Differentially Private Feature Selection via Stability Arguments, and the Robustness of the Lasso

Adam Smith
asmith@cse.psu.edu
Pennsylvania State University

Abhradeep Thakurta
azg161@cse.psu.edu
Pennsylvania State University

Abstract

We design differentially private algorithms for statistical model selection. Given a data set and a large, discrete collection of “models”, each of which is a family of probability distributions, the goal is to determine the model that best “fits” the data. This is a basic problem in many areas of statistics and machine learning.

We consider settings in which there is a well-defined answer, in the following sense: Suppose that there is a *nonprivate* model selection procedure f, which is the reference to which we compare our performance. Our differentially private algorithms output the correct value $f(\mathcal{D})$ whenever f is *stable* on the input data set $\mathcal{D}$. We work with two notions, *perturbation* stability and *subsampling* stability.

We give two classes of results: generic ones, that apply to any function with discrete output set; and specific algorithms for the problem of sparse linear regression. The algorithms we describe are efficient and in some cases match the optimal *nonprivate* asymptotic sample complexity.

Our algorithms for sparse linear regression require analyzing the stability properties of the popular LASSO estimator. We give sufficient conditions for the LASSO estimator to be robust to small changes in the data set, and show that these conditions hold with high probability under essentially the same stochastic assumptions that are used in the literature to analyze convergence of the LASSO.

1 Introduction

Model selection is a basic problem in machine learning and statistics. Given a data set and an collection of “models”, where each model is normally a family of probability distributions, the goal is to determine the model that best “fits” the data in some sense. The choice of model could reflect a measure of complexity, such as the number of components in a mixture model, or a choice about which aspects of the data appear to be most relevant, such as the set of features used for a regression model.

In sparse linear regression problems, for example, each entry in the data set consists of a p-dimensional real *feature vector* $\boldsymbol{x}$ and real-valued *response* (or *label*) y. The overall goal is to find a parameter vector $\theta \in \mathbb{R}^p$ such that $\langle \boldsymbol{x}_i, \theta \rangle \approx y_i$ for all n data points $(\boldsymbol{x}_i, y_i)$. When p is much larger than n, the problem is underdetermined and so solutions to this problem won't necessarily generalize well. A common approach is to look for a vector $\boldsymbol{\theta}$ with at most s nonzero entries (where $s \ll n$) that labels the data set well. Each set of at most s positions defines a model and, for a specific model, the problem simplifies to textbook linear regression. The model selection problem is to decide which of the roughly $\binom{p}{s}$ subsets to consider.

In this paper we investigate the possibility of carrying out sophisticated model selection algorithms without leaking significant information about individuals entries in the data set. This is critical when the information in the data set is sensitive, for example if it consists of financial records or health data. Our algorithms satisfy *differential privacy* [8, 5], which essentially ensures that adding or removing an individual's data from a data set will have little effect on the inferences made about them based on an algorithm's output [5, 9].

Formally, there is no reason to separate model selection from the fitting of a specific distribution of the data once the model is selected—either way, one is trying to select a best fit from among a class of probability distributions. However, the separation into two phases survives for (at least) two reasons: First, individual models are often parameterized by a finite-dimensional real vector, and so fitting a particular model to the data is a continuous optimization task. In contrast, the set of models is typically discrete, and the corresponding optimization problems tend to have a very different feel. Second, the model selection step is typically much more computationally expensive.

The question, then, is how well differentially private algorithms can do at model selection, and how to design algorithms that are computationally efficient. The challenge is to select from a large number of possible models using as few resources (samples and running time) as possible.

Our Contributions. We consider the setting in which there is a "well-defined" answer, in the following sense: Suppose that there is nonprivate model selection procedure f, which is the reference to which we compare our performance. Our algorithms output the correct value $f(\mathcal{D})$ whenever f is *stable* on the input data set $\mathcal{D}$. We work with two notions, *perturbation* stability and *subsampling* stability.

We give two classes of results: generic ones, that apply to any function; and specific algorithms for the problem of sparse linear regression. The algorithms we describe are efficient and in some cases match the optimal asymptotic sample complexity for nonprivate algorithms.

Our algorithms for sparse linear regression require analyzing the stability properties of the popular LASSO estimator. We give sufficient conditions for the LASSO estimator to be robust to small changes in the data set, and show that these conditions hold with high probability under essentially the same stochastic assumptions that are used in the literature to analyze convergence of the LASSO. This analysis may be of independent interest.

Differential privacy. Our algorithms take as input a data set $\mathcal{D} \in U^*$ that is a list of elements in a universe U. The algorithms we consider are all symmetric in their inputs, so we may equivalently view the data as a multi-set in U. We say multi-sets $\mathcal{D}$ and $\mathcal{D}'$ are *neighbors* if $|\mathcal{D} \triangle \mathcal{D}'| = 1$. More generally, the *distance* between two data sets is the size of their symmetric difference, which equals the minimum number of entries that need to be added to or removed from $\mathcal{D}$ to obtain $\mathcal{D}'$.

Definition 1 (Differential privacy [8, 7]). *A randomized algorithm $\mathcal{A}$ is (ϵ, δ)-differentially private if for every two neighboring datasets $\mathcal{D}$ and $\mathcal{D}'$ in U^* (that is, with $|\mathcal{D} \triangle \mathcal{D}'| = 1$), and for all events $\mathcal{O} \subseteq Range(\mathcal{A})$ the following holds:*

$$\Pr[\mathcal{A}(\mathcal{D}) \in \mathcal{O}] \leq e^{\epsilon} \Pr[\mathcal{A}(\mathcal{D}') \in \mathcal{O}] + \delta\,.$$

This definition is meaningful roughly when ϵ is at most a small constant (say $1/10$) and δ is significantly less than $1/n$ (see [12] for a discussion).

The next two sections describe our contributions in more detail.

1.1 Generic Algorithms For Stable Functions

We give two simple, generic transformations that, given any function f and parameters $\epsilon, \delta > 0$, return a (ϵ, δ)-differentially private algorithm (see Definition 1) that is correct whenever f is sufficiently stable on a particular input $\mathcal{D}$. The two algorithms correspond to different notions of stability. In both cases, the correctness guarantees do not have any dependence on the size of the range of f, only on the privacy parameters ϵ and δ. In the context of model selection, this implies that there is no dependency on the number of models under consideration.

- *Perturbation Stability*: We say that f is *stable* on $\mathcal{D}$ if f takes the value $f(\mathcal{D})$ on all of the neighbors of $\mathcal{D}$ (and *unstable* otherwise). We give an algorithm $\mathcal{A}_{dist}$ that, on input $\mathcal{D}$, outputs $f(\mathcal{D})$ with high probability if $\mathcal{D}$ is at distance at least $\frac{2\log(1/\delta)}{\epsilon}$ from the nearest *un*stable data set. Unfortunately, the algorithm $\mathcal{A}_{dist}$ is not efficient, in general.
- *Subsampling stability*: We say f is q-subsampling stable on $\mathcal{D}$ if $f(\hat{D}) = f(\mathcal{D})$ with probability at least $3/4$ when $\hat{D}$ is a random subsample from $\mathcal{D}$ which includes each entry independently with probability q. We give an algorithm $\mathcal{A}_{samp}$ that, on input $\mathcal{D}$, outputs $f(\mathcal{D})$ with high probability whenever f is q-subsampling stable for $q = \frac{\epsilon}{32\log(1/\delta)}$. The running time of $\mathcal{A}_{samp}$ dominated by running f about $1/q^2$ times; hence it is efficient whenever f is.

  This result has an clean statistical interpretation: Given a collection of models, let the sample complexity of model selection be the minimum number of samples (over nonprivate algorithms) from a distribution in one of the models needed to select the correct model with probability at least $2/3$. Then the sample complexity needed for differentially private model selection increases by a problem-independent factor of $O(\log(1/\delta)/\epsilon)$.

Technique: Proxies for the distance to instability. The idea behind the first algorithm comes from the work of Dwork and Lei [6] on private parametric estimation. If we were somehow given a *promise* that f is stable on $\mathcal{D}$, we could release $f(\mathcal{D})$ without violating differential privacy. The issue is that stability itself can change between neighboring data sets, and so stating that f is stable on $\mathcal{D}$ may violate differential privacy. The solution implicit in [6] (specifically, in their algorithms for estimating interquartile distance and the median) is to instead look at the *distance* to the nearest unstable instance. This distance changes by at most 1 between neighboring data sets, and so one can release a noisy version of the distance privately, and release $f(\mathcal{D})$ when that noisy estimate is sufficiently high. Developing this simple idea leads to the algorithm $\mathcal{A}_{dist}$.

The difficulty with this approach is that it requires computing the distance to the nearest unstable instance explicitly. We observe that if one can compute a *lower bound* $\hat{d}(\mathcal{D})$ on the distance to the nearest unstable instance, and if $\hat{d}$ does not change much between neighboring data sets, then one can release a noisy version of $\hat{d}$ differentially privately, and release $f(\mathcal{D})$ when the noisy estimate is sufficiently high. *The challenge, then, is to efficiently compute useful proxies for the distance to the nearest unstable input.*

We obtain our algorithm for subsampling-stable functions by giving an efficient distance bound for a bootstrapping-based model selector $\hat{f}(\mathcal{D})$ that outputs the most commonly occurring value of f in a set of about $1/\epsilon^2$ random subsamples taken from the input $\mathcal{D}$. The approach is inspired by the "sample and aggregate" framework of Nissim et al. [18]. However, our analysis allows working with much larger subsamples than those in previous work [18, 25, 13]. In our context, the analysis from previous work would lead to a polynomial blowup in sample complexity (roughly, squaring the number of samples needed nonprivately), whereas our result increases the sample complexity by a small factor.

1.2 Feature Selection for Sparse Linear Regression and Robustness of the LASSO

Our second class of results concerns feature selection for sparse linear regression. Recall the task is to select a set of s out of p features to be used for linear regression. This problem provides an interesting challenge for model selection problems since the number of distinct models to be considered is enormous ($\binom{p}{s}$, where we want to make s as large as possible without losing statistical validity).

Given a data set of n entries $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}$, let $X \in \mathbb{R}^{n \times p}$ be the matrix with rows $\boldsymbol{x}_i$ and $\boldsymbol{y} \in \mathbb{R}^p$ be the column vector with entries y_i. Suppose that the data set satisfies a linear system

$$\boldsymbol{y} = X\boldsymbol{\theta}^* + \boldsymbol{w} \tag{1}$$

where $\boldsymbol{\theta}^*$ is a parameter vector (in $\mathbb{R}^p$) to be estimated, and $\boldsymbol{w} \in \mathbb{R}^{n\times 1}$ is an error vector whose entries are assumed to be small. We say a vector is s-sparse if has at most s nonzero entries. The problem we consider is: assuming that $\boldsymbol{\theta}^*$ is s-sparse, *under what conditions can we recover the support of $\boldsymbol{\theta}^*$ while satisfying differential privacy?*

The nonprivate version of this problem has been studied extensively in the literature on high-dimensional statistics and compressed sensing. Several works [30, 27, 17] have shown that $n = O(s \log p)$ samples suffice to recover the support of $\boldsymbol{\theta}^*$, assuming the data are drawn i.i.d. from one of a fairly large class of probability distributions. Moreover, this bound is known to be asymptotically tight [27, 20].

Differentially private algorithms for sparse regression were first considered in our recent work (Kifer et al. [13]). They gave feature selection procedures that require $n \gg s\log(p)\cdot(\min\{s, \log p\})$ samples to recover the support of $\boldsymbol{\theta}^*$. Matching the optimal sample complexity (under reasonable assumptions) was left as an open problem.

We give two efficient algorithms that approach the optimal sample complexity $s \log p$.

- Our results on subsampling stability imply immediately that one can get efficient differentially private algorithms with sample complexity $O(\frac{\log(1/\delta)}{\epsilon} s \log p)$ under the same stochastic assumptions used in nonprivate upper bounds. This is already a significant improvement over the previous work [13]. However, it retains the multiplicative dependence on $\log(1/\delta)/\epsilon$.
- We also give explicit estimators for the distance to instability of a popular technique for sparse regression known as the Lasso (as well as a more robust variant). This allows us to get efficient differentially private algorithms with *optimal* sample complexity $O(s \log p)$ — removing the dependence on ϵ and δ —when p is very large. More specifically, we derive algorithms with sample complexity $n = O(\max\{s \log p, k^2 s^4/\log(p), ks^{3/2}\})$, where $k = \log(1/\delta)/\epsilon$. This is $O(s\log p)$ when $s < \frac{\log^{2/3} p}{k^{2/3}}$ and $k < \log^{2/3} p$. Note that it is interesting to come up with good model selectors even when s is constant, since p may be very large.

Techniques: Stability and Robustness of the LASSO. The (efficient, nonprivate) upper bounds on feature selection for sparse linear regression derive mostly from analyses of a popular approach known as the Lasso. The idea is to find an estimate $\hat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}^*$ which is sparse and which minimizes some norm of the estimated error $\hat{\boldsymbol{w}} = \boldsymbol{y} - X\hat{\boldsymbol{\theta}}$. This is done by penalizing the usual mean squared error loss with some multiple of the L_1 norm of $\boldsymbol{\theta}$:

$$\hat{\boldsymbol{\theta}}(\mathcal{D}) = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \frac{\Lambda}{n}\|\boldsymbol{\theta}\|_1 \tag{2}$$

The consistency properties of the Lasso are well-studied: a variety of assumptions on the data, when $n = \omega(s \log p)$, the estimate $\hat{\boldsymbol{\theta}}$ is known to converge to $\boldsymbol{\theta}^*$ in the L_2 norm [27, 17]. Moreover, if the entries of $\boldsymbol{\theta}^*$ are bounded away from zero, $\hat{\boldsymbol{\theta}}$ will have the same support as $\boldsymbol{\theta}^*$ [27].

We extend these results to show that, *under essentially the same assumptions*, the support of $\hat{\boldsymbol{\theta}}$ does not change when a small number of data points are changed. Other work on LASSO robustness captures different properties. (See Section 1.3 below.) Our analysis requires significantly refining the "primal-dual" construction technique of Wainwright [27]. The idea is to show that an optimal solution to (2) for data set $\mathcal{D}'$ which is "near" $\mathcal{D}$ can be transformed into an optimal solution for $\mathcal{D}$. This involves analyzing how the KKT conditions on the subgradient of the nondifferentiable loss function in (2) change as the data varies.

Significantly, we also use the primal-dual analysis to give an *efficient* and smooth estimator for the distance from a given data set $\mathcal{D}$ to the nearest unstable data set. The estimator essentially uses the subgradient of the regularized loss (2) to measure how big a change would be needed to one of the zero entries of $\hat{\boldsymbol{\theta}}$ to "jump" away from zero. This is delicate because changing the data set changes both the minimizer and the geometry of the loss function. The efficient distance estimator gives us the differentially private feature selector with optimal sample complexity.

Assumptions. As mentioned, our analyses of stability make various assumptions about the data. First, it is important to note that the assumptions are made only for the utility analysis. *The privacy guarantees are unconditional.* Second, we distinguish between "fixed data" assumptions, which give deterministic conditions the data set for a given algorithm to perform well, and "stochastic" assumptions, which give conditions on a distribution from which the data are drawn i.i.d. We analyze the Lasso's robustness under essentially the same assumptions (fixed-data and stochastic) used in previous work to analyze consistency. The difference is that we require certain constants to be larger, leading to a constant factor increase in sample complexity.

1.3 Prior Work on Learning and Stability

The relationship between learning, statistics and stability has been studied in the learning theory literature (e.g., Rogers and Wagner [21]) and in robusts statistics (e.g., Huber [10]) for over thirty years. Many variants of stability have been studied, and the literature is too vast to survey here. We highlight only the most relevant works.

The main difference with our work is that the learning literature focuses on algorithms whose output lies in a metric space; stability measures how much the output changes under various models of perturbation, and the focus is on settings where some change in unavoidable even for very "nice" data sets [3, 23, 2, 19]. Several papers on privacy have sought to exploit such stability properties for privacy purposes [6, 22, 4]. In contrast, we look at settings where some discrete structure may remain unchanged under perturbations. This is effectively a stronger assumption, leading to tighter sample complexity bounds.

Both notions of peturbation that we conisder have been studied previously, namely robustness to changes in the input data set $\mathcal{D}$ [3, 23, 28, 6] and stability to subsampling or resampling from the training data set $\mathcal{D}$ [24, 1, 16, 15].

Robustness to small changes in the input data was studied both to provide resilience to outliers and noise (as in robust statistics) as well as to get good generalization error [3, 23]. One consequence of these works is that if a learning algorithm f satisfies our notion of stability, then it generalizes well. Perhaps the most relevant work in this line is by Xu et al. [28], who study the L_2 robustness of Lasso-like estimators to small perturbations, and show that *uniform* stability (in which the set of selected features changes by only small steps between *any pairs* of neighbors) is impossible for algorithms with sparse output. Finally, Lee et al. [14] look at Huberized versions of the LASSO with the goal of providing robustness, but no not provide formal consistency or convergence guarantees.

Stability under subsampling and resampling were also studied extensively in the prior work [24, 1, 16, 15]. In particular, they were used for model selection and clustering [15]. Again, their notion of stability is weaker than ours.

2 Stability and Privacy

Consider a function $f : U^* \to \mathcal{R}$ from data sets to a range $\mathcal{R}$. We assume that the range $\mathcal{R}$ is finite, for simplicity.

Definition 2. *A function $f : U^* \to \mathcal{R}$ is k-*stable on input $\mathcal{D}$ *if adding or removing any k elements from $\mathcal{D}$ does not change the value of f, that is, $f(D) = f(D')$ for all $\mathcal{D}'$ such that $|\mathcal{D} \triangle \mathcal{D}'| \leq k$. We say f is* stable *on $\mathcal{D}$ if it is (at least) 1-stable on $\mathcal{D}$, and* unstable *otherwise.*

The distance to instability *of a data set $\mathcal{D} \in U^*$ with respect to a function f is the number of elements that must be added to or removed from $\mathcal{D}$ to reach an data set that is not stable. Note that $\mathcal{D}$ is k-stable if and only if its distance to instability is at least k.*

A First Attempt. For any function f, there is a differentially private algorithm $\mathcal{A}_{dist}$ that outputs $f(\mathcal{D})$ whenever $\mathcal{D}$ is sufficiently stable. It follows the lines of more general approaches from previous work [6, 11] that calibrate noise to differentially private estimates of local sensitivity. The algorithm is not efficient, in general, but it is very simple: On input $\mathcal{D}$ and parameters $\epsilon, \delta > 0$, $\mathcal{A}_{dist}$ computes the distance d from $\mathcal{D}$ to the nearest unstable instance, and add $\mathsf{Lap}(1/\epsilon)$ noise to get an estimate $\tilde{d}$ of d. Finally, if $\tilde{d} > \frac{\log(1/\delta)}{\epsilon}$, then it releases $f(\mathcal{D})$, otherwise it outputs a special symbol $\perp$.

Proposition 3. *For every function f:*

(1) $\mathcal{A}_{dist}$ is (ϵ, δ)-differentially private.

(2) For all $\beta > 0$: if f is $\frac{\log(1/\delta)+\log(1/\beta)}{\epsilon}$-stable on $\mathcal{D}$, then $A(\mathcal{D}) = f(\mathcal{D})$ with probability at least $1-\beta$.

For brevity, we defer the proof of this proposition to Section A.1.This result based on distance is the best possible, in the following sense: if there are two data sets $\mathcal{D}_1$ and $\mathcal{D}_2$ for which $\mathcal{A}$ outputs different values $f(\mathcal{D}_1)$ and $f(\mathcal{D}_2)$, respectively, with at least constant probability, then the distance from $\mathcal{D}_1$ to $\mathcal{D}_2$ must be $\Omega(\log(1/\delta)/\epsilon)$.

However, there are two problems with this straightforward approach. First, the algorithm is not efficient, in general, since it may require searching all data sets within distance up to d from $\mathcal{D}$ (this may not even be implementable at all if U is infinite). Second, the model selection algorithm given to us may not be stable on the instances of interest.

More Robust Functions, and Efficient Proxies for Distance. We remedy these problems by (a) modifying the functions to obtain a more stable function $\hat{f}$ that equals f on "nice" inputs, and (b) designing efficient, private estimators for the distance to instability with respect to $\hat{f}$.

We combine these two goals into a single definition: we are looking for a pair of functions $\hat{f}, \hat{d}$ that act as proxies for f and the stability of f, respectively. We measure the usefulness of the pair by a set N of "nice" inputs on which this pair allows us to release the actual value f.

Definition 4. *Given $f : U^* \to \mathcal{R}$, a pair of functions $\hat{f} : U^* \to \mathcal{R}$, $\hat{d} : U^* \to \mathbb{R}$ are* proxies *for f and its stability which are accurate on a set N (which depends on parameters ϵ, δ) if the following hold:*

1. For all $\mathcal{D}$: $\hat{d}(\mathcal{D}) \leq$ (dist. of $\mathcal{D}$ to instability of $\hat{f}$).

2. $GS_{\hat{d}} \leq 1$

3. For all $\mathcal{D} \in N$: $f(\mathcal{D}) = \hat{f}(\mathcal{D})$ and $\hat{d}(\mathcal{D}) \geq 2\log(1/\delta)/\epsilon$.

One can use such a proxy by adding Laplace noise to $\hat{d}$ and releasing $\hat{f}(\mathcal{D})$ whenever the noisy version of $\hat{d}$ is sufficiently large. The resulting mechanism will be (ϵ, δ)-differentially private and, on all inputs $\mathcal{D} \in N$, will release $f(\mathcal{D})$ with probability at least $1 - \delta$.

For every function f, one can get a valid proxy by letting $\hat{f} = f$ and letting $\hat{d}(\mathcal{D})$ be the distance to instability of $\mathcal{D}$ w.r.t. f. The set N of good instances for this pair is exactly the set of inputs $\mathcal{D}$ on which f is $\frac{2\log(1/\delta)}{\epsilon}$-stable. As mentioned above, the main problem is computational efficiency.

Given a function f, the goal is to find proxies $(\hat{f}, \hat{d})$ that are efficient (ideally, as efficient as evaluating f alone) and have as large as possible a set N of good inputs.

2.1 From Sampling Stability to Stability

We give a generic construction that takes any function f and produces a pair functions $(\hat{f}, \hat{d})$ that are efficient—they take essentially the same time to evaluate as f—and are accurate for data sets on which the original f is *subsampling stable*.

Definition 5 (Subsampling stability)**.** *Given a data set $\mathcal{D} \in U^*$, let $\hat{\mathcal{D}}$ be a random subset of $\mathcal{D}$ in which each element appears independently with probability q. We say f* is q-subsampling stable on input $\mathcal{D} \in U^*$ *if $f(\hat{\mathcal{D}}) = f(\mathcal{D})$ with probability at least $3/4$ over the choice of $\hat{\mathcal{D}}$.*

The algorithm $\mathcal{A}_{samp}$ (Algorithm 1) uses bootstrapping to create a modified function $\hat{f}$ that equals $f(\mathcal{D})$ and is far from unstable on a given $\mathcal{D}$ whenever f is subsampling stable on $\mathcal{D}$. The output of $\hat{f}(\mathcal{D})$ is the mode (most frequently occurring value) in the list $F = (f(\hat{\mathcal{D}}_1), ..., f(\hat{\mathcal{D}}_m))$ where the $\hat{\mathcal{D}}_i$'s are random subsamples of size about $\epsilon n / \log(1/\delta)$. The distance estimator $\hat{d}$ is, up to a scaling factor, the difference between the frequency of the mode and the next most frequent value in F. Following the generic template in the previous section, the algorithm $\mathcal{A}_{samp}$ finally adds Laplace noise to $\hat{d}$ and outputs $\hat{f}(\mathcal{D})$ if the noise distance estimate is sufficiently high.

We summarize the properties of $\mathcal{A}_{samp}$ below.

Theorem 6.

1. *Algorithm $\mathcal{A}_{samp}$ is (ϵ, δ)-differentially private.*
2. *If f is q-subsampling stable on input $\mathcal{D}$ where $q = \frac{\epsilon}{32\log(1/\delta)}$, then algorithm $\mathcal{A}_{samp}(\mathcal{D})$ outputs $f(\mathcal{D})$ with probability at least $1 - 3\delta$.*
3. *If f can be computed in time $T(n)$ on inputs of length n, then $\mathcal{A}_{samp}$ runs in expected time $O(\frac{\log n}{q^2})(T(qn) + n)$.*

Note that the utility statement here is an input-by-input guarantee; f need not be subsampling stable on all inputs. *Importantly, there is no dependence on the size of the range $\mathcal{R}$.* In the context of model selection, this means that one can efficiently satisfy differential privacy with a modest blow-up in sample complexity (about $\log(1/\delta)/\epsilon$) whenever there is a particular model that gets selected with reasonable probability.

Previous work in data privacy has used the idea of bootstrapping or subsampling to convert from various forms of subsampling stability to some sort of stability [18, 6, 25, 13]. The main advantage of the version we present here is that size of the subsamples is quite large: our algorithm requires a blowup in sample complexity of about $\log(1/\delta)/\epsilon$, independent of the size of the output range $\mathcal{R}$, as opposed to previous algorithms that had blowups polynomial in n and some measure of "dimension" of the output.

The following lemma provides the key to analyzing our approach. The main observation is that the stability of the $mode$ is a function of the difference between the frequency of the mode and the next most

Algorithm 1 $\mathcal{A}_{samp}$: Bootstrapping for Subsampling-Stable f

```
Require: dataset: $\mathcal{D}$, function $f : U^* \rightarrow \mathcal{R}$, privacy parameters $\epsilon, \delta > 0$.
$q \leftarrow \frac{\epsilon}{32\log(1/\delta)}$, $m \leftarrow \frac{\log(n/\delta)}{q^2}$.
repeat
    Subsample $m$ data sets $\hat{\mathcal{D}}_1, ..., \hat{\mathcal{D}}_m$ from $\mathcal{D}$, where $\hat{\mathcal{D}}_i$ includes each position of $\mathcal{D}$ independently w.p. $q$.
until each position of $\mathcal{D}$ appears in at most $2mq$ sets $\hat{\mathcal{D}}_i$
Compute $F = \langle f(\hat{\mathcal{D}}_1), \cdots, f(\hat{\mathcal{D}}_m)\rangle$.
For each $r \in \mathcal{R}$, let $count(r) = \#\{i : f(\hat{\mathcal{D}}_i) = r\}$.
$\hat{d} \leftarrow (count_{(1)} - count_{(2)})/(2mq)$ where $count_{(1)}, count_{(2)}$ are the two highest counts from the previous step.
$\tilde{d} \leftarrow \hat{d} + \mathsf{Lap}(\frac{1}{\epsilon})$.
if $\tilde{d} > \log(1/\delta)/\epsilon$ then
    Output $\hat{f}(\mathcal{D}) = mode(F)$.
else
    Output $\perp$.
```

frequent element. The lemma roughly says that if f is subsampling stable on $\mathcal{D}$, then $\mathcal{D}$ is far from unstable w.r.t. $\hat{f}$ (not necessarily w.r..t f), and moreover one can estimate the distance to instability of $\mathcal{D}$ *efficiently* and privately. Proof of this lemma is deferred to Section A.2 for brevity.

Lemma 7. *Fix $q \in (0, 1)$. Given $f : U^* \rightarrow \mathcal{R}$, let $\hat{f} : U^* \rightarrow \mathcal{R}$ be defined as $\hat{f}(\mathcal{D}) = mode(f(\hat{\mathcal{D}}_1), ..., f(\hat{\mathcal{D}}_m))$ where each $\hat{\mathcal{D}}_i$ includes elements of $\mathcal{D}$ independently w.p. q and $m = \log(1/\delta)/q^2$. Let $\hat{d}(\mathcal{D}) = (count_{(1)} - count_{(2)})/(4mq)$. Fix a data set $\mathcal{D}$. Let E be the event that no position of $\mathcal{D}$ is included in more than $2mq$ of the subsets $\hat{\mathcal{D}}_i$.*

(1) E occurs with probability at least $1 - \delta$.

(2) Conditioned on E, the pair $(\hat{f}, \hat{d})$ are a good proxy for f and its stability (that is, $\hat{d}$ lower bounds the stability of $\hat{f}$ on $\mathcal{D}$, and $\hat{d}$ has global sensitivity 1).

(3) If f is q-subsampling stable on $\mathcal{D}$, then with probability at least $1 - \delta$ over the choice of subsamples, we have $\hat{f}(\mathcal{D}) = f(\mathcal{D})$, and $\hat{d}(\mathcal{D}) \geq 1/16q$.

The events in (2) and (3) occur simultaenously with probability at least $1 - 2\delta$.

Theorem 6 follows from the lemma by noting that for small enough q, the function d, which acts as an efficient proxy for stability, will be large enough that even after adding Laplace noise one can tell that $\hat{f}$ is stable on instance $\mathcal{D}$, and release f.

3 Consistency and Stability of Sparse Regression using LASSO

Recall the linear system in (1). A common approach for estimating the underlying parameter vector $\boldsymbol{\theta}^*$ is via least-squared regression, where the loss function in the regression problem is $\hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2$, where the data set $\mathcal{D} = (\boldsymbol{y}, X)$ and has size n. When the dimensionality of the problem (p) is larger than the data set size (n), a common approach is to add an L_1 penalty term to the loss function to encourage selection of sparse minimizers. This formulation is commonly called LASSO *(Least Absolute Shrinkage and Selection Operator)* [26]. The formal optimization problem corresponding to the current formulation is given in (3). Here $\mathcal{C} \subseteq \mathbb{R}^p$ is some fixed convex set and Λ is some regularization parameter.

$$\hat{\boldsymbol{\theta}}(\mathcal{D}) = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \frac{\Lambda}{n}\|\boldsymbol{\theta}\|_1 \tag{3}$$

[27, 17] showed that under certain "niceness" conditions on the data set $\mathcal{D} = (\boldsymbol{y}, X)$ and the underlying parameter vector $\boldsymbol{\theta}^*$, the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ equals the support of $\boldsymbol{\theta}^*$ and $\|\hat{\boldsymbol{\theta}}(\mathcal{D}) - \boldsymbol{\theta}^*\|_2$ goes down to zero as n goes to infinity. This property is often referred to as *consistency*. In this paper we revisit the consistency assumptions (in [27]) for LASSO and relate the two different sets of assumptions sufficient for consistency, namely, *fixed data* and *stochastic* assumptions. Additionally, we weaken the *fixed data* assumptions that are sufficient for consistency.

An important property of any algorithm is the *stability* of its output with respect any changes in its input data. In this paper we study the stability properties of the support of the minimizer of a LASSO program. We follow a very strong notion of stability where the changes in the data set can be addition or removal of any constant (k) number of entries. At a high level, we show that *almost* under the same set of "niceness" conditions for consistency one can also guarantee stability.

In this work we study the consistency and stability properties of LASSO (and one of its variants) in two different settings: i) *fixed data setting* where the data set $\mathcal{D}$ is deterministic, and ii) *stochastic* where the dataset $\mathcal{D}$ is drawn from some underlying distribution. The general flavor of our results in this section is that we first prove the consistency and stability properties in the fixed data setting and then show one particular stochastic setting which satisfies the fixed data assumptions with high probability. The fixed data assumptions are given in Assumption 8 below.

Assumption 8 (Typical system)**.** *Data set $(X_{n\times p}, \boldsymbol{y}_{n\times 1})$ and parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$ are (s, Ψ, σ, Φ)-*TYPICAL *if there exists a $\boldsymbol{w} \in \mathbb{R}^p$ such that $\boldsymbol{y} = X\boldsymbol{\theta}^* + \boldsymbol{w}$ and*

(1) ***Column normalization:*** *$\forall j, \|c_j\|_2 \le \sqrt{n}$, where c_j is the j-th column of X.*

(2) ***Bounded parameter vector:*** *$\|\boldsymbol{\theta}^*\|_0 \le s$ and all nonzero entries of $\boldsymbol{\theta}^*$ have absolute value in $(\Phi, 1 - \Phi)$.*

(3) ***Incoherence****: Let Γ be the support of $\boldsymbol{\theta}^*$.*
$\|(X_{\Gamma^c}{}^T X_\Gamma)(X_\Gamma{}^T X_\Gamma)^{-1} sign(\boldsymbol{\theta}^*)\|_\infty < \frac{1}{4}$. Here $\Gamma^c = [p] - \Gamma$ is the complement of Γ; X_Γ is the matrix formed by the columns of X whose indices are in Γ; and $sign(\boldsymbol{\theta}^*) \in \{-1, 1\}^{|\Gamma|}$ is the vector of signs of the nonzero entries in $\boldsymbol{\theta}^*$.

(4) ***Restricted Strong Convexity:*** *The minimum eigenvalue of $X_\Gamma{}^T X_\Gamma$ is at least Ψn.*

(5) ***Bounded Noise:*** *$\|X_{\Gamma^c}^T V\boldsymbol{w}\|_\infty \le 2\sigma\sqrt{n\log p}$, where $V = \mathbb{I}_{n\times n} - X_\Gamma(X_\Gamma{}^T X_\Gamma)^{-1} X_\Gamma{}^T$ is the projector on to the complement of the column space of X_Γ.*

3.1 Consistency of LASSO Estimator

Under a strengthened version of the *fixed data conditions* above (Assumption 8), [27] showed that one can correctly recover the exact support of the parameter vector $\boldsymbol{\theta}^*$ and moreover the estimated parameter vector $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is close to $\boldsymbol{\theta}^*$ in the L_2 metric. Theorem 9 restates the result of [27] in the context of this paper. We note that the result of [27] holds even under this weaker assumption (Assumption 8).

Theorem 9 (Modified Theorem 1 of [27])**.** *Let* $\Lambda = 4\sigma\sqrt{n \log p}$*. If there exists a* $\boldsymbol{\theta}^*$ *such that* $(X, \boldsymbol{y}, \boldsymbol{\theta}^*)$ *is* (s, Ψ, σ, Φ)-$\mathsf{TYPICAL}$ *with* $\Phi = \frac{16\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}$*, then* $\|\hat{\boldsymbol{\theta}}(\mathcal{D}) - \boldsymbol{\theta}^*\|_2 \leq \frac{8\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}$*. Moreover, the support of* $\hat{\boldsymbol{\theta}}(\mathcal{D})$ *and* $\boldsymbol{\theta}^*$ *are same.*

Along with the fixed data setting, [27] considered the *stochastic setting* where the rows of the design matrix X are drawn from $\mathcal{N}(0, \frac{1}{4}\mathbb{I}_p)$ and the noise vector $\boldsymbol{w}$ is drawn independently from a mean zero sub-Gaussian distribution with variance σ^2. They showed that with high probability, under such a setting and choosing $\Lambda = 4\sigma\sqrt{n\log p}$, one has $\|\hat{\boldsymbol{\theta}}(\mathcal{D}) - \boldsymbol{\theta}^*\|_2 \leq \frac{8\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}$ and the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\boldsymbol{\theta}^*$ are same. The analysis of [27] in the stochastic setting relies on a different set of arguments compared to the arguments for the fixed data setting in Theorem 9. Moreover, it is not clear apriori if *any* stochastic setting satisfies the fixed data setting conditions with high probability. In the following theorem, we connect the stochastic and the fixed data setting, i.e., we show that under the stochastic setting considered above, with high probability, the data set $\mathcal{D} = (\boldsymbol{y}, X)$ satisfies the fixed data conditions. It should be mentioned here that [27] considered a more general distribution $\mathcal{N}(0, \Sigma)$ (with a specific class of covariance matrices). But for the purposes of brevity, we stick to the simpler $\mathcal{N}(0, \frac{1}{4}\mathbb{I}_p)$ distribution. The proof of this theorem is provided in Section B.1.

Theorem 10 (Stochastic Consistency)**.** *Let* $\Lambda = 4\sigma\sqrt{n\log p}$ *and* $n = \omega(s\log p)$*. If each row of the design matrix* X *be drawn i.i.d. from* $\mathcal{N}(0, \frac{1}{4}\mathbb{I}_p)$ *and each entry of the noise vector* $\boldsymbol{w}$ *be drawn i.i.d. from a mean zero sub-Gaussian distribution with variance* σ^2*, then there exists a constant* Ψ *such that with probability at least* $3/4$*, the data set* $\mathcal{D} = (\boldsymbol{y}, X)$ *obtained via* (1) *and under permissible choices of* $\boldsymbol{\theta}^*$ *in Assumption 8,* $(\boldsymbol{y}, X, \boldsymbol{\theta}^*)$ *satisfies* (s, Ψ, σ, Φ)-$\mathsf{TYPICAL}$ *with* $\Phi = \frac{16\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}$.

3.2 Normalization

Recall the linear system defined in (1). In the rest of the paper, we assume following normalization bounds on the data set $\mathcal{D} = (\boldsymbol{y}, X)$ and the underlying parameter vector $\boldsymbol{\theta}^*$. We assume that $\boldsymbol{\theta}^*$ is from the convex set $\mathcal{C} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_\infty \leq 1\}$. Often we restrict the convex set $\mathcal{C}$ to a support Γ (represented by $\mathcal{C}_\Gamma$). The set $\mathcal{C}_\Gamma$ is set of all vectors in $\mathcal{C}$ whose coordinates are zero outside Γ. Notice that since $\mathcal{C}$ is convex, $\mathcal{C}_\Gamma$ is also convex.

We assume that each entry of the design matrix X has absolute value at most one, i.e., $\|X\|_{\max} = \max_{1\leq i\leq n, 1\leq j\leq p} |X_{i,j}| \leq 1$, and additionally we assume that the response vector $\boldsymbol{y}$ has L_∞-norm at most s, i.e., $\|\boldsymbol{y}\|_\infty \leq s$. Notice that bounding the L_∞-norm of $\boldsymbol{y}$ is without loss of generality, since when the design matrix X and the parameter vector $\boldsymbol{\theta}^*$ are bounded as above, bounding $\boldsymbol{y}$ will only decrease the noise. In case the data set $\mathcal{D} = (\boldsymbol{y}, X)$ does not satisfy the above bound we *normalize* the data set to enforce such a bound. By *normalizing* we mean scaling down each data entry individually, so that they satisfy the above bound. For clarity of exposition, in the rest of the paper we define the universe of data sets U^* to be sets of entries from this domain and we will assume this normalization to be implicit in all the algorithms we state (unless mentioned otherwise).

3.3 Stability of LASSO Estimator in the Fixed Data Setting

In Section 3.1 we saw that under certain "niceness" conditions (Assumption 8) on the data set $\mathcal{D} = (\boldsymbol{y}, X)$, with suitable choice of regularization parameter Λ, one can ensure that the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ equals the support of $\boldsymbol{\theta}$. Moreover, $\|\hat{\boldsymbol{\theta}}(\mathcal{D}) - \boldsymbol{\theta}^*\|_2$ goes down to zero as $n \to \infty$ as long as $n = \omega(s\log p)$. In this section we ask the following question: *"Under what (further) assumptions on the data set* $\mathcal{D}$ *and the* *parameter vector $\boldsymbol{\theta}^*$, the support of the minimizer $\hat{\boldsymbol{\theta}}(\mathcal{D})$ does not change even if a constant k number of entries from the domain U are either added or removed from $\mathcal{D}$?"*

We answer this question in two different settings. In the first setting we analyze the stability properties of the original LASSO program in (3) where we show that under assumptions very similar to the one for consistency, the support of the minimizer $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is also stable. In the second setting we huberize the LASSO program (i.e., transform the program to make sure that the gradient of the objective function is always bounded.) This enables us to get better stability guarantees without compromising on the correctness of support selection.

3.3.1 Stability of unmodified LASSO

We show that under Assumption 8, the support of the minimizer $\hat{\boldsymbol{\theta}}(\mathcal{D})$ in (3) does not change even if k data entries are removed or added to $\mathcal{D}$ as long as $n = \omega(s \log p, \frac{s^4k^2}{\log p}, ks^{3/2})$. We call this property *k-stability* (Definition 2). Moreover, the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ equals the support of underlying parameter vector $\boldsymbol{\theta}^*$ (see (1)) and $\|\hat{\boldsymbol{\theta}}(\mathcal{D}) - \boldsymbol{\theta}^*\|_2$ goes down to zero as $n \to \infty$. It is important to note that Assumption 8 in particular is satisfied by a random Gaussian design matrix X and a sub-gaussian noise vector $\boldsymbol{w}$. We will discuss the stochastic setting for stability in Section 3.5.

The main stability theorem for LASSO is given in Theorem 11. For the purpose of clarity, we defer the complete proof of the stability theorem to Section B.2.1. The correctness follows directly from Theorem 9. It is important to note that our stability theorem bypasses the impossibility result of [29]. In their work, [29] showed that under worst case assumptions, minimizer of the LASSO program (i.e., $\hat{\boldsymbol{\theta}}(\mathcal{D})$ in (3)) does not have a stable support, i.e., the support changes with changing one entry in $\mathcal{D}$. Since, we work with stronger assumptions, the impossibility result does not apply to us.

Theorem 11 (Stability of unmodified LASSO)**.** *Fix $k \geq 1$. Suppose $s \leq \sqrt{\frac{\sigma n^{1/2} \log^{1/2} p}{2k(1/\Psi+1)}}$ and $\Lambda = 4\sigma\sqrt{n \log p}$. If there exists a $\boldsymbol{\theta}^*$ such that $(X, \boldsymbol{y}, \boldsymbol{\theta}^*)$ is (s, Ψ, σ, Φ)-*TYPICAL *with $\Phi = \max\left\{\frac{16\sigma}{\Psi}\sqrt{\frac{s \log p}{n}}, \frac{8ks^{3/2}}{\Psi n}\right\}$ (for the data set $\mathcal{D} = (\boldsymbol{y}, X)$ from U^*), then $\hat{\boldsymbol{\theta}}(\mathcal{D})$ has k-stable support.*

Proof sketch. For any data set $\mathcal{D}'$ differing in at most k entries from $\mathcal{D}$, we construct a vector $\boldsymbol{v}$ which has the same support as $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and then argue that $\boldsymbol{v} = \hat{\boldsymbol{\theta}}(\mathcal{D}')$, i.e., $\boldsymbol{v}$ is indeed the true minimizer of the LASSO program on $\mathcal{D}'$. The novelty in the proof goes in constructing the vector $\boldsymbol{v}$.

Let $\hat{\Gamma}$ be the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$. We obtain the vector $\boldsymbol{v}$ by minimizing the objective function $\hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}') + \Lambda\|\boldsymbol{\theta}\|_1$ restricted to the convex set $\mathcal{C}_{\hat{\Gamma}}$. Recall that all the vectors in $\mathcal{C}_{\hat{\Gamma}}$ have support in $\hat{\Gamma}$. Using the consistency result from Theorem 9 and a claim that shows that the L_2 distance between $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\boldsymbol{v}$ is small, we conclude that the support of $\boldsymbol{v}$ equals $\hat{\Gamma}$. By showing that under the assumptions of the theorem, the objective function at $\boldsymbol{v}$ has a zero sub-gradient, we conclude that $\boldsymbol{v} = \hat{\boldsymbol{\theta}}(\mathcal{D}')$.

We should mention here that a similar line of argument was used in the proof of Theorem 9 by [27] to argue consistency of LASSO estimators. Here we use it to argue stability of the support.

3.3.2 Stability of huberized LASSO

In this section, we modify the LASSO program of (3) in the following (4) to have better stability properties when $s = \Omega(\log n)$. The main idea is to huberize the loss function in order to control the gradient of the loss. Before providing the exact details of the huberization, we provide a toy example below to make the presentation clear.

Consider a simple quadratic function $f(x) = \frac{1}{2}x^2$ and a maximum gradient constraint of $\alpha \in \mathbb{R}$. One way to modify the function such that it satisfies the gradient constraint is by replacing $f(x)$ with the following.

$$\hat{f}(x) = \begin{cases} \alpha x - \frac{\alpha^2}{2} & \text{if } x > \alpha \\ \alpha x - \frac{\alpha^2}{2} & \text{if } x < -\alpha \\ \frac{1}{2}x^2 & \text{otherwise} \end{cases}$$

The two main properties of $\hat{f}$ are: i) it is continuously differentiable and ii) its gradient is always bounded by α. We will perform a similar transformation to the loss function for linear regression to control its gradient.

Recall that the loss function for linear regression is given by $\hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}) = \frac{1}{2n}\sum_{i=1}^{n}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle)^2$, where y_i is the i-th entry of the vector $\boldsymbol{y}$ and $\boldsymbol{x}_i$ is the i-th row of the design matrix X. We denote the function $(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta}\rangle)^2$ by $\ell(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i)$. Consider the following huberization of the loss function ℓ. For any given $y \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^p$, $\hat{\ell}(\boldsymbol{\theta}; y, \boldsymbol{x})$ is defined as follows. (Here s denotes the number of non-zero entries in the underlying parameter vector $\boldsymbol{\theta}^*$ in the linear system defined in (1).)

$$\hat{\ell}(\boldsymbol{\theta}; y, \boldsymbol{x}) = \begin{cases} 5\sqrt{s\log n}(y - \langle \boldsymbol{x}, \boldsymbol{\theta}\rangle) - 12.5 s\log n & \text{if } (y - \langle \boldsymbol{x}, \boldsymbol{\theta}\rangle) > 5\sqrt{s\log n} \\ -5\sqrt{s\log n}(y - \langle \boldsymbol{x}, \boldsymbol{\theta}\rangle) - 12.5 s\log n & \text{if } (y - \langle \boldsymbol{x}, \boldsymbol{\theta}\rangle) < -5\sqrt{s\log n} \\ \frac{1}{2}(y - \langle \boldsymbol{x}, \boldsymbol{\theta}\rangle)^2 & \text{otherwise} \end{cases}$$

$$\tilde{\boldsymbol{\theta}}(\mathcal{D}) = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \frac{1}{n}\sum_{i=1}^{n}\hat{\ell}(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i) + \frac{\Lambda}{n}\|\boldsymbol{\theta}\|_1 \tag{4}$$

In this section we show the correctness (Theorem 12) and stability property (Theorem 13) of $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ under Assumption TYPICAL (Assumption 8).

Theorem 12 (Correctness of huberized LASSO)**.** *Let $\Lambda = 4\sigma\sqrt{n\log p}$, let $\mathcal{D} = (\boldsymbol{y}, X)$ be a data set from U^* and $n = \omega(s\log p)$. If there exists a $\boldsymbol{\theta}^*$ such that for each row $\boldsymbol{x}_i$ in the design matrix X, $|\langle \boldsymbol{x}_i, \boldsymbol{\theta}^*\rangle| \leq 2\sqrt{s\log n}$, $(\boldsymbol{y}, X, \boldsymbol{\theta}^*)$ is (s, Ψ, σ, Φ)-*TYPICAL *with $\Phi = \frac{16\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}$, then the support of $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ matches the support of $\boldsymbol{\theta}^*$ and moreover $\|\tilde{\boldsymbol{\theta}}(\mathcal{D}) - \boldsymbol{\theta}^*\|_\infty \leq \frac{8\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}$.*

We defer the proof of this theorem till Section B.2.2. In the proof of Theorem 12 we show that under the assumptions of the theorem, the region where the unconstrained minimizer of the huberized LASSO estimator lies, the huberized loss function and the unmodified loss functions are the same. In Theorem 13 we show that as long as the data set size $n = \omega(s\log p, \frac{s^3k^2\log n}{\log p}, ks\sqrt{\log n})$, the support of $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ does not change even if a constant number (k) of data entries from U are removed or added in $\mathcal{D}$. The proof structure of Theorem 13 is same as the proof structure of Theorem 11 for the unmodified LASSO. For the purpose of brevity we defer the proof of Theorem 13 till Section B.2.2.

Theorem 13 (Stability of huberized LASSO)**.** *Fix $k > 1$. Under assumptions of Theorem 12 and $n = \omega(s\log p, \frac{s^3k^2\log n}{\log p})$, $(\boldsymbol{y}, X, \boldsymbol{\theta}^*)$ is (s, Ψ, σ, Φ)-*TYPICAL *with $\Phi = \max\left\{\frac{16\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}, \frac{20ks\sqrt{\log n}}{\Psi n}\right\}$, then $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ has a k-stable support.*

Function	Instantiation (Parameters: s, Λ, Ψ)	Threshold (t_i)	Slack (Δ_i)
$g_1(\mathcal{D})$	negative of the $(s+1)^{\text{st}}$ largest absolute value of $n \bigtriangledown \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}); \mathcal{D})$	$-\frac{\Lambda}{2}$	$\frac{6s^2}{\Psi}$
$g_2(\mathcal{D})$	minimum eigenvalue of $X_{\hat{\Gamma}}{}^T X_{\hat{\Gamma}}$	$2\Psi n$	s
$g_3(\mathcal{D})$	minimum absolute value of the non-zero entries in $\hat{\boldsymbol{\theta}}(\mathcal{D}) \times n$	$\frac{8s^{3/2}}{\Psi}$	$\frac{4s^{3/2}}{\Psi}$
$g_4(\mathcal{D})$	negative of the max. absolute value of the non-zero entries in $\hat{\boldsymbol{\theta}}(\mathcal{D}) \times n$	$\frac{8s^{3/2}}{\Psi} - n$	$\frac{4s^{3/2}}{\Psi}$

Table 1: Instantiation of the four test functions

3.4 Efficient Test for k-stability

In Section 3.3 we saw that under Assumption 8 and under proper asymptotic setting of the size of the data set (n) with respect to the parameters $s, \log p$ and k, both the unmodified LASSO in (3) and the huberized LASSO in (4) have k-stable support for their minimizers $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ respectively. An interesting question that arises is *"can we efficiently test the stability of the support of the minimizer, given a LASSO instance?"* In this section we design efficiently testable proxy conditions which allow us to test for k-stability of the support of a LASSO minimizer. For the ease of exposition, we present the results in the context of unmodified LASSO instance only. The result for the huberized LASSO follows analogously.

The main idea in designing the proxy conditions is to define a set of four test functions $g_1, \cdots, g_4$ (with each $g_i : U^* \to \mathbb{R}$) that have the following properties: i) For a given data set $\mathcal{D}$ from U^* and given set of thresholds $t_1, \cdots, t_4$, if each $g_i(\mathcal{D}) > t_i$, then adding or removing any one entry in $\mathcal{D}$ does not change the support of the minimizer $\hat{\boldsymbol{\theta}}(\mathcal{D})$. In other words, the minimizer $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is 1-stable. ii) Let $\Delta_1, \cdots, \Delta_4$ be a set of *slack* values. If each $g_i(\mathcal{D}) > t_i + (k-1)\Delta_i$, then the support of the minimizer $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is k-stable. In Table 1 we define the test functions (in the notation of LASSO from (3)) and the corresponding thresholds (t_i) and the slacks (s_i). There s refers to the sparsity parameter and $(s+1)^{\text{st}}$ largest absolute value of $n \bigtriangledown \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}); \mathcal{D})$ refers the $(s+1)$-st maximum absolute value of the coordinates from the vector $n \bigtriangledown \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}); \mathcal{D}) = -X^T(\boldsymbol{y} - \langle X, \hat{\boldsymbol{\theta}}(\mathcal{D})\rangle)$.

Design intuition. The main intuitions that govern the design on the proxy conditions in Table 1 are as follows. i) One needs to make sure that gradients of the loss function along the directions not in the support of the minimizer are sufficiently smaller than Λ/n, so that changing k data entries do not increase gradient beyond Λ/n, otherwise that particular coordinate will become non-zero. ii) Along the directions in the support of the minimizer, one needs to make sure that the objective function has sufficient strong convexity, so that changing k data entries do not move the minimizer along that direction too far. iii) On data sets where the minimizer has stable support, the *local sensitivity* [18] of the proxy conditions at $\mathcal{D}$ should be small. By local sensitivity we mean the amount by which the value of a proxy condition changes when one entry is added or removed from the data set $\mathcal{D}$.

Theorem 14 shows that the g_i's (with their corresponding thresholds t_i and slacks Δ_i) are efficiently testable proxy conditions for the k-stability of the support of the minimizer $\hat{\boldsymbol{\theta}}(\mathcal{D})$. For the purposes of brevity, we defer the proof of this theorem till Section B.2.3. Next in Theorem 16 we show that if the data set $\mathcal{D} = (\boldsymbol{y}, X)$ satisfies a slight strengthening of Assumption 8 (see Assumption 15), then for all $i \in \{1, \cdots, 4\}$, $g_i(\mathcal{D}) > t_i + (k-1)\Delta_i$. This ensures that the proxy conditions are almost as good as the *fixed data conditions* in Assumption 8. In Section 3.5 we analyze a stochastic setting where Assumption 15 is satisfied with high probability.

Theorem 14 (k-stability (proxy version)). *Let $\mathcal{D}$ be a data set from U^*. If $g_i(\mathcal{D}) > t_i + (k-1)\Delta_i$ for all $i \in \{1, \cdots, 4\}$ and $\Lambda > \frac{16ks^2}{\Psi}$, then $\hat{\boldsymbol{\theta}}(\mathcal{D})$ has k-stable support.*

Assumption 15 (Super-typical system)**.** *Data set $(X_{n\times p}, \boldsymbol{y}_{n\times 1})$ and parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$ are (s,Ψ,σ,Φ,k)-*STRONGLY-TYPICAL *if there exists a $\boldsymbol{w} \in \mathbb{R}^p$ such that $\boldsymbol{y} = X\boldsymbol{\theta}^* + \boldsymbol{w}$ and*

(1) $(\boldsymbol{y}, X, \boldsymbol{\theta}^*)$ *is* (s,Ψ,σ,Φ)-TYPICAL.

(2) ***Restricted Strong Convexity:*** *The minimum eigenvalue of $X_\Gamma{}^T X_\Gamma$ is at least $\hat{\Psi}n$, where $\hat{\Psi}n = 2\Psi n + (k-1)s$.*

(3) ***Bounded Noise:*** *For any set Γ of size s, $\|X_{\Gamma^c}^T V\boldsymbol{w}\|_\infty \leq 2\sigma\sqrt{n\log p} - 6(k-1)s^2/\Psi$, where $V = \mathbb{I}_{n\times n} - X_\Gamma(X_\Gamma{}^T X_\Gamma)^{-1}X_\Gamma{}^T$ is the projector on to the complement of the column space of X_Γ.*

Theorem 16. *Let $\mathcal{D} = (\boldsymbol{y}, X)$ be a data set from U^* and let $\Lambda = 4\sigma\sqrt{n\log p}$. If there exists a $\boldsymbol{\theta}^*$ such that $(\boldsymbol{y}, X, \boldsymbol{\theta}^*)$ is (s,Ψ,σ,Φ,k)-*STRONGLY-TYPICAL *with $\Phi = \max\left\{\frac{16\sigma}{\Psi}\sqrt{\frac{s\log p}{n}},\ \frac{16ks^{3/2}}{\Psi n}\right\}$, then $g_i > t_i + (k-1)\Delta_i$ for all $i \in \{1,\cdots,4\}$.*

The proof of this theorem follows using an intuition very similar to that used in the proof of Theorem 11. For the sake of clarity, we defer the proof till Section B.2.3.

3.5 Stability of LASSO in Stochastic Setting

In Section 3.3 we saw two variants of LASSO (*unmodified* and *huberized*) and a set of conditions (from Theorems 11 and 13) under which we argued that if the data set size n is sufficiently large compared to $(s,k,\log p)$, then the minimizers of the two LASSO programs ((3) and (4)) are k-stable. In Section 3.4 we saw a strengthened set of assumptions (in Theorem 16) which implies that under these assumptions, the data set $\mathcal{D}$ will pass the efficient k-stability test designed in Section 3.4.

In this section we will see one specific stochastic setting for the data set $\mathcal{D} = (\boldsymbol{y}, X)$, where the set of conditions (in Theorems 11, 13 and 16) are satisfied with high probability. The specific stochastic setting we consider here is the same we considered for consistency in Section 3.1. Consider each row of the design matrix X is drawn i.i.d. from $\mathcal{N}(0, \frac{1}{4}\mathbb{I}_p)$ and the entries in the noise vector $\boldsymbol{w}$ is drawn i.i.d. from a mean zero sub-Gaussian distribution with variance σ^2.

Analysis of unmodified LASSO in stochastic setting. In order to make sure that Theorem 11 is applicable in the stochastic setting, we need to ensure two things: i) the data set $\hat{\mathcal{D}} = (\hat{\boldsymbol{y}}, \hat{X})$ that gets used in Theorem 11 is from the domain U^*, and ii) $(\hat{\boldsymbol{y}}, \hat{X}, \boldsymbol{\theta}^*)$ satisfy (s,Ψ,σ,Φ,k)-STRONGLY-TYPICAL. This in particular implies that $(\hat{\boldsymbol{y}}, \hat{X}, \boldsymbol{\theta}^*)$ satisfy (s,Ψ,σ,Φ)-TYPICAL.

Given the data set $\mathcal{D} = (\boldsymbol{y}, X)$ drawn from the distribution mentioned above, we first divide each entry in the design matrix X by $\sqrt{\log(ns)}$, where s is the sparsity parameter of the parameter vector $\boldsymbol{\theta}^*$. If the absolute value of any entry in X after dividing by $\sqrt{\log(ns)}$ exceeds 1, then just round it to -1 or 1 (whichever is closer). Call this design matrix $\hat{X}$. Similarly, if the absolute value of any entry in $\boldsymbol{y}$ exceeds s, then round it to $-s$ or s whichever is closer. By union bound and the tail property of Gaussian distribution it follows that once each entry of the design matrix X is divided by $\sqrt{\log(ns)}$, with high probability (i.e., with probability at least $1 - e^{-4}$) none of the columns which are in the support of $\boldsymbol{\theta}^*$ gets truncated. Conditioned on this event, with probability at least $15/16$, the design matrix $\hat{X}$ satisfies *column normalization* condition and *restricted strong convexity* condition in Assumption 15 with parameter Ψ' (as long as $n = \omega(ks\log n)$), where $\Psi' = \Psi/\sqrt{\log(ns)}$ and Ψ is the restricted strong convexity parameter corresponding to random Gaussian design matrix. Also by similar arguments as in the proof of Theorem

10, it follows that as long as $n = \omega(s \log p \log n, k^2 s^4 / \log p)$, with probability at least $7/8$, the *incoherence* and *bounded noise* conditions are satisfied. Thus, we have the following stochastic analogue of Theorem 11. We do not need to argue about the truncation of the entries in $\boldsymbol{y}$, since the truncation can be viewed as reducing the noise $\boldsymbol{w}$.

Corollary 17. *Fix $k \geq 1$. Let $\Lambda = 4\sigma\sqrt{n \log p}$ and $n = \omega(s \log p \log n, \frac{s^4 k^2 \log n}{\log p}, ks \log n)$. There exists a constant Ψ such that under the assumptions $\|\boldsymbol{\theta}^*\|_0 \leq s$, and the absolute value of any non-zero entry of $\boldsymbol{\theta}^*$ is in $(\Phi, 1 - \Phi)$ (for $\Phi = \max\left\{\frac{16\sigma}{\Psi}\sqrt{\frac{s \log n \log p}{n}}, \frac{16ks^{3/2}\sqrt{\log n}}{\Psi n}\right\}$), with probability at least $3/4$, the tuple $(\hat{\boldsymbol{y}}, \hat{X}, \boldsymbol{\theta}^*)$ satisfy $(s, \Psi, \sigma, \Phi, k)$-*STRONGLY-TYPICAL *assumption.*

The above theorem implies that as long as $\frac{n}{\log n} = \omega(s \log p, \frac{s^4 k^2}{\log p}, ks^{3/2})$, with high probability the support of the minimizer $\hat{\boldsymbol{\theta}}(\hat{\mathcal{D}})$ is k-stable. The analysis for huberized LASSO is analogous and is omitted for brevity.

4 Private Support Selection for Sparse Linear Regression

In Section 2 we designed a generic framework to transform a model selection function $f : U^* \to \mathcal{R}$ (which takes the data set $\mathcal{D}$ and outputs a model $\Gamma \in \mathcal{R}$) to an efficient differentially private algorithm for model selection if we have proxy functions $\hat{f}$and $\hat{d}$ for f and its distance to instability (see Definition 4). In this section we use this framework for support selection in sparse linear regression. Note that in the context of linear regression (with sparsity parameter s for the underlying parameter vector $\boldsymbol{\theta}^*$) one can view the space of all possible models $\mathcal{R}$ to be all the $\binom{p}{s}$ sets of coordinates from $[p]$. Once a support of size s is chosen, one can restrict the linear regression problem to the set of s-coordinates chosen and then use algorithms (*e.g.,* objective perturbation) for private linear regression from [13] to obtain a parameter vector $\boldsymbol{\theta}^{\text{priv}}$ such that $\hat{\mathcal{L}}(\boldsymbol{\theta}^{\text{priv}}; \mathcal{D}) - \hat{\mathcal{L}}(\boldsymbol{\theta}^*; \mathcal{D})$ scales as $O\left(\frac{s^2 \log(1/\delta)}{n\epsilon}\right)$. For more details, see Theorem 9 in [13].

The main challenge in designing a private support selection algorithm is to come up with effective proxies $\hat{f}$ and $\hat{d}$ for a given support selection function f. In the following two sections we design two different sets of proxies $\hat{f}$ and $\hat{d}$. Later we compare the sample complexities of the algorithms corresponding to both. In the current discussion, we set f to be the function that returns the support of the minimizer of unmodified LASSO program in (3). Although the results in this section can be easily extended to the huberized LASSO program in (4), we do not present it for brevity.

4.1 Support Selection via Sampling Stability

We use the same $\hat{f}$ and $\hat{d}$ used in Lemma 7 and use Algorithm 1 for support selection. In the current context, the non-private model selection function f in Algorithm 1 is the function that returns the support of the minimizer of unmodified LASSO program in (3). By Theorem 6, the output is always (ϵ, δ)-differentially private. In order to argue that Algorithm 1 outputs the correct support, we make the following assumption (Assumption 18) about the data set $\mathcal{D}$ and the parameter vector $\boldsymbol{\theta}^*$. Under this assumption, we obtain the following utility guarantee (Corollary 19) for the support selection algorithm as a corollary to Theorem 6.

Assumption 18. *[(s, Ψ, σ, Φ)-Sub-sampled TYPICAL] Let $\hat{\mathcal{D}}$ be a random subset of $\mathcal{D} = (\boldsymbol{y}, X)$ in which each element appears independently with probability $q = \frac{\epsilon}{32 \log(1/\delta)}$. The data set $\hat{\mathcal{D}}$ and parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$ satisfy*
(s, Ψ, σ, Φ)-TYPICAL with probability at least $3/4$.

It is important to note that the above assumption is satisfied by the stochastic setting in Section 3.5 with high probability.

Corollary 19 (Utility)**.** *Let $\Lambda = 8\sigma\sqrt{nq\log p}$ where $q = \frac{\epsilon}{32\log(1/\delta)}$. If there exists a θ^* such that the data set $\mathcal{D} = (\boldsymbol{y}, X)$ and $\boldsymbol{\theta}^*$ satisfy Assumption 18 (Assumption (s, Ψ, σ, Φ)-*Sub-sampled TYPICAL*) with $\Phi \geq \frac{16\sigma}{\Psi}\sqrt{\frac{s\log p}{nq}}$, then w.p. at least $1 - 3\delta$, the current instantiation of Algorithm 1 outputs the correct support of $\boldsymbol{\theta}^*$.*

In order to analyze the sample complexity for support selection implied by the corollary above, first note that in expectation the sub-sampled data set $\hat{\mathcal{D}}$ will be of size nq, where n is the size of the original data set $\mathcal{D}$. Therefore, by sub-sampling we have blown up the sample complexity by a factor of $1/q$ with respect to the non-private sample complexity implied by Theorem 9. Hence, the sample complexity for consistent support selection, implied by the above corollary, is $(s\log p)/q$.

4.2 Support Selection via Stability of LASSO

In Section 3.3.1 we analyzed the stability properties of the unmodified LASSO program in (3). Moreover, in Section 3.4 we designed an efficient test for k-stability via defining four proxy conditions $g_1, \cdots, g_4$ in Table 1. In this section we transform it to a differentially private algorithm for outputting the support. In the language of Section 2 (and using the notation of Section 3.4) the proxy functions $\hat{f}$ and $\hat{d}$ we define are: i) $\hat{d}(\mathcal{D}) = \max\left\{\min_i \frac{g_i(\mathcal{D})-t_i}{\Delta_i} + 1, 0\right\}$ and ii) $\hat{f}(\mathcal{D})$ equals the support of the minimizer of the LASSO program in (3) when the data set $\mathcal{D}$ is stable, and $\perp$ o.w.

Lemma 20. *If $\Lambda > \frac{16s^2}{\Psi}$ (in (3)), then the proxy functions $\hat{f}$ and $\hat{d}$ defined above satisfy Definition 4.*

Proof. In Theorem 14 we saw that if for all i, $g_i(\mathcal{D}) > t_i + (k-1)\Delta_i$, then the data set $\mathcal{D}$ is k-stable. This straight away implies that if $\hat{d}(\mathcal{D}) > k$, then the data set $\mathcal{D}$ is k-stable w.r.t. the support of the minimizer.

To complete the proof, we need to show that the global sensitivity of $\hat{\mathcal{D}}$ is at most one. When $\hat{d}(\mathcal{D})$ is greater than or equal to zero, changing one entry in $\mathcal{D}$ changes $\hat{d}(\mathcal{D})$ by at most one, since one can show that in such a case each g_i changes by at most Δ_i. (See Claims 35, 36, and 37 in Section B.2.3.) Now since $\hat{d}$ cannot be negative, global sensitivity of $\hat{d}$ is at most one. □

With Lemma 20 in hand, the algorithm for support selection follows from Section 2. Add $Lap(1/\epsilon)$ noise to $\hat{d}(\mathcal{D})$ and then test if it is greater than $\log(1/\delta)/\epsilon$. If the answer is "yes", then output the exact support of the minimizer $\hat{\boldsymbol{\theta}}(\mathcal{D})$. By Proposition 3, the above algorithm is (ϵ, δ)-differentially private. Moreover, whenever $\hat{d}(\mathcal{D})$ is greater than $2\log(1/\delta)/\epsilon$, the algorithm outputs $f(\mathcal{D})$ with probability $1-\delta$. We obtain the following corollary.

Corollary 21. *Let $\mathcal{D} = (\boldsymbol{y}, X)$ be a data set from U^* and let $\Lambda = 4\sigma\sqrt{n\log p}$. If there exists a $\boldsymbol{\theta}^*$ such that $(\boldsymbol{y}, X, \boldsymbol{\theta}^*)$ is $(s, \Psi, \sigma, \Phi, k)$-*STRONGLY-TYPICAL* with $\Phi = \max\left\{\frac{16\sigma}{\Psi}\sqrt{\frac{s\log p}{n}},\ \frac{16ks^{3/2}}{\Psi n}\right\}$, where $k = 2\log(1/\delta)/\epsilon$, then the above algorithm outputs the correct support of $\boldsymbol{\theta}^*$ with probability at least $1-\delta$.*

Hence, it directly follows that the sample complexity for consistent support selection is $(s\log p, k^2s^4/\log p, ks^{3/2})$, where $k = \log(1/\delta)/\epsilon$. It is important to note that the assumption in the above corollary is satisfied by the stochastic setting in Section 3.5 with high probability.

Comparing to the sample complexity obtained in Section 4.1, we find that when the sparsity parameter s is greater than $\log^2 p$, the *sampling* based approach has better sample complexity. When s is small (i.e., $s < \frac{\log^{2/3} p}{k^{1/3}}$), the *stability of LASSO* based approach has better sample complexity.

Note on optimal sample complexity. [27] mentioned that in the stochastic setting (i.e., the setting in Section 3.5) any non-private algorithm for consistent recovery of the support of $\boldsymbol{\theta}^*$ will have sample complexity of at least $s \log p$. (See Section D in [27] for a detailed discussion.) Comparing to the sample complexities of our private algorithms we see that our *sampling* based algorithm matches the non-private lower bound on sample complexity up to factors in ϵ and $\log(1/\delta)$. Similarly, when $s < \frac{log^{2/3}p}{k^{2/3}}$ and $k < \log^{2/3} p$, the *stability of LASSO* based algorithm matches the sample complexity lower bound (without any dependence on the privacy parameters ϵ and δ).

References

[1] F. R. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *ICML*, 2008.

[2] S. Ben-David, U. Von Luxburg, and D. Pál. A sober look at clustering stability. *Learning Theory*, 2006.

[3] O. Bousquet and A. Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2: 499 – 526, 2002.

[4] K. Chaudhuri and D. Hsu. Convergence rates for differentially private statistical estimation. In *ICML*, 2012.

[5] C. Dwork. Differential privacy. In *ICALP*, 2006.

[6] C. Dwork and J. Lei. Differential privacy and robust statistics. In *STOC*, 2009.

[7] C. Dwork, K. Kenthapadi, F. Mcsherry, I. Mironov, and M. Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 2006.

[8] C. Dwork, F. McSherry, K. Nissim, and A. Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006.

[9] S. R. Ganta, S. P. Kasiviswanathan, and A. Smith. Composition attacks and auxiliary information in data privacy. In *KDD*, 2008.

[10] P. Huber. *Robust Statistics*. Wiley, 1981.

[11] V. Karwa, S. Raskhodnikova, A. Smith, and G. Yaroslavtsev. Private analysis of graph structure. *PVLDB*, 2011.

[12] S. P. Kasiviswanathan and A. Smith. A note on differential privacy: Defining resistance to arbitrary side information. *CoRR*, arXiv:0803.39461 [cs.CR], 2008.

[13] D. Kifer, A. Smith, and A. Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *COLT*, 2012.

[14] Y. Lee, S. N. MacEachern, and Y. Jung. Regularization of case-specific parameters for robustness and efficiency. Technical report, Statistics Deepartment, Ohio State University, April 2011.

[15] M. Meilă. The uniqueness of a good optimum for k-means. In *ICML*, pages 625–632, 2006.

[16] N. Meinshausen and P. Bhlmann. High dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 2006.

[17] S. Negahban, P. Ravikumar, M. J. Wainwright, and B. Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *NIPS*, 2009.

[18] K. Nissim, S. Raskhodnikova, and A. Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, 2007.

[19] A. Rakhlin and A. Caponnetto. Stability of k-means clustering. In *NIPS*, 2007.

[20] G. Raskutti, M. J. Wainwright, and B. Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. 2011.

[21] W. Rogers and T. Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, 1978.

[22] B. I. P. Rubinstein, P. L. Bartlett, L. Huang, and N. Taft. Learning in a large function space: Privacy-preserving mechanisms for svm learning. *CoRR*, abs/0911.5708, 2009.

[23] S. Shalev-Shwartz, O. Shamir, N. Srebro, and K. Sridharan. Learnability, stability and uniform convergence. *The Journal of Machine Learning Research*, 2010.

[24] J. Shao. Bootstrap model selection. *Journal of the American Statistical Association*, 1996.

[25] A. Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *STOC*, 2011.

[26] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1996.

[27] M. J. Wainwright. Sharp thresholds for high-dimensional and noisy recovery of sparsity using ℓ_1-constrained quadratic programs. In *IEEE Transactions on Information Theory*, 2006.

[28] H. Xu, C. Caramanis, and S. Mannor. Robust regression and lasso. *Information Theory, IEEE Transactions on*, 2010.

[29] H. Xu, C. Caramanis, and S. Mannor. Sparse algorithms are not stable: A no-free-lunch theorem. In *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012.

[30] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 2007.

A Stability and Privacy

A.1 Proof of Proposition 3

Proof of part (1). Note that Algorithm $\mathcal{A}_{dist}$ can have only two possible outputs: $\perp$ or $f(\mathcal{D})$. We show that for each of the outputs, the differential privacy condition holds. Firstly, since the true distance d can change by at most one if one entry is removed (added) from (to) the data set $\mathcal{D}$, therefore, by the following theorem (*Laplace mechanism*) from [8], the variable $\tilde{d}$ (in Algorithm $\mathcal{A}_{dist}$) satisfies $(\epsilon, 0)$-differential privacy.

Theorem 22 (Laplace Mechanism [8]). *Let $f : U^* \to \mathbb{R}$ be a function (with U^* being the domain of data sets). If for any pair of data sets $\mathcal{D}$ and $\mathcal{D}'$ with symmetric difference at most one, $|f(\mathcal{D}) - f(\mathcal{D}')| \leq 1$, then the output $\mathcal{A}(\mathcal{D}) = f(\mathcal{D}) + Lap\left(\frac{1}{\epsilon}\right)$ is $(\epsilon, 0)$-differentially private.*

Since we have shown $\tilde{d}$ is $(\epsilon, 0)$-differentially private, it follows that for any pair of data sets $\mathcal{D}$ and $\mathcal{D}'$ differing in one entry, differential privacy condition holds for the output $\perp$, i.e.,

$$\Pr[\mathcal{A}_{dist}(\mathcal{D}) = \perp] \leq e^{\epsilon} \Pr[\mathcal{A}_{dist}(\mathcal{D}') = \perp]$$

Notice that by the tail property of Laplace distribution, it follows that if $\tilde{d} > \frac{\log(1/\delta)}{\epsilon}$, then with probability at least $1-\delta$ the actual distance d is greater than zero. Define the event E equal to be true, if the noise $\mathsf{Lap}(1/\epsilon)$ is greater than $\frac{1}{\epsilon}\log(1/\delta)$. Then, we have,

$$\begin{aligned}\Pr[\mathcal{A}_{dist}(\mathcal{D}) = f(\mathcal{D})] &\leq \Pr[\mathcal{A}_{dist}(\mathcal{D}) = f(\mathcal{D}) \wedge \bar{E}] + \Pr[E] \\ &\leq \Pr[\mathcal{A}_{dist}(\mathcal{D}') = f(\mathcal{D}) \wedge \bar{E}] + \delta \\ &\leq \Pr[\mathcal{A}_{dist}(\mathcal{D}') = f(\mathcal{D})] + \delta\end{aligned}$$

Thus, we can conclude that Algorithm $\mathcal{A}_{dist}$ is (ϵ, δ)-differentially private.

□

Proof of Part (2). By the tail property of Laplace distribution, if the true distance d is at least $\frac{1}{\epsilon}(\log(1/\delta) + \log(1/\beta))$, then with probability at least $1-\beta$, the noisy distance $\tilde{d}$ is greater than $\frac{1}{\epsilon}\log(1/\delta)$. Hence with probability at least $1-\beta$, $f(\mathcal{D})$ is output. □

A.2 Proof of Lemma 7

Proof. Proof of part (1) of the lemma follows by a direct application of Chernoff-Hoeffding's bound. To prove part (2), notice that conditioned on the event E adding or removing one entry in the original data set changes any of the counts $count_{(r)}$ by at most $2mq$. Therefore, $count_{(1)} - count_{(2)}$ changes by at most $4mq$. This in turn means that $\hat{d}(\mathcal{D})$ changes by at most one for any $\mathcal{D}$ and hence have global sensitivity of one. This also implies that $\hat{d}$ lower bounds the stability of $\hat{f}$ on $\mathcal{D}$. To prove part (3), notice that when $\hat{d}(D) \geq 1/16q$, it implies that $count_{(1)} - count_{(2)} \geq m/4$. Thus, if we bound the probability of the highest bin having count less than $5/8m$ by $1-\delta$, then we are done. Recall that in expectation the highest bin has count at least $3/4m$. Now the remaining proof follows directly via the application of Chernoff-Hoeffding's bound. □

B Consistency and Stability of Sparse Linear Regression via LASSO

B.1 Consistency of LASSO Estimator

Proof of Theorem 10 (Stochastic Consistency). In the following we show that each of the Conditions 1, 3, 4, and 5 in Assumption 8 are satisfied with probability at least $15/16$. By union bound over the failure probabilities of these events, this will straightaway imply Theorem 10.

- **Column normalization condition:** Since we assumed $n = \Omega(s \log p)$, by tail bound over the norm of random Gaussian vectors, with probability at least $15/16$, the *column normalization condition* is satisfied.
- **Restricted strong convexity (RSC):** By Proposition 1 from [20], it directly follows that there exists a constant Ψ such that with probability at least $15/16$ the minimum eigenvalue of $X_\Gamma^T X_\Gamma$ is at least Ψn.
- **Incoherence:** Let us represent the vector $(X^T{}_\Gamma X_\Gamma)\text{sign}(\boldsymbol{\theta}^*)$ to be $\boldsymbol{u}$. Recall that by definition $\|\text{sign}(\boldsymbol{\theta}^*)\|_\infty \leq 1$. Hence, by the RSC property above, $\|\boldsymbol{u}\|_2 \leq \frac{\sqrt{s}}{\Psi n}$, which implies that $\|\boldsymbol{u}\|_\infty \leq \frac{\sqrt{s}}{\Psi n}$.

Let $\boldsymbol{a_i}$ be the i-th column of the matrix X_{Γ^c} and $\boldsymbol{b_i}$ be the i-th column of the matrix X_Γ Now for any row $j \in [p-s]$,

$$|\left({X^T}_{\Gamma^c} X_\Gamma \boldsymbol{u}\right)_j| = \left|\sum_{i\in[s]} u_i \langle \boldsymbol{a_j}, \boldsymbol{b_i}\rangle\right| = \left|\langle \boldsymbol{a_j}, \sum_{i\in[s]} u_i \boldsymbol{b_i}\rangle\right| \tag{5}$$

Notice that $\sum_{i\in[s]} u_i \boldsymbol{b_i} = X_\Gamma \boldsymbol{u}$. Therefore, $\| \sum_{i\in[s]} u_i \boldsymbol{b_i}\|_2 \leq |\text{largest singular value of } X_\Gamma| \cdot \|\boldsymbol{u}\|_2$. It is well known from random matrix theory that with probability at least $1 - e^{-n}$, the largest singular value of X_Γ is at most $\sqrt{n}$. Therefore, it follows that $\| \sum_{i\in[s]} u_i \boldsymbol{b_i}\|_2 \leq \frac{1}{\Psi}\sqrt{\frac{s}{n}}$. Since $\boldsymbol{a_j} \sim \mathcal{N}(0, \frac{1}{4}\mathbb{I}_p)$, $\left|\langle \boldsymbol{a_j}, \sum_{i\in[s]} u_i \boldsymbol{b_i}\rangle\right|$ in (5) is sub-Gaussian with standard deviation at most $\frac{1}{\Psi}\sqrt{\frac{s}{n}}$. Therefore by the tail property of sub-Gaussian random variables, with probability at most $\frac{1}{p}$, $\left|\langle \boldsymbol{a_j}, \sum_{i\in[s]} u_i \boldsymbol{b_i}\rangle\right| \leq \frac{1}{\Psi}\sqrt{\frac{s\log p}{n}}$. Taking union bound over all the possible columns in X_{Γ^c}, as long as $n = \omega(s \log p)$, we obtain the required *incoherence* condition with probability at least $15/16$.

- **Bound $\|X_{\Gamma^c}^T V \boldsymbol{w}\|_\infty \leq 2\sigma\sqrt{n \log p}$:** From the column normalization condition, we know that with probability at least $15/16$ each column of X_{Γ^c} has L_2-norm of at most $\sqrt{n}$. Let $\boldsymbol{\tilde{a}_i}$ be the random variable for the $i \in [p-s]$-th entry of the vector $X_{\Gamma^c}^T V \boldsymbol{w}$. Notice that (over the randomness of $\boldsymbol{w}$) $\tilde{a}_i$ is sub-Gaussian with standard deviation at most $\sigma\sqrt{n}$. Therefore, using the tail property of sub-Gaussian random variables and taking an union bound over all the columns of X_{Γ^c}, with probability at least $15/16$, we get the required bound $\|X_{\Gamma^c}^T V \boldsymbol{w}\|_\infty \leq 2\sigma\sqrt{n \log p}$.

□

B.2 Stability of LASSO Estimator in the Fixed Data Setting

B.2.1 Proof of Theorem 11 (Stability of unmodified LASSO)

Proof of Theorem 11 follows directly from the following two lemmas and a claim (Lemmas 23 and 24 and Claim 25). The main idea is to show that under Assumption (s, Ψ, σ, Φ)-TYPICAL with $\Phi = \max\left\{\frac{16\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}, \frac{8ks^{3/2}}{\Psi n}\right\}$, changing k entries in $\mathcal{D}$ does not change the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$.

Lemma 23. *Under the assumptions of Theorem 11 if $\hat{\Gamma}$ is the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\hat{\boldsymbol{\theta}}(\mathcal{D})_{\hat{\Gamma}} = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}_{\hat{\Gamma}}} \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \frac{\Lambda}{n}\|\boldsymbol{\theta}\|_1$, then $\hat{\boldsymbol{\theta}}(\mathcal{D})_{\hat{\Gamma}}$ equals $\hat{\boldsymbol{\theta}}(\mathcal{D})$.*

For the ease of notation, we denote $\hat{\boldsymbol{\theta}}(\mathcal{D})_{\hat{\Gamma}}$ by $\boldsymbol{z}$.

Lemma 24. *Let $\mathcal{D}' = (\boldsymbol{y}', X')$ be a data set formed by inserting (removing) k entries in the data set $\mathcal{D}$ from the domain U and let $\boldsymbol{z}' = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}_{\hat{\Gamma}}} \frac{1}{2|\mathcal{D}'|}\|\boldsymbol{y}' - X'\boldsymbol{\theta}\|_2^2 + \frac{\Lambda}{|\mathcal{D}'|}\|\boldsymbol{\theta}\|_1$. Under assumptions of Lemma 23, $\boldsymbol{z}' = \hat{\boldsymbol{\theta}}(\mathcal{D}')$, where $\hat{\boldsymbol{\theta}}(\mathcal{D}') = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \frac{1}{2|\mathcal{D}'|}\|\boldsymbol{y}' - X'\boldsymbol{\theta}\|_2^2 + \frac{\Lambda}{|\mathcal{D}'|}\|\boldsymbol{\theta}\|_1$.*

To prove the above lemma, we use a proof technique which was developed by [27] under the name of *primal-dual construction* and was used to argue consistency in non-private sparse linear regression.

Claim 25. *Under assumptions of Lemma 24, $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\hat{\boldsymbol{\theta}}(\mathcal{D}')$ have the same support.*

In the following we provide the proofs of the above two lemmas and the claim.

Proof of Lemma 23. In order to prove this lemma, we first prove that the minimizer $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is unique. We use Theorem 9 (which is a modified version of Theorem 1 from [27]) to prove the above claim.

Since from Theorem 9 we have $\|\hat{\boldsymbol{\theta}}(\mathcal{D}) - \boldsymbol{\theta}^*\|_\infty \leq \Phi$, it follows that $\hat{\boldsymbol{\theta}}(\mathcal{D})$ lies in the interior of the set $\mathcal{C}$. This in turn implies that the objective function $\frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \frac{\Lambda}{n}\|\boldsymbol{\theta}\|_1$ has a sub-gradient of zero at $\hat{\boldsymbol{\theta}}(\mathcal{D})$. Additionally, notice that by assumption, the objective function restricted to the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is strongly convex, since the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\boldsymbol{\theta}^*$ are same. These two observations along with the fact that the gradient of the objective function just outside $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is at least Λ (on the subspace orthogonal to the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$) imply that the gradient of the objective function just outside $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is strictly greater than zero. Hence, $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is the unique minimizer.

By the restricted strong convexity property of the objective function, $\frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \frac{\Lambda}{n}\|\boldsymbol{\theta}\|_1$ has an unique minimizer $\hat{\boldsymbol{\theta}}(\mathcal{D})_{\hat{\Gamma}}$ in $\mathcal{C}_{\hat{\Gamma}}$. Now, if $\hat{\boldsymbol{\theta}}(\mathcal{D})_{\hat{\Gamma}}$ does not equal $\hat{\boldsymbol{\theta}}(\mathcal{D})$, then it contradicts that $\hat{\boldsymbol{\theta}}(\mathcal{D}) = \arg\min\limits_{\boldsymbol{\theta}\in\mathcal{C}} \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \frac{\Lambda}{n}\|\boldsymbol{\theta}\|_1$. □

Proof of Lemma 24. For the ease of notation, we fix the following: i) $\hat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{D}) = \frac{1}{n}\sum_{i=1}^n \ell(\boldsymbol{\theta}; d_i)$, where $d_i = (y_i, \boldsymbol{x}_i)$, y_i is the i-th entry of $\boldsymbol{y}$ and $\boldsymbol{x}_i$ is the i-th row of X, ii) we denote $\hat{\boldsymbol{\theta}}(\mathcal{D})_{\hat{\Gamma}}$ by $\boldsymbol{z}$. Also, since by Theorem 9, $\hat{\Gamma}$ equals the support of $\boldsymbol{\theta}^*$ (i.e., Γ^*), we fix $\hat{\Gamma} = \Gamma^*$.

Let $\boldsymbol{z}' = \arg\min\limits_{\boldsymbol{\theta}\in\mathcal{C}_{\Gamma^*}} \hat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{D}') + \frac{\Lambda}{n+k}\|\boldsymbol{\theta}\|_1$. W.l.o.g. assume that $\mathcal{D}'$ has k entries more than $\mathcal{D}$ and call these entries $\alpha_1, \cdots, \alpha_k$. (The analysis for the case when $\mathcal{D}'$ has k entries less than $\mathcal{D}$ follows analogously.) In the following claim we show that $\boldsymbol{z}'$ does not differ too much from $\boldsymbol{z}$ in the L_2-metric.

Claim 26. $\|\boldsymbol{z} - \boldsymbol{z}'\|_2 \leq \frac{4ks^{3/2}}{\Psi n}$.

Proof. By restricted strong convexity of $\hat{\mathcal{L}}$ at $\boldsymbol{z}$ in a ball (in the subspace formed by the support set Γ^*) of radius $\frac{2k\zeta}{\Psi n}$ around it, we have the following.

$$\begin{aligned}
n\hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D}) + \Lambda\|\boldsymbol{z}'\|_1 &\geq n\hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D}) + \Lambda\|\boldsymbol{z}\|_1 + \frac{\Psi n}{2}\|\boldsymbol{z}' - \boldsymbol{z}\|_2^2 \\
\Rightarrow \left((n+k)\hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D}') - \sum_{i=1}^k \ell(\boldsymbol{z}';\alpha_i)\right) + \lambda\|\boldsymbol{z}'\|_1 &\geq \left((n+k)\hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D}') - \sum_{i=1}^k \ell(\boldsymbol{z};\alpha_i)\right) \\
&\quad + \Lambda\|\boldsymbol{z}\|_1 + \frac{\Psi n}{2}\|\boldsymbol{z}' - \boldsymbol{z}\|_2^2 \\
\Rightarrow \frac{\Psi n}{2}\|\boldsymbol{z} - \boldsymbol{z}'\|_2^2 &\leq \sum_{i=1}^k |\ell(\boldsymbol{z};\alpha_i) - \ell(\boldsymbol{z}';\alpha_i)|
\end{aligned}$$

The last inequality follows from the fact that $\hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D}') \leq \hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D}')$. Now, by mean value theorem for any data entry d, $|\ell(\boldsymbol{z}; d) - \ell(\boldsymbol{z}'; d)| \leq \|\bigtriangledown \ell(\boldsymbol{z}''; d)\|_2\|\boldsymbol{z} - \boldsymbol{z}'\|_2$, where $\boldsymbol{z}''$ is some vector in $\mathcal{C}_{\Gamma^*}$. By assumption, $\|\bigtriangledown \ell(\boldsymbol{z}''; d)\|_2 \leq 2s^{3/2}$.

Hence, it follows that $\|\boldsymbol{z} - \boldsymbol{z}'\|_2 \leq \frac{4ks^{3/2}}{\Psi n}$. □

Now using Claim 27 below, we conclude that $\boldsymbol{z}'$ is indeed the unique minimizer in $\mathcal{C}$ which minimizes $\hat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{D}') + \frac{\Lambda}{n+k}\|\boldsymbol{\theta}\|_1$.

Claim 27. *If $\Lambda = 4\sigma\sqrt{\log p}$, then z' is the unique minimizer of* $\arg\min\limits_{\boldsymbol{\theta}\in\mathcal{C}} \hat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{D}') + \frac{\Lambda}{n+k}\|\boldsymbol{\theta}\|_1$.

Proof. By assumption, $\|\boldsymbol{\theta}^*\|_\infty \leq 1 - \max\left\{\frac{16\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}, \frac{8ks^{3/2}}{\Psi n}\right\}$. Also from Theorem 9, we know that $\|\boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}}(\mathcal{D})\|_\infty \leq \frac{8\sigma}{\Psi}\sqrt{\frac{\log p}{n}}$. Using the bound obtained in Claim 26, we conclude that $\boldsymbol{z}'$ lie in the interior of the set $\mathcal{C}$. Hence, along any direction $i \in \Gamma^*$ there exist a sub-gradient of the objective function at $\boldsymbol{z}'$ whose slope is zero. In the following we analyze the sub-gradients of the objective functions along directions $i \in [p] - \Gamma^*$.

For any direction $i \in [p] - \Gamma^*$ we have,

$$
\begin{aligned}
(n+k)\bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D}')_i &= n\bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D})_i + n(\bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D})_i - \bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D})_i) + \sum_{j=1}^{k}\bigtriangledown\ell(\boldsymbol{z}';\alpha_j)_i \\
\Rightarrow |(n+k)\bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D}')_i| &\leq \underbrace{|n\bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D})_i|}_{A} + \underbrace{|n(\bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D})_i - \bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D}))_i|}_{B} + \underbrace{\left|\sum_{j=1}^{k}\bigtriangledown\ell(\boldsymbol{z}';\alpha_j)_i\right|}_{C} \qquad (6)
\end{aligned}
$$

We will bound each of the terms (A, B and C) on the right individually in order to show that $A+B+C < \Lambda$. This will imply that $\boldsymbol{z}'$ is the minimizer of the objective function $\hat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{D}') + \frac{\Lambda}{n+k}\|\boldsymbol{\theta}\|_1$ when restricted to the convex set $\mathcal{C}$. The uniqueness follows from the restricted strong convexity of the objective function in the directions in Γ^*.

Bound term $A \leq \frac{\Lambda}{2}$ in (6): Notice that term A is equal to $|X^T(\boldsymbol{y} - X\boldsymbol{z})|_i$. We have argued in the proof of Lemma 23, that $\boldsymbol{z}$ lies in the interior of the convex set $\mathcal{C}$. Now since $\boldsymbol{z}$ is the minimizer of $\frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \frac{\Lambda}{2n}\|\boldsymbol{\theta}\|_1$, therefore

$$
\frac{1}{n}\begin{bmatrix} X_{\Gamma^*}{}^T X_{\Gamma^*} & X_{\Gamma^*}{}^T X_{\Gamma^{*c}} \\ X_{\Gamma^{*c}}{}^T X_{\Gamma^*} & X_{\Gamma^{*c}}{}^T X_{\Gamma^*} \end{bmatrix}\begin{bmatrix} \boldsymbol{z}_{|\Gamma^*} - \boldsymbol{\theta}^*_{|\Gamma^*} \\ 0 \end{bmatrix} + \frac{1}{n}\begin{bmatrix} X_{\Gamma^*}{}^T \\ X^T_{\Gamma^{*c}} \end{bmatrix}\boldsymbol{w} + \frac{\Lambda}{n}\begin{bmatrix} \boldsymbol{v}_{|\Gamma^*} \\ \boldsymbol{v}_{|\Gamma^{*c}} \end{bmatrix} = 0 \qquad (7)
$$

Here $\Gamma^{*c} = [p] - \Gamma^*$ and for any vector $\boldsymbol{\theta} \in \mathbb{R}^p$, $\boldsymbol{\theta}_{|\Gamma^*}$ is the vector formed by the coordinates of $\boldsymbol{\theta}$ which are in Γ^*. Additionally, the vector $\boldsymbol{v}$ is a sub-gradient of $\|\cdot\|_1$ at $\boldsymbol{z}$. From (7) we have the following.

$$
\begin{aligned}
&(X_{\Gamma^*}{}^T X_{\Gamma^*})(\boldsymbol{z}_{|\Gamma^*} - \boldsymbol{\theta}^*{}_{\Gamma^*}) + X_{\Gamma^*}{}^T\boldsymbol{w} + \Lambda\boldsymbol{v}_{|\Gamma^*} = 0 && (8) \\
&\Leftrightarrow (\boldsymbol{z}_{|\Gamma^*} - \boldsymbol{\theta}^*{}_{|\Gamma^*}) = -(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1}X_{\Gamma^*}{}^T\boldsymbol{w} - \Lambda(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1}\boldsymbol{v}_{|\Gamma^*} && (9)
\end{aligned}
$$

In the above expression $\boldsymbol{v}_{|\Gamma^*} \in \{-1,1\}^{|\Gamma^*|}$, since for all $i \in \Gamma^*$, we have $|\boldsymbol{z}_i| > 0$, where $\boldsymbol{z}_i$ is the i-th coordinate of $\boldsymbol{z}$. Now note that $\boldsymbol{v}_{|\Gamma^{*c}} \in [-1,1]^{p-|\Gamma^*|}$. Therefore, if we bound each of the coordinates of $\boldsymbol{v}_{|\Gamma^{*c}}$ to be in $[-\frac{1}{2},\frac{1}{2}]$, we can conclude that for $i \in \Gamma^{*c}$, $|X^T(\boldsymbol{y} - X\boldsymbol{z})_i| \leq \frac{\Lambda}{2}$.

Combining (7) and (9), we have the following.

$$\begin{aligned}
(X_{\Gamma^{*c}}^T X_{\Gamma^*})(\boldsymbol{z}_{\Gamma^*} - \theta^*_{\Gamma^*}) + X_{\Gamma^{*c}}{}^T \boldsymbol{w} + \Lambda \boldsymbol{v}_{\Gamma^{*c}} &= 0 \\
\Leftrightarrow \boldsymbol{v}_{\Gamma^{*c}} &= \frac{1}{\Lambda}\left((X_{\Gamma^{*c}}{}^T X_{\Gamma^*})(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1} X_{\Gamma^*}{}^T \boldsymbol{w} - X_{\Gamma^{*c}}^T \boldsymbol{w} \right. \\
&\quad \left. - \Lambda (X_{\Gamma^{*c}}{}^T X_{\Gamma^*})(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1} \boldsymbol{v}_{\Gamma^*} \right) \\
&= -(X_{\Gamma^{*c}}{}^T X_{\Gamma^*})(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1} \boldsymbol{v}_{\Gamma^*} \\
&\quad - \frac{X_{\Gamma^{*c}}{}^T}{\Lambda} \left(\mathbb{I}_{n\times n} - X_{\Gamma^*}(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1} X_{\Gamma^*}^T \right) \boldsymbol{w} \\
\Leftrightarrow \|\boldsymbol{v}_{\Gamma^{*c}}\|_\infty &\leq \|(X_{\Gamma^{*c}}{}^T X_{\Gamma^*})(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1} \boldsymbol{v}_{\Gamma^*}\|_\infty \\
&\quad + \frac{1}{\Lambda} \|X_{\Gamma^{*c}}{}^T \left(\mathbb{I}_{n\times n} - X_{\Gamma^*}(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1} X_{\Gamma^*}^T \right) \boldsymbol{w}\|_\infty \\
&= \|(X_{\Gamma^{*c}}{}^T X_{\Gamma^*})(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1} \boldsymbol{v}_{\Gamma^*}\|_\infty + \\
&\quad + \frac{1}{\Lambda} \|X_{\Gamma^{*c}}{}^T V \boldsymbol{w}\|_\infty \qquad (10)
\end{aligned}$$

In the above expression $V = \left(\mathbb{I}_{n\times n} - X_{\Gamma^*}(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1} X_{\Gamma^*}^T \right)$ is a projection matrix. Applying the bounds from Bullets 3 and 5 from Assumption TYPICAL (Assumption 8), we have $\|\boldsymbol{v}_{\Gamma^c}\|_\infty < \frac{1}{2}$. From this it directly follows that for all $i \in \Gamma^{*c}$, $|X^T(\boldsymbol{y} - X\boldsymbol{z})_i| < \frac{\Lambda}{2}$.

Bound on term $B \leq \frac{4ks^2}{\Psi}$ in (6): The term B is upper bounded by $\|X^T X(\boldsymbol{z}' - \boldsymbol{z})\|_\infty$. Since by assumption on the domain of data entries U every column of X has L_2-norm of at most $\sqrt{n}$, it follows that every entry of the matrix $X^T X$ is at most n. Also note that $(\boldsymbol{z} - \boldsymbol{z}')$ has only s-non-zero entries. Therefore, $\|X^T X(\boldsymbol{z}' - \boldsymbol{z})\|_\infty \leq n\sqrt{s}\|\boldsymbol{z} - \boldsymbol{z}'\|_2$. From Claim 26 we already know that $\|\boldsymbol{z} - \boldsymbol{z}'\|_2 \leq \frac{4ks^{3/2}}{\Psi n}$. With this we get the relevant bound on B.

Bound on term $C \leq 2ks^{3/2}$ in (6): By the definition of $\ell(\boldsymbol{z}; \alpha_j)$ (where $\alpha_j = (y, \boldsymbol{x})$ is as defined in (6)), we have $\bigtriangledown \ell(\boldsymbol{z}; \alpha_j) = -\boldsymbol{x}(y - \langle \boldsymbol{x}, \boldsymbol{z} \rangle)$. Using the assumed bounds on y and $\|\boldsymbol{x}\|_2$, we bound $| \bigtriangledown \ell(\boldsymbol{z}; \alpha_j)_i|$ by $2s^{3/2}$. Now, it directly follows that the term C is bounded by $2ks^{3/2}$.

Now to complete the proof of Claim 27, we show that $A + B + C < \Lambda$. From the bounds on A, B and C above, we have $A + B + C \leq \frac{\Lambda}{2} + \frac{4ks^2}{\Psi} + 2ks^{3/2}$. Recall, that $\Lambda = 4\sigma\sqrt{n \log p}$. By assumption on s, it now follows that $A + B + C < \Lambda$. □

This concludes the proof of Lemma 24. □

To complete the proof of Theorem 11 (utility guarantee), all that is left is to prove Claim 25.

Proof of Claim 25. We need to show that the supports of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\hat{\boldsymbol{\theta}}(\mathcal{D}')$ are the same. From Lemma 24 it directly follows that $\mathrm{supp}(\hat{\boldsymbol{\theta}}(\mathcal{D}')) \subseteq \mathrm{supp}(\hat{\boldsymbol{\theta}}(\mathcal{D}))$. To prove equality, we provide the following argument.

From Theorem 9 we know that $\|\hat{\boldsymbol{\theta}}(\mathcal{D}) - \boldsymbol{\theta}^*\|_\infty \leq \frac{8\sigma}{\Psi}\sqrt{\frac{s \log p}{n}}$. Additionally, by assumption the absolute value of the minimum non-zero entry of $\boldsymbol{\theta}^*$ is at least $\Phi = \max\left\{ \frac{16\sigma}{\Psi}\sqrt{\frac{s \log p}{n}}, \frac{8ks^{3/2}}{\Psi n} \right\}$. This means that the absolute value of the minimum non-zero entry of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is at least $\frac{4ks^{3/2}}{\Psi n}$. Recall that in Claim 26 we showed

$\|\hat{\boldsymbol{\theta}}(\mathcal{D}) - \hat{\boldsymbol{\theta}}(\mathcal{D}')\|_\infty \leq \frac{4ks^{3/2}}{\Psi n}$. From this we can conclude that every coordinate where $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is non-zero, $\hat{\boldsymbol{\theta}}(\mathcal{D}')$ is also non-zero.

Hence, $\text{supp}(\hat{\boldsymbol{\theta}}(\mathcal{D}')) = \text{supp}(\hat{\boldsymbol{\theta}}(\mathcal{D}))$. This concludes the proof. □

B.2.2 Proofs of Theorems 12 (Correctness Theorem) and 13 (Stability Theorem) for huberized LASSO

Proof of Theorem 12 (Correctness Theorem). We first show that the support of $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ in (4) will be the same as the output of LASSO in (3), i.e., the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ in (3) is same as the support of $\tilde{\boldsymbol{\theta}}(\mathcal{D})$. Moreover, we show that the minimizer $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ equals $\hat{\boldsymbol{\theta}}(\mathcal{D})$.

Claim 28. *$\tilde{\boldsymbol{\theta}}(\mathcal{D})$ equals $\hat{\boldsymbol{\theta}}(\mathcal{D})$.*

Proof. In order to prove this claim, we invoke Theorem 1 from [27] (see Theorem 9). Notice for all the rows $\boldsymbol{x}_i$ of X, by assumption $|\langle \boldsymbol{x}_i, \boldsymbol{\theta}^* \rangle| \leq 2\sqrt{s \log n}$. By triangle inequality we have

$$\begin{aligned} |\langle \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}(\mathcal{D}) \rangle| &\leq |\langle \boldsymbol{x}_i, \boldsymbol{\theta}^* \rangle| + |\langle \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}(\mathcal{D}) - \boldsymbol{\theta}^* \rangle| \\ &\leq 2\sqrt{s \log n} + \sqrt{\frac{s^2 \log p}{n}} \end{aligned}$$

The last inequality follows from the bound $\|\hat{\boldsymbol{\theta}}(\mathcal{D}) - \boldsymbol{\theta}^*\|_2$ (see Theorem 9). Since, we assumed $n = \omega(s \log p)$, it follows that for all the rows $\boldsymbol{x}_i$ (with $i \in [n]$), $|\langle \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}(\mathcal{D}) \rangle| \leq 3\sqrt{s \log n}$. Therefore the following are true for all $i \in [n]$: $-\boldsymbol{x}_i(y_i - \langle \boldsymbol{x}_i, \hat{\boldsymbol{\theta}}(\mathcal{D}) \rangle) = \bigtriangledown \hat{\ell}(\hat{\boldsymbol{\theta}}(\mathcal{D}); y_i, \boldsymbol{x}_i)$. This property straight away implies that $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is the minimizer of the objective function in (4). To show that $\tilde{\boldsymbol{\theta}}(\mathcal{D}) = \hat{\boldsymbol{\theta}}(\mathcal{D})$, now all we need to show is that $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is the *unique* minimizer of the objective function in (4). This is true because at $\hat{\boldsymbol{\theta}}(\mathcal{D})$ in a ball of radius $r \to 0$, the function $\hat{\ell}(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i)$ equals the function $\frac{1}{2}(y_i - \langle \boldsymbol{x}_i, \boldsymbol{\theta} \rangle)^2$ for all $i \in [n]$. Hence, from the proof Lemma 23 since $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is the unique minimizer of (3), it follows that $\tilde{\boldsymbol{\theta}}(\mathcal{D}) = \hat{\boldsymbol{\theta}}(\mathcal{D})$. □

To conclude the proof of Theorem 12, we invoke Theorem 1 from [27]. For completeness purposes we provide it in Theorem 9. □

Proof of Theorem 13 (Stability Theorem). Since, in huberized LASSO we intend to get a better dependence on the data set size n, we weaken the constraint on the maximum and minimum allowable values of $\boldsymbol{\theta}^*$. We assume that $\|\boldsymbol{\theta}^*\|_\infty \leq 1 - \max\left\{\frac{16\sigma}{\Psi}\sqrt{\frac{s \log p}{n}}, \frac{20ks\sqrt{\log n}}{\Psi n}\right\}$ and the absolute value of every non-zero entry of $\boldsymbol{\theta}^*$ is at least $\max\left\{\frac{16\sigma}{\Psi}\sqrt{\frac{s \log p}{n}}, \frac{20ks\sqrt{\log n}}{\Psi n}\right\}$. Similar to the stability proof for LASSO (Theorem 11), we prove the stability guarantee via Lemma 29 and 30, and Claim 31.

Lemma 29. *Under assumptions of Theorem 12, if $\hat{\Gamma}$ is the support of $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ and $\tilde{\boldsymbol{\theta}}(\mathcal{D})_{\hat{\Gamma}} = \arg\min\limits_{\boldsymbol{\theta} \in \mathcal{C}_{\hat{\Gamma}}} \frac{1}{n}\sum_{i=1}^{n} \hat{\ell}(\boldsymbol{\theta}; (y_i, X_i)) + \frac{\Lambda}{n}\|\boldsymbol{\theta}\|_1$, then $\tilde{\boldsymbol{\theta}}(\mathcal{D})_{\hat{\Gamma}}$ equals $\tilde{\boldsymbol{\theta}}(\mathcal{D})$.*

For the ease of notation, we denote $\tilde{\boldsymbol{\theta}}(\mathcal{D})_{\hat{\Gamma}}$ by $\boldsymbol{z}$.

Lemma 30. *Let $\mathcal{D}' = (\boldsymbol{y}', X')$ be a data set formed by inserting (removing) k entries in $\mathcal{D}$ (which are from the domain U) and let $\boldsymbol{z}' = \arg\min\limits_{\boldsymbol{\theta} \in \mathcal{C}_{\hat{\Gamma}}} \frac{1}{|\mathcal{D}'|}\sum_{i=1}^{|\mathcal{D}'|} \hat{\ell}(\boldsymbol{\theta}; (y'_i, X'_i)) + \frac{\Lambda}{|\mathcal{D}'|}\|\boldsymbol{\theta}\|_1$. Under assumptions of Lemma 29, $\boldsymbol{z}' = \tilde{\boldsymbol{\theta}}(\mathcal{D}')$, where $\tilde{\boldsymbol{\theta}}(\mathcal{D}') = \arg\min\limits_{\boldsymbol{\theta} \in \mathcal{C}} \frac{1}{|\mathcal{D}'|}\sum_{i=1}^{|\mathcal{D}'|} \hat{\ell}(\boldsymbol{\theta}; (y'_i, X'_i)) + \frac{\Lambda}{|\mathcal{D}'|}\|\boldsymbol{\theta}\|_1$.*

To prove the above lemma, we use a proof technique which was developed by [27] under the name of *primal-dual construction* and was used to argue consistency in non-private sparse linear regression.

Claim 31. *Under assumptions of Lemma 30, $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ and $\tilde{\boldsymbol{\theta}}(\mathcal{D}')$ have the same support.*

In the following we provide the proofs of the above two lemmas and the claim. The proof of Lemma 29 is exactly the same for Lemma 23 in Section B.2.1 and hence omitted here.

Proof of Lemma 30. For the ease of notation, we fix the following: i) $\hat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{D}) = \frac{1}{n}\sum_{i=1}^{n}\hat{\ell}(\boldsymbol{\theta};d_i)$, where $d_i = (y_i, \boldsymbol{x}_i)$, y_i is the i-th entry of $\boldsymbol{y}$ and $\boldsymbol{x}_i$ is the i-th row of X, ii) we denote $\tilde{\boldsymbol{\theta}}(\mathcal{D})_{\hat{\Gamma}}$ by $\boldsymbol{z}$. Also, since by Theorem 12, $\hat{\Gamma}$ equals the support of $\boldsymbol{\theta}^*$ (i.e., Γ^*), we fix $\hat{\Gamma} = \Gamma^*$.

Let $\boldsymbol{z}' = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}_{\Gamma^*}} \hat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{D}') + \frac{\Lambda}{n+k}\|\boldsymbol{\theta}\|_1$. W.l.o.g. assume that $\mathcal{D}'$ has k entries more than $\mathcal{D}$ and call these entries $\alpha_1, \cdots, \alpha_k$. (The analysis for the case when $\mathcal{D}'$ has k entries less than $\mathcal{D}$ follows analogously.) In the following claim we show that $\boldsymbol{z}'$ does not differ too much from $\boldsymbol{z}$ in the L_2-metric.

Claim 32. $\|\boldsymbol{z} - \boldsymbol{z}'\|_2 \leq \frac{10ks\sqrt{\log n}}{\Psi n}$.

Proof. By restricted strong convexity of $\hat{\mathcal{L}}$ at $\boldsymbol{z}$ in a ball (in the subspace formed by the support set Γ^*) of radius $\frac{2k\zeta}{\Psi n}$ around it, we have the following.

$$\begin{aligned}
n\hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D}) + \Lambda\|\boldsymbol{z}'\|_1 &\geq n\hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D}) + \Lambda\|\boldsymbol{z}\|_1 + \frac{\Psi n}{2}\|\boldsymbol{z}' - \boldsymbol{z}\|_2^2 \\
\Rightarrow \left((n+k)\hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D}') - \sum_{i=1}^{k}\ell(\boldsymbol{z}';\alpha_i)\right) + \lambda\|\boldsymbol{z}'\|_1 &\geq \left((n+k)\hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D}') - \sum_{i=1}^{k}\ell(\boldsymbol{z};\alpha_i)\right) \\
&\quad + \Lambda\|\boldsymbol{z}\|_1 + \frac{\Psi n}{2}\|\boldsymbol{z}' - \boldsymbol{z}\|_2^2 \\
\Rightarrow \frac{\Psi n}{2}\|\boldsymbol{z} - \boldsymbol{z}'\|_2^2 &\leq \sum_{i=1}^{k}|\ell(\boldsymbol{z};\alpha_i) - \ell(\boldsymbol{z}';\alpha_i)|
\end{aligned}$$

The last inequality follows from the fact that $\hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D}') \leq \hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D}')$. Now, by mean value theorem for any data entry d, $|\ell(\boldsymbol{z};d) - \ell(\boldsymbol{z}';d)| \leq \|\bigtriangledown\ell(\boldsymbol{z}'';d)\|_2\|\boldsymbol{z} - \boldsymbol{z}'\|_2$, where $\boldsymbol{z}''$ is some vector in $\mathcal{C}_{\Gamma^*}$. Therefore, $\|\bigtriangledown\ell(\boldsymbol{z}'';d)\|_2 \leq 2s\sqrt{\log n}$.

Hence, it follows that $\|\boldsymbol{z} - \boldsymbol{z}'\|_2 \leq \frac{10ks\sqrt{\log n}}{\Psi n}$. □

Now using Claim 33 below, we conclude that $\boldsymbol{z}'$ is indeed the unique minimizer in $\mathcal{C}$ which minimizes $\hat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{D}') + \frac{\Lambda}{n+k}\|\boldsymbol{\theta}\|_1$.

Claim 33. *z' is the unique minimizer of* $\arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \hat{\mathcal{L}}(\boldsymbol{\theta};\mathcal{D}') + \frac{\Lambda}{n+k}\|\boldsymbol{\theta}\|_1$.

Proof. By assumption, $\|\boldsymbol{\theta}^*\|_\infty \leq 1 - \max\left\{\frac{16\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}, \frac{20ks\sqrt{\log n}}{\Psi n}\right\}$. Also from Theorem 9, we know that $\|\boldsymbol{\theta}^* - \tilde{\boldsymbol{\theta}}(\mathcal{D})\|_\infty \leq \frac{8\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}$. Using the bound obtained in Claim 32, we conclude that $\boldsymbol{z}'$ lie in the interior of the set $\mathcal{C}$. Hence, along any direction $i \in \Gamma^*$ there exist a sub-gradient of the objective function at $\boldsymbol{z}'$ whose slope is zero. In the following we analyze the sub-gradients of the objective functions along directions $i \in [p] - \Gamma^*$.

For any direction $i \in [p] - \Gamma^*$ we have,

$$(n+k) \triangledown \hat{\mathcal{L}}(\boldsymbol{z}'; \mathcal{D}')_i = n \triangledown \hat{\mathcal{L}}(\boldsymbol{z}; \mathcal{D})_i + n(\triangledown\hat{\mathcal{L}}(\boldsymbol{z}'; \mathcal{D})_i - \triangledown\hat{\mathcal{L}}(\boldsymbol{z}; \mathcal{D})_i) + \sum_{j=1}^{k} \triangledown\ell(\boldsymbol{z}'; \alpha_j)_i$$

$$\Rightarrow |(n+k) \triangledown \hat{\mathcal{L}}(\boldsymbol{z}'; \mathcal{D}')_i| \leq \underbrace{|n \triangledown \hat{\mathcal{L}}(\boldsymbol{z}; \mathcal{D})_i|}_{A} + \underbrace{|n(\triangledown\hat{\mathcal{L}}(\boldsymbol{z}'; \mathcal{D})_i - \triangledown\hat{\mathcal{L}}(\boldsymbol{z}; \mathcal{D}))_i|}_{B} + \underbrace{\left|\sum_{j=1}^{k} \triangledown\ell(\boldsymbol{z}'; \alpha_j)_i\right|}_{C} \tag{11}$$

We will bound each of the terms (A, B and C) on the right individually in order to show that $A+B+C < \Lambda$. This will imply that $\boldsymbol{z}'$ is the minimizer of the objective function $\hat{\mathcal{L}}(\boldsymbol{\theta}; \mathcal{D}') + \frac{\Lambda}{n+k}\|\boldsymbol{\theta}\|_1$ when restricted to the convex set $\mathcal{C}$. The uniqueness follows from the restricted strong convexity of the objective function in the directions in Γ^*.

Bound term $A \leq \frac{\Lambda}{2}$ in (11): Notice that term A is equal to $|X^T(\boldsymbol{y} - X\boldsymbol{z})|_i$. We have argued in the proof of Lemma 23, that $\boldsymbol{z}$ lies in the interior of the convex set $\mathcal{C}$. Now since $\boldsymbol{z}$ is the minimizer of $\frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \frac{\Lambda}{2n}\|\boldsymbol{\theta}\|_1$, therefore

$$\frac{1}{n}\begin{bmatrix} X_{\Gamma^*}{}^T X_{\Gamma^*} & X_{\Gamma^*}{}^T X_{\Gamma^{*c}} \\ X_{\Gamma^{*c}}{}^T X_{\Gamma^*} & X_{\Gamma^{*c}}{}^T X_{\Gamma^*} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_{|\Gamma^*} - \boldsymbol{\theta}^*_{|\Gamma^*} \\ 0 \end{bmatrix} + \frac{1}{n}\begin{bmatrix} X_{\Gamma^*}{}^T \\ X_{\Gamma^{*c}}^T \end{bmatrix} \boldsymbol{w} + \frac{\Lambda}{n}\begin{bmatrix} \boldsymbol{v}_{|\Gamma^*} \\ \boldsymbol{v}_{|\Gamma^{*c}} \end{bmatrix} = 0 \tag{12}$$

Here $\Gamma^{*c} = [p] - \Gamma^*$ and for any vector $\boldsymbol{\theta} \in \mathbb{R}^p$, $\boldsymbol{\theta}_{|\Gamma^*}$ is the vector formed by the coordinates of $\boldsymbol{\theta}$ which are in Γ^*. Additionally, the vector $\boldsymbol{v}$ is a sub-gradient of $\|\cdot\|_1$ at $\boldsymbol{z}$. From (12) we have the following.

$$(X_{\Gamma^*}{}^T X_{\Gamma^*})(\boldsymbol{z}_{|\Gamma^*} - \boldsymbol{\theta^*}_{\Gamma^*}) + X_{\Gamma^*}{}^T\boldsymbol{w} + \Lambda\boldsymbol{v}_{|\Gamma^*} = 0 \tag{13}$$

$$\Leftrightarrow (\boldsymbol{z}_{|\Gamma^*} - \boldsymbol{\theta^*}_{|\Gamma^*}) = -(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1}X_{\Gamma^*}{}^T\boldsymbol{w} - \Lambda(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1}\boldsymbol{v}_{|\Gamma^*} \tag{14}$$

In the above expression $\boldsymbol{v}_{|\Gamma^*} \in \{-1, 1\}^{|\Gamma^*|}$, since for all $i \in \Gamma^*$, we have $|\boldsymbol{z}_i| > 0$, where $\boldsymbol{z}_i$ is the i-th coordinate of $\boldsymbol{z}$. Now note that $\boldsymbol{v}_{|\Gamma^{*c}} \in [-1, 1]^{p-|\Gamma^*|}$. Therefore, if we bound each of the coordinates of $\boldsymbol{v}_{|\Gamma^{*c}}$ to be in $[-\frac{1}{2}, \frac{1}{2}]$, we can conclude that for $i \in \Gamma^{*c}$, $|X^T(\boldsymbol{y} - X\boldsymbol{z})_i| \leq \frac{\Lambda}{2}$.

Combining Equations 12 and 14, we have the following.

$$\begin{aligned}
(X_{\Gamma^{*c}}^T X_{\Gamma^*})(\boldsymbol{z}_{\Gamma^*} - \theta^*_{\Gamma^*}) + X_{\Gamma^{*c}}{}^T\boldsymbol{w} + \Lambda\boldsymbol{v}_{\Gamma^{*c}} &= 0 \\
\Leftrightarrow \boldsymbol{v}_{\Gamma^{*c}} &= \frac{1}{\Lambda}\big((X_{\Gamma^{*c}}{}^T X_{\Gamma^*})(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1}X_{\Gamma^*}{}^T\boldsymbol{w} - X_{\Gamma^{*c}}^T\boldsymbol{w} \\
&\quad - \Lambda(X_{\Gamma^{*c}}{}^T X_{\Gamma^*})(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1}\boldsymbol{v}_{\Gamma^*}\big) \\
&= -(X_{\Gamma^{*c}}{}^T X_{\Gamma^*})(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1}\boldsymbol{v}_{\Gamma^*} \\
&\quad -\frac{X_{\Gamma^{*c}}{}^T}{\Lambda}\left(\mathbb{I}_{n\times n} - X_{\Gamma^*}(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1}X_{\Gamma^*}^T\right)\boldsymbol{w} \\
\Leftrightarrow \|\boldsymbol{v}_{\Gamma^{*c}}\|_\infty &\leq \|(X_{\Gamma^{*c}}{}^T X_{\Gamma^*})(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1}\boldsymbol{v}_{\Gamma^*}\|_\infty \\
&\quad +\frac{1}{\Lambda}\|X_{\Gamma^{*c}}{}^T\left(\mathbb{I}_{n\times n} - X_{\Gamma^*}(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1}X_{\Gamma^*}^T\right)\boldsymbol{w}\|_\infty \\
&= \|(X_{\Gamma^{*c}}{}^T X_{\Gamma^*})(X_{\Gamma^*}{}^T X_{\Gamma^*})^{-1}\boldsymbol{v}_{\Gamma^*}\|_\infty + \\
&\quad +\frac{1}{\Lambda}\|X_{\Gamma^{*c}}{}^T V\boldsymbol{w}\|_\infty
\end{aligned}$$

In the above expression $V = \left(\mathbb{I}_{n\times n} - X_{\Gamma^*}({X_{\Gamma^*}}^T X_{\Gamma^*})^{-1}X_{\Gamma^*}^T\right)$ is a projection matrix. Applying the bounds from Bullets 3 and 5 from Assumption TYPICAL (Assumption 8), we have $\|\boldsymbol{v}_{\Gamma^c}\|_\infty < \frac{1}{2}$. From this it directly follows that for all $i \in \Gamma^{*c}$, $|X^T(\boldsymbol{y} - X\boldsymbol{z})_i| < \frac{\Lambda}{2}$.

Bound on term $B \leq \frac{10ks^{3/2}\sqrt{\log n}}{\Psi}$ **in** (11)**:** The term B is upper bounded by $\|X^TX(\boldsymbol{z}' - \boldsymbol{z})\|_\infty$. First notice that since by Assumption (s, Ψ, σ, Φ)-TYPICAL every column of X has L_2-norm of at most $\sqrt{n}$. Hence, it follows that every entry of the matrix X^TX is at most n. Also note that $(\boldsymbol{z} - \boldsymbol{z}')$ has only s-non-zero entries. Therefore, $\|X^TX(\boldsymbol{z}' - \boldsymbol{z})\|_\infty \leq n\sqrt{s}\|\boldsymbol{z} - \boldsymbol{z}'\|_2$. From Claim 32 we already know that $\|\boldsymbol{z} - \boldsymbol{z}'\|_2 \leq \frac{10ks\sqrt{\log n}}{\Psi n}$. With this we get the relevant bound on B.

Bound on term $C \leq 10ks\sqrt{\log n}$ **in** (11)**:** By the definition of $\ell(\boldsymbol{z}; \alpha_j)$ (where $\alpha_j = (y, \boldsymbol{x})$ is as defined in (11)), we have $\bigtriangledown\ell(\boldsymbol{z}; \alpha_j) = -\boldsymbol{x}(y - \langle\boldsymbol{x}, \boldsymbol{z}\rangle)$. From the assumed bounds on y and $\|\boldsymbol{x}\|_2$ in Section 3.2, we bound $|\bigtriangledown \ell(\boldsymbol{z}; \alpha_j)_i|$ by $10s^2\sqrt{\log n}$. Now, it directly follows that the term C is bounded by $10ks\sqrt{\log n}$.

Now to complete the proof of Claim 33, we show that $A + B + C < \Lambda$. From the bounds on A, B and C above, we have $A + B + C \leq \frac{\Lambda}{2} + \frac{10ks^{3/2}\sqrt{\log n}}{\Psi} + 10ks\sqrt{\log n}$. Recall, that $\Lambda = 4\sigma\sqrt{n\log p}$. By assumption on s, it now follows that $A + B + C < \Lambda$. □

This concludes the proof of Lemma 30. □

To complete the proof of Theorem 13 (utility guarantee), all is left is to provide the proof for Claim 31.

Proof of Claim 31. We need to show that the supports of $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ and $\tilde{\boldsymbol{\theta}}(\mathcal{D}')$ are the same. From Lemma 24 it directly follows that supp($\tilde{\boldsymbol{\theta}}(\mathcal{D}')$) $\subseteq$ supp($\tilde{\boldsymbol{\theta}}(\mathcal{D})$). To prove equality, we provide the following argument.

From Theorem 9 we know that $\|\tilde{\boldsymbol{\theta}}(\mathcal{D}) - \boldsymbol{\theta}^*\|_\infty \leq \frac{8\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}$. Additionally, by assumption the absolute value of the minimum non-zero entry of $\boldsymbol{\theta}^*$ is at least $\Phi = \max\left\{\frac{16\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}, \frac{20ks\sqrt{\log n}}{\Psi n}\right\}$. This means that the absolute value of the minimum non-zero entry of $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ is at least $\frac{10ks\sqrt{\log n}}{\Psi n}$. Recall that in Claim 32 we showed $\|\tilde{\boldsymbol{\theta}}(\mathcal{D}) - \tilde{\boldsymbol{\theta}}(\mathcal{D}')\|_\infty \leq \frac{10ks\sqrt{\log n}}{\Psi n}$. From this we can conclude that every coordinate where $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ is non-zero, $\tilde{\boldsymbol{\theta}}(\mathcal{D}')$ is also non-zero.

Hence, supp($\tilde{\boldsymbol{\theta}}(\mathcal{D}')$) = supp($\tilde{\boldsymbol{\theta}}(\mathcal{D})$). This concludes the proof. □

□

B.2.3 Proofs of Theorems 14 (k-stability (proxy version)) and 16 (STRONGLY-TYPICAL $\Rightarrow$ k-stability (proxy version))

Proof of Theorem 14

Proof of Theorem 14 (k-stability (proxy version)). The proof of this theorem directly follows from Lemma 34 and Claims 35, 36, and 37 below. We prove these statements after stating them.

Lemma 34. *If $g_i(\mathcal{D}) > t_i$ for all $i \in \{1, \cdots, 4\}$ and $\Lambda > \frac{16s^2}{\Psi}$, then changing one entry in $\mathcal{D}$ does not change the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$.*

In the following three lemmas we bound the local sensitivity (i.e., the amount by which the value of $g_i(\mathcal{D})$ changes when an entry is added or removed from $\mathcal{D}$) of the test functions $g_1, \cdots, g_4$ on a data set $\mathcal{D}$ when $g_i(\mathcal{D}) > t_i$ for all $i \in \{1, \cdots, 4\}$.

Claim 35. *Following the definition in Table 1, if $g_i(\mathcal{D}) > t_i$ for all $i \in \{1, \cdots, 4\}$ and $\Lambda > \frac{16s^2}{\Psi}$, then for any neighboring dataset $\mathcal{D}'$ (*i.e.*, having one entry more (less) compared to $\mathcal{D}$),*

$$\left|g_1(\mathcal{D}) - g_1(\mathcal{D}')\right| \leq \frac{6s^2}{\Psi} = \Delta_1$$

Claim 36. *If $g_i(\mathcal{D}) > t_i$ for all $i \in \{1, \cdots, 4\}$ and $\Lambda > \frac{16s^2}{\Psi}$, then for any neighboring dataset $\mathcal{D}'$ (*i.e.*, having one entry more (less) compared to $\mathcal{D}$),*

$$\left|g_2(\mathcal{D}) - g_2(\mathcal{D}')\right| \leq s = \Delta_2$$

Claim 37. *If $g_i(\mathcal{D}) > t_i$ for all $i \in \{1, \cdots, 4\}$ and $\Lambda > \frac{16s^2}{\Psi}$, then for any neighboring dataset $\mathcal{D}'$ (*i.e.*, having one entry more (less) compared to $\mathcal{D}$),*

$$n\|\hat{\boldsymbol{\theta}}(\mathcal{D}) - \hat{\boldsymbol{\theta}}(\mathcal{D}')\|_\infty \leq n\|\hat{\boldsymbol{\theta}}(\mathcal{D}) - \hat{\boldsymbol{\theta}}(\mathcal{D}')\|_2 \leq \frac{4s^{3/2}}{\Psi} = \Delta_3 = \Delta_4$$

Proof of Lemma 34. We prove the lemma via the following three claims (Claims 38, 39 and 40).

Claim 38. *If $g_i(\mathcal{D}) > t_i$ for all $i \in \{1, \cdots, 4\}$ and $\Lambda > \frac{16s^2}{\Psi}$, if $\hat{\Gamma}$ is the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\hat{\boldsymbol{\theta}}(\mathcal{D})_{\hat{\Gamma}} = \arg\min_{\boldsymbol{\theta} \in \mathcal{C}_{\hat{\Gamma}}} \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \frac{\Lambda}{n}\|\boldsymbol{\theta}\|_1$ (where $\mathcal{C}_{\hat{\Gamma}} \subseteq \mathcal{C}$ is the convex subset of $\mathcal{C}$ restricted to support in $\hat{\Gamma}$), then $\hat{\boldsymbol{\theta}}(\mathcal{D})_{\hat{\Gamma}}$ equals $\hat{\boldsymbol{\theta}}$.*

Claim 39. *Let $\mathcal{D}' = (\boldsymbol{y}', X'))$ be a data set formed by inserting (removing) one entry in $\mathcal{D}$. Let $z' = \arg\min_{\boldsymbol{\theta} \in \mathcal{C}_{\hat{\Gamma}}} \frac{1}{2|\mathcal{D}'|}\|\boldsymbol{y}' - X'\boldsymbol{\theta}\|_2^2 + \frac{\Lambda}{|\mathcal{D}'|}\|\boldsymbol{\theta}\|_1$. Then, if $g_i(\mathcal{D}) > t_i$ for all $i \in \{1, \cdots, 4\}$ and $\Lambda > \frac{16s^2}{\Psi}$, then $\boldsymbol{z}' = \hat{\boldsymbol{\theta}}(\mathcal{D}')$, where $\hat{\boldsymbol{\theta}}(\mathcal{D}') = \arg\min_{\boldsymbol{\theta} \in \mathcal{C}} \frac{1}{2|\mathcal{D}'|}\|\boldsymbol{y}' - X'\boldsymbol{\theta}\|_2^2 + \frac{\Lambda}{|\mathcal{D}'|}\|\boldsymbol{\theta}\|_1$.*

Claim 40. *If $g_i(\mathcal{D}) > t_i$ for all $i \in \{1, \cdots, 4\}$ and $\Lambda > \frac{16s^2}{\Psi}$, then $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\hat{\boldsymbol{\theta}}(\mathcal{D}')$ have the same support.*

The proof of these claims follow directly from the proofs of Lemmas 23, 24 and Claim 25 respectively. □

Proof of Claim 35. W.l.o.g. we assume that the dataset $\mathcal{D}'$ has one entry more than $\mathcal{D}$ (call this entry d_{new}). First note that if $g_i(\mathcal{D}) > t_i$ for all $i \in \{1, \cdots, 4\}$, then $(s+1)$-th coordinate of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is zero. Additionally, note that by Lemma 34 the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\hat{\boldsymbol{\theta}}(\mathcal{D}')$ is the same. We now need to bound the following.

$$(n+1) \bigtriangledown \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}'); \mathcal{D}') = n \bigtriangledown \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}); \mathcal{D}) + n(\bigtriangledown \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}'); \mathcal{D}) - \bigtriangledown \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}); \mathcal{D})) + \bigtriangledown \ell(\hat{\boldsymbol{\theta}}(\mathcal{D}'); d_{new}) \quad (15)$$

For any $i \in [p] - \hat{\Gamma}$ (where $\hat{\Gamma}$ is the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$), by triangle inequality the following is true.

$$\left|(n+1) \bigtriangledown \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}'); \mathcal{D}')_i - n \bigtriangledown \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}); \mathcal{D})_i\right| \leq \underbrace{n\left|(\bigtriangledown \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}'); \mathcal{D})_i - \bigtriangledown \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}); \mathcal{D})_i)\right|}_{B} + \underbrace{\left|\bigtriangledown \ell(\hat{\boldsymbol{\theta}}(\mathcal{D}'); d_{new})_i\right|}_{C} \quad (16)$$

We can bound each of this terms (B and C) individually.

Bound on term $B \leq \frac{4s^2}{\Psi}$ in (16)**:** The term B is upper bounded by $\|X^T X(\hat{\boldsymbol{\theta}}(\mathcal{D}') - \hat{\boldsymbol{\theta}}(\mathcal{D}))\|_\infty$. First notice that by definition every column of X has L_2-norm of at most $\sqrt{n}$. Thus it follows that every entry of the matrix $X^T X$ is at most n. Also note that $(\hat{\boldsymbol{\theta}}(\mathcal{D}') - \hat{\boldsymbol{\theta}}(\mathcal{D}))$ has only s-non-zero entries. Therefore, $\|X^T X(\hat{\boldsymbol{\theta}}(\mathcal{D}') - \hat{\boldsymbol{\theta}}(\mathcal{D}))\|_\infty \leq n\sqrt{s}\|\hat{\boldsymbol{\theta}}(\mathcal{D}') - \hat{\boldsymbol{\theta}}(\mathcal{D})\|_2$. From Claim 26 we already know that $\|\hat{\boldsymbol{\theta}}(\mathcal{D}') - \hat{\boldsymbol{\theta}}(\mathcal{D})\|_2 \leq \frac{4s^{3/2}}{\Psi}$. With this we get the relevant bound.

Bound on term $C \leq 2s^{3/2}$ in (16)**:** By the definition of $\ell(\hat{\boldsymbol{\theta}}(\mathcal{D}); \alpha_j)$ (where $\alpha_j = (y, \boldsymbol{x})$ is as defined in (6)), we have $\triangledown \ell(\hat{\boldsymbol{\theta}}(\mathcal{D}); \alpha_j) = -\boldsymbol{x}(y - \langle \boldsymbol{x}, \hat{\boldsymbol{\theta}}(\mathcal{D})\rangle)$. From the assumed bounds on y and $\|\boldsymbol{x}\|_2$, we bound $|\triangledown \ell(\hat{\boldsymbol{\theta}}(\mathcal{D}); \alpha_j)_i|$ by $2s^{3/2}$. Now, it directly follows that the term C is bounded by $2s^{3/2}$. □

Proof of Claim 36. From Lemma 34 we know that the minimizers $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\hat{\boldsymbol{\theta}}(\mathcal{D}')$ share the same support. Additionally, since if $g_i(\mathcal{D}) > t_i$ for all $i \in \{1, \cdots, 4\}$, we know that the the size of the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is less than or equal to s.

Now to prove Lemma 36, all we need to show is that restricted to any support Φ of size s, the minimum eigenvalue of the Hessian of $\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}); \mathcal{D})$ does not change by more than s when the dataset $\mathcal{D}$ is changed to a neighboring one $\mathcal{D}'$. Since, we are only concerned with linear regression, the Hessian of the loss function $\hat{\mathcal{L}}(\cdot; \mathcal{D})$ evaluated at any point is $X^T X$, where X is the design matrix. W.l.o.g. if we assume that $\mathcal{D}'$ has one entry more than $\mathcal{D}$ (and call that entry $d_{new} = (y, \boldsymbol{x})$, where $y \in \mathbb{R}$ and $\boldsymbol{x} \in \mathbb{R}^p$, then the Hessian of $\hat{\mathcal{L}}(\cdot; \mathcal{D}')$ at any point is given by $X^T X + \boldsymbol{x}\boldsymbol{x}^T$.

Representing the minimum eigenvalue of a matrix A as $\lambda(A)$ and A_Φ as the matrix formed by columns from the set Φ, we have the following.

$$\begin{aligned}\left|g_2(\mathcal{D}) - g_2(\mathcal{D}')\right| &= \left|\lambda({X_{\hat{\Gamma}}}^T X_{\hat{\Gamma}}) - \lambda({X_{\hat{\Gamma}}}^T X_{\hat{\Gamma}} + \boldsymbol{x}_{\hat{\Gamma}}{\boldsymbol{x}_{\hat{\Gamma}}}^T)\right| \\ &\leq \text{max. eigenvalue}(\boldsymbol{x}_{\hat{\Gamma}}{\boldsymbol{x}_{\hat{\Gamma}}}^T) \leq s\end{aligned}$$

The first inequality follows from Weyl's inequalities. This completes the proof. □

Proof of Claim 37. From Lemma 34 we know that the unique minimizers $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\hat{\boldsymbol{\theta}}(\mathcal{D}')$ share the same support.

Now, from Claim 26, it follows that $\|\hat{\boldsymbol{\theta}}(\mathcal{D}) - \hat{\boldsymbol{\theta}}(\mathcal{D}')\|_2 \leq \frac{4s^{3/2}}{\Psi n}$. This in turn implies that $\|\hat{\boldsymbol{\theta}}(\mathcal{D}) - \hat{\boldsymbol{\theta}}(\mathcal{D}')\|_\infty \leq \frac{4s^{3/2}}{\Psi n}$ since L_∞-norm is less than or equal to L_2-norm. □

□

Proof of Theorem 16

*Proof of Theorem 16 (*STRONGLY-TYPICAL *$\Rightarrow$ k-stability (proxy version)).* From Assumption $(s, \Psi, \sigma, \Phi, k)$-STRONGLY-TYPICAL, it directly follows that $g_2(\mathcal{D}) > t_2 + (k-1)\Delta_2$. To argue about $g_3(\mathcal{D})$ and $g_4(\mathcal{D})$, notice that by Theorem 9 it follows that the absolute value of any non-zero entry of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is in $\left(\frac{2(4+(k-1))s^{3/2}}{\Psi n}, 1 - \frac{2(4+(k-1))s^{3/2}}{\Psi n}\right)$. Hence, $g_3(\mathcal{D}) > t_3 + (k-1)\Delta_3$ and $g_4(\mathcal{D}) > t_3 + (k-1)\Delta_4$. To complete the proof, all we need to argue is about $g_1(\mathcal{D})$. Using similar proof technique of Claim 27 (more precisely (10)) and the *bounded noise* condition from Assumption $(s, \Psi, \sigma, \Phi, k)$-STRONGLY-TYPICAL (i.e., $\|X_{\Gamma^c}^T V\boldsymbol{w}\|_\infty \leq 2\sigma\sqrt{n \log p} - 6(k-1)s^2/\Psi$) it follows that $g_1(\mathcal{D}) > t_1 + (k-1)\Delta_1$. □