regression problems, sensitivity sampling is a successful and widely applied technique for constructing coresets. However, extending these methods to LASSO typically requires coreset size to scale with $O(\mathcal{G}d)$, where d is the VC dimension and $\mathcal{G}$ is the total sensitivity, following existing generalization bounds. A key challenge in improving upon this general bound lies in the difficulty of capturing the sparse and localized structure of the function space induced by the ℓ_1 penalty in LASSO objective. To address this, we first provide an empirical process-based method of sensitivity sampling for LASSO, localizing the procedure by decomposing the functional space into independent spaces, which leads to tighter estimation error. By carefully leveraging the geometric properties of these localized spaces, we establish tight empirical process bounds on the required coreset size. These techniques enable us to achieve a coreset of size $\tilde{O}(\epsilon^{-2}d \cdot ((\log d)^3 \cdot \min\{1, \log d/\lambda^2\} + \log(1/\delta)))$, which ensures a $(1 \pm \epsilon)$-approximation for any $\epsilon, \delta \in (0,1)$ and $\lambda > 0$. Furthermore, we give a lower bound showing that any algorithm achieving a $(1+\epsilon)$-approximation must select at least $\Omega(\frac{d \log d}{\epsilon^2})$ rows in the regime where $\lambda = O(d^{-1/2})$. Empirical experiments show that our proposed algorithm is at least 4 times faster than the existing LASSO solver and more than 9 times faster on half of the datasets, while ensuring high solution quality and sparsity.

1 INTRODUCTION

In machine learning and regression analysis, sparse models have been extensively studied over the past decades. These models typically address issues such as sparse regression (Natarajan, 1995), variable selection (Zou & Hastie, 2005), and multicollinearity (Altelbany, 2021), aiming to improve model interpretability and computational efficiency by reducing the number of features. One of the most widely used methods for solving sparse models is the Least Absolute Shrinkage and Selection Operator (LASSO), which is first introduced in (Tibshirani, 1997). The core idea of LASSO regression is to apply an ℓ_1-norm penalty, ensuring sparsity by shrinking some coefficients to zero. Therefore, in practice, LASSO is widely applied in sparse models due to its effectiveness in enabling both variable selection and regularization with improved interpretability and prevented overfitting issues. The formal definition of the LASSO regression is given as follows.

LASSO Regression Problem. Given an $n \times d$ matrix A, an n-dimension vector b, and a regularization parameter $\lambda > 0$, the goal of LASSO problem is to find a d-dimension vector x that minimizes $\|Ax - b\|_2^2 + \lambda\|x\|_1$, where $\|Ax - b\|_2^2$ is the residual sum of squares, and the $\|x\|_1$ denotes the sum of the absolute values of the entries in x.

Although LASSO regression has been extensively studied over the past decade, the efficiency of LASSO algorithms in handling large-scale data still heavily depends on the number of rows n of the input matrix A. Specifically, the running time of existing algorithms, such as coordinate descent (Friedman et al., 2010), ISTA (Daubechies et al., 2004), and FISTA (Beck & Teboulle, 2009), is typically $O(nT \cdot \text{poly}(d))$, where T denotes the number of iterations. However, for datasets with a large number of samples n, LASSO may suffer from scalability issues. Therefore, developing row subsampling methods for LASSO regression is crucial for improving solving efficiency.

Among the vast literature on large-scale regression tasks, coreset techniques have played major roles in data subsampling. These algorithms aim to construct a weighted subset of the rows from both A and b, forming a compact representation that effectively approximates the original regression problem with strong theoretical guarantees. Along this line of research, several coreset construction algorithms have been proposed for the ℓ_p linear regression (Clarkson, 2005; Drineas et al., 2006; Dasgupta et al., 2009; Cohen & Peng, 2015; Woodruff & Yasuda, 2024; 2023; Munteanu & Omlor, 2024). In regularized regression tasks, Avron et al. (2017) constructed a coreset of size $O(\frac{sd_\lambda(A)+\log(1/\epsilon)}{\epsilon}\log\frac{sd_\lambda(A)}{\epsilon})$ for ridge regression, where $sd_\lambda(A) \leq d$ denotes the statistical dimension of the matrix A. Moreover, the works in (Kacham & Woodruff, 2020) introduced deterministic algorithms for coreset construction and explored a streaming model for this problem. Curtin et al. (2019) provided a logistic regression coreset with size $O(d\sqrt{n})$. Chhaya et al. (2020) proposed a coreset based on sensitivity sampling for the norm based regularized regression problem $\|Ax-b\|_p^p + \lambda\|x\|_p^p$ with $p \geq 2$. In a related recent work, Chhaya et al. (2020) studied a modified LASSO problem by constructing a coreset for the objective $\|Ax-b\|_2^2 + \lambda\|x\|_1^2$. However, the regularization term $\lambda\|x\|_1^2 = \lambda(\sum_i |x_i|)^2$, due to its quadratic nature, introduces cross terms among the x_i values. This may lead to solutions with substantially more nonzero coefficients than standard LASSO, thereby preventing it from promoting sparsity in the same way as the ℓ_1 norm and weakening its sparsity-inducing effect. To the best of our knowledge, there are currently no relevant theoretical results on coreset construction for standard LASSO, which motivates our work on developing such a coreset.

Coreset for LASSO. Let $A \in \mathbb{R}^{n\times d}$ be a matrix and $b \in \mathbb{R}^n$. Define $S \in \mathbb{R}^{n\times n}$ as a diagonal matrix, where each row $i \in [n]$ of both A and b is independently sampled with probability p_i. Let m denote the number of sampled rows. If row i is selected, set $S_{i,i} = 1/\sqrt{mp_i}$, and set $S_{i,i} = 0$ otherwise. We say that S defines an (ϵ, δ)-coreset for the LASSO problem if, with probability at least $1-\delta$, for all $x \in \mathbb{R}^d$ and $\lambda > 0$, the following holds

$$\|S(Ax-b)\|_2^2 + \lambda\|x\|_1 \in (1\pm\epsilon)\left(\|Ax-b\|_2^2 + \lambda\|x\|_1\right),$$

where $\epsilon \in (0,1)$. The coreset size is defined as the number of non-zeros entries on the diagonal of S, i.e., the number of sampled rows m.

Sensitivity sampling (Feldman & Langberg, 2011; Chhaya et al., 2020; Woodruff & Yasuda, 2023) has been extensively studied in regression without regularization, where rows are sampled in proportion to their importance in regression objective. A common challenge in directly applying sensitivity sampling to LASSO lies in bounding the generalization error under ℓ_1-regularized objective using standard empirical process tools. In the general framework of sensitivity sampling, Braverman et al. (2016) showed that, given sensitivity scores $\{\varrho_i\}_{i=1}^n$, a $(1\pm\epsilon)$-approximate coreset typically requires size $\tilde{O}\left(\frac{\mathcal{G}d}{\epsilon^2}\right)$ [1], where $\mathcal{G}$ is the sum of the sensitivity scores and d denotes the VC dimension of the given problem. This bound arises from applying a union bound to worst-case ϵ-net methods and variance analysis. Consequently, directly applying traditional analysis to LASSO leads to large coreset sizes, which can limit scalability in high-dimensional settings. To address this, empirical process techniques and chaining methods have been proposed to reduce the $\mathcal{G}d$ bound (Cohen et al., 2015; Woodruff & Yasuda, 2023; Munteanu & Omlor, 2024; Bansal et al., 2024). However, integrating empirical process theory with LASSO regression requires addressing the sparse and localized structure of the parameter space induced by the ℓ_1-penalty. In particular, the functional space $\Omega = \{x \in \mathbb{R}^d \mid h(x) + p(x) \leq R\}$, defined for a fixed radius $R > 0$, is determined by the residual term $h(x) = \|Ax-b\|_2^2$ and the penalty term $p(x) = \lambda\|x\|_1$ in the objective function. The interaction between the residual and penalty terms results in a highly complex geometry for Ω, complicating the standard empirical process analysis. Additionally, the non-smooth boundary introduced by the ℓ_1-penalty lead to large error bounds when applying the Bernstein inequality and ϵ-net analysis in (Chhaya et al., 2020). Therefore, developing a sensitivity sampling method that constructs a coreset smaller than $\tilde{O}(\mathcal{G}d)$ remains a key challenge for LASSO solvers.

1.1 Our Contribution

In this paper, we aim to improve upon existing standard bounds for LASSO coresets, which often lead to large sizes due to the application of union-bound–based ϵ-net methods. The main difficulty arises from the intricate structure of the function space introduced by both the residual error

[1] We write $\tilde{O}(f(n))$ to denote $O(f(n)\cdot \text{poly}\log f(n))$.

and the ℓ_1 regularization term. This complexity makes it difficult to directly apply standard empirical process techniques for sensitivity sampling. To address this issue, we propose a localized empirical process method that reformulates the sensitivity scores and sampling error in a more tractable way. Specifically, we define a weighted Gaussian-based empirical process for the coreset loss and decompose the overall function space into two independent components: the residual space and the ℓ_1 penalty space. Each of these components has lower complexity than the original space Ω, allowing for tighter bounds on Gaussian diameter and metric entropy within each component. By carefully applying symmetrization techniques and leveraging the geometric properties of these localized spaces, we derive upper bounds on the localized Gaussian diameter and metric entropy. These bounds allow us to control the sampling error and construct a coreset of size $\tilde{O}(\epsilon^{-2}d \cdot ((\log d)^3 \cdot \min\{1, \log d/\lambda^2\} + \log(1/\delta)))$, achieving a $(1 \pm \epsilon)$-approximation for any $\epsilon, \delta \in (0, 1)$ and $\lambda > 0$.

To complement our upper bound analysis, we establish a matching lower bound on the coreset size for LASSO regression via an information-theoretic method. By reducing the problem to a classical sparse recovery setting, we show that any estimator achieving $(1+\epsilon)$-approximation from the coreset must access a minimum number of rows to achieve sparse recovery task. In particular, in the regime where $\lambda = O(\frac{1}{\sqrt{d}})$, corresponding to the case where the number of nonzero entries can be large, we prove that the number of required rows is at least $\Omega(\frac{d}{\epsilon^2} \log(d))$. Our coreset size matches the lower bound up to polylogarithmic factors in the dimension d. Empirical experiments show that our proposed algorithm is at least 4 times faster than the direct LASSO solver and more than 9 times faster on half of the datasets, while preserving high solution quality. Notably, on a dataset with 8 million samples, our method completes in only 15 minutes.

1.2 Other Related Work

LASSO regression has been widely studied to perform various sparse models, such as variable selection (Tibshirani, 1997; Hans, 2010) and compressed sensing (Angelosante et al., 2009), which was first introduced in (Tibshirani, 1996). Many optimization algorithms have been developed for LASSO, including the fast iterative shrinkage-thresholding algorithm (Beck & Teboulle, 2009), coordinate descent algorithm (Friedman et al., 2010), smooth ℓ_1 algorithm (Schmidt et al., 2007), and path following algorithm (Tibshirani & Taylor, 2011). LASSO regression uses ℓ_1-regularization to relax the sparsity penalty (typically denoted by $\|x\|_0$), which is NP-hard (Natarajan, 1995). However, tuning the regularization parameter often leads to high computational costs. To address this, several methods have been proposed. Friedman et al. (2010) provided a "glmnet" package using coordinate descent method for LASSO solving. Obozinski & Bach (2012) proposed a stochastic variant that improves convergence via random selection. Wang et al. (2025) accelerated hyperparameter tuning using Markov resampling. To the best of our knowledge, there currently exists no coreset construction method for the LASSO task.

Sensitivity sampling is a well-studied technique for coreset construction in both theory and practice. It was first introduced by (Agarwal et al., 2004), and has since been widely applied to various problems, including clustering (Feldman & Langberg, 2011; Braverman et al., 2022; Bansal et al., 2024), linear regression (Drineas et al., 2006; Woodruff & Yasuda, 2024; 2023; Munteanu & Omlor, 2024), and matrix approximation (Dasgupta et al., 2009; Cohen et al., 2015). For the ordinary least squares regression, (Drineas et al., 2006) proposed a coreset algorithm based on the well-known statistical leverage score sampling. Dasgupta et al. (2009) extended this line of work to ℓ_p linear regression using well-conditioned basis method. More recently, a tight framework for constructing coresets for unregularized regression was developed by (Woodruff & Yasuda, 2023; 2024; Munteanu & Omlor, 2024), leveraging chaining techniques from empirical process theory.

Sensitivity sampling techniques have been extensively studied for regularized regression problems. For logistic regression, sensitivity-based sampling has been successfully applied in a series of works Munteanu et al. (2018); Curtin et al. (2019); Mai et al. (2021); Munteanu & Omlor (2024). In particular, Munteanu & Omlor (2024) recently provided a strong coreset of size $\tilde{O}(\mu d/\epsilon^2)$ based on the Lewis weight sampling Parulekar et al. (2021), where μ captures the complexity of the input data distribution. For ridge regression, Avron et al. (2017) pioneered the use of coreset techniques by showing that a weak coreset of size $\tilde{O}(sd_\lambda(A)/\epsilon^2)$ suffices to achieve a $(1 + \epsilon)$-approximation. Kacham & Woodruff (2020) developed the optimal deterministic coreset constructions for multi-response ridge regression. Their method selects $O(sd_\lambda(A)/\epsilon)$ rows and achieves a $(1+\epsilon)$-approximation, with matching lower bounds that establish the tightness of the dependence on $sd_\lambda(A)$. In the regime where $n \gg d$, the statistical dimension $d_\lambda(A)$ satisfies $sd_\lambda(A) \le \text{rank}(A) \le d$, and increase in regularization parameter λ can lead to lead to smaller coreset sizes for ridge regression.

In the broader context of norm-regularized regression, Chhaya et al. (2020) considered the coreset construction for ℓ_p regularized regression problems of the form $\|Ax-b\|_p^p + \lambda\|x\|_p^p$, where $p \ge 1$. They provided a strong coreset of $\tilde{O}(\frac{d^{p+1}}{\epsilon^2 \cdot (1+\lambda/\|A'\|_{(p)}^p)})$ based on the sensitivity sampling techniques. Moreover, they first showed that when $r \ne s$, no strong coreset can be smaller than the optimal coreset size for the unregularized term $\|Ax-b\|_p^r$. The result applies in particular to the LASSO, where $p=r=2$ and $q=s=1$. To address the LASSO objective, Chhaya et al. (2020) proposed a modified formulation in which the regularization term $\|x\|_1$ is replaced with $\|x\|_1^2$, enabling the use of ridge regression coreset techniques Avron et al. (2017) to construct a coreset of size $\tilde{O}(sd_\lambda(A)/\epsilon^2)$. However, this modification introduces cross terms among the components of x, which may weaken the sparsity-inducing effect of the standard ℓ_1 regularization. In this paper, the proposed coreset for standard LASSO objective has size $\tilde{O}(\epsilon^{-2}d \cdot ((\log d)^3 \cdot \min\{1, \log d/\lambda^2\} + \log(1/\delta)))$, which preserves the $\tilde{O}(d/\epsilon^2)$ bound when λ approaches to 0 or ∞. In addition, sketching-based methods using randomized projections have also been applied to the LASSO problem in recent Mai et al. (2023). Designing sensitivity sampling methods for constructing coresets for LASSO remains an interesting open problem.

2 Preliminaries

Given a positive integer n, let $[n] = \{1, 2, \dots, n\}$. For a d-dimensional vector $x \in \mathbb{R}^d$, the ℓ_p-norm of x is $\|x\|_p = (\sum_{i=1}^d x_i^p)^{1/p}$. For an $n \times d$ matrix A, the induced p-norm is $\|A\|_{(p)}$, which is defined as $\|A\|_{(p)} = \sup_{x\ne 0, x\in\mathbb{R}^d} \frac{\|Ax\|_p}{\|x\|_p}$. The ℓ_2-norm (or spectral norm) $\|A\|_{(2)}$ corresponds to the maximum singular value of A. For a matrix $A \in \mathbb{R}^{n\times d}$, the ℓ_p norm of A is $\|A\|_p = (\sum_{i=1}^n \sum_{j=1}^d A_{ij}^p)^{1/p}$, and the Frobenius norm of A is $\|A\|_F = (\sum_{i=1}^n \sum_{j=1}^d A_{ij}^2)^{1/2}$. Let $A_{i:}$ be the i-th row of A, and let A_{ij} be the entry in the i-th row and j-th column of A. Let $A^\top$ be the transport matrix of the matrix A. The Singular Value Decomposition (SVD) of matrix A is $A = U\Sigma V^\top$, where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{d\times d}$ are orthogonal matrices, and $\Sigma \in \mathbb{R}^{n\times d}$ is a diagonal matrix containing the singular values $\sigma_1, \dots, \sigma_r$, where $r \le \min\{n, d\}$. For a vector $x \in \mathbb{R}^n$ and weight vector $w \in \mathbb{R}^n_{\ge 0}$, the weighted ℓ_p-norm is $\|x\|_{w,p} = (\sum_{i=1}^n w_i|x_i|^p)^{1/p}$, and the weighted ℓ_∞ norm is $\|x\|_{w,\infty} = \max_{i\in[n]} |x_i|$. An ϵ-net for a set K in a metric space (X, d) is a subset $T \subseteq K$ such that for every point $x \in K$, there exists $y \in T$ with $d(x, y) \le \epsilon$. Given a parameter $\lambda > 0$, we define the statistical dimension of a matrix A as

$$sd_\lambda(A) = \sum_{i=1}^r \frac{1}{1+\lambda/\sigma_i^2},$$

where r denotes the rank of A. For any vector $x \in \mathbb{R}^d$, let $\text{supp}(x) = \{i \in [d] \mid x_i \ne 0\}$ denote its support, and write $|\text{supp}(x)|$ for the number of nonzero coordinates.

ℓ_2 **Leverage Scores.** The ℓ_2-norm leverage score of the i-th row of matrix A is $\tau_{i,2}(A) = \sup_{x\in\mathbb{R}^d} \frac{\|A_{i:}^\top x\|_2^2}{\|Ax\|_2^2}$. Alternatively, the leverage scores can be expressed as $\tau_{i,2}(A) = \|e_i^\top U\|_2^2$, where $U \in \mathbb{R}^{n\times d}$ is an orthonormal basis for the column space of A (Cohen et al., 2015). Therefore, the sum of the ℓ_2 leverage scores is satisfies $\sum_{i=1}^d \tau_{i,2}(A) = d$.

3 Sensitivity Sampling for LASSO regression

In this section, we propose a sensitivity sampling algorithm for LASSO regression, called LASSO-Sens. The main goal is to derive a better upper bound on the coreset size using empirical process methods applied to the LASSO objective. The primary technical challenge lies in handling the in-

Algorithm 1 LASSO-Sens

Input: a matrix $A \in \mathbb{R}^{n\times d}$, a vector $b \in \mathbb{R}^n$, a regularized parameter λ, the over-sampling parameter α, the coreset size T, a set of approximate sensitivity scores $\{\varrho_i\}_{i=1}^n$, and a parameter $\epsilon > 0$

Output: a set of indices Q, and a weight vector $w \in \mathbb{R}^n_{\geq 0}$

1: Initialize an empty set Q, and let w be an n-dimensional zero vector.
2: Initialize the total sensitivity $\mathcal{G} = 0$.
3: **for** $i \leftarrow 1, 2, \ldots, n$ **do**
4: Compute the sampling probability for the i-row $p_i = \min\{1, \alpha(\varrho_i + \frac{1}{n})\}$.
5: Update $\mathcal{G} = \mathcal{G} + \varrho_i$.
6: **end for**
7: **for** $t \leftarrow 1, 2, \ldots, T$ **do**
8: Sample a row index $i \in [n]$ with probability p_i, and set the weight $w_i = 1/\sqrt{p_i}$.
9: $Q \leftarrow Q \cup \{i\}$.
10: **end for**
11: **return** Q and w.

teraction between the residual loss and the ℓ_1 penalty, as standard empirical process techniques typically rely on analyzing the ratio between them, which is difficult to handle and may lead to weaker coreset size bounds. To address this issue, we provide a localization method for coreset within the empirical process framework, which decouples the problem into two components over localized regions. This allows us to analyze the empirical process in a localized space involving only a single term. Over these localized sets we develop a weighted Gaussian empirical-process framework and derive upper bounds on the Gaussian diameter, covering numbers, and metric entropy. These ingredients yield a coreset of size $\tilde{O}(\frac{d}{\epsilon^2} \cdot ((\log d)^3 \min\{1, \frac{\log d}{\lambda^2}\} + \log(1/\delta)))$, which nearly matches the lower bound in the regime $\lambda = O(1/\sqrt{d})$. The detailed algorithm for constructing the LASSO coreset is given in Algorithm 1.

In sensitivity sampling, the sensitivity score of the i-th row for LASSO objective is defined as

$$\varrho_i = \sup_{x\in\mathbb{R}^d} \frac{\|(Ax-b)_i\|_2^2 + \lambda\frac{\|x\|_1}{n}}{\|Ax-b\|_2^2 + \lambda\|x\|_1}, \tag{1}$$

where $\lambda > 0$. The definition of ϱ_i is to capture the worst-case contribution to the LASSO objective, with the regularization term $\lambda\|x\|_1$ ensuring that each row contributes equally to sampling. Bounding the score ϱ_i by the ℓ_2 leverage score τ_i with an additive $1/n$ in this paper is straightforward; see formal details Section A.1 in Appendix.

The LASSO-Sens algorithm mainly consists of a sampling procedure for coreset construction. We initialize an empty set of indices Q and a zero vector w. Then, we calculate the sampling probability $p_i = \min\{1, \alpha(\varrho_i + \frac{1}{n})\}$ and update the total sensitivity $\mathcal{G}$ by adding ϱ_i, where α represents the over-sampling parameter. Next, the algorithm then randomly selects a row index $i \in [n]$ with probability p_i, assigns the weight of the i-th row to $1/\sqrt{p_i}$, and updates the set of indices to $Q = Q \cup \{i\}$. By repeating this sampling process T times, Algorithm 1 returns the final set of row indices Q and the corresponding weight vector w.

Before providing the theoretical guarantees for the coreset, we first present an equivalent transformation of the LASSO objective and its sensitivity scores. Let $A' = [A \ \ -b] \in \mathbb{R}^{n\times(d+1)}$ be the matrix obtained by concatenating A and b, and $x' = [x\ 1]$ be the vector obtained by concatenating x with 1. Using A' and x', the original objective function $\min_x \|Ax-b\|_2^2 + \lambda\|x\|_1$ is rewritten as

$$\min_{x'\in\mathbb{R}^{d+1}, x'_{d+1}=1} \|A'x'\|_2^2 + \lambda\|x'\|_1.$$

Thus, we reformulate the sensitivity score ϱ_i as

$$\varrho_i = \sup_{x'\in\mathbb{R}^{d+1}, x'_{d+1}=1} \frac{\|(A'x')_i\|_2^2 + \frac{\lambda}{n}\|x'\|_1}{\|A'x'\|_2^2 + \lambda\|x'\|_1} > 0.$$

3.1 Sampling Error Analysis

In this subsection, we develop a localized empirical process framework to analyze the sampling error introduced by sensitivity sampling in the LASSO objective. To achieve this, we decompose the function space into the residual and penalty components, and localize our study to their intersection. This separation enables us to independently bound the Gaussian complexity and metric entropy of each component using a combination of weighted chaining techniques. By constructing multi-scale ϵ-nets and applying concentration inequalities for Gaussian processes, we establish an upper bound on the coreset size that controls the sampling error.

We now analyze the sampling error introduced by sensitivity sampling. Let $\{p_i\}_{i=1}^n$ denote the sampling probabilities associated with each row of the augmented matrix A'. Define the sampling and rescaling matrix $S \in \mathbb{R}^{n\times n}$ as

$$S = w^\top \Psi, \text{ where } \Psi = \text{diag}(\psi_1, \dots, \psi_n), \psi_i = \begin{cases} 1, & \text{with probability } p_i \\ 0, & \text{otherwise} \end{cases}, \tag{2}$$

and w is a vector of rescaling weights. The matrix Ψ is diagonal with m nonzero entries in expectation. Let $\mathcal{T} = \{x \mid x \in \mathbb{R}^{d+1}, x_{d+1} \neq 0\}$, and let $\Omega = \{x \mid x \in \mathcal{T}, \|A'x\|_2^2 + \lambda\|x\|_1 = 1\}$ be the unit ball of the LASSO objective. Then, we define the sampling error $\mathcal{E}$ over the domain Ω as

$$\begin{aligned}\mathcal{E} &= \sup_{x'\in\Omega} \left| \|SA'x'\|_2^2 + \lambda\|x'\|_1 - (\|A'x'\|_2^2 + \lambda\|x'\|_1) \right| \\ &= \sup_{x'\in\Omega} \left| \|SA'x'\|_2^2 - \|A'x'\|_2^2 \right|.\end{aligned}$$

Our goal is to bound $\mathcal{E}$ by ϵ, leading to the inequality

$$\|SA'x'\|_2^2 + \lambda\|x'\|_1 \leq (1 \pm \epsilon) \left(\|A'x'\|_2^2 + \lambda\|x'\|_1 \right)$$

for every $x' \in \Omega$. To bound $\mathcal{E}$ using the chaining method (Cohen & Peng, 2015; Koltchinskii, 2001; Hu et al., 2022), we analyze the moments of $\mathcal{E}$ with the symmetrization technique, which allows us to construct a Gaussian reduction as follows. (A detailed proof of Lemma 1 is given in Appendix Lemma 3.)

Lemma 1. *Let $A' \in \mathbb{R}^{n\times(d+1)}$, let S be a random sampling matrix, and let Q denote the set of the sampled rows from A'. For $\lambda > 0$ and integer $l \geq 2$, the following inequality holds*

$$\mathbb{E}_S |\mathcal{E}|^l \leq (2\pi)^{l/2} \mathbb{E}_S \mathbb{E}_{g\sim\mathcal{N}(0,I_n)} \sup_{x\in\Omega} \left| \sum_{i\in Q} g_i w_i \left|(A_{i:}x)\right|^2 \right|^l,$$

where $g \sim \mathcal{N}(0, I_n)$ represents a Gaussian vector with independent entries.

We bound the sampling error $\mathcal{E}$ by analyzing the associated Gaussian process, as described in Lemma 1. To handle higher-order moments on $\mathcal{E}$, we apply a moment bound from Woodruff & Yasuda (2023), which uses Dudley's tail inequality for Gaussian processes. Consequently, we obtain the following inequality on the sampling error

$$\mathbb{E}_S[|\mathcal{E}|^l] \leq (C\mathcal{M}_\mathcal{E})^l (\mathcal{M}_\mathcal{E}/\mathcal{D}) + O(\sqrt{l}\mathcal{D})^l, \tag{3}$$

where C is an absolute constant, $\mathcal{M}_\mathcal{E}$ denotes the metric entropy of the Gaussian process, and $\mathcal{D}$ is the Gaussian diameter. (The detailed definitions of $\mathcal{M}_\mathcal{E}$ and $\mathcal{D}$ are provided in the following.)

By appropriately choosing the parameter l and bounding both the metric entropy $\mathcal{M}_\mathcal{E}$ and the Gaussian diameter $\mathcal{D}$ of the Gaussian process, we can ensure that $\mathbb{E}_S[|\mathcal{E}|^l] \leq \epsilon$, which leads to a sufficiently small coreset size m. We now decompose the unit ball $\mathcal{L} = \{x \mid x \in \mathcal{T}, \|A'x\|_2^2 + \lambda\|x\|_1 \leq 1\}$, which arises from the residual term $\|A'x\|_2$ and the ℓ_1 penalty. (The proof of Lemma 2 is provided in Appendix Lemma 4.)

Lemma 2. *Let $A' \in \mathbb{R}^{n\times(d+1)}$ be a matrix and $\lambda > 0$. Define the sets $\Omega = \{x \mid x \in \mathcal{T}, \|A'x\|_2^2 + \lambda\|x\|_1 \leq 1\}$ and $\mathcal{L} = \{x \mid x \in \mathcal{T}, \|A'x\|_2^2 \leq 1$ and $\|x\|_1 \leq \frac{1}{\lambda}\}$. Then, it holds that $\Omega \subseteq \mathcal{L}$.*

Define $B_2(A') = \{x \mid x \in \mathcal{T}, \|A'x\|_2^2 \leq 1\}$ as the unit ball in the residual space, and $B_1(\frac{1}{\lambda}) = \{x \mid x \in \mathcal{T}, \|x\|_1 \leq \frac{1}{\lambda}\}$ as the unit ball in the ℓ_1-penalty space. By Lemma 2, we have

$$\mathcal{L} \subseteq \mathcal{L}_{A'} = B_2(A') \cap B_1(\frac{1}{\lambda}).$$

This allows us to proceed with bounding both the Gaussian diameter $\mathcal{D}$ and the metric entropy $\mathcal{M}_{\mathcal{E}}$ for the convex sets $B_2(A')$ and $B_1\left(\frac{1}{\lambda}\right)$, respectively. Let $M = \Psi A'$, where $\Psi \in \mathbb{R}^{n\times n}$ is a diagonal sampling matrix. In this formulation, each nonzero row of M corresponds to a selected row of A'.

We start by bounding the Gaussian diameter $\mathcal{D}$ by relaxing the pseudo-metric d_X using the maximum ℓ_2 leverage score and λ. Define the convex set $\mathcal{L}_M = \{y = Mx \mid x \in \mathcal{L}_{A'}\}$. Let $\tau = \sup_{x'\in\mathcal{L}_M} \|Mx'\|_{2,\infty}^2$ be the maximum of ℓ_2 leverage score. Next, we prove that the diameter $\mathcal{D}(\mathcal{L}_M)$ with d_X is bounded as the following inequality. (Detailed proof of Lemma 3 is given in Appendix Lemma 5.)

Lemma 3. *Let $M \in \mathbb{R}^{m\times(d+1)}$, and let w be the weight vector. Define the pseudo-metric*

$$d_X(y, y') = \left(\mathbb{E}_{g\sim\mathcal{N}(0,I_n)} \left|\sum_{i=1}^m g_i w_i |y_i|^2 - \sum_{i=1}^m g_i w_i |y_i'|^2\right|^2\right)^{1/2}$$

for any $y, y' \in \mathcal{L}_M$. Then, the diameter $\mathcal{D}(\mathcal{L}_M)$ with respect to d_X is bounded by

$$\mathcal{D}(\mathcal{L}_M) \leq O(\tau \cdot \sqrt{\log\left(d(\lambda^2 \wedge 1)\right)} \wedge (\lambda\sqrt{d})).$$

To obtain a precise bound for the Gaussian process over $\mathcal{L}_M$, we apply the chaining method to construct a sequence of t-nets at varying scales $t > 0$, which capture the convex structure of $\mathcal{E}$ on $\mathcal{L}_M$. Utilizing this chaining method, we can derive a bound on $\mathbb{E}_S|\mathcal{E}|^l$ via the covering numbers of the sequence of t-nets. Thus, we aim to bound the minimal number of weighted unit ℓ_p (or ℓ_∞) balls required to cover the convex set $\mathcal{L}_M$ for $p \in [1, \infty)$. We define the weighted unit ball of the residual space $B_{w,2}(M)$ as $B_{w,2}(M) = \{y = Mx \mid x \in \mathcal{T}, \|Mx\|_{w,2}^2 \leq 1\}$, and define $\mathcal{L}_{w,M} = B_{w,2}(M) \cap B_1(1/\lambda)$. Let $G = 1 + \mathcal{E} = 1 + \sup_{x'\in\mathcal{L}_M} \left|\|SA'x'\|_2^2 - \|A'x'\|_2^2\right|$.

To bound the metric entropy entropy of the convex set $B_{w,2}(M)$, we first define the weighted unit $\ell_{w,p}$-ball as $B_{w,p}(M) = \{x \mid x \in \mathcal{T}, \|Mx\|_{w,p}^2 \leq 1\}$. Let $\mathcal{T}_p$ denote the t-net of $B_{w,2}(M)$ with respect to the weighted ℓ_p-norm, i.e., a finite subset of $B_{w,2}(M)$ such that every point in $B_{w,2}(M)$ is within distance t (measured in $\|\cdot\|_{w,p}$) from some point in $\mathcal{T}_p$. We define $N(B_{w,2}(M), \|\cdot\|_{w,p}, t)$ as the minimal cardinality of such a set $\mathcal{T}_p$, and the metric entropy of $B_{w,2}(M)$ w.s.t the weighted ℓ_p-norm is then defined as $\log N(B_{w,2}(M), \|\cdot\|_{w,p}, t)$. (Detailed definitions are provided in Appendix, Definitions 11 and 12.)

Lemma 4 ((Munteanu & Omlor, 2024), slightly modified)**.** *Let $2 \leq p < \infty$, and let $M \in \mathbb{R}^{m\times(d+1)}$ be an orthonormal matrix with a weight vector $w \in \mathbb{R}^m_{\geq 0}$. Then, the following inequalities hold* $\log N(B_{w,2}(M), \|\cdot\|_{w,p}, t) \leq O(1)\frac{m^{2/p} p\cdot\tau}{t^2}$ *and* $\log N(B_{w,2}(M), \|\cdot\|_{w,\infty}, t) \leq O(1)\frac{\log m\cdot\tau}{t^2}$.

For bounding the metric entropy of the convex set $B_1\left(\frac{1}{\lambda}\right)$, we aim to bound the number of unit B_∞-balls needed to cover the B_1-ball. Specifically, the covering process can be decomposed into two steps: first, cover the B_2-ball using B_∞-balls, and second, cover the B_1-ball using B_2-balls. The B_1 ball has a unique geometric structure, with a large portion of its volume concentrated near its center, as pointed out in (Vershynin, 2018). This concentration implies that fewer small-radius balls are required to cover B_1, compared to naive volume-based estimates. While a straightforward volumetric argument yields a worst-case covering number of $O((1+\frac{1}{\epsilon})^d)$, this bound can be quite loose. To obtain a tighter estimate, we leverage the Sudakov Minoration inequality (Vershynin, 2018), which provides an upper bound on the covering number $N(B_1, B_\infty, t)$ with respect to the ℓ_∞ norm and covering radius t. (Detailed proof of Lemma 5 is given in Appendix Lemma 13.)

Lemma 5. *Let $p \geq 1$ be a parameter, and let $B_p = \{x \in \mathbb{R}^d : \|x\|_p \leq 1\}$ be the unit ball for the ℓ_p norm. Then,* $\log N(B_1, B_\infty, t) \leq O(\frac{\log d}{t})$.

To bound the metric entropy of these t-nets, we need to calculate the following integral

$$\mathcal{M}_{\mathcal{E}} \leq \int_0^\infty \sqrt{\log N(\mathcal{L}_{w,M}, d_X, t)}\, \mathrm{d}t.$$

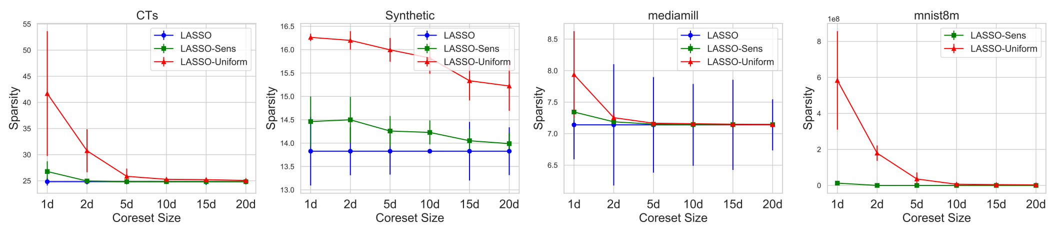
For diameters $t > \mathcal{D}(\mathcal{L}_{w,M})$, the covering number satisfies $\log N(\mathcal{L}_{w,M}, d_X, t) = 0$, which implies that any single vector $y \in \mathcal{L}_{w,M}$ serves as a t-net. Therefore, we only need to focus on the case where the diameter t lies within the interval $[0, \mathcal{D}(\mathcal{L}_{w,M})]$. We derive the following inequality, whose proof provided in Appendix Lemma 19.

Lemma 6. *Let $M \in \mathbb{R}^{m\times(d+1)}$ be a matrix and λ be a positive parameter. Then, the metric entropy $\mathcal{M}_{\mathcal{E}}$ of $\mathcal{L}_{w,M}$ satisfies*

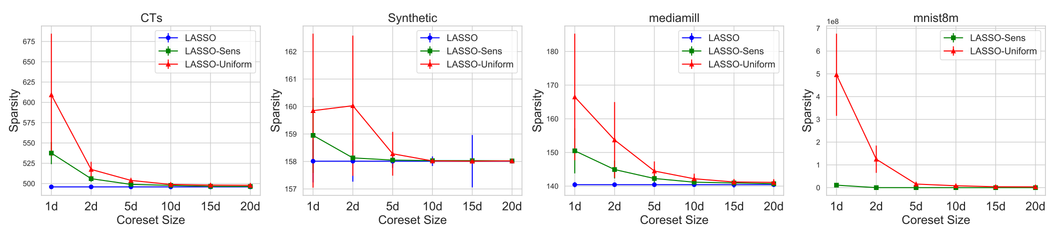
$$\int_0^\infty \sqrt{\log N(\mathcal{L}_{w,M}, d_X, t)}\, dt \le O(G \cdot \sqrt{\tau \cdot \log m} \log d \cdot \min\{1, \frac{\sqrt{\log d}}{\lambda}\}),$$

where τ is the maximum weighted ℓ_2-leverage score of M.

We now present the main result, which provides a bound on the coreset size required to guarantee that $\mathbb{E}|\mathcal{E}|^l \le \epsilon$. (The proof is provided in Appendix, Theorem 22.)



(a) $\lambda = 0.5$



(b) $\lambda = 10$

Figure 1: LASSO regression loss comparison across varying coreset sizes for $\lambda = \{0.5, 10\}$.

Table 1: Comparison results of loss, runtime, and sparsity on CTs dataset (n = 53,500, d = 386) for varying coreset sizes at $\lambda = 0.5$.

Metrics	Algorithms	Coreset Sizes					
		$1d$	$2d$	$5d$	$10d$	$15d$	$20d$
Loss	LASSO	**24.83±0.51**					
	LASSO-Sens	26.77±1.95	24.96±0.26	24.84±0.01	**24.83±0.01**	**24.83±0.01**	**24.83±0.01**
	LASSO-Uniform	41.69±11.93	30.75±4.11	25.85±1.48	25.28±0.32	25.21±0.26	25.06±0.27
Time (s)	LASSO	691.72					
	LASSO-Sens	5.80	8.12	10.28	16.67	23.10	37.37
	LASSO-Uniform	6.84	8.23	9.85	18.86	26.73	42.03
Sparsity	LASSO	315					
	LASSO-Sens	379	348	330	**313**	**312**	**317**
	LASSO-Uniform	379	359	336	319	320	**317**

Theorem 7. *Let $A' \in \mathbb{R}^{n\times(d+1)}$ be an input matrix, S be a random sampling matrix, and let $\varepsilon, \delta \in (0,1)$ and $\lambda > 0$ be a parameter. If $\alpha = \tilde{O}\left(\frac{1}{\epsilon^2} \cdot \left(\log\left(d \log(1/\delta)\right)(\ln d)^2 \cdot \min\left\{1, \frac{\log d}{\lambda^2}\right\} + \ln\left(1/\delta\right)\right)\right)$ and for all $i \in [n]$ it holds that*

$$p_i \ge \min\{1, \alpha(\tau_{i,2}(A') + \frac{1}{n})\},$$

where $\tau_{i,2}(A')$ denotes the ℓ_2 leverage score of the i-th row of A'. Then, with failure probability at most δ, it holds that, $\forall x \in \mathbb{R}^{d+1}, x_{d+1} = 1$,

$$\|SA'\hat{x}\|_2^2 + \lambda\|\hat{x}\|_1 \le (1 \pm \epsilon)(\|A'x\|_2^2 + \lambda\|x\|_1),$$

and the coreset size is at most $m = \tilde{O}\left(\frac{d(\log d)^3}{\epsilon^2} \cdot \min\{1, \frac{\log d}{\lambda^2}\} + \frac{d}{\epsilon^2}\log\frac{1}{\delta}\right)$.

To establish a lower bound on the coreset size m, we utilize a reduction from the support recovery for sparse recovery problem. We consider the task of recovering the support of a sparse vector x^*, and apply information-theoretic techniques for LASSO regression problem. Our analysis shows that, under certain conditions, any algorithm achieving a $(1+\epsilon)$-approximation from the coreset requires at least $\Omega(\frac{d}{\epsilon^2}\log d)$ rows. Since Mai et al. (2023) pointed out the lack of scale-invariance in the LASSO objective, we normalize the inputs by assuming $\|A\|_2 \le 1$ and $\|b\|_2 \le 1$. Detailed proofs are provided in Appendix B.4.

Lemma 8. *Let $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, and $\lambda \in (0,1)$. Assume that $\|A\|_2 \le 1$ and $\|b\|_2 \le 1$. Let S be a diagonal sampling matrix with m non-zero entries. Suppose there exists an estimator that returns $\tilde{x} = \arg\min_{x\in\mathbb{R}^d} \|SAx - Sb\|_2^2 + \lambda\|x\|_1$ satisfies*

$$\|A\tilde{x} - b\|_2^2 + \lambda\|\tilde{x}\|_1 \le (1+\epsilon)\cdot \min_{x\in\mathbb{R}^d}(\|Ax - b\|_2^2 + \lambda\|x\|_1).$$

Then, the coreset size m must satisfy

$$m = \begin{cases} \Omega(\frac{\log d}{\lambda^2\cdot\epsilon^2}), & \text{if } \lambda = \Omega(\frac{1}{\sqrt{d}}) \\ \Omega(\frac{d}{\epsilon^2}\log d), & \text{if } \lambda = O(\frac{1}{\sqrt{d}}) \end{cases}.$$

4 Experiments

In this section, we compare three algorithms for solving the LASSO regression problem: direct optimization using the full dataset, and solving LASSO on subsamples selected via sensitivity sampling and uniform sampling, respectively. All experiments are conducted on a machine with 72 Intel Xeon Gold 6230 CPUs and 340 GB of memory, and all implementations are executed in MATLAB 2017A.

Datasets. We evaluate the three algorithms on 4 datasets: Synthetic ($n = 10,000, d = 200$), mediamill ($n = 30,993, d = 120$), CTs ($n = 53,500, d = 386$), mnist8m ($n = 8,000,000, d = 784$). The synthetic dataset is generated by constructing a matrix $A \in \mathbb{R}^{10000\times 200}$, where a small number of rows have high leverage scores. This construction follows the method described in (Chhaya et al., 2020). The resulting matrix is designed to exhibit a non-uniform leverage score distribution while maintaining a well-conditioned structure. For all datasets, the response vector is defined as $b = Ax + 10^{-5}\cdot\frac{\|b\|_2}{\|e\|_2}\cdot e$, where $x \in \{0,1\}^d$ is a randomly generated sparse vector, and e is a noise vector. All datasets used in our experiments are publicly available at: `https://archive.ics.uci.edu/datasets` and `https://www.csie.ntu.edu.tw/~cjlin/libsvmtools/datasets/`.

Algorithms. In our experimental evaluation, we compare the following three algorithms:

- LASSO. The standard LASSO regression is solved using the FISTA method as described in (Beck & Teboulle, 2009).
- LASSO-Sens. Our proposed approach (see Algorithm 1), which first constructs a coreset via sensitivity-based sampling and solves the LASSO problem on the coreset using FISTA.
- LASSO-Uniform. A baseline that first uniformly samples rows from the input data and then applies FISTA to solve the LASSO problem on the sampled data.

Methodology. We evaluate algorithm performance using the loss function $f(x) = \|Ax - b\|_2^2 + \lambda\|x\|_1$, where a lower value of loss indicates a better solution. To evaluate the sparsity of the solution, we follow the method in (Chhaya et al., 2020) by setting any entry of x with an absolute value less than 10^{-6} to 0, and we count the remaining nonzero entries. Our experiments test three methods: LASSO, LASSO-Sens, and LASSO-Uniform on four datasets. To ensure a fair comparison, we test each algorithm 10 times and report the average loss, runtime, and sparsity. To compare the performance of different sampling strategies, we run LASSO-Sens and LASSO-Uniform across a range of coreset sizes and regularization parameters, with the values $\lambda \in \{0.5, 1, 5, 10\}$. Specifically, the coreset size is selected from $\{1, 2, 5, 10, 15, 20\} \times d$ for each dataset.

Results for the LASSO Regression. As shown in Figures 1 and 2 (see Appendix), the LASSO-Sens algorithm achieves loss values that closely match those of the exact LASSO solver as the coreset size increases, particularly for $\lambda = 0.5$ and $\lambda = 10$. The comparison of performance metrics across four datasets under varying values of λ and coreset sizes are reported in Table 1 and Appendix Tables 2-5, including average loss, standard deviation, runtime, and solution sparsity. On the Synthetic, Mediamill, and CTs datasets, LASSO-Sens is at least 4 times faster than LASSO, and up to 18 times faster on CTs. On mnist8m dataset, LASSO-Sens obtains a feasible solution within 15 minutes, whereas the standard LASSO solver fails to return a solution even after 48 hours. Furthermore, the LASSO-Sens algorithm consistently outperforms the LASSO-Uniform in terms of both accuracy and sparsity on mnist8m dataset. At a coreset size of $10d$, the sparsity of the solutions produced by LASSO-Sens closely matches that of the exact LASSO solver across all datasets. These experimental results show the sensitivity sampling in accelerating the LASSO regression process while preserving high-quality and the sparsity of solutions. All of these findings, together with our Theorem 8, confirm the effectiveness of sensitivity sampling for LASSO regression.

5 Conclusion

In this paper, we propose the first coreset construction method for LASSO regression via sensitivity sampling algorithm. Directly applying existing coreset techniques for regularized regression to LASSO yields a coreset size bound of $\tilde{O}(\mathcal{G}d/\epsilon^2)$. To achieve a smaller coreset, we propose an empirical process analysis that addresses the complex functional space arising from the interaction between the residual error and ℓ_1-penalty in LASSO, thereby achieving a coreset of size $\tilde{O}(\epsilon^{-2}d \cdot ((\log d)^3 \cdot \min\{1, \log d/\lambda^2\} + \log(1/\delta)))$. An interesting future direction is to study how our method can be extended to the elastic net and other regression problems involving more complex regularization.

Ethics Statement

This work does not involve any human subjects, sensitive data, or other ethical concerns as outlined in the ICLR Code of Ethics.

Reproducibility Statement

This paper is committed to ensuring the reproducibility of our work. To ensure the completeness of the comparison, we have provided detailed descriptions of our proposed method and its components in Section 4 of the main paper. To enable accurate replication, we clearly specify all hyperparameters, training procedures, and evaluation protocols in the same section. Additional implementation results, including figures and tables, are provided in the Appendix.

References

Pankaj K Agarwal, Sariel Har-Peled, and Kasturi R Varadarajan. Approximating extent measures of points. *Journal of the ACM*, 51(4):606–635, 2004.

Shady Altelbany. Evaluation of ridge, elastic net and lasso regression methods in precedence of multicollinearity problem: a simulation study. *Journal of Applied Economics and Business Studies*, 5 (1):131–142, 2021.

Daniele Angelosante, Georgios B Giannakis, and Emanuele Grossi. Compressed sensing of time-varying signals. In *Proceedings of the 16th International Conference on Digital Signal Processing*, pp. 1–8, 2009.

Haim Avron, Kenneth L Clarkson, and David P Woodruff. Sharper bounds for regularized data fitting. In *Proceedings of the 17th Annual Conference on Approximation, Randomization, and Combinatorial Optimization: Algorithms and Techniques.*, pp. 27:1–27:22, 2017.

Nikhil Bansal, Vincent Cohen-Addad, Milind Prabhu, David Saulpic, and Chris Schwiegelshohn. Sensitivity sampling for k-means: Worst case and stability optimal coreset bounds. In *Proceeding of the 65th IEEE Annual Symposium on Foundations of Computer Science*, pp. 1707–1723, 2024.

Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal on Imaging Sciences*, 2(1):183–202, 2009.

Pierre C Bellec. Localized gaussian width of m-convex hulls with applications to lasso and convex aggregation. *Bernoulli: official journal of the Bernoulli Society for Mathematical Statistics and Probability*, 25(4A):3016–3040, 2019.

Lindenstrauss J Bourgain J and Milman V. Approximation of zonoids by zonotopes. *Acta Math*, 162:73–141, 1989.

Vladimir Braverman, Dan Feldman, Harry Lang, Adiel Statman, and Samson Zhou. New frameworks for offline and streaming coreset constructions. *arXiv preprint arXiv:1612.00889*, 2016.

Vladimir Braverman, Vincent Cohen-Addad, H-C Shaofeng Jiang, Robert Krauthgamer, Chris Schwiegelshohn, Mads Bech Toftrup, and Xuan Wu. The power of uniform sampling for coresets. In *Proceedings of the 63rd IEEE Annual Symposium on Foundations of Computer Science*, pp. 462–473, 2022.

Bernd Carl. Inequalities of bernstein-jackson-type and the degree of compactness of operators in banach spaces. In *Annales de l'institut Fourier*, volume 35, pp. 79–118, 1985.

Rachit Chhaya, Anirban Dasgupta, and Supratim Shit. On coresets for regularized regression. In *Proceedings of the 37th International Conference on Machine Learning*, pp. 1866–1876, 2020.

Kenneth L Clarkson. Subgradient and sampling algorithms for ℓ_1 regression. In *Proceedings of the 16th Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 257–266, 2005.

Michael B Cohen and Richard Peng. ℓ_p row sampling by lewis weights. In *Proceedings of the 47th Annual ACM Symposium on Theory of Computing*, pp. 183–192, 2015.

Michael B Cohen, Yin Tat Lee, Cameron Musco, Christopher Musco, Richard Peng, and Aaron Sidford. Uniform sampling for matrix approximation. In *Proceedings of the 6th Innovations in Theoretical Computer Science Conference*, pp. 181–190, 2015.

Ryan R Curtin, Sungjin Im, Ben Moseley, Kirk Pruhs, and Alireza Samadian. On coresets for regularized loss minimization. *arXiv preprint arXiv:1905.10845*, 2019.

Anirban Dasgupta, Petros Drineas, Boulos Harb, Ravi Kumar, and Michael W Mahoney. Sampling algorithms and coresets for ℓ_p regression. *SIAM Journal on Computing*, 38(5):2060–2078, 2009.

Ingrid Daubechies, Michel Defrise, and Christine De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57(11):1413–1457, 2004.

Petros Drineas, Michael W Mahoney, and Shan Muthukrishnan. Sampling algorithms for ℓ_2 regression and applications. In *Proceedings of the 70th Annual ACM-SIAM Symposium on Discrete algorithm*, pp. 1127–1136, 2006.

Dan Feldman and Michael Langberg. A unified framework for approximating and clustering data. In *Proceedings of the 43rd annual ACM symposium on Theory of computing*, pp. 569–578, 2011.

Jerome Friedman, Trevor Hastie, and Rob Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010.

Chris Hans. Model uncertainty and variable selection in bayesian lasso regression. *Statistics and Computing*, 20(2):221–229, 2010.

Lunjia Hu, Charlotte Peale, and Omer Reingold. Metric entropy duality and the sample complexity of outcome indistinguishability. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, pp. 515–552, 2022.

Lingxiao Huang, Zhize Li, Jialin Sun, and Haoyu Zhao. Coresets for vertical federated learning: Regularized linear regression and k-means clustering. *Proceedings of the 36th Annual Conference on Neural Information Processing Systems*, 35:29566–29581, 2022.

Praneeth Kacham and David Woodruff. Optimal deterministic coresets for ridge regression. In *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, pp. 4141–4150, 2020.

Vladimir Koltchinskii. Rademacher penalties and structural risk minimization. *IEEE Transactions on Information Theory*, 47(5):1902–1914, 2001.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Michael W Mahoney et al. Randomized algorithms for matrices and data. *Foundations and Trends® in Machine Learning*, 3(2):123–224, 2011.

Tung Mai, Cameron Musco, and Anup Rao. Coresets for classification–simplified and strengthened. pp. 11643–11654, 2021.

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.

Alexander Munteanu and Simon Omlor. Optimal bounds for ℓ_p sensitivity sampling via ℓ_2 augmentation. In *Proceedings of the 45th International Conference on Machine Learning*, pp. 36769–36796, 2024.

Alexander Munteanu, Chris Schwiegelshohn, Christian Sohler, and David Woodruff. On coresets for logistic regression. *Proceedings of the 32nd Annual Conference on Neural Information Processing Systems*, 31, 2018.

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

Guillaume Obozinski and Francis Bach. Convex relaxation for combinatorial penalties. *arXiv preprint arXiv:1205.1240*, 2012.

Aditya Parulekar, Advait Parulekar, and Eric Price. L1 regression with lewis weights subsampling. In *Proceedings of the 24th International Workshop on Approximation Algorithms for Combinatorial Optimization Problems and the 25th International Workshop on Randomization and Computation*, pp. 49:1–49:21, 2021.

Mert Pilanci and Martin J Wainwright. Randomized sketches of convex programs with sharp guarantees. *IEEE Transactions on Information Theory*, 61(9):5096–5115, 2015.

Mark Schmidt, Glenn Fung, and Rómer Rosales. Fast optimization methods for ℓ_1 regularization: A comparative study and two new approaches. In *Proceedings of the 11th European Conference on Machine Learning*, pp. 286–297, 2007.

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

Robert Tibshirani. The lasso method for variable selection in the cox model. *Statistics in Medicine*, 16(4):385–395, 1997.

Ryan J Tibshirani and Jonathan Taylor. The solution path of the generalized lasso. *The Annals of Statistics*, 39(3):1335, 2011.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge University Press, 2018.

Martin J Wainwright. Information-theoretic limits on sparsity recovery in the high-dimensional and noisy setting. *IEEE transactions on Information theory*, 55(12):5728–5741, 2009.

Wei Wang, Martin J Wainwright, and Kannan Ramchandran. Information-theoretic limits on sparse signal recovery: Dense versus sparse measurement matrices. *IEEE Transactions on Information Theory*, 56(6):2967–2979, 2010.

Yuhang Wang, Bin Zou, Jie Xu, Chen Xu, and Yuan Yan Tang. Alr-ht: A fast and efficient lasso regression without hyperparameter tuning. *Neural Networks*, 181:106885, 2025.

David P Woodruff. Sketching as a tool for numerical linear algebra. *Foundations and Trends® in Theoretical Computer Science*, 10(1–2):1–157, 2014.

David P Woodruff and Taisuke Yasuda. Coresets for multiple ℓ_p regression. In *Proceedings of the 41st International Conference on Machine Learning*, pp. 53202–53233, 2024.

P David Woodruff and Taisuke Yasuda. Sharper bounds for ℓ_p sensitivity sampling. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 37238–37272, 2023.

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 APPENDIX

A.1 MISSING PROOF OF SENSITIVITY SCORES

In this subsection, we provide an upper bound on the sensitivity score ρ_i using the ℓ_2 leverage score and a fixed term $1/n$. While this result is not fundamentally new (see, e.g., (Mahoney et al., 2011; Woodruff, 2014; Chhaya et al., 2020)), we slightly extend the well-conditioned basis method to the LASSO objective.

Definition 1. (ℓ_2 Well-Conditioned Basis.) Given a matrix $A \in \mathbb{R}^{n\times d}$, we define a $(\sqrt{d}, 1, 2)$ well-conditioned basis for A such that $\|U\|_2 \le \sqrt{d}$, and $\forall x \in \mathbb{R}^d, \|x\|_2 \le \|Ux\|_2$, where $U \in \mathbb{R}^{n\times d}$ is the orthogonal matrix obtained from SVD of A.

Lemma 2. Let $A' \in \mathbb{R}^{n\times(d+1)}$, and let $\lambda > 0$ be a regularized parameter. Then, the estimated sensitivity score $\hat{\varrho}_i$ satisfies $\hat{\varrho}_i = 2\tau_{i,2}(A') + \frac{1}{n} \ge \varrho_i$, where $\tau_{i,2}(A')$ denotes the ℓ_2 leverage score of the i-th row of A'. All sensitivity scores $\hat{\varrho}_i$ can be computed in time $O(\text{nnz}(A')\log n + d^3 \log(n/d) \log d)$, where $\text{nnz}(A')$ denotes the number of non-zero entries in A'. Moreover, the total sensitivity is bounded as $\mathcal{G} \le 2d + 3$.

Proof. Let $A' = UV$, where $U \in \mathbb{R}^{n\times(d+1)}$ is a $(\sqrt{d+1}, 1, 2)$-well-conditioned basis for A'. Denote the i-th row of A' as $A'_i = u_i^\top V$, where $u_i^\top$ is the i-th row of U. For any $x' \in \mathbb{R}^{d+1}$, define $z = Vx'$, so that $A'x' = Uz$. Let $T = \{x' \in \mathbb{R}^{d+1} : x'_{d+1} = 1\}$. Then, we obtain

$$\varrho_i = \sup_{x'\in T} \frac{|A'_i x'|^2 + \frac{\lambda}{n}\|x'\|_1}{\|A'x'\|_2^2 + \lambda\|x'\|_1} = \sup_z \frac{|u_i^T z|^2 + \frac{\lambda}{n}\|V^{-1}z\|_1}{\|Uz\|_2^2 + \lambda\|V^{-1}z\|_1} \le \sup_z \frac{|u_i^T z|^2}{\|Uz\|_2^2} + \frac{1}{n} \le \tau_{i,2}(A') + \frac{1}{n}.$$

Thus, the total sensitivity satisfies $\mathcal{G} = \sum_{i=1}^n \varrho_i \le \sum_{i=1}^n \left(\tau_{i,2}(A') + \frac{1}{n}\right) \le d + 2$, where $\tau_{i,2}(A') = \|u_i\|_2^2$ denotes the ℓ_2 leverage score of the i-th row. Furthermore, by extending Lemma 8 of (Cohen et al., 2015), the approximate leverage score $\hat{\tau}_{i,2}(A') \le 2\tau_{i,2}(A')$ can be computed in time $O(\text{nnz}(A')\log n + d^3 \log d \log(n/d))$. Substituting this into the bound yields $\varrho_i \le 2\tau_{i,2}(A') + \frac{1}{n}$ and $\mathcal{G} \le 2d + 3$. □

B OMITTED PROOFS OF SAMPLING ERROR ANALYSIS

In this section, we reduce the empirical process associated with the sampling error $\mathcal{E}$ to a Gaussian process using the symmetrization technique. The sampling error $\mathcal{E}$ is defined on the set $\Omega = \left\{x \in \mathcal{T} \;\middle|\; \|A'x\|_2^2 + \lambda\|x\|_1 = 1\right\}$, where $\mathcal{T} = \left\{x \in \mathbb{R}^{d+1} \;\middle|\; x_{d+1} = 1\right\}$. To analyze the functional complexity, we consider the larger set $\mathcal{T}' = \left\{x \in \mathbb{R}^{d+1} \;\middle|\; x_{d+1} \ne 0\right\}$, since any $x \in \mathcal{T}$ can be obtained by scaling an element of $\mathcal{T}'$. Specifically, for each $x \in \mathcal{T}$, there exists a scalar c and an $x' \in \mathcal{T}'$ such that $x = c \cdot x'$. This inclusion implies that $\mathcal{T} \subseteq \mathcal{T}'$. Consequently, we define the extended domain $\Omega' = \left\{x \in \mathcal{T}' \;\middle|\; \|A'x\|_2^2 + \lambda\|x\|_1 = 1\right\}$ and focus our subsequent analysis on this set using tools from Gaussian process theory, particularly those developed for unregularized regression in Woodruff & Yasuda (2023).

Lemma 3. Let $A' \in \mathbb{R}^{n\times(d+1)}$, let S be a random sampling matrix, and let Q denote the set of the sampled rows from A'. For $\lambda > 0$ and integer $l \ge 2$, the following inequality holds

$$\mathbb{E}_S |\mathcal{E}|^l \le (2\pi)^{l/2} \mathbb{E}_S \mathbb{E}_{g\sim\mathcal{N}(0,I_n)} \sup_{x\in\Omega} \left| \sum_{i\in Q} g_i w_i \left|(A_{i:}x)\right|^2 \right|^l,$$

where $g \sim \mathcal{N}(0, I_n)$ represents a Gaussian vector with independent entries.

Proof. We consider the simple convex function $|a+b|^l$ for $a, b \in \mathbb{R}$, where $l > 1$ is a positive number. Given a random sampling matrix S, the linearity of expectation implies

$$\mathbb{E}\left[\|SA'x\|_2^2 + \lambda\|x\|_1\right] = \|A'x\|_2^2 + \lambda\|x\|_1$$

for any vector $x \in \mathbb{R}^{d+1}$. Next, without loss of generality, we assume that $\|A'x\|_2^2 + \lambda\|x\|_1 = 1$; otherwise, we can rescale x by a constant to satisfy this condition.

We now analyze the following quantity

$$\mathcal{E} = \mathbb{E}_S \sup_{\|A'x\|_2^2+\lambda\|x\|_1=1, x\in\mathcal{T}'} \left|\|SA'x\|_2^2 + \lambda\|x\|_1 - 1\right|^l$$

Let S' be an independently copy of S. Applying Jensen inequality, we have

$$\begin{aligned}
&\mathbb{E}_S \sup_{\|A'x\|_2^2+\lambda\|x\|_1=1, x\in\mathcal{T}'} |\|SA'x\|_2^2 + \lambda\|x\|_1 - (\|A'x\|_2^2 + \lambda\|x\|_1)|^l \\
&= \mathbb{E}_S \sup_{\|A'x\|_2^2+\lambda\|x\|_1=1, x\in\mathcal{T}'} \left|\|SA'x\|_2^2 - \|A'x\|_2^2 + 0\right|^l \\
&= \mathbb{E}_S \sup_{\|A'x\|_2^2+\lambda\|x\|_1=1, x\in\mathcal{T}'} |\|SA'x\|_2^2 - \|A'x\|_2^2 + \mathbb{E}_{S'}(\|A'x\|_2^2 - \|S'A'x\|_2^2)|^l \\
&\leq \mathbb{E}_{S,S'} \sup_{\|A'x\|_2^2+\lambda\|x\|_1=1, x\in\mathcal{T}'} \left|\||SA'x\|_2^2 - \mathbb{E}_{S'}\|S'A'x\|_2^2\right|^l \\
&\leq \mathbb{E}_{S,S'} \sup_{\|A'x\|_2^2+\lambda\|x\|_1=1, x\in\mathcal{T}'} \left|\||SA'x\|_2^2 - \|S'A'x\|_2^2\right|^l.
\end{aligned}$$

Using a standard symmetrization argument (Vershynin, 2018), we obtain

$$\begin{aligned}
&\mathbb{E}_{S,S'} \sup_{\|A'x\|_2^2+\lambda\|x\|_1=1, x\in\mathcal{T}'} \left|\|SA'x\|_2^2 - \|S'A'x\|_2^2\right|^l \\
&\leq 2^l \mathbb{E}_{S,\epsilon} \sup_{\|A'x\|_2^2+\lambda\|x\|_1=1, x\in\mathcal{T}'} \left|\sum_{i\in Q} \epsilon_i w_i |(A'_{i:})x|^2\right| \\
&\leq 2^l (\pi/2)^{l/2} \mathbb{E}_{S,g} \sup_{\|A'x\|_2^2+\lambda\|x\|_1=1, x\in\mathcal{T}'} \left|\sum_{i\in Q} g_i w_i |(A'_{i:})x|^2\right|^l,
\end{aligned}$$

where $\epsilon \sim \{\pm 1\}^n$ are independent Rademacher variables in the first inequality, and the second inequality follows from the Rademacher–Gaussian comparison theorem (Ledoux & Talagrand, 1991) with $g \sim \mathcal{N}(0, I_n)$ a standard Gaussian vector in $\mathbb{R}^n$. □

We now provide a detailed analysis of the localization of the empirical process over the residual space $B_2(A') = \{x \mid x \in \mathcal{T}', \|A'x\|_2 \leq 1\}$ and the ℓ_1-penalty space $B_1(1/\lambda) = \{x \mid x \in \mathcal{T}', \|x\|_1 \leq 1\}$. In the following lemma, we show that the set Ω' is contained in the intersection of these two sets.

Lemma 4. Let $A' \in \mathbb{R}^{n\times(d+1)}$ be a matrix and $\lambda > 0$. Define the sets $\Omega = \{x \mid x \in \mathcal{T}, \|A'x\|_2^2 + \lambda\|x\|_1 \leq 1\}$ and $\mathcal{L} = \{x \mid x \in \mathcal{T}, \|A'x\|_2^2 \leq 1 \text{ and } \|x\|_1 \leq \frac{1}{\lambda}\}$. Then, it holds that $\Omega \subseteq \mathcal{L}$.

Proof. Let the vector $x \in \Omega$. By the definition of the set Ω, we have

$$\|A'x\|_2^2 + \lambda\|x\|_1 = 1.$$

Since $\|A'x\|_2^2$ is non-negative, we can derive

$$\|A'x\|_2^2 = 1 - \lambda\|x\|_1.$$

By the equation $\|A'x\|_2^2 = 1 - \lambda\|x\|_1$, it follows that

$$1 - \lambda\|x\|_1 \geq 0 \quad \Rightarrow \quad \lambda\|x\|_1 \leq 1 \quad \Rightarrow \quad \|x\|_1 \leq \frac{1}{\lambda}.$$

Next, from the equation $\|A'x\|_2^2 + \lambda\|x\|_1 \leq 1$, we can express $\|A'x\|_2^2$ as follows

$$\|A'x\|_2^2 = 1 - \lambda\|x\|_1.$$

Since we have already shown that $\|x\|_1 \leq \frac{1}{\lambda}$, we have

$$\|A'x\|_2^2 = 1 - \lambda\|x\|_1 \leq 0.$$

Therefore, we obtain

$$\|A'x\|_2^2 \leq 1 \rightarrow \|A'x\|_2 \leq 1.$$

In summary, for any $x \in \Omega$, the conditions $\|A'x\|_2 \leq 1$ and $\|x\|_1 \leq \frac{1}{\lambda}$ are satisfied. Therefore, we conclude that $\Omega \subseteq \mathcal{L}$. □

B.1 Bounding the Gaussian Diameter

We start by bounding the Gaussian diameter $\mathcal{D}$ with respect to the pseudo-metric d_X.

Lemma 5. Let $M \in \mathbb{R}^{m\times(d+1)}$, and let w be the weight vector. Define the pseudo-metric

$$d_X(y, y') = \left(\mathbb{E}_{g\sim\mathcal{N}(0,I_n)} \left|\sum_{i=1}^{m} g_i w_i |y_i|^2 - \sum_{i=1}^{m} g_i w_i |y_i'|^2\right|^2\right)^{1/2}$$

for any $y, y' \in \mathcal{L}_M$. Then, the diameter $\mathcal{D}(\mathcal{L}_M)$ with respect to d_X is bounded by

$$\mathcal{D}(\mathcal{L}_M) \leq O(\sqrt{\tau} \cdot \sqrt{\log\left(d(\lambda^2 \wedge 1)\right)} \wedge (\lambda\sqrt{d})).$$

Proof. We aim to bound the Gaussian diameter $\mathcal{D}(\mathcal{L}_M)$ under the pseudo-metric d_X. A standard result (see e.g., (Vershynin, 2018, Proposition 7.5.4)) implies that for any convex set T,

$$D_T \leq \sqrt{2\pi}\mathcal{W}(T),$$

where $w(T) := \mathbb{E}_{g\sim\mathcal{N}(0,I)}\left[\sup_{t\in T}\langle g, t\rangle\right]$ is the Gaussian width of T. Therefore, it suffices to bound $\mathcal{W}(\mathcal{L}_M)$.

We observe that $\mathcal{L}_M$ is the image of a convex set under a linear mapping. Specifically, we define the set

$$\hat{\mathcal{L}} = \{x \in \mathbb{R}^{d+1} \mid \|x\|_1 \leq \frac{1}{\lambda}, \|Mx\|_2 \leq 1\},$$

where $M \in \mathbb{R}^{m\times d}$. Then, $\mathcal{L}_M = w^\top M\mathcal{L}$.

According to the definition of Gaussian width, we have

$$\begin{aligned}
\mathcal{W}(\mathcal{L}_M) &= \mathbb{E}_{g\sim\mathcal{N}(0,I_m)}\left[\sup_{x\in\hat{\mathcal{L}}}\langle g, w^\top Mx\rangle\right] \\
&= \mathbb{E}_{g\sim\mathcal{N}(0,I_m)}\left[\sup_{x\in\hat{\mathcal{L}}}\langle wM^\top g, x\rangle\right] \\
&\leq \|w^\top M\|_2 \cdot \mathbb{E}_{g\sim\mathcal{N}(0,I_{d+1})}\left[\sup_{x\in\hat{\mathcal{L}}}\langle g, x\rangle\right] \\
&= \|M\|_{w,2} \cdot \mathcal{W}(\hat{\mathcal{L}}).
\end{aligned}$$

Now we bound $\mathcal{W}(\hat{\mathcal{L}})$. Let $B_2(M) = \{x \mid x \in \mathbb{R}^{d+1}, \|Mx\|_2 \leq 1\}$. Then, it follows that

$$\mathcal{W}(\hat{\mathcal{L}}) = \mathcal{W}(B_2(M) \cap \frac{1}{\lambda}B_1) \leq \|M\| \cdot \mathcal{W}(B_2 \cap \frac{1}{\lambda}B_1) = \|M\| \cdot \mathcal{W}(\lambda \cdot B_2 \cap B_1),$$

where B_1, B_2 are the unit balls in ℓ_1 and ℓ_2 norms, respectively.

Applying the localized Gaussian width bound (see e.g., (Bellec, 2019, Proposition 1)), we obtain

$$w\left(B_1 \cap \lambda \cdot B_2\right) \leq C \cdot \left(\sqrt{\log\left(2d \cdot (\lambda^2 \wedge 1)\right)} \ \wedge\ \lambda \cdot \sqrt{d}\right),$$

for some universal constant C.

Therefore, we obtain

$$\mathcal{W}(\mathcal{L}_M) \leq \|M\|_{w,2}\mathcal{W}(\hat{\mathcal{L}}) \leq C\|M\|_{w,2}^2 \cdot \left(\sqrt{\log\left(2d \cdot (\lambda^2 \wedge 1)\right)} \ \wedge\ \lambda \cdot \sqrt{d}\right).$$

Recall that the maximum weighted ℓ_2 leverage score is defined as $\tau := \|M\|_{w,2,\infty}^2$. Finally, applying $\mathcal{D}(\mathcal{L}_M) \leq \sqrt{2\pi} \cdot \mathcal{W}(\mathcal{L}_M)$, we conclude that

$$\mathcal{D}(\mathcal{L}_M) \leq O(\tau\sqrt{\log\left(d(\lambda^2 \wedge 1)\right)} \wedge (\lambda\sqrt{d})).$$

□

B.2 Bounding the Metric Entropy

In this subsection, we establish an upper bound for the metric entropy $\mathcal{M}_{\mathcal{E}}$ of the space $\mathcal{L}_{w,M}$. To estimate this entropy, we first provide detailed definitions of covering numbers and metric entropy.

Definition 6. Let d_X be a pesudo-metric on $\mathbb{R}^d$. Given a vector $x \in \mathbb{R}^d$ and $t \geq 0$, we define the d_X-ball of radius t centered at x as $B_X(x,t) = \{x' \in \mathbb{R}^d : d_X(x,x') \leq t\}$.

Definition 7. Let $K, T \subseteq \mathbb{R}^d$ be two convex bodies. The covering number $N(K,T)$ represents the minimum number of copies of T required to cover K

$$N(K,T) = \min\{k \in \mathbb{N} : \exists \{x_i\}_{i=1}^k, K \subseteq \bigcup_{i=1}^k (x_i + T)\}.$$

Let d_X be a pseudo-metric and $t > 0$ a scalar. The covering number of a set K with respect to d_X and radius t is denoted by $N(K, d_X, t) = N(K, B_X(0,t))$, where $B_X(0,t)$ is the d_X-ball of radius t centered at the origin. The metric entropy is given by $\mathcal{M}_{\mathcal{E}} = \log N(K, d_X, t)$.

We now apply the standard tool, Dual Sudakov Minoration (Bourgain J & V., 1989), to bound the covering numbers in both the residual space and the ℓ_1-penalty space. The following theorem provides an upper bound on the covering numbers of the Euclidean unit ball within a metric space by using ℓ_p-norm balls with radius $t > 0$.

Definition 8. The Levy mean of ℓ_p is defined as

$$M_p = \frac{\mathbb{E}_{g \in \mathcal{N}(0,I_d)} \|g\|_p}{\mathbb{E}_{g \in \mathcal{N}(0,I_d)} \|g\|_2}.$$

Theorem 9. (Dual Sudakov Minoration) Let $\|\cdot\|_p$ be a norm, and let $B_2 \subseteq \mathbb{R}^d$ denote the Euclidean unit ball, defined as $B_2 = \{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}$. Then,

$$\log N(B_2, \|\cdot\|_p, t) \leq O(d) \frac{M_p^2}{t^2}.$$

Lemma 10. (Woodruff & Yasuda (2023), slightly modified) Let $q \geq 2$, let $M \in \mathbb{R}^{m \times (d+1)}$ be a matrix, and let $w \in \mathbb{R}^m$ be a wight vector. Then, for a standard Gaussian vector $g \sim \mathcal{N}(0, I_{d+1})$, it holds that

$$\mathbb{E}_{g \sim N(0, I_{d+1})} [\|Mg\|_{w,q}] \leq m^{1/q} \cdot \sqrt{q\tau},$$

and

$$\mathbb{E}_{g \sim N(0, I_{d+1})} [\|g\|_2] \leq \sqrt{d+1}.$$

We now focus on the ℓ_1-penalty space for the Gaussian process. To bound the metric entropy of the set $B_1(1/\lambda)$ using the unweighted ℓ_∞-ball, we decompose the process into two steps: covering the Euclidean unit ball with B_∞, and covering B_1 using the Euclidean unit ball. We define the unweighted ℓ_p (including ℓ_∞) unit ball as $B_p = \{x \mid x \in \mathbb{R}^{d+1}, \|x\|_p \leq 1\}$. The following lemma provides a bound for the first step.

Lemma 11. (Woodruff & Yasuda, 2023) Let $p \geq 2$ and let B_p be the unit ball for the ℓ_p norm. Then,

$$\log N(B_2, B_\infty, t) \leq O(1) \frac{\log d}{(t/2)^2}.$$

Since the B_1 ball has a non-smooth geometric structure, a substantial portion of its volume is concentrated near its center. This concentration implies that fewer smaller balls are needed to effectively cover the unit ball. A directly application of the ϵ-net argument typically yields a general bound of $O((1 + \frac{1}{t})^{d+1})$ in the worst-case. To obtain the better bound by utilizing this concentration, we use the Sudakov Minoration inequality (Vershynin, 2018), specifically for the non-smooth B_1 ball, as follows.

Theorem 12. Let $\mathcal{K}$ be a convex body in $\mathbb{R}^d$, and let $N(\mathcal{K}, B_2, t)$ denote the covering number of balls of radius t required to cover $\mathcal{K}$. Then, for any $t > 0$,

$$\sqrt{\log N(\mathcal{K}, B_2, t)} \leq C \cdot \frac{\mathcal{W}(\mathcal{K})}{t},$$

where C is an absolute constant, and $\mathcal{W}(\mathcal{K}) = \mathbb{E}\left[\sup_{x\in\mathcal{K}}\langle g, x\rangle\right]$ represents the Gaussian width with respect to a standard Gaussian vector $g \sim \mathcal{N}(0, I_d)$.

Lemma 13. Let $p \geq 1$ be a parameter, and let $B_p = \{x \mid x \in \mathbb{R}^d, \|x\|_p \leq 1\}$ be the unit ball for the ℓ_p norm. Then,

$$\log N(B_1, B_\infty, t) \leq O(\frac{\log d}{t}).$$

Proof. To bound the covering numbers of B_1 by B_∞, we first cover B_1 by B_2, and then use Lemma 11 to cover B_2 by B_∞.

Define the Guassian width of B_1 as $\mathcal{W}(B_1) = \mathbb{E}(\sup_{t\in B_1} g't)$, where $g \in \mathbb{R}^{d+1}$ is a standard Gaussian vector. By applying the Hölder inequality, $g't \leq |g't| \leq \|g\|_\infty \cdot \|t\|_1$. Thus, we can bound the Gaussian width by

$$\begin{aligned}\mathcal{W}(B_1) = \mathbb{E}\left(\sup_{t\in B_1} g't\right) &\leq \mathbb{E}\left(\sup_{t\in B_1} \|g\|_\infty \cdot \|t\|_1\right) \\ &= \mathbb{E}\left(\max_j |g_j|\right).\end{aligned}$$

To bound $\mathbb{E}(\max_j |g_j|)$, note that $\max_j |g_j| = \max(\max_j g_j, \max_j -g_j)$. For the vector g and a positive parameter $u \geq 0$, we can derive an upper bound for $\mathbb{E}(\max_j |g_j|)$ as follows

$$\begin{aligned}\exp(u\mathbb{E}(\max_i g_i)) &\leq \mathbb{E}\exp(u \cdot \max_i g_i) \\ &= \mathbb{E}(\max_i \exp(ug_i)) \\ &\leq \sum_{i=1}^{d+1} \mathbb{E}(\exp(ug_i)) \\ &\leq (d+1) \cdot \exp(u^2/2).\end{aligned}$$

where the first inequality follows from the Jensen inequality and utilizes the moment generating function of a Gaussian distribution.

Thus, we get

$$\mathbb{E}(\max_i V_i) \leq \frac{\log(d+1)}{u} + \frac{u\sigma^2}{2}.$$

Minimizing w.r.t u by choosing $u = \sqrt{2\log(d+1)}$, we obtain

$$\mathbb{E}(\max_i V_i) \leq \sigma\sqrt{2\log(d+1)}.$$

Since the fact that $\max_j |g_j| = \max(\max_j g_j, \max_j -g_j)$, we have

$$\mathbb{E}\left(\max_j |g_j|\right) \leq \sqrt{2\log 2(d+1)} \leq \sqrt{4\log(d+1)}.$$

Consequently, we have that the Gaussian average of the ℓ_1-ball is $\mathcal{W}(B_1) \leq \sqrt{2\log(d+1)}$. By Theorem 12, we can bound the covering number

$$N(B_1, B_2, t) \leq \exp(\frac{C^2 \cdot 4\log d}{t^2}) = d^{4C^2/t^2},$$

where C is a constant.

Thus, the metric entropy of covering B_1 by tB_2 ball is at most

$$\log N(B_1, B_2, t) \leq \log\left(d^{4C^2/t^2}\right) \leq O(\frac{\log d}{t^2}).$$

Using the above inequality and Lemma 11, we obtain

$$\begin{aligned}\log N(B_1, B_\infty, t) &\leq \log N(B_1, B_2, \gamma) + \log N(\lambda B_2, B_\infty, t) \\ &\leq \log N(B_1, B_2, \gamma) + \log N(B_2, B_\infty, t/\gamma) \\ &\leq O(1)\frac{\log d}{\gamma^2} + O(1)\frac{(\log d)}{(t/\gamma)^2}\end{aligned}$$

for any $\gamma \in [1, t]$. Choosing γ, we obtain

$$\log N(B_1, B_\infty, t) \leq O(\frac{\log d}{t}).$$

□

Lemma 14. (Munteanu & Omlor, 2024) Let $M \in \mathbb{R}^{m\times d}$ and let $w \in \mathbb{R}^m_{\geq 0}$ be a non-negative weight vector corresponding to the rows of M. Then, for any $1 \leq r \leq q$ and any $t \geq 0$,

$$N(B_{1,r}(M), B_{1,q}(M), t) \geq N(B_{w,r}(M), B_{w,q}(M), t).$$

We give two upper bounds for the covering numbers in the ℓ_1-penalty space based on the radius t. For larger radii ($t > t_0$), the covering number scales with $(1/t)^2$, indicating a quadratic increase as t decreases. Conversely, for smaller radii ($t \leq t_0$), the covering number grows logarithmically with $1/t$.

Lemma 15. Let $M \in \mathbb{R}^{m\times(d+1)}$ be an orthogonal matrix, and let $\lambda > 0$. Define the set $B_\infty(M) = \{x \mid x \in \mathcal{T}', \|Mx\|_\infty \leq 1\}$ as the unit ball in the ℓ_∞-norm mapped by M. Let $H = \max_{1\leq i\leq m} \|e_i^T M\|_\infty$, where $e_i \in \mathbb{R}^m$ is the i-th standard basis vector. Let $t_0 = O(H\sqrt{\frac{\log d}{m}})$. Then, the following bounds on the metric entropy hold for all $t > 0$

$$\log N(B_1(1/\lambda), B_\infty(M), t) \leq O(H)\frac{\log d \cdot \log m}{\lambda^2 t^2},$$

and

$$\log N(B_1(1/\lambda), B_\infty(M), t) \leq O(m \log(1 + \frac{t_0}{t\lambda}) + \log m).$$

Proof. Given $\delta > 0$, we define the scaled convex set $\delta B_1(1)$ as $\delta B_1(1) = \{\delta x \mid x \in \mathcal{T}', \|x\|_1 \leq 1\}$. For any $y \in \delta B_1(1)$, there exists $x \in \mathbb{R}^{d+1}$ such that $\|x\|_1 \leq 1$ and $y = \delta x$. Then, $\|y\|_1 = \|\delta x\|_1 = \delta\|x\|_1$. Conversely, suppose $y \in \mathbb{R}^{d+1}$ satisfies $\|y\|_1 \leq \delta$. Define $x = \frac{y}{\delta}$ (for $\delta > 0$), then $\|x\|_1 = \left\|\frac{y}{\delta}\right\|_1 = \frac{1}{\delta}\|y\|_1 \leq 1$, so $x \in B_1(1)$ and $y = \delta x \in \delta B_1(1)$.

Hence, we conclude $\delta B_1(1) = \{y \mid y \in \mathbb{R}^{d+1}, \|y\|_1 \leq \delta\}$, and $\frac{1}{\delta}B_1(1) = B_1\left(\frac{1}{\delta}\right)$.

Now, we aim to prove that $\log N\left(\frac{1}{\lambda}B_1, B_\infty, t\right) = \log N\left(B_1, B_\infty, \lambda t\right)$. Define $K = \{x \mid x \in \mathbb{R}^{d+1}, \|x\|_1 \leq \frac{1}{\lambda}\}$. For any $x \in K$, we have $\|x\|_1 \leq \frac{1}{\lambda}$, and hence $\|x\|_\infty \leq \frac{1}{\lambda}$.

Covering $x \in 1/\lambda B_1$ with t-balls in the ℓ_∞-norm is equivalent to covering B_1 with λt-balls due to scaling. Therefore, we obtain

$$\log N\left(\frac{1}{\lambda}B_1, B_\infty, t\right) = \log N\left(B_1, B_\infty, \lambda t\right). \tag{4}$$

Next, we define the set $H_m = \{x \mid x \in \mathbb{R}^{d+1}, \max_{1\leq i\leq m} |\langle x, M_{i:}\rangle\| \leq 1\}$ and let $\|\cdot\|_{H_m}$ be the associated quasi-norm on $\mathbb{R}^{d+1}$. Define the linear operator $F : \ell_1^m \to \mathbb{R}^{d+1}$ by $Fe_i = M_{i:}$. Then, the covering number of using H_n to cover B_1 satisfies

$$N(B_1, H_m, t) = N(F^* B_1, B_\infty^m, t),$$

where $B_\infty^m = \{x \mid x \in \mathbb{R}^m, \|x\|_\infty \leq 1\}$. By the Bernstein-Jackson-type inequality (Carl, 1985), for the embedding $\ell_1^m \to \ell_\infty^d$, the metric entropy satisfies $\log N(B_1, H_m, t) \leq$

$O(H)(\frac{\log(1+m/l)\cdot\log(1+d/l)}{l})^{1/2}$, where $l = \arg\inf\{\epsilon > 0, N(B_1, H_m, \epsilon) \leq 2^l\}$. Let $t_0 = O(H\sqrt{\frac{\log d}{m}})$. Then, for $t > t_0$, we have

$$\log N(B_1, H_m, t) \leq O(\frac{H \cdot \log d \cdot \log m}{t^2}).$$

By applying Lemma 5, for $t < t_0$, we obtain

$$\begin{aligned}\log N(B_1, H_m, t) &\leq \log N(B_1, H_m, t_0) + \log N(t_0 H_m, H_m, t)\\ &\leq O(\frac{H^2}{t_0^2}\log d \cdot \log m) + m\log(1+\frac{t_0}{t})\\ &\leq O(m\log(1+\frac{t_0}{t}) + \log m).\end{aligned}$$

Finally, using equation 4, for the case that $t > t_0$, we have

$$\begin{aligned}\log N(B_1(\frac{1}{\lambda}), B_\infty(M), t) &= \log N(\frac{1}{\lambda}B_1, B_\infty(M), t)\\ &= \log N(\frac{1}{\lambda}B_1, B_\infty(M), t)\\ &= \log N(B_1, B_\infty(M), \lambda t)\\ &\leq \log N(B_1, H_m, \lambda t)\\ &\leq O(H)\frac{\log d \cdot \log m}{\lambda^2 t^2}.\end{aligned}$$

For the case $t < t_0$, we similarly obtain $\log N(B_1(\frac{1}{\lambda}), B_\infty(M), t) \leq O(m\log(1+\frac{t_0}{t\lambda}) + \log m)$. □

In the following lemma, we present two different upper bounds on the metric entropy of the intersection between the residual space $B_{w,2}(M)$ and the ℓ_1-penalty space $B_1(1/\lambda)$. Specifically, we employ the weighted ℓ_∞ unit ball to cover both the weighted ball $B_{w,2}(M)$ and the unweighted ball B_1 with the same radius. We then provide bounds for two cases: when the radius t is larger than t_0, and when t is smaller than t_0. Let $\tau = \sup_{x' \in \mathcal{L}_{w,M}} \|Mx\|^2_{w,2,\infty}$ be the maximum of ℓ_2 leverage score, and define $G = 1 + \mathcal{E} = 1 + \sup_{x' \in \mathcal{L}} \left|\|SA'x'\|_2^2 - \|A'x'\|_2^2\right|$.

Lemma 16. Let $\lambda > 0$, and let $\mathcal{L}_{w,M} = B_{w,2}(M) \cap B_1(1/\lambda)$. For any $t \in (0, 1]$, the metric entropy of $\mathcal{L}_{w,M}$ with respect to the metric d_X satisfies the following bounds: for $t < t_0$,

$$\log N(\mathcal{L}_{w,M}, d_X, t) \leq \min\{O(d\log\frac{Gm}{t}), O(m\log(1+\frac{Gt_0}{t\lambda}) + \log m)\},$$

and for $t > t_0$,

$$\log N(\mathcal{L}_{w,M}, d_X, t) \leq O(\frac{\tau G^2 \log m}{t^2} \cdot \min\{1, \frac{\log d}{\lambda^2}\}),$$

where $t_0 = \tau\sqrt{\frac{\log d}{m}}$.

Proof. For all $y, y' \in B_{w,2}(M)$, we have $d_X(y, y') \leq 2\|y - y'\|_{w,\infty}$. Next, we define the matrix $M_w \in \mathbb{R}^{m\times(d+1)}$ such that each rows of M_w is obtained by multiplying the corresponding entry of the weight vector w by the respective row of the matrix M, i.e., $(M_w)_i = \sqrt{w_i} \cdot M_i$. Since $w_i = 1/p_i$ represents the weight of the i-th row and p_i is the sampling probability, we have $w_i \geq 1$ for all i. Then, the convex body $B_{w,2}(M,)$ is contained within $B_2(M_w, G)$, since

$$\begin{aligned}B_{w,2}(M) &= \left\{y \in \text{range}(M) : \sum_{i=1}^m w_i y_i^2 \leq 1\right\}\\ &\subseteq \left\{y \in \text{range}(M) : \sum_{i=1}^m w_i^{1/2} y_i^2 \leq 1\right\} = B_2(M_w).\end{aligned}$$

Thus, for any $t > 0$, we have

$$\log N(B_{w,2}(M,G), d_X, t/G) \leq \log N(B_2(M_w), 2\|\cdot\|_{w,\infty}, t/G)$$
$$= \log N\left(B_2(M_w), B_\infty(M_w), t/2G\right).$$

By Lemma 4 and Lemma 14, the following inequality holds

$$\log N(B_{w,2}(M), d_X, t) \leq O(d \log \frac{Gm}{t}).$$

Furthermore, by a slight adaptation to Lemma 4, we also have $\log N(B_{w,2}(M), d_X, t) \leq O(\log m \frac{G^2 \tau}{t^2})$.

Let $H = \max_{1 \leq i \leq m} \|e_i^T M\|_\infty$ be the maximum row-wise ℓ_∞-norm of matrix M. By the inequality $\|x\|_\infty \leq \|x\|_2$ for any vector x, it follows that $H = \max_{1 \leq i \leq m} \|e_i^T M\|_\infty \leq \max_{1 \leq i \leq m} \|e_i^T M\|_2 \leq \tau$. Consequently, applying Lemma 15, we obtain the following bounds on the metric entropy $\log N(B_1(1/\lambda), B_\infty(M_w), t/2)) \leq O(\tau \frac{\log d \cdot \log m}{\lambda^2 t^2})$ for $t > t_0$, and the inequality $\log N(B_1(1/\lambda), B_\infty(M_w), t/2)) \leq O(m \log(1 + \frac{t_0}{t\lambda}) + \log m)$ for $t < t_0$, where $t_0 = O(\tau \sqrt{\frac{\log d}{m}})$.

Next, we consider the metric entropy on the $\mathcal{L}_{w,M}$

$$\log N(\mathcal{L}_{w,M}, d_X, t) \leq \log N(\mathcal{L}_{w,M}, 2\|\cdot\|_{w,\infty}, \frac{t}{G})$$
$$= \log N(B_{w,2}(M) \cap B_1(1/\lambda), 2\|\cdot\|_{w,\infty}, \frac{t}{G})$$
$$\leq \min\{N(B_{w,2}(M), 2\|\cdot\|_{w,\infty}, \frac{t}{2G}), N(B_1(1/\lambda), 2\|\cdot\|_{w,\infty}, \frac{t}{2G})\}.$$

Combining the above inequalities, we conclude

$$\log N(\mathcal{L}_{w,M}, d_X, t) \leq \min\{O(d \log \frac{Gm}{t}), O(m \log(1 + \frac{Gt_0}{t\lambda}) + \log m)\},$$

for $t < t_0$, and

$$\log N(\mathcal{L}_{w,M}, d_X, t) \leq O(\tau G^2) \min\{O(\frac{\log m}{t^2}), \frac{\log d \cdot \log m}{\lambda^2 t^2}\},$$

for $t > t_0$. □

B.3 Computing the Entropy Integral

In this subsection, we bound the integral metric entropy of these t-nets using the following Dudley inequality (Vershynin, 2018) for Gaussian processes.

Theorem 17. (Dudley inequality,(Vershynin, 2018)) Let $(X(t))_{t \in T}$ be a standard Gaussian process defined on a measurable space with a pseudo-metric d_X. Then, it holds that

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \leq C \int_0^\infty \sqrt{\log N(T, d_X, t)}\, dt,$$

where T is a convex set, C is an absolute constant, and X_t is the standard Gaussian vector at $t \in T$.

Lemma 18. (Woodruff & Yasuda, 2023) Let $0 < \delta \leq 1$ and C be a positive constant. Then,

$$\int_0^\delta \sqrt{\log \frac{C}{t}}\, dt \leq \delta \left(\sqrt{\log \frac{C}{\delta}} + \frac{C\sqrt{\pi}}{2}\right).$$

Lemma 19. Let $M \in \mathbb{R}^{m \times (d+1)}$ be orthonormal and λ be a positive parameter. Then, the metric entropy of $\mathcal{L}_{w,M}$ satisfies

$$\int_0^\infty \sqrt{\log N(\mathcal{L}_{w,M}, d_X, t)}\, dt \leq O(G\sqrt{\tau \cdot \log m} \log d \cdot \min\{1, \frac{\sqrt{\log d}}{\lambda}\}),$$

where τ is the maximum weighted ℓ_2-leverage score of M.

Proof. Note that it suffices to integrate the entropy integral from 0 to the diameter $\mathcal{D} = \mathrm{diam}(\mathcal{L}_{w,M})$, because for $t > \mathcal{D}$, the entropy is zero. Let $t_0 = \tau\sqrt{\frac{\log d}{m}}$, and let t' be a radii with $t' \in [t_0, \mathcal{D}]$. For small radii $t < t'$, we use the first bound of Lemma 16 as follows

$$\log N(\mathcal{L}_{w,M}, d_X, t) \le \min\{O(d \log \frac{Gm}{t}), O(m \log(1 + \frac{Gt_0}{t\lambda}) + \log m)\}.$$

By Lemma 18, the entropy integral is bounded by

$$\begin{aligned}\int_0^{t'} \sqrt{\log N(\mathcal{L}_{w,M}, d_X, t)}\, dt &\le \min\{\int_0^{t'} \sqrt{O(d \log \frac{Gm}{t})}\, dt, \int_0^{t'} \sqrt{O(m \log(1 + \frac{Gt_0}{t\lambda}) + \log m)\, dt}\} \\ &= \min\{O(\sqrt{d}) \int_0^{t'} \sqrt{\log \frac{Gm}{t}}\, dt, O(\sqrt{m}) \int_0^{t'} \sqrt{\log(1 + \frac{Gt_0}{\lambda t})}\, dt\} \\ &\le \min\{O(t' \cdot \sqrt{d \log \frac{Gm}{t'}}), O(\frac{Gt_0}{\lambda}\sqrt{m \log m} \cdot t')\} \\ &\le O(t') \min\{\sqrt{d \log \frac{Gm}{t'}}, \frac{Gt_0}{\lambda}\sqrt{m \log m}\} \\ &\le O(t') \min\{\sqrt{d \log \frac{Gm}{t'}}, \frac{G\tau}{\lambda}\frac{\log d}{m}\sqrt{m \log m}\} \\ &\le O(t') \min\{\sqrt{d \log \frac{Gm}{t'}}, \frac{G\tau \log d}{\lambda}\}.\end{aligned}$$

On the other hand, for large radii $t > t'$, we use the second bound of Lemma 16 (in Appendix), which gives

$$\log N(\mathcal{L}_{w,M}, d_X, t) \le O(\frac{\tau G^2 \log m}{t^2} \cdot \min\{1, \frac{\log d}{\lambda^2}\}).$$

Combining these inequalities, we obtain

$$\begin{aligned}\int_{t'}^{\mathcal{D}} \sqrt{\log N(\mathcal{L}_{w,M}, d_X, t)}\, dt &\le O(1)\sqrt{\tau G^2 \log m} \cdot \min\{1, \frac{\sqrt{\log d}}{\lambda}\} \int_{t'}^{\mathcal{D}} \frac{1}{t}\, dt \\ &= O(1)\sqrt{\tau G^2 \log m} \cdot \min\{1, \frac{\sqrt{\log d}}{\lambda}\} \log\left(\frac{G}{t'}\right).\end{aligned}$$

Applying Lemma 3 and choosing the radius $t' = G\sqrt{\tau/d}$, we get

$$\begin{aligned}\int_0^{\infty} \sqrt{\log N(\mathcal{L}_{w,M}, d_X, t)}\, dt &\le \int_0^{t'} \sqrt{\log N(\mathcal{L}_{w,M}, d_X, t)}\, dt \\ &\quad + \int_{t'}^{\mathcal{D}} \sqrt{\log N(\mathcal{L}_{w,M}, d_X, t)}\, dt \\ &\le O(G) \min\{\sqrt{\log m \log d}, \frac{G\sqrt{\tau}}{\lambda}\} \\ &\quad + O(1)\sqrt{\tau G^2 \log m} \cdot \min\{1, \frac{\sqrt{\log d}}{\lambda}\} \log d \\ &\le O(G\sqrt{\tau \cdot \log m} \log d \cdot \min\{1, \frac{\sqrt{\log d}}{\lambda}\}).\end{aligned}$$

□

Lemma 20. (Woodruff & Yasuda, 2023) Let $A' \in \mathbb{R}^{n \times (d+1)}$ and $\lambda > 0$. Let $\Lambda = \sup_{\|A'x\|_2^2 + \lambda\|x\|_1 \le 1, x \in \mathbb{R}^{d+1}} \left|\sum_{i=1}^n g_i |[A'x](i)|^2\right|$. Given a convex set L, let $\mathcal{M}_{\mathcal{E}}$ be the metric entropy of L, and let $\mathcal{D}$ be the Guassian width indexed by L. Then,

$$\mathbb{E}_{g \sim \mathcal{N}(0, I_n)}[|\Lambda|^l] \le (2\mathcal{M}_{\mathcal{E}})^l (\mathcal{M}_{\mathcal{E}}/\mathcal{D}) + O(\sqrt{l}\mathcal{D})^l.$$

Lemma 21. Let $A' \in \mathbb{R}^{n\times(d+1)}$. Let S be a sampling matrix such that, with probability at least $3/4$,

$$\|SA'x\|_2^2 = (1 \pm 1/2)\|A'x\|_2^2$$

simultaneously for every $x \in \mathbb{R}^{d+1}$. Then, with probability at least $1/2$,

$$\Pr\{\mathcal{G}(SA') \le 8\mathcal{G}(A')\} \ge \frac{1}{2}.$$

Proof. We have

$$\begin{aligned}
\mathcal{G}(SA') &= \sum_{i=1}^{n} \sup_{SA'x\neq 0} \frac{|(SA')_i x|^2 + \frac{\lambda}{n}\|x\|_1}{\|SA'x\|_2^2 + \lambda\|x\|_1} \\
&\le \sum_{i=1}^{n} \sup_{SA'x\neq 0} \frac{S_{ii}^2\left(|A'_i x|^2 + \frac{\lambda}{n}\|x\|_1\right)}{\|SA'x\|_2^2 + \lambda\|x\|_1} \\
&= \sum_{i=1}^{n} \sup_{SA'x\neq 0} \frac{S_{ii}^2\left(|A'_i x|^2 + \frac{\lambda}{n}\|x\|_1\right)}{\|A'x\|_2^2 + \lambda\|x\|_1} \frac{\|A'x\|_2^2 + \lambda\|x\|_1}{\|SA'x\|_2^2 + \lambda\|x\|_1} \\
&= \sum_{i=1}^{n} \sup_{SA'x\neq 0} S_{ii}^2 \varrho_i(A') \sup_{SA'x\neq 0} \frac{\|A'x\|_2^2 + \lambda\|x\|_1}{\|SA'x\|_2^2 + \lambda\|x\|_1}.
\end{aligned}$$

We are guaranteed that

$$\Pr\{\sup_{SA'x\neq 0} \frac{\|A'x\|_2^2 + \lambda\|x\|_1}{\|SA'x\|_2^2 + \lambda\|x\|_1} \le 2\} \ge \frac{3}{4}.$$

On the other hand, we have that

$$\mathbb{E}\sum_{i=1}^{n} S_{i,i}^2 \varrho_i(A') = \sum_{i=1}^{n} \mathbb{E}[S_{i,i}^2]\varrho_i(A') = \mathcal{G}(A').$$

By Markov's inequality,

$$\Pr\{\sum_{i=1}^{n} S_{ii}^2 \varrho_i(A') \le 4\mathcal{G}(A')\} \ge \frac{3}{4}.$$

Combining the above inequalities, we conclude

$$\Pr\{\mathcal{G}(SA') \le 8\mathcal{G}(A')\} \ge \frac{1}{2}.$$

□

In the following theorem, we present the main result provides a bound on the l-th moment of the sampling error $\mathbb{E}|\mathcal{E}|^l$.

Theorem 22. Let $A' \in \mathbb{R}^{n\times(d+1)}$ be an input matrix, S be a random sampling matrix, and let $\varepsilon, \delta \in (0,1)$ and $\lambda > 0$ parameters. If $\alpha = \tilde{O}\left(\frac{1}{\epsilon^2}\cdot\left(\log\left(d\log(\delta^{-1})\right)(\ln d)^2 \cdot \min\left\{1, \frac{\log d}{\lambda^2}\right\} + \ln\delta^{-1}\right)\right)$ and for all $i \in [n]$ it holds that

$$p_i \ge \min\{1, \alpha(\tau_{i,2}(A') + \frac{1}{n})\},$$

where $\tau_{i,2}(A')$ denotes the ℓ_2 leverage score of the i-th row of A'. Then, with failure probability at most δ, it holds that, $\forall x \in \mathbb{R}^{d+1}, x_{d+1} = 1$,

$$\|SA'\hat{x}\|_2^2 + \lambda\|\hat{x}\|_1 \le (1\pm\epsilon)(\|A'x\|_2^2 + \lambda\|x\|_1),$$

and the coreset size is at most

$$m = \tilde{O}\left(\frac{d(\log d)^3}{\epsilon^2}\cdot\min\{1, \frac{\log d}{\lambda^2}\} + \frac{d}{\epsilon^2}\log\frac{1}{\delta}\right).$$

Proof. By the construction of the sampling matrix S, for any $i \in [n]$, the sampling probability satisfies $0 \leq p_i \leq 1$, and the corresponding sampling weight is $S_{ii} = 1/p_i \geq 1$. This implies that $\mathbb{E}(\|SAx\| + \lambda\|x\|_1) = \|A'x\| + \lambda\|x\|_1$ for $\lambda > 0$ and any vector x. We set $\alpha = O(\frac{\sqrt{l}}{\epsilon} \cdot \min\{\log(d(\lambda^3 \wedge 1)), \lambda^2 d\})$, where $\sqrt{l} = O\left(\frac{(\log d/\delta)^2 \cdot \min\{1, \frac{\log d}{\lambda^2}\} + \epsilon \cdot \log(1/\delta)}{\epsilon \cdot \min\{\log(d(\lambda^2 \wedge 1)), \lambda^2 d\}}\right)$ denotes the maximum number of finite moments of the sampling error $\mathcal{E}$. To bound the coreset size m, let X_i be the indicator random variable that represents whether the i-th row is included in S. Applying Lemma 1 in Appendix, we get

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} p_i = \alpha\left(1 + \sum_{i=1}^{n}(2\tau_i + \frac{1}{n})\right) = \alpha\,(2 + 2d) \leq 4\alpha d.$$

Similarly, we can derive the lower bound of m as follows

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} p_i \geq \alpha\left(\sum_{i=1}^{n} 2\tau_i\right) = 2\alpha d.$$

By applying the Chernoff inequality, we have

$$m = \sum_{i=1}^{n} X_i \leq 2 \cdot \mathbb{E}\left(\sum_{i=1}^{n} X_i\right) \leq 8\alpha d$$

with failure probability at most $2\exp\left(-\mathbb{E}(\sum_{i=1}^{n} X_i)/3\right) \leq 2\exp(-\frac{2\alpha d}{3}) \leq \delta$.

Applying Lemma 1, the analysis of the empirical process associated with the sampling error can be reduced to a Gaussian process. Specifically, we obtain

$$\begin{aligned}&\mathbb{E}_S \sup_{\|A'x\|_2^2+\lambda\|x\|_1=1, x\in\mathcal{T}'} |\|SA'x\|_2^2 - \|A'x\|_2^2|^l \\ &\leq (2\pi)^{l/2}\mathbb{E}_{S,g} \sup_{\|A'x\|_2^2+\lambda\|x\|_1=1, x\in\mathcal{T}'} \left|\sum_{i\in Q} g_i w_i |A_i' x|^2\right|^l,\end{aligned}$$

where $l > 1$ is an integer, w_i denotes the weight for sampling the i-row, Q the indices of non-zero diagonal entries in S, and $g \sim \mathcal{N}(0, I_m)$ is a standard Gaussian vector.

Next, we define $\Lambda = \sup_{\|A'x\|_2^2+\lambda\|x\|_1=1, x\in\mathcal{T}'} \left(\sum_{i\in S} g_i w_i |A_i'x|^2\right)$ for the random sampling matrix S. To further bound the quantity Λ, we utilize Lemma 20 to relate Λ to the metric entropy $\mathcal{M}_{\mathcal{E}}$ and the diameter $\mathcal{D}$ of geometric body resulted by the Gaussian process. This gives us the following bound

$$\mathbb{E}_{g\sim\mathcal{N}(0,I_m)}[|\Lambda|^l] \leq (2\mathcal{M}_{\mathcal{E}})^l(\mathcal{M}_{\mathcal{E}}/\mathcal{D}) + O(\sqrt{l}\mathcal{D})^l$$

for a fixed l.

Let $M = SA'$ denote an m-row submatrix of A', and let w represent the weight vector corresponding to each row of M. Next, we bound the maximum weight leverage score $\tau = \sup_{\|Mx\|_{w,2}=1, i\in[m]} w_i|M_i x|^2$.

We set the number of samples m at least $\tilde{O}(d + \log(1/\delta))$ using ℓ_2 leverage scores sampling method (Cohen et al., 2015), which achieves $\|S'A'x\| \leq (1 \pm \frac{1}{2})\|A'x\|_2^2$ for fixed sampling matrix S' with probability at least $1 - \delta$. By applying above inequality and the definition of sampling probability, we have $\tau \leq 8/\alpha$.

According to Lemma 6, by choosing the constants for α sufficiently large, we obtain a bound on the metric entropy

$$\begin{aligned}&O(G\tau^{1/2}(\log m)^{1/2}\log d \cdot \min\{1, \frac{\sqrt{\log d}}{\lambda}\}) \\ &\leq O(G\alpha^{-1/2}(\log m)^{1/2}\log d \cdot \min\{1, \frac{\sqrt{\log d}}{\lambda}\}) \\ &\leq G\epsilon/8 := \mathcal{M}_{\mathcal{E}}.\end{aligned}$$

By Lemma 3, we derive a bound on the diameter $O(\tau \cdot \sqrt{\log{(d(\lambda^2 \wedge 1))}} \wedge (\lambda\sqrt{d})) \leq O((1/\alpha) \cdot \sqrt{\log{(d(\lambda^2 \wedge 1))}} \wedge (\lambda\sqrt{d})) \leq \frac{\epsilon}{2\sqrt{l}} := \mathcal{D}$.

By combining the bounds on the metric entropy $\mathcal{M}_{\mathcal{E}}$ and the diameter $\mathcal{D}$, we ensure the sampling error $\mathbb{E}_{g\sim\mathcal{N}(0,I_m)}[|\Lambda|^l] \leq \epsilon^l\delta$. Since the sampling error $\mathcal{E} = \sup_{x'\in\mathcal{L}_M} \left|\|SA'x'\|_2^2 - \|A'x'\|_2^2\right|$, we have $\mathcal{E}^l \leq 3^l\varepsilon^l\delta$, which yields $\mathbb{E}|\mathcal{E}|^l \leq (3\varepsilon)^l\delta$. By using Markov inequality, we have $\mathcal{E} \leq 3\varepsilon$ with probability at least $1-\delta$. □

B.4 Omitted Proofs of Lower Bound for Coreset Size

In this section, we provide the lower bound of the coreset size for LASSO regression, using a standard information-theoretic approach (Wang et al., 2010; Wainwright, 2009; Parulekar et al., 2021; Mai et al., 2023) based on Fano's inequality and KL divergence computations. Here, we start by constructing the hard instance for the k-sparse supports.

Let $\mathcal{C} \subset \{0,1\}^d$ be a set of k-sparse binary vectors (i.e., each vector has exactly k non-zeros entries), such that $|\mathcal{C}| = N \geq (d/k)^k$ and for any two distinct vectors $c^{(i)}, c^{(j)} \in \mathcal{C}$ satisfy $|supp(c^{(i)}) \cap supp(c^{(j)})| \leq Ck$ for some constant $C \in (0,1)$. Such a codebook can be constructed using standard techniques from coding theory. For each codeword $c^{(i)} \in \mathcal{C}$, we define

$$v^{(i)} = \left[1, \frac{\epsilon}{\sqrt{k}}c^{(i)}\right] \in \mathbb{R}^{d+1}.$$

Let $G \in \mathbb{R}^{m\times d}$ be a matrix with i.i.d standard Gaussian entries. Define the data matrix $Z^i = G(I + v^{(i)}v^{(i)\top})^{1/2}$. Then, each row $z_j^i \sim \mathcal{N}(0, I + v^{(i)}v^{(i)\top})$, and the data distribution is

$$\mathbb{P}_i = \mathcal{N}(0, I + v^{(i)}v^{(i)\top}).$$

We show that exact support recovery is impossible with fewer measurements than those suggested by the information-theoretic lower bound, given the input distribution.

Lemma 23. Let $\epsilon \in (0,1)$, and let $v \in \mathbb{R}^{d+1}$ be the vector with $v = (1, \frac{1}{\sqrt{k}}c)$, where c is a codeword uniformly chosen from $\mathcal{C}$. Let P_i be the multivariate Gaussian distribution with covariance $I + vv^\top$. Then, for any estimator attempting to recover a k-sparse vector c, with at least $1/2$ probability, the number of samples m must satisfy

$$m \geq \Omega(\frac{k\log(d/k)}{\epsilon^2}).$$

Proof. Let $\mathbb{P}_i, \ldots, \mathbb{P}_N$ be the distributions constructed above. By the Fano's inequality, we have

$$\Pr[error] \geq 1 - \frac{\frac{1}{N^2}\sum_{i\neq j} D_{\mathrm{KL}}(\mathbb{P}_i||\mathbb{P}_j) + \log 2}{\log(N-1)},$$

where $D_{\mathrm{KL}}(\mathbb{P}_i||\mathbb{P}_j)$ denotes the Kullback-Leibler divergence between distributions $\mathbb{P}_i$ and $\mathbb{P}_j$.

To ensure the error probability is less than $1/2$, it suffices to ensure

$$\frac{1}{N^2}\sum_{i\neq j} D_{\mathrm{KL}}(\mathbb{P}_i||\mathbb{P}_j) \leq \frac{1}{4}\log N.$$

According the definition of $\mathbb{P}_i$, the KL divergence between two such distributions is

$$D_{\mathrm{KL}}(\mathbb{P}_i||\mathbb{P}_j) = m \cdot D_{\mathrm{KL}}(\mathcal{N}(0,\Sigma_i)||\mathcal{N}(0,\Sigma_j)),$$

where $\Sigma_i = I + v^{(i)}v^{(i)\top}$. Using the formula for KL divergence between zero-mean Gaussian distribution, we have

$$D_{\mathrm{KL}}(\mathcal{N}(0,\Sigma_i)||\mathcal{N}(0,\Sigma_j)) = \frac{1}{2}(tr(\Sigma_j^{-1}\Sigma_i) - d + \log\frac{\det\Sigma_j}{\det\Sigma_i})$$

Since Σ_i is a rank-1 perturbation, by the Sherman-Morrison Formula, we apply $\det(\Sigma_i) = 1 + \|v^{(i)}\|_2^2$ and $\Sigma_i^{-1} = I - \frac{v^{(i)}v^{(i)\top}}{1+\|v^{(i)}\|_2^2}$. Thus, we obtain

$$D_{\mathrm{KL}}(\mathbb{P}_i||\mathbb{P}_j) = \frac{n}{2}(\|v^{(i)}\|_2^2 - \frac{\|v^{(j)}\|_2^2 + (v^{(j)\top}v^{(i)})^2}{1+\|v^{(i)}\|_2^2} + \log\frac{1+\|v^{(j)}\|_2^2}{1+\|v^{(i)}\|_2^2})$$

For any $i \in [N]$, it holds that $\|v^{(i)}\|_2^2 = 1 + \epsilon^2$. Similarity, for any $i \neq j$, the inner product satisfies

$$(v^{(j)\top}v^{(i)})^2 = \left(1 + \frac{\epsilon^2}{k}\langle c^{(i)}, c^{(j)}\rangle\right)^2 \leq (1 + C\epsilon^2)^2.$$

Plugging into the expression for KL divergence, we get

$$D_{\mathrm{KL}}(\mathbb{P}_i||\mathbb{P}_j) \leq O(\epsilon^2 m).$$

Let $\log N = \Theta(k\log(d/k))$. Applying the Fano's inequality, it holds that

$$\Pr[error] \geq 1 - \frac{mC\epsilon^2 + \log 2}{\log N}.$$

To ensure $\Pr[error] \leq 1/2$, we have

$$C\epsilon^2 \cdot m \geq \frac{1}{2}\log N = \Omega(k\log(d/k)) \to m \geq \Omega(\frac{k\log(d/k)}{\epsilon^2}).$$

□

Our proof method, while differing in approach from previous work (Wainwright, 2009; Mai et al., 2023) that focuses on sketching algorithms, is based on similar ideas. In particular, by analyzing the coreset algorithm on a constructed hard instance, we establish a lower bound on the sample size required by any algorithm to achieve a $(1+\epsilon)$-approximation on the constructed hard instance.

Lemma 24. Let $A \in \mathbb{R}^{n\times d}$, $b \in \mathbb{R}^n$, and $\lambda \in (0, 1)$. Assume that $\|A\|_2 \leq 1$ and $\|b\|_2 \leq 1$. Let S be a diagonal sampling matrix with m non-zero entries. Suppose there exists an estimator that returns $\tilde{x} = \arg\min_{x\in\mathbb{R}^d} \|SAx - Sb\|_2^2 + \lambda\|x\|_1$ satisfies

$$\|A\tilde{x} - b\|_2^2 + \lambda\|\tilde{x}\|_1 \leq (1+\epsilon) \cdot \min_{x\in\mathbb{R}^d}(\|Ax - b\|_2^2 + \lambda\|x\|_1).$$

Then, the coreset size m must satisfy

$$m = \begin{cases} \Omega(\frac{\log d}{\lambda^2\cdot\epsilon^2}), & \text{if } \lambda = \Omega(\frac{1}{\sqrt{d}}) \\ \Omega(\frac{d}{\epsilon^2}\log d), & \text{if } \lambda = O(\frac{1}{\sqrt{d}}) \end{cases}.$$

Proof. We prove this result using a similar approach to that in Theorem 13 (Mai et al., 2023). We will take $[b\ A] \sim \frac{1}{\sqrt{n}}G(I + vv^\top)^{1/2}$, where v is a codeword in set $\mathcal{C}$. Let S be a sampling matrix that selects m rows of A and b. Since only the row indices selected by S affect the coreset, and the weights can be absorbed into the analysis via rescaling, we may, without loss of generality, assume that all non-zero diagonal entries of S are equal to 1. Under this assumption, the compressed matrix $SG(I + vv^\top)^{1/2}$ has the same distribution as $G(I + vv^\top)^{1/2}$.

By the concentration properties of Gaussian matrices (see Exercise 4.7.3 in (Vershynin, 2018)), with high probability, the LASSO objective satisfies

$$\|Ax - b\|_2^2 + \lambda\|x\|_1 \approx 1 + \|x\|_2^2 + (1 - \epsilon c^T x)^2 + \lambda\|x\|_1 =: L(x),$$

where $v = (1\ c)$. Since $L(x)$ is a 1-strongly convex function, we get

$$L(\hat{x}) \geq L(x^*) + \|\hat{x} - x^*\|_2^2.$$

for any $\hat{x}$, where x^* is the minimizer of $L(x)$.

Fix $\epsilon = 1/2$, we set $\lambda = \frac{1}{2\sqrt{k}}$. Here, it holds that $x^* = c/5$ and $L(x^*) \approx 2$. Suppose there exist a estimator algorithm satisfies $L(\hat{x}) \leq (1 + c_1)L(x^*)$ for a sufficiently small c_1. Then we have

$$(1 + c_1)L(x^*) \geq L(x^*) + \|\hat{x} - x^*\|_2^2,$$

which means the gap $\|\hat{x} - x^*\|_2^2 \leq 2 \cdot c_1$.

Choosing a small enough constant c_1, we can recover $\mathrm{supp}(v)$. By Lemma 23 (in Appendix), if $\frac{1}{4\lambda^2} = o(d)$, the required lower bound of $\Omega(\frac{1}{\lambda^2\epsilon^2}\log d)$ on the coreset size; if $\frac{1}{4\lambda^2} = \Theta(d)$, the required size is at least $\Omega(\frac{d}{\epsilon^2}\log d)$. □

C Complementary Experiments

C.1 Experiments on Sketching Algorithm and LASSO-Sens

In this section, we present experimental results comparing the performance of our proposed LASSO-Sens algorithm with a sketching-based algorithm for solving LASSO regression. We also acknowledge recent advances in sketching for LASSO, such as the work by Mai et al. (2023), which utilizes random projections to accelerate the optimization process.

To ensure a fair comparison, we follow the same experimental setup used in Section 4, conducting experiments on four datasets with identical coreset sizes and regularization parameters. We evaluate algorithm performance in terms of loss, runtime, and sparsity. For the sketching-based algorithm, we set the number of sketching rows equal to the coreset size and run each experiment 10 times, reporting the average results.

As shown in Tables 6,7, 8, and 9, the proposed LASSO-Sens algorithm consistently achieves lower loss values than the sketching method on both small- and large-scale datasets. On the large-scale mnist8m dataset, LASSO-Sens is up to 10 times faster than the sketching algorithm when the coreset size is set to $m = \{15, 20\} \times d$. Moreover, on the Synthetic and mnist8m datasets, the sparsity of the LASSO-Sens solution is highly lower than that of the sketching algorithm. Overall, the experimental results show that the proposed algorithm achieves lower regression loss and sparsity, particularly on large-scale dataset.

C.2 Experiments on Sensitivity Sampling for Standard and Modified LASSO Objectives

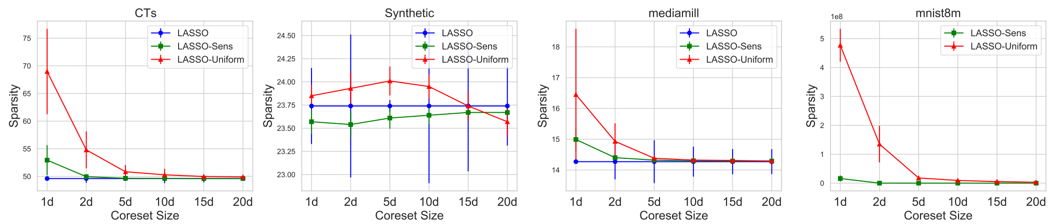
In this section, we compare the performance of the sensitivity sampling algorithm on both the standard LASSO objective and the modified LASSO objective proposed in Chhaya et al. (2020), which takes the form $\|Ax - b\|_2^2 + \lambda\|x\|_1^2$. In Section 4, we used the FISTA algorithm to solve the standard LASSO problem, as it leverages the proximal operator of the ℓ_1 norm. However, this solver is not applicable to the modified LASSO formulation, which involves a squared ℓ_1 regularization term and lacks an efficient proximal operator. As a result, directly comparing the two objectives under our original framework would be unfair.

To ensure a fair comparison, we follow the methodology of Chhaya et al. (2020), which utilizes the global optimization toolbox from MATLAB. Specifically, we use the `patternsearch` solver to address both standard and modified LASSO problems. In our experiments, the solver parameters are set as follows: `MaxFunctionEvaluations` $= 1,000,000$, and `MaxIterations` $= 25,000$. To quantify the approximation quality of the coreset solution, we utilize the LASSO objective function as the evaluation metric. The experiments are conducted on a machine equipped with an Intel(R) Core(TM) i7-9700 CPU and 16 GB of RAM, and the implementation is executed MATLAB R2021.

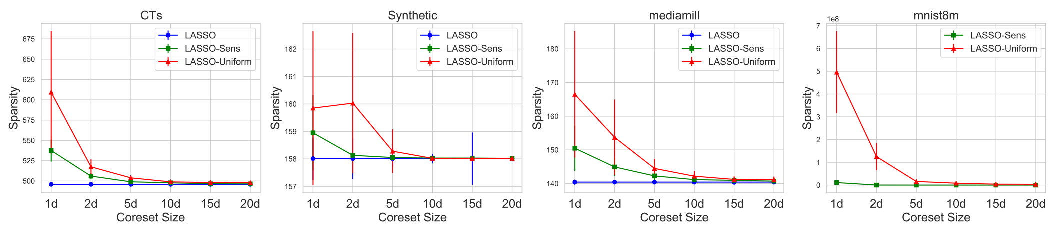
We first use Algorithm 1 to construct the coreset, and then apply the `patternsearch` solver to solve both objective functions on the coreset samples. The experiments are conducted on synthetic datasets using the same coreset sizes and regularization parameters λ as in Section 4. Each experiment is repeated 10 times, and we report the average results. As shown in Table 10, the sensitivity sampling algorithm for standard LASSO achieve lower sparsity compare to modified objective. Meanwhile, the computational time required by the two objectives is comparable.

D Use of Large Language Models (LLMs)

No large language models were used in the ideation or writing of this paper.



(a) $\lambda = 1$

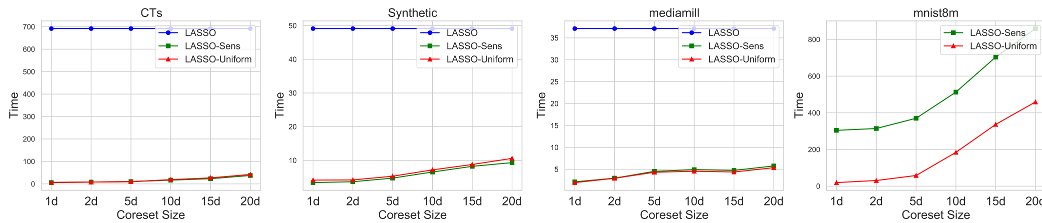


(b) $\lambda = 5$

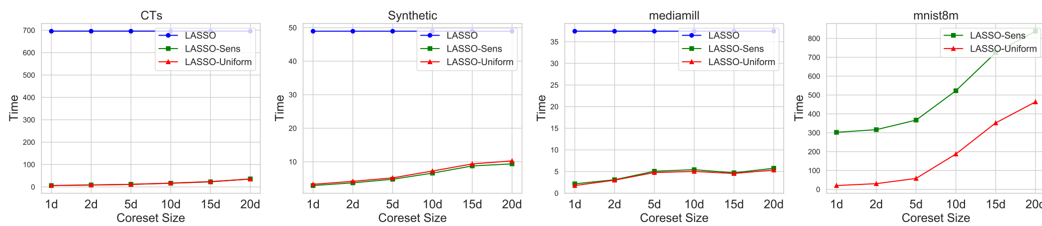
Figure 2: Comparison results of LASSO regression loss across varying coreset sizes

Table 2: Comparison results of loss, runtime, and sparsity on CTs dataset (n = 53,500, d = 386) for varying coreset sizes at $\lambda = \{1, 5, 10\}$.

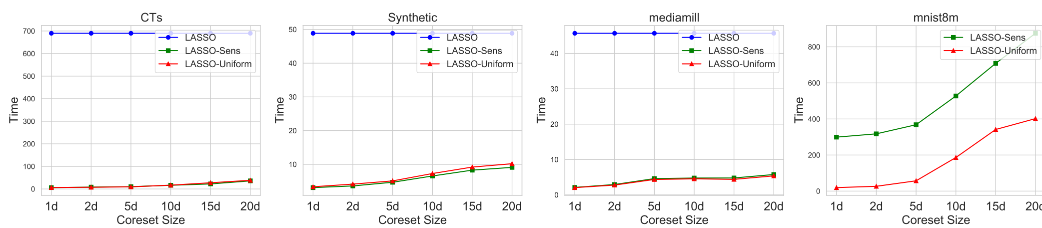
Lambda	Metrics	Algorithms	Coreset Sizes					
			1d	2d	5d	10d	15d	20d
$\lambda = 1$	Loss	LASSO			**49.61±0.06**			
		LASSO-Sens	52.94±2.73	49.95±0.31	49.66±0.02	49.64±0.01	49.63±0.01	49.63±0.01
		LASSO-Uniform	68.98±7.72	54.80±3.34	50.84±1.21	50.29±1.09	49.98±0.29	49.90±0.29
	Time (s)	LASSO			695.79			
		LASSO-Sens	6.08	8.76	11.59	16.40	22.58	35.79
		LASSO-Uniform	6.41	8.29	11.24	17.07	23.73	34.98
	Sparsity	LASSO			229			
		LASSO-Sens	325	243	226	221	226	227
		LASSO-Uniform	320	251	223	219	221	211
$\lambda = 5$	Loss	LASSO			**247.94±0.68**			
		LASSO-Sens	267.49±10.93	251.30±2.29	248.69±0.49	248.34±0.29	248.19±0.19	248.04±0.19
		LASSO-Uniform	307.27±38.10	258.08±4.63	251.28±2.26	249.22±1.15	248.68±0.71	248.83±0.54
	Time (s)	LASSO			689.77			
		LASSO-Sens	6.02	8.22	10.24	16.38	22.40	35.91
		LASSO-Uniform	6.45	7.67	9.45	16.98	27.48	38.40
	Sparsity	LASSO			166			
		LASSO-Sens	185	167	158	160	162	160
		LASSO-Uniform	192	179	163	155	158	161
$\lambda = 10$	Loss	LASSO			**495.91±0.77**			
		LASSO-Sens	537.53±13.62	506.01±3.84	498.99±1.92	497.75±1.46	496.58±0.88	496.45±0.64
		LASSO-Uniform	609.27±75.35	517.43±9.40	503.88±3.16	498.92±2.36	498.10±1.78	497.87±1.77
	Time (s)	LASSO			693.29			
		LASSO-Sens	5.95	9.18	10.38	16.71	22.41	35.96
		LASSO-Uniform	5.92	8.31	10.23	18.53	28.86	42.11
	Sparsity	LASSO			162			
		LASSO-Sens	172	160	154	159	155	153
		LASSO-Uniform	181	165	157	158	156	155



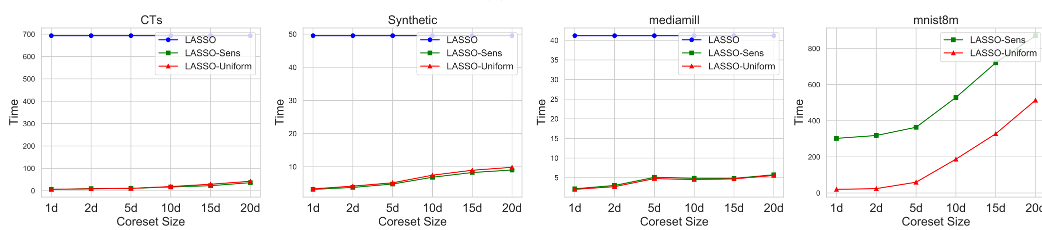
(a) $\lambda = 0.5$



(b) $\lambda = 1$



(c) $\lambda = 5$



(d) $\lambda = 10$

Figure 3: Comparison results of running time across varying coreset sizes for different λ values

Table 3: Comparison results of loss, runtime, and sparsity on Synthetic dataset ($n = 10,000, d = 200$) for varying coreset sizes at $\lambda = \{0.5, 1, 5, 10\}$.

Lambda	Metrics	Algorithm	Coreset Sizes 1d	2d	5d	10d	15d	20d
$\lambda = 0.5$	Loss	LASSO	**13.82±0.78**					
		LASSO-Sens	14.46±0.54	14.50±0.49	14.26±0.32	14.23±0.26	14.05±0.25	13.99±0.23
		LASSO-Uniform	16.26±0.08	16.20±0.20	16.00±0.25	15.82±0.34	15.34±0.42	15.22±0.53
	Time (s)	LASSO	49.11					
		LASSO-Sens	3.40	3.65	4.73	6.54	8.22	9.32
		LASSO-Uniform	4.14	4.19	5.29	7.17	8.77	10.59
	Sparsity	LASSO	41					
		LASSO-Sens	41	34	35	36	38	39
		LASSO-Uniform	28	28	28	30	29	31
$\lambda = 1$	Loss	LASSO	23.74±0.60					
		LASSO-Sens	**23.57±0.13**	**23.54±0.14**	**23.61±0.11**	**23.64±0.10**	**23.67±0.07**	23.67±0.06
		LASSO-Uniform	23.85±0.12	23.93±0.18	24.01±0.16	23.95±0.13	23.74±0.15	**23.57±0.16**
	Time (s)	LASSO	48.94					
		LASSO-Sens	2.91	3.72	4.79	6.62	8.76	9.38
		LASSO-Uniform	3.28	4.17	5.20	7.25	9.36	10.27
	Sparsity	LASSO	28					
		LASSO-Sens	32	29	29	28	28	28
		LASSO-Uniform	28	28	28	30	30	32
$\lambda = 5$	Loss	LASSO	**83.42±0.34**					
		LASSO-Sens	83.63±0.23	83.46±0.03	83.43±0.01	**83.42±0.00**	**83.42±0.00**	**83.42±0.00**
		LASSO-Uniform	84.71±1.56	85.25±1.87	84.35±0.95	83.55±0.48	83.56±0.22	**83.42±0.01**
	Time (s)	LASSO	48.86					
		LASSO-Sens	3.10	3.60	4.69	6.55	8.28	9.10
		LASSO-Uniform	3.35	4.14	5.09	7.28	9.19	10.19
	Sparsity	LASSO	28					
		LASSO-Sens	32	28	28	28	28	28
		LASSO-Uniform	28	28	28	28	28	28
$\lambda = 10$	Loss	LASSO	**158.01±0.72**					
		LASSO-Sens	158.95±1.37	158.13±0.06	158.05±0.03	158.03±0.01	158.03±0.01	158.02±0.01
		LASSO-Uniform	159.85±2.80	160.03±2.55	158.28±0.80	158.02±0.01	**158.01±0.01**	158.02±0.01
	Time (s)	LASSO	49.54					
		LASSO-Sens	3.17	3.76	4.79	6.83	8.26	9.01
		LASSO-Uniform	3.32	4.15	5.16	7.45	8.94	9.83
	Sparsity	LASSO	28					
		LASSO-Sens	28	27	27	28	28	28
		LASSO-Uniform	27	28	28	28	28	28

Table 4: Comparison results of loss, runtime, and sparsity on mediamill dataset ($n = 30,993, d = 120$) for varying coreset sizes at $\lambda = \{0.5, 1, 5, 10\}$.

Lambda	Metrics	Algorithms	Coreset Sizes					
			1d	2d	5d	10d	15d	20d
$\lambda = 0.5$	Loss	LASSO	**7.13±0.69**					
		LASSO-Sens	7.34±0.13	7.19±0.02	7.15±0.00	7.15±0.00	7.14±0.01	7.14±0.00
		LASSO-Uniform	7.94±0.69	7.25±0.07	7.17±0.02	7.16±0.01	7.15±0.01	7.15±0.01
	Time (s)	LASSO	37.12					
		LASSO-Sens	2.13	2.97	4.53	4.91	4.74	5.76
		LASSO-Uniform	1.96	2.94	4.32	4.55	4.38	5.36
	Sparsity	LASSO	47					
		LASSO-Sens	44	44	44	45	45	46
		LASSO-Uniform	39	42	42	45	45	44
$\lambda = 1$	Loss	LASSO	**14.27±0.53**					
		LASSO-Sens	14.99±0.39	14.40±0.08	14.32±0.03	14.30±0.02	14.29±0.01	14.29±0.01
		LASSO-Uniform	16.45±2.14	14.93±0.59	14.38±0.08	14.32±0.05	14.31±0.03	14.29±0.02
	Time (s)	LASSO	37.43					
		LASSO-Sens	2.16	3.10	5.07	5.43	4.71	5.76
		LASSO-Uniform	1.76	3.01	4.77	5.03	4.55	5.36
	Sparsity	LASSO	49					
		LASSO-Sens	40	43	44	45	44	45
		LASSO-Uniform	41	41	44	45	43	45
$\lambda = 5$	Loss	LASSO	**70.97±0.61**					
		LASSO-Sens	78.09±3.52	73.50±1.15	71.68±0.50	71.37±0.35	71.26±0.20	71.07±0.15
		LASSO-Uniform	87.79±12.04	78.56±5.31	73.24±1.89	71.98±1.06	71.51±0.61	71.55±0.58
	Time (s)	LASSO	45.70					
		LASSO-Sens	2.11	2.93	4.60	4.76	4.78	5.75
		LASSO-Uniform	2.05	2.77	4.36	4.54	4.39	5.37
	Sparsity	LASSO	45					
		LASSO-Sens	32	39	39	39	41	40
		LASSO-Uniform	31	35	38	39	38	39
$\lambda = 10$	Loss	LASSO	**140.43±0.46**					
		LASSO-Sens	150.49±6.70	144.93±2.66	142.27±1.36	141.17±0.63	140.95±0.52	140.72±0.44
		LASSO-Uniform	166.49±18.80	153.72±11.25	144.49±2.85	142.19±1.51	141.23±0.90	141.12±0.95
	Time (s)	LASSO	41.19					
		LASSO-Sens	2.16	3.01	5.08	4.85	4.81	5.74
		LASSO-Uniform	2.01	2.70	4.76	4.52	4.69	5.56
	Sparsity	LASSO	40					
		LASSO-Sens	31	32	34	34	36	38
		LASSO-Uniform	28	30	33	33	35	36

Table 5: Comparison results of loss, runtime, and sparsity on mnist8m datasets ($n = 8,000,000, d = 784$) for varying coreset sizes at $\lambda = \{0.5, 1, 5, 10\}$. If an algorithm fails to output a solution within 48 hours, the metrics are marked as $48 > h$.

Lambda	Metrics	Algorithms	Coreset Sizes					
			1d	2d	5d	10d	15d	20d
$\lambda = 0.5$	Loss	LASSO	$48 > h$					
		LASSO-Sens	**1.27E7 ± 1.02E7**	**3.30E4 ± 7.71E3**	**1.64E4 ± 4.03E2**	**1.45E4 ± 3.06E2**	**1.43E4 ± 1.80E2**	**1.37E4 ± 1.16E2**
		LASSO-Uniform	$5.83E8 \pm 2.73E8$	1.79E8 ± 4.25E7	3.59E7 ± 3.63E7	6.53E6 ± 2.90E6	4.46E6 ± 2.85E6	3.00E6 ± 1.09E6
	Time (s)	LASSO	$48 > h$					
		LASSO-Sens	304.32	314.23	370.11	512.73	703.91	859.22
		LASSO-Uniform	19.39	30.61	58.05	184.22	336.58	459.11
	Sparsity	LASSO	$48 > h$					
		LASSO-Sens	780	776	770	770	763	760
		LASSO-Uniform	704	712	725	728	735	735
$\lambda = 1$	Loss	LASSO	$48 > h$					
		LASSO-Sens	**1.61E7 ± 1.04E7**	**3.46E4 ± 8.18E3**	**1.68E4 ± 5.27E2**	**1.48E4 ± 2.05E2**	**1.41E4 ± 1.38E2**	**1.38E4 ± 7.54E1**
		LASSO-Uniform	4.77E8 ± 5.68E7	1.35E8 ± 6.33E7	1.80E7 ± 3.20E6	9.35E6 ± 4.36E6	5.37E6 ± 2.21E6	3.04E6 ± 1.13E6
	Time (s)	LASSO	$48 > h$					
		LASSO-Sens	302.37	316.26	366.99	522.60	724.51	838.05
		LASSO-Uniform	20.25	30.3	58.01	187.43	352.15	463.57
	Sparsity	LASSO	$48 > h$					
		LASSO-Sens	778	774	768	763	765	758
		LASSO-Uniform	705	709	731	737	737	729
$\lambda = 5$	Loss	LASSO	$48 > h$					
		LASSO-Sens	**1.39E7 ± 4.50E6**	**2.87E4 ± 4.21E3**	**1.72E4 ± 4.57E2**	**1.49E4 ± 2.52E2**	**1.45E4 ± 1.88E2**	**1.42E4 ± 1.49E2**
		LASSO-Uniform	5.98E8 ± 7.60E7	1.23E8 ± 3.26E7	2.15E7 ± 7.61E6	9.22E6 ± 2.22E6	3.99E6 ± 1.09E6	3.05E6 ± 9.08E5
	Time (s)	LASSO	$48 > h$					
		LASSO-Sens	299.24	317.45	368.02	527.51	708.16	875.08
		LASSO-Uniform	19.32	26.66	57.07	186.32	341.58	401.96
	Sparsity	LASSO	$48 > h$					
		LASSO-Sens	780	776	770	767	765	763
		LASSO-Uniform	691	714	729	735	730	741
$\lambda = 10$	Loss	LASSO	$48 > h$					
		LASSO-Sens	**1.10E7 ± 8.22E6**	**9.47E4 ± 1.27E5**	**1.73E4 ± 1.06E2**	**1.55E4 ± 1.54E2**	**1.49E4 ± 3.28E1**	**1.47E4 ± 8.80E1**
		LASSO-Uniform	4.96E8 ± 1.81E8	1.25E8 ± 5.99E7	1.58E7 ± 5.71E6	8.34E6 ± 1.86E6	3.99E6 ± 9.91E5	3.50E6 ± 6.16E5
	Time (s)	LASSO	$48 > h$					
		LASSO-Sens	302.45	318.28	363.87	528.66	720.44	870.80
		LASSO-Uniform	19.95	24.2	59.64	187.25	327.57	513.41
	Sparsity	LASSO	$48 > h$					
		LASSO-Sens	781	777	769	769	765	763
		LASSO-Uniform	700	713	730	736	728	737

Table 6: Comparison of the sketching algorithm and LASSO-Sens on the CTs dataset ($n = 53{,}500,\ d = 386$).

Lambda	Metrics	Algorithms	Coreset Sizes					
			1d	2d	5d	10d	15d	20d
$\lambda = 0.5$	Loss	**LASSO-Sens**	**24.11±0.30**	23.99±0.01	**23.98±0.00**	**23.97±0.00**	**23.97±0.00**	**23.97±0.00**
		Sketching	24.13±0.03	**23.99±0.00**	**23.98±0.00**	**23.97±0.00**	**23.97±0.00**	**23.97±0.00**
	Time (s)	**LASSO-Sens**	7.21	7.18	9.66	13.58	17.68	22.62
		Sketching	5.91	6.61	10.49	16.05	21.37	27.46
	Sparsity	**LASSO-Sens**	383	359	322	317	314	316
		Sketching	382	360	327	319	313	309
$\lambda = 1$	Loss	**LASSO-Sens**	**48.10±0.75**	**47.97±0.02**	**47.93±0.00**	**47.93±0.00**	**47.93±0.00**	**47.92±0.00**
		Sketching	48.17±0.06	47.98±0.01	47.94±0.00	**47.93±0.04**	**47.93±0.02**	47.93±0.03
	Time (s)	**LASSO-Sens**	6.83	7.26	9.95	13.38	18.51	22.27
		Sketching	6.11	6.74	10.40	16.12	21.54	27.23
	Sparsity	**LASSO-Sens**	345	241	219	215	217	215
		Sketching	327	243	222	214	214	216
$\lambda = 5$	Loss	**LASSO-Sens**	**242.48±1.45**	**240.20±0.62**	**239.77±0.11**	**239.65±0.06**	**239.61±0.05**	**239.61±0.04**
		Sketching	243.89±1.67	240.39±0.14	239.86±0.05	239.68±0.05	239.65±0.05	239.64±0.02
	Time (s)	**LASSO-Sens**	7.13	7.28	9.90	13.85	18.00	21.96
		Sketching	6.12	6.79	10.20	16.36	21.52	26.74
	Sparsity	**LASSO-Sens**	179	168	156	154	158	157
		Sketching	174	166	159	155	159	158
$\lambda = 10$	Loss	**LASSO-Sens**	**495.54±2.59**	**481.58±1.01**	**479.82±0.69**	**479.18±0.22**	**479.17±0.14**	**479.00±0.23**
		Sketching	495.48±2.57	482.36±1.07	479.98±0.24	479.52±0.14	479.33±0.17	479.28±0.05
	Time (s)	**LASSO-Sens**	7.21	7.29	9.77	13.34	18.26	21.74
		Sketching	6.09	7.05	10.41	16.14	21.51	26.85
	Sparsity	**LASSO-Sens**	169	153	157	153	155	153
		Sketching	170	159	149	149	156	153

Table 7: Comparison of the sketching algorithm and LASSO-Sens on the Synthetic dataset ($n = 10,000, d = 200$).

Lambda	Metrics	Algorithm	Coreset Sizes					
			1d	2d	5d	10d	15d	20d
$\lambda = 0.5$	Loss	**LASSO-Sens**	**14.24±0.41**	**14.02±0.30**	**13.27±0.59**	**12.83±0.43**	**12.93±0.26**	**13.12±0.10**
		Sketching	25.82±4.57	17.15±0.81	14.12±0.51	13.52±0.20	13.39±0.12	13.19±0.16
	Time (s)	**LASSO-Sens**	4.23	4.77	5.55	6.19	7.60	8.21
		Sketching	3.84	4.96	5.72	6.38	7.93	9.21
	Sparsity	**LASSO-Sens**	40	34	38	9	40	40
		Sketching	101	106	113	111	111	108
$\lambda = 1$	Loss	**LASSO-Sens**	**21.41±0.17**	**21.42±0.22**	**21.53±0.06**	**21.47±0.14**	**21.57±0.12**	**21.61±0.06**
		Sketching	31.52±1.92	24.89±0.22	22.90±0.18	22.31±0.08	22.23±0.03	22.14±0.05
	Time (s)	**LASSO-Sens**	4.03	4.98	4.88	6.34	7.67	7.99
		Sketching	4.02	4.99	5.14	6.51	7.85	9.11
	Sparsity	**LASSO-Sens**	35	29	28	27	26	26
		Sketching	96	95	94	92	90	90
$\lambda = 5$	Loss	**LASSO-Sens**	**66.78±0.08**	**66.77±0.01**	**66.76±0.00**	**66.76±0.00**	**66.76±0.00**	**66.76±0.00**
		Sketching	72.24±1.33	69.02±0.43	67.58±0.19	67.09±0.05	67.00±0.06	66.88±0.02
	Time (s)	**LASSO-Sens**	3.90	5.40	4.81	6.39	7.64	8.04
		Sketching	3.82	5.29	5.18	6.64	7.84	9.20
	Sparsity	**LASSO-Sens**	29	25	24	24	24	24
		Sketching	80	78	70	58	55	49
$\lambda = 10$	Loss	**LASSO-Sens**	**122.99±0.28**	**122.87±0.09**	**122.81±0.02**	**122.82±0.01**	**122.82±0.01**	**122.83±0.01**
		Sketching	127.49±0.81	124.60±0.35	123.48±0.16	123.02±0.04	122.96±0.02	122.92±0.03
	Time (s)	**LASSO-Sens**	4.13	5.30	4.83	6.48	7.63	8.07
		Sketching	3.95	5.20	5.11	6.65	7.71	9.16
	Sparsity	**LASSO-Sens**	30	25	24	24	24	24
		Sketching	65	59	47	37	33	27

Table 8: Comparison of the sketching algorithm and LASSO-Sens on the mediamill dataset ($n = 30,993, d = 120$).

Lambda	Metrics	Algorithms	Coreset Sizes					
			1d	2d	5d	10d	15d	20d
$\lambda = 0.5$	Loss	**LASSO-Sens**	**8.40±0.43**	**8.19±0.04**	**8.17±0.02**	**8.15±0.01**	**8.15±0.00**	**8.15±0.00**
		Sketching	8.54±0.15	8.25±0.04	8.18±0.01	8.16±0.00	**8.15±0.00**	**8.15±0.00**
	Time (s)	**LASSO-Sens**	2.57	3.37	4.69	4.94	4.69	5.81
		Sketching	2.14	3.28	4.59	5.07	5.13	5.82
	Sparsity	**LASSO-Sens**	58	60	62	63	61	61
		Sketching	57	60	62	60	62	63
$\lambda = 1$	Loss	**LASSO-Sens**	**16.77±0.23**	**16.34±0.16**	**16.27±0.03**	**16.24±0.02**	**16.23±0.02**	**16.22±0.01**
		Sketching	17.16±0.45	16.46±0.07	16.29±0.04	16.25±0.01	16.25±0.01	16.24±0.01
	Time (s)	**LASSO-Sens**	2.47	3.68	4.68	4.91	4.48	5.82
		Sketching	2.13	3.07	4.61	4.91	5.30	5.78
	Sparsity	**LASSO-Sens**	57	58	59	61	57	60
		Sketching	54	59	62	59	61	60
$\lambda = 5$	Loss	**LASSO-Sens**	**83.62±4.37**	**81.00±0.93**	**80.30±0.18**	**79.95±0.18**	**80.07±0.08**	**80.01±0.09**
		Sketching	87.07±1.31	82.66±0.93	80.57±0.37	80.24±0.11	80.26±0.17	80.20±0.11
	Time (s)	**LASSO-Sens**	2.46	3.45	4.71	4.93	4.49	5.82
		Sketching	2.26	3.30	4.64	5.02	5.20	5.90
	Sparsity	**LASSO-Sens**	44	51	51	51	51	51
		Sketching	46	49	49	53	51	52
$\lambda = 10$	Loss	**LASSO-Sens**	**163.26±8.88**	**160.69±5.25**	**159.24±0.35**	**158.89±0.50**	**158.91±0.34**	**158.65±0.40**
		Sketching	177.47±5.03	163.25±0.83	160.31±0.97	159.24±0.37	159.08±0.26	159.03±0.14
	Time (s)	**LASSO-Sens**	2.56	3.45	4.65	4.90	4.47	5.81
		Sketching	2.11	3.23	4.50	4.80	5.31	5.83
	Sparsity	**LASSO-Sens**	40	44	48	51	49	49
		Sketching	41	46	51	49	49	49

Table 9: Comparison of the sketching algorithm and LASSO-Sens on the mnist8m datasets ($n = 8,000,000, d = 784$).

Lambda	Metrics	Algorithms	Coreset Sizes					
			1d	2d	5d	10d	15d	20d
$\lambda = 0.5$	Loss	**LASSO-Sens** Sketching	**1.27E7±1.02E7** 4.27E7±5.20E5	**3.30E4±7.71E3** 4.62E4±3.90E4	**1.64E4±4.03E2** 1.67E4±1.58E3	**1.45E4±3.06E2** 1.48E4±1.05E2	**1.43E4±1.80E2** 1.43E4±1.55E3	**1.37E4±1.16E2** 1.41E4±6.27E2
	Time (s)	**LASSO-Sens** Sketching	304.32 533.75	314.23 1000.12	370.11 2484.28	512.73 5048.47	703.91 7635.30	859.22 10265.84
	Sparsity	**LASSO-Sens** Sketching	780 783	776 779	770 770	770 773	763 771	760 764
$\lambda = 1$	Loss	**LASSO-Sens** Sketching	**1.61E7±1.04E7** 3.67E7±1.43E6	3.46E4±8.18E3 2.18E4±4.26E3	**1.68E4±5.27E2** 1.73E4±9.62E2	**1.48E4±2.05E2** 1.48E4±1.29E3	**1.41E4±1.38E2** 1.43E4±2.82E3	**1.38E4±7.54E1** 1.40E4±9.60E2
	Time (s)	**LASSO-Sens** Sketching	302.37 499.51	316.26 990.37	366.99 2475.33	522.60 5055.59	724.51 7678.16	838.05 10277.15
	Sparsity	**LASSO-Sens** Sketching	778 783	774 776	768 775	763 776	765 767	758 764
$\lambda = 5$	Loss	**LASSO-Sens** Sketching	**1.39E7±4.50E6** 1.58E7±5.96E5	**2.87E4±4.21E3** 2.90E4±2.26E3	**1.72E4±4.57E2** 1.72E4±2.53E3	**1.49E4±2.52E2** 1.53E4±2.29E3	**1.45E4±1.88E2** 1.47E4±1.04E3	**1.42E4±1.49E2** 1.45E4±2.14E3
	Time (s)	**LASSO-Sens** Sketching	299.24 498.59	317.45 994.44	368.02 2477.81	527.51 5033.69	708.16 7690.62	875.08 10277.57
	Sparsity	**LASSO-Sens** Sketching	780 783	776 779	770 774	767 770	765 770	763 768
$\lambda = 10$	Loss	**LASSO-Sens** Sketching	**1.10E7±8.22E6** 1.55E7±2.22E5	9.47E4±1.27E5 2.85E4±3.36E3	**1.73E4±1.06E2** 1.79E4±1.17E3	**1.55E4±1.54E2** 1.58E4±6.86E2	**1.49E4±3.28E1** 1.53E4±1.16E3	**1.47E4±8.80E1** 1.48E4±1.92E3
	Time (s)	**LASSO-Sens** Sketching	302.45 502.08	318.28 994.11	363.87 2483.74	528.66 5034.16	720.44 7666.31	870.80 10231.89
	Sparsity	**LASSO-Sens** Sketching	781 782	777 777	769 774	769 762	765 763	763 767

Table 10: Comparison of sensitivity sampling applied to standard and modified LASSO objectives on Synthetic dataset ($n = 10{,}000,\ d = 200$).

Lambda	Metrics	Algorithms	Coreset Sizes					
			1d	2d	5d	10d	15d	20d
$\lambda = 0.5$	Relative error	**LASSO** modified LASSO	**4.06E5±5.61E4** 4.10E5±3.33E4	**2.00E5±2.54E4** 2.16E5±3.15E4	**9.88E4±1.35E4** 9.97E4±1.42E4	**3.02E4±7.42E3** 3.54E4±1.07E4	1.69E4±4.61E3 1.65E4±4.09E3	1.27E4±1.66E3 1.21E4±1.84E3
	Time (s)	**LASSO** modified LASSO	25.85 26.26	29.26 29.78	37.15 37.97	47.91 48.39	60.96 61.33	74.55 75.62
	Sparsity	**LASSO** modified LASSO	200 200					
$\lambda = 1$	Relative error	**LASSO** modified LASSO	**3.28E5±3.09E4** 3.39E5±4.80E4	1.94E5±3.14E4 1.73E5±1.56E4	**7.39E4±2.67E4** 7.52E4±3.10E4	2.15E4±3.53E3 2.13E4±1.81E3	**1.32E4±1.06E3** 1.46E4±3.53E3	**1.12E4±3.67E3** 1.14E4±2.54E3
	Time (s)	**LASSO** modified LASSO	25.85 26.06	29.05 29.21	36.98 37.38	47.96 48.56	60.90 62.13	74.55 76.29
	Sparsity	**LASSO** modified LASSO	200 200	200 200	199 199	197 198	197 198	197 198
$\lambda = 5$	Relative error	**LASSO** modified LASSO	**2.94E5±2.24E4** 3.14E5±2.38E4	**1.60E5±1.30E4** 1.65E5±3.15E4	**4.22E4±5.26E3** 5.49E4±1.46E4	**1.47E4±2.28E3** 1.77E4±3.61E3	**6.26E3±2.62E3** 9.66E3±2.09E3	**4.30E3±1.17E3** 7.23E3±6.02E2
	Time (s)	**LASSO** modified LASSO	25.74 26.15	29.03 29.29	37.20 37.48	47.78 48.66	60.88 61.06	74.50 76.17
	Sparsity	**LASSO** modified LASSO	200 200	200 200	197 199	197 198	196 197	192 197
$\lambda = 10$	Relative error	**LASSO** modified LASSO	**2.55E5±3.39E4** 2.71E5±4.12E4	**1.43E5±6.77E3** 1.61E5±1.90E4	**2.42E4±2.89E3** 4.70E4±8.35E3	**1.33E4±2.36E3** 1.62E4±1.74E3	**6.82E3±9.08E2** 1.13E4±2.03E3	**5.23E3±8.21E2** 7.58E3±2.88E3
	Time (s)	**LASSO** modified LASSO	25.70 25.80	29.19 29.24	37.15 37.49	48.07 48.50	61.15 61.27	75.83 78.09
	Sparsity	**LASSO** modified LASSO	197 200	192 197	191 186	185 188	176 182	173 177