JMLR: Workshop and Conference Proceedings vol 30 (2013) 1–32

Differentially Private Model Selection via Stability Arguments and the Robustness of the Lasso

Adam Smith ASMITH@CSE.PSU.EDU
Department of Computer Science and Engineering
Pennsylvania State University

Abhradeep Thakurta B-ABHRAG@MICROSOFT.COM
Stanford University and Microsoft Research Silicon Valley Campus

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 $\mathcal{D}$ and a discrete collection of models $M_1, M_2, \ldots$, where each model is a family of probability distributions $M_i = \{p_{i,\theta}\}_{\theta \in \Theta_i}$, 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 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* (Dwork et al., 2006b; Dwork, 2006), which 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 (Dwork, 2006; Ganta et al., 2008).

We give both general results on model selection and specific results on a prototypical problem in high-dimensional statistics, namely *sparse linear regression*: each entry in the data set consists of a p-dimensional real *feature vector* $\boldsymbol{x}$ and real-valued *response* (or *label*) y. The goal is to find a parameter vector $\theta \in \mathbb{R}^p$ with few nonzero entries such that $\langle \boldsymbol{x}_i, \theta \rangle \approx y_i$ for all n data points $(\boldsymbol{x}_i, y_i)$. Such sparse regression vectors provide predictive relationships that generalize well to unseen data and are relatively easy to interpret. Specifically, 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" (say constant). We say a vector is s-sparse if has at most s nonzero entries. Assuming that $\boldsymbol{\theta}^*$ is s-sparse for a given parameter s, under can conditions can we recover the support of $\boldsymbol{\theta}^*$ while minimizing the information leaked about individual data points?

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 Dwork et al. (2006b,a)) *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 Kasiviswanathan and Smith (2008) for a discussion). In Section A we review some of the basic concepts associated with differential privacy, which also act as basic building blocks for our algorithms.

1.1. Our Contributions

Generic Transformations. We give two transformations that take a nonprivate model selection procedure f and produce a private procedure $\mathcal{A}$ with similar *sample complexity* and *computational efficiency*.

Our transformations work whenever there is a "well-defined" output, 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 either one of two notions, *perturbation* stability and *subsampling* stability, both of which have been studied in the machine learning literature. Roughly, a function f is *perturbation* stable on a data set $\mathcal{D}$ if it takes the value $f(\mathcal{D})$ on all the neighbors of $\mathcal{D}$. A function is f is *subsampling* stable on $\mathcal{D}$, if for a random subsample $\hat{\mathcal{D}}$ from $\mathcal{D}$ we have $f(\mathcal{D}) = f(\hat{D})$ with good probabiliy.

An important implication of our results is that if the nonprivate model selection algorithm f is *consistent* on n i.i.d. samples from a distribution P, then the private analogue $\mathcal{A}$ is consistent on $n' = O(n \frac{\log(1/\delta)}{\epsilon})$ samples. Here, *consistency* means that f selects a particular model for P with

reasonable probability (say, at least $2/3$). The statistical literature has considered a wide array of model selection procedures, many of which are proven consistent under very general assumptions; see Claeskens and Hjort (2008, Ch. 4) for a textbook introduction.

The main new technical idea in our generic transformations is to identify *efficient* ways to privately test if the algorithm f is subsampling stable on a given input. For this, we adapt the "sample-aggregate" framework of Nissim et al. (2007).

Sparse Regression Algorithms. We give new algorithms for sparse linear regression which improve on previous efforts by Kifer, Smith, and Thakurta (2012). With very sparse models (roughly, when s is smaller than $\log p$) our new algorithms match the optimal *nonprivate* sample complexity.

Our algorithms are based on the popular LASSO technique of L_1 penalized regression. Our analyses work under the same assumptions used in the literature to analyze nonprivate feature selection algorithms (Wainwright, 2006; Zhao and Yu, 2007). A particularly clean implication of our results is that with Gaussian data and noise, if there exists an s-sparse vector $\boldsymbol{\theta}^*$ that labels the data well, then our algorithms recover the support of $\boldsymbol{\theta}^*$ with n' samples (with good probability), when

$$n' = \Omega^*(\min\{ks \log p, \max\{s \log p, \tfrac{k^2 s^4}{\log p}\}\}) \quad \text{and } k = \tfrac{\log(1/\delta)}{\epsilon}\,.$$

For comparison, the nonprivate sample complexity is $\Theta^*(s \log p)$ (Wainwright, 2006) (the Ω^* and Θ^* notation hide logarithmic factors in n). In particular, our algorithms match the nonprivate lower bound when $s < \big((\log p)/k\big)^{2/3}$. Our algorithms improve the previous work by Kifer et al. (2012) in all ranges of parameters. Even with constant ϵ and δ, the efficient algorithms of Kifer et al. (2012) have sample complexity $\Omega(s^2 \log^2 p)$, whereas ours scales as $O^*(s \log p)$.

Stability and Robustness of the LASSO. Our algorithms for sparse linear regression require analyzing the stability properties of the popular LASSO estimator that we feel are of independent interest. The LASSO minimizes the usual mean squared error loss penalized with (a multiple of) the L_1 norm of $\boldsymbol{\theta}$:

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

The consistency properties of the Lasso are well-studied: under 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 (Wainwright, 2006; Negahban et al., 2009). Moreover, if the entries of $\boldsymbol{\theta}^*$ are bounded away from zero, $\hat{\boldsymbol{\theta}}$ will have the same support as $\boldsymbol{\theta}^*$ (Wainwright, 2006).

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 1.2 below.) Our analysis requires significantly refining the "primal-dual" construction technique of Wainwright (2006). 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.

Efficient Tests for LASSO Stability. Significantly, we 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 private feature selector with optimal sample complexity.

Most results on the LASSO's convergence require *restricted strong convexity* (RSC) of the loss function (Wainwright, 2006; Negahban et al., 2009). Evaluating the RSC parameter is known to be NP-hard. Our efficient test circumvents this — we show that one need only bound the strong convexity of the loss function for the set of coordinates actually recovered by the LASSO.

1.2. Previous Work

Private Model Selection. We are not aware of any prior work on general private model selection. Zhou et al. (2009) studied the specific problem of private sparse regression in low dimension, where $p \ll n$; their techniques do not give meaningful results when $p \geq n$. Kifer et al. (2012) gave the first results on private sparse regression in high dimension. They designed a computationally efficient algorithm, implicitly based on *subsampling stability*, for support recovery using the LASSO estimator. In this work we significantly extend and improve on the results of Kifer et al. (2012) (see "Our Contributions" for a comparison). Our algorithm based on subsampling stability is inspired by that of Kifer et al. (2012), but is based on a more sophisticated and sample-efficient test of stability. Our algorithm based on perturbation stability is considerably different in flavor; it is inspired by the work of Dwork and Lei (2009) on robust statistics and privacy (see Section) 2 for a discussion) but requires new results on the robustness of the LASSO.

Nonprivate Sparse Regression. We draw on a rich literature studying (nonprivate) sparse high-dimensional regression (e.g. Donoho (2000); Zhao and Yu (2007); Wainwright (2006); Negahban et al. (2009)). Several works (Zhao and Yu, 2007; Wainwright, 2006; Kim et al., 2012) show that $n = \omega(s \log p)$ samples suffice for consistent support recovery under a variety of distributions. In this work we show that under the same set of assumptions considered by Wainwright (2006), the support of $\boldsymbol{\theta}^*$ is also stable. In contrast, Xu et al. (2010) study the L_2-stability of LASSO-like estimators to small perturbations and show that *uniform* stability (in which the set of selected features change by only small steps between *any pairs* of neighbors) is impossible for algorithms with sparse output; some assumptions on the input are this necessary. Finally, Lee et al. (2011) look at Huberization of the LASSO with the goal of providing robustness, but did not provide formal consistency or convergence guarantees.

Note that the notion of support recovery requires a *unique* "best" underlying model for the data. Most analyses of the LASSO's convergence assume such an underlying model. However, if one only requires low generalization error (as opposed to exact model selection), then one can work with weaker assumptions (Juditsky and Nemirovski, 2000; Greenshtein and Ritov, 2004).

Stability and Learning. The general relationship between learning, statistics and stability has been studied in the learning theory literature (e.g., Rogers and Wagner (1978)) and in robusts statistics (e.g., Huber (1981)) for over thirty years. Many variants of stability have been studied, and the literature is too vast to survey in detail here. A variant of the notion of *perturbation stability* that we consider have been studied previously (Bousquet and Elisseeff, 2002; Shalev-Shwartz et al., 2010; Xu et al., 2010; Dwork and Lei, 2009). One consequence of these works is that if a learning algorithm f satisfies our notion of stability, then it generalizes well. The notion of stability to subsampling or resampling from the training data set has also been studied (Shao, 1996; Bach, 2008; Meinshausen and Buehlmann, 2010; Meilă, 2006). In particular, stability under resampling was proposed as a criterion for model selection in clustering.

2. Stability and Privacy

We give two simple, generic transformations (Sections 2.1 and 2.2) that, given any function f and parameters $\epsilon, \delta > 0$, return a (ϵ, δ)-differentially private algorithm 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. Both of the following notions have been studied extensively in machine learning (see Previous Work).

- *Perturbation Stability*: We say that f is *perturbation 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 $f(\mathcal{D})$ w.h.p. whenever f is q-subsampling stable for $q = \frac{\epsilon}{32\log(1/\delta)}$. The running time of $\mathcal{A}_{samp}$ is dominated by running f about $1/q^2$ times; hence it is efficient whenever f is.

  As mentioned earlier, this result has a clean statistical interpretation: if f is a *consistent* model selection procedure on n samples, then $\mathcal{A}_f$ is consistent on $n' = O(n\frac{\log(1/\delta)}{\epsilon})$ samples.

Throughout this section we will use two notions that quantify perturbation stability:

Definition 2 (k-stability) *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.*

2.1. From Perturbation Stability to Privacy

The idea behind the first algorithm comes from the work of Dwork and Lei (2009) on private parametric estimation. For basic terminology related to differential privacy, see Appendix A. 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 Dwork and Lei (2009) (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 one 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 $\mathcal{A}_{dist}$.

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 (Dwork and Lei, 2009; Karwa et al., 2011) 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 adds $\mathsf{Lap}(1/\epsilon)$ 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 $1 - \beta$.*

The lemma is proved in Appendix C.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 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 (possibly randomized) function $\hat{f}$ that equals f on "nice" inputs with high probability, and (b) designing efficient, private estimators for the distance to instability with respect to $\hat{f}$. To that end, we define the proxy for distance to instability for any function $\hat{f}$.

Definition 4 *Given $\hat{f} : U^* \to \mathcal{R}$, a function $\hat{d} : U^* \to \mathbb{R}$ is a* proxy for the distance to instability *of $\hat{f}$ if: (1) For all $\mathcal{D}$: $\hat{d}(\mathcal{D}) \leq$ (dist. of $\mathcal{D}$ to instability of $\hat{f}$), and (2) the global sensitivity of $\hat{d}$ is at most 1.*

One can use such a proxy by adding Laplace noise $Lap(\frac{1}{\epsilon})$ to $\hat{d}$ and releasing $\hat{f}(\mathcal{D})$ whenever the noisy version of $\hat{d}$ is sufficiently large (at least $(\log(1/\delta)/\epsilon)$. The resulting mechanism will be (ϵ, δ)-differentially private and it will release $\hat{f}(\mathcal{D})$ with high probability on instances for which $\hat{f}$ is at least $\frac{2\log(1/\delta)}{\epsilon}$-stable. Given a function f, the goal is to find proxies $(\hat{f}, \hat{d})$ that are efficient (ideally, as efficient as evaluating f) and have a large set of "nice" inputs where $\hat{f} = f$.

2.2. From Sampling Stability to Stability

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. (2007). However, our analysis allows working with much larger subsamples than those in previous work (Nissim et al., 2007; Smith, 2011; Kifer et al., 2012). 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. Apart from providing the sample and aggregate based algorithm, Kifer et al. (2012) also provided a computationally inefficient model selection algorithm based on the *exponential mechanism* of McSherry and Talwar (2007). The exponential mechanism based algorithm works under weaker set of assumptions and the sample

complexity suffers from a polynomial blowup. It is perceivable that the sample complexity bound can be improved under stronger assumptions.

Our generic construction 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. For the proof of this theorem, see Section C.2.

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.

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

Require: dataset: $\mathcal{D}$, function $f : U^* \to \mathcal{R}$, 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 $count(r) = \#\{i : f(\hat{\mathcal{D}}_i) = r\}$.

$\hat{d} \leftarrow (count_{(1)} - count_{(2)})/(4mq) - 1$ where $count_{(1)}, count_{(2)}$ are the two highest counts.

$\tilde{d} \leftarrow \hat{d} + \mathsf{Lap}(\frac{1}{\epsilon})$.

if $\tilde{d} > \log(1/\delta)/\epsilon$, **then** Output $\hat{f}(\mathcal{D}) = mode(F)$, **else** Output $\perp$.

Previous works in data privacy have used the idea of bootstrapping or subsampling to convert from various forms of subsampling stability to some sort of stability (Nissim et al., 2007; Dwork and Lei, 2009; Smith, 2011; Kifer et al., 2012). The main advantage of the version we present here is that the 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.

3. Consistency and Stability of Sparse Regression using LASSO

The results in this section are about the consistency and stability of feature selection for sparse linear regression using the LASSO estimator (eq. (2)). We consider two questions: (1) Assuming that $\boldsymbol{\theta}^*$ is s-sparse, under what conditions can we obtain a consistent estimator $\hat{\boldsymbol{\theta}}(\mathcal{D})$ (i.e., the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ equals the support of $\boldsymbol{\theta}^*$ and $\|\hat{\boldsymbol{\theta}}(\mathcal{D}) - \boldsymbol{\theta}^*\|_2$ goes to zero as n goes to infinity), and (2) under what conditions is the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is perturbation-stable?

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 below.

This assumption is *weaker* than that of previous work; see Appendix B for a detailed discussion of the assumption.

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

(2) ***Bounded parameter vector:*** *$\|\boldsymbol{\theta}^*\|_0 \leq s$ 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:*** *$\|X_{\Gamma^c}^T V \boldsymbol{w}\|_\infty \leq 2\sigma\sqrt{n \log p}$, 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 (mildly) strengthened version of the *fixed data conditions* above (Assumption 1), Wainwright (2006) 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 7 restates the result of Wainwright (2006) in the context of this paper. We note that the result of Wainwright (2006) holds even under this weaker assumption (Assumption 1).

Theorem 7 (Modified Theorem 1 of Wainwright (2006)) *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)-*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, we consider the 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 from $\mathcal{N}(0, \sigma^2\mathbb{I}_n)$. We show that in such a setting Assumption 1 (Typical system) holds with high-probability. The formal theorem statement and its proof is provided in Section D.1.

3.2. Notation and Data Normalization

We assume that the underlying parameter vector $\boldsymbol{\theta}^*$ is from the convex set $\mathcal{C} = \{\boldsymbol{\theta} : \|\boldsymbol{\theta}\|_\infty \leq 1\}$. The set $\mathcal{C}_\Gamma$ is set of all vectors in $\mathcal{C}$ whose coordinates are zero outside a set $\Gamma \subseteq [p]$. We assume that each entry of the design matrix X has absolute value of at most one, and additionally we assume that the response vector $\boldsymbol{y}$ has L_∞-norm at most s, i.e., $\|\boldsymbol{y}\|_\infty \leq s$. In case the data set $\mathcal{D} = (\boldsymbol{y}, X)$ does not satisfy the above bound, we *normalize* the data set by scaling down each data entry individually, so that they satisfy the above bound. In the rest of the paper we define the universe of data sets U^* to be sets of entries from this domain (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 1 and suitable choice of regularization parameter Λ) the LASSO estimator is *consistent*. 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 show that under Assumption 1, the support of the minimizer $\hat{\boldsymbol{\theta}}(\mathcal{D})$ in (2) does not change even if k data entries are removed or added to $\mathcal{D}$ as long as $n = \omega(\max\{s\log p, \frac{s^4k^2}{\log p}, ks^{3/2}\})$. 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 $n \to \infty$.

The main stability theorem for LASSO is given in Theorem 8. For the purpose of clarity, we defer the complete proof of the stability theorem to Section D.2.1. The correctness follows directly from Theorem 7.

Theorem 8 (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 $\mathcal{D} = (\boldsymbol{y}, X)$), 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 main novelty is the construction of 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 $\mathcal{C}_{\hat{\Gamma}}$. Using the consistency result from Theorem 7 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 7 by Wainwright (2006) to argue consistency of LASSO estimators. Here we use it to argue stability of the support. In Section D.2.2 we show that one can obtain better stability properties if the loss function in the LASSO program is huberized.

Function	Instantiation (Parameters: s, Λ, Ψ)	Threshold (t_i)	Slack (Δ_i)
$g_1(\mathcal{D})$	$-$ ($(s+1)^{\text{st}}$ largest absolute value of $n \bigtriangledown \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}); \mathcal{D})$)	$-\frac{\Lambda}{2}$	$\frac{12s^2}{\Psi}$
$g_2(\mathcal{D})$	min. eigenvalue of ${X_{\hat{\Gamma}}}^T X_{\hat{\Gamma}}$	$2\Psi n$	s
$g_3(\mathcal{D})$	$n\times$(min. absolute value of the non-zero entries in $\hat{\boldsymbol{\theta}}(\mathcal{D})$)	$\frac{8s^{3/2}}{\Psi}$	$\frac{4s^{3/2}}{\Psi}$
$g_4(\mathcal{D})$	$-n\times$(max. absolute value of the non-zero entries in $\hat{\boldsymbol{\theta}}(\mathcal{D})$)	$\frac{8s^{3/2}}{\Psi} - n$	$\frac{4s^{3/2}}{\Psi}$

Table 1: Instantiation of the four test functions

3.4. Efficient Test for k-stability

In Section 3.3 we saw that under Assumption 1 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 (2) and the huberized LASSO in (9) 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.

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 (2)) 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 $n \bigtriangledown \hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}); \mathcal{D}) = -X^T(\boldsymbol{y} - X\hat{\boldsymbol{\theta}}(\mathcal{D}))$.

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* (Nissim et al., 2007) 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 9 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 D.2.3. Next in Theorem 10 we show that if the data set $\mathcal{D} = (\boldsymbol{y}, X)$ satisfies a slight strengthening of Assumption 1 (see Assumption 2), 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 1. In Section 3.5 we analyze a stochastic setting where Assumption 2 is satisfied with high probability.

Theorem 9 (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 $i \in \{1, \cdots, 4\}$, $\Lambda > \frac{16ks^2}{\Psi}$ and $s \leq \sqrt{\frac{\sigma n^{1/2} \log^{1/2} p}{2k(1/\Psi + 1)}}$, then $\hat{\boldsymbol{\theta}}(\mathcal{D})$ has k-stable support.*

Assumption 2 (Strongly-typical system) *Data set $(X_{n\times p}, \boldsymbol{y}_{n\times 1})$ and parameter vector $\boldsymbol{\theta}^* \in \mathbb{R}^p$ are $(s, \Psi, \sigma, \Phi, k)$-*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, $\|X_{\Gamma^c}^T V \boldsymbol{w}\|_\infty \leq 2\sigma\sqrt{n\log p} - 12(k-1)s^2/\Psi$, where $V = \mathbb{I}_{n\times n} - X_\Gamma({X_\Gamma}^T X_\Gamma)^{-1}{X_\Gamma}^T$ projects on to the complement of the column space of X_Γ.*

Theorem 10 *Let $\mathcal{D} = (\boldsymbol{y}, X)$ be a data set from U^*, $\Lambda = 4\sigma\sqrt{n\log p}$ and suppose $s \leq \sqrt{\frac{\sigma n^{1/2}\log^{1/2} p}{2k(1/\Psi+1)}}$. If there exists a $\boldsymbol{\theta}^*$ such that $(\boldsymbol{y}, X, \boldsymbol{\theta}^*)$ is $(s, \Psi, \sigma, \Phi, k)$-*Strongly-Typical *with $\Phi = \max\left\{\frac{16\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}, \frac{16ks^{3/2}}{\Psi n}\right\}$, 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 8. We defer the proof to Section D.2.3.

3.5. Stability of LASSO in Stochastic Setting

In this section we will see one specific stochastic setting for the data set, where the support of $\hat{\boldsymbol{\theta}}$ is k-stable with high-probability. This will in turn mean that the set of conditions in Theorems 8 and 10 are satisfied with high probability.

Assumption 3 (Normalized Gaussian Data) *Given a parameter vector $\boldsymbol{\theta}^*$, suppose each row of a matrix X' is drawn i.i.d. from $\mathcal{N}(0, \frac{1}{4}\mathbb{I}_p)$ and the entries of a noise vector $\boldsymbol{w}'$ are i.i.d. from $\mathcal{N}(0, \sigma^2)$, and set $\boldsymbol{y}' = X'\boldsymbol{\theta}^* + \boldsymbol{w}'$. Now let the design matrix X be the matrix formed by first dividing each entry of X' by $\sqrt{\log(ns)}$ and then rounding each entry to the interval [-1,1], and let $\boldsymbol{y}$ be the response vector formed by first dividing each entry of $\boldsymbol{y}'$ by $\sqrt{\log(ns)}$ and then rounding to $[-s, s]$. The data set is the pair $\mathcal{D} = (X, \boldsymbol{y})$.*

Remark: We normalize the Gaussian design matrix X and the corresponding response vector $\boldsymbol{y}$ in order to argue perturbation stability. The scaling by $\sqrt{\log(ns)}$ ensures that the columns in X which are in the support of $\boldsymbol{\theta}^*$ do not get rounded (with high probability), simplifying our analysis. It is unlikely that this renormalization is necessary.

The following proposition is proved in Appendix D.2.4.

Proposition 11 *Fix $k \geq 1$. Let $\Lambda = 4\sigma\sqrt{n\log p}$ and $n = \omega(s\log p\log n, \frac{s^4k^2\log n}{\log p}, ks\log n)$. For any arbitrarily small positive constant Φ, if $\|\boldsymbol{\theta}^*\|_0 \leq s$, and the absolute value of any non-zero entry of $\boldsymbol{\theta}^*$ is in $(\Phi, 1-\Phi)$, then with probability at least $3/4$, the support of the minimizer $\hat{\boldsymbol{\theta}}(\mathcal{D})$ (in (2)) is k-stable, where the data set $\mathcal{D}$ is distributed as in Assumption 3.*

4. Private Support Selection for Sparse Linear Regression

In this section we use our generic differentially private model selection algorithm (from Section 2) 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 the set $[p]$. Once a support of size s is chosen, one can restrict the regression problem to the set of s-coordinates chosen and then use algorithms (*e.g.,* objective perturbation) for private linear regression from Kifer et al. (2012) 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 $O\left(\frac{s^2 \log(1/\delta)}{n\epsilon}\right)$.

4.1. Support Selection via Sampling Stability

We use use Algorithm 1 for support selection. In the current context, the non-private model selection function f in Algorithm 1 returns the support of the minimizer of unmodified LASSO program in (2). If the support has cardinality greater than s, then just pick the first s coordinates. By Theorem 6, the output is always (ϵ, δ)-differentially private.

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 $\mathcal{N}(0, \sigma^2 \mathbb{I}_n)$. In this stochastic setting, we obtain the following utility theorem. For the corresponding deterministic version of this theorem, see Section E.1.

Theorem 12 (Stochastic utility) *Let $\Lambda = \sigma\sqrt{nk \log p}$, $n = \omega(ks \log p)$ and $\delta < 1/100$. For any arbitrarily small positive constant Φ, if $\|\boldsymbol{\theta}^*\|_0 \leq s$, and the absolute value of any non-zero entry of $\boldsymbol{\theta}^*$ is in $(\Phi, 1-\Phi)$, then under the stochastic model in the preceding paragraph, with probability at least $3/4$, Algorithm 1 outputs the correct support of $\boldsymbol{\theta}^*$. Here $k = \log(1/\delta)/\epsilon$.*

The sample complexity implied by the above theorem is $s \log p \log(1/\delta)/\epsilon$. Compared to the optimal sample complexity of $s \log p$, we have a blow up by a factor of $\log(1/\delta)\epsilon$.

4.2. Support Selection via Stability of LASSO

In Section 3.4 we designed an efficient test for k-stability. In this section we transform it into a differentially private algorithm for outputting the support.

Algorithm Description. In the language of Section 2 (and using the notation of Section 3.4) let the function for the distance be $\hat{d}(\mathcal{D}) = \max\left\{\min_i \frac{g_i(\mathcal{D}) - t_i}{\Delta_i} + 1, 0\right\}$. Using Lemma 13 below, we show that $\hat{d}$ is the proxy for the distance to instability of the minimizer $\hat{\boldsymbol{\theta}}$ for the LASSO program (in(2)). See Section D.2.5 for proof.

Lemma 13 *If $\Lambda > \frac{16s^2}{\Psi}$ and $s \leq \sqrt{\frac{\sigma n^{1/2} \log^{1/2} p}{2k(1/\Psi + 1)}}$, then the function $\hat{d}$ is a proxy for the distance to instability of the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ (in (2)), i.e., for all $\mathcal{D} \in U^*$, $\hat{d}(\mathcal{D})$ is the lower bound on the distance to instability of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and global sensitivity of $\hat{d}$ is at most one.*

Now, the algorithm for support selection directly 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 $\hat{\boldsymbol{\theta}}(\mathcal{D})$.

Analysis. 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 state the utility guarantee of this Algorithm in the stochastic setting considered in Assumption 3. For the corresponding deterministic version of this theorem, see Section E.2.

Theorem 14 (Stochastic utility) *Let $\Lambda = \sigma\sqrt{n\log p}$, $\frac{n}{\log n} = \omega(s\log p, \frac{s^4k^2}{\log p}, ks)$ and $\delta < 1/100$. For any arbitrarily small positive constant Φ, if $\|\boldsymbol{\theta}^*\|_0 \leq s$, and the absolute value of any non-zero entry of $\boldsymbol{\theta}^*$ is in $(\Phi, 1-\Phi)$, then under the stochastic setting in Assumption 3, with probability at least $3/4$, the above algorithm run on the data set $\mathcal{D} = (\boldsymbol{y}, X)$ outputs the correct support of $\boldsymbol{\theta}^*$. Here $k = \log(1/\delta)/\epsilon$.*

With the two private feature selection algorithms (from Sections 4.1 and 4.2) in hand, the overall sample complexity scales as $O^*(\min(ks\log p, \max(s\log p, ks^4/\log p, k^2s^3))$. A simple calculation shows there are in fact three distinct regimes, based on how $\log p$ relates to k and s:

$$n' = \begin{cases} O^*(s\log p) & \text{if} \sqrt{k^2s^3} < \log p \\ O^*(k^2s^4/\log p) & \text{if } \sqrt{ks^3} \leq \log p \leq \sqrt{k^2s^3} \\ O^*(ks\log p) & \text{if } \log p < \sqrt{ks^3} \end{cases} \quad \text{where } k = \frac{\log(1/\delta)}{\epsilon}\,.$$

Getting a private algorithm whose complexity scales as the optimal $s\log p$ (without the extra factor of $k = \frac{\log(1/\delta)}{\epsilon}$) over a larger parameter range remains an interesting open problem.

Acknowledgments

We are grateful to comments from Jing Lei, Cynthia Dwork, Moritz Hardt and Pradeep Ravikumar. A.S. and A.T. (while at Penn State) were supported by NSF awards #0941553 and #0747294.

References

Francis R. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *ICML*, 2008.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2:499 – 526, 2002.

Gerda Claeskens and Nils Lid Hjort. *Model selection and model averaging*. Cambridge University Press, 2008.

David L. Donoho. Aide-memoire. high-dimensional data analysis: The curses and blessings of dimensionality, 2000.

Cynthia Dwork. Differential privacy. In *ICALP*, 2006.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *STOC*, 2009.

Cynthia Dwork, Krishnaram Kenthapadi, Frank Mcsherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *EUROCRYPT*, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *TCC*, 2006b.

Srivatsava Ranjit Ganta, Shiva Prasad Kasiviswanathan, and Adam Smith. Composition attacks and auxiliary information in data privacy. In *KDD*, 2008.

Eitan Greenshtein and Ya'acov Ritov. Persistence in high-dimensional linear predictor selection and the virtue of overparametrization. *Bernoulli*, 2004.

Peter Huber. *Robust Statistics*. Wiley, 1981.

Anatoli Juditsky and Arkadii Nemirovski. Functional aggregation for nonparametric regression. *The Annals of Statistics*, 2000.

Vishesh Karwa, Sofya Raskhodnikova, Adam Smith, and Grigory Yaroslavtsev. Private analysis of graph structure. *PVLDB*, 2011.

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

Daniel Kifer, Adam Smith, and Abhradeep Thakurta. Private convex empirical risk minimization and high-dimensional regression. In *COLT*, 2012.

Yongdai Kim, Sunghoon Kwon, and Hosik Choi. Consistent model selection criteria on high dimensions. *Journal of Machine Learning Research*, 2012.

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

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *FOCS*, 2007.

M. Meilă. The uniqueness of a good optimum for k-means. In *ICML*, 2006.

Nicolai Meinshausen and Peter Buehlmann. Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2010.

Sahand Negahban, Pradeep Ravikumar, Martin J. Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. In *NIPS*, 2009.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *STOC*, 2007.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. 2011.

WH Rogers and TJ Wagner. A finite sample distribution-free performance bound for local discrimination rules. *The Annals of Statistics*, pages 506–514, 1978.

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

Jun Shao. Bootstrap model selection. *Journal of the American Statistical Association*, 1996.

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *STOC*, 2011.

Martin 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.

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

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

Shuheng Zhou, Katrina Ligett, and Larry Wasserman. Differential privacy with compression. In *ISIT*, 2009.

Appendix A. Preliminaries

In this section we review some of the concepts commonly used in the differential privacy literature. Some of these concepts form the basic building blocks for the model selection algorithms we describe in this paper.

Global sensitivity (Dwork et al., 2006b). For a given domain of data sets U^* and a function $f : U^* \rightarrow \mathbb{R}$, the global sensitivity of f (represented by GS_f) refers to the maximum change that $f(\mathcal{D})$ can have for any data set $\mathcal{D} \in U^*$ when one data entry is added or removed from $\mathcal{D}$.

$$GS_f = \max_{\mathcal{D}, \mathcal{D}' \in U^*, |\mathcal{D} \Delta \mathcal{D}'| = 1} |f(\mathcal{D}) - f(\mathcal{D}')|$$

Local sensitivity (Nissim et al., 2007). For a given data set $\mathcal{D} \in U^*$, local sensitivity (represented by $LS_f(\mathcal{D})$) refers to the maximum change that $f(\mathcal{D})$ can have if any one data entry is added or removed from $\mathcal{D}$. It is trivial to show that global sensitivity always upper bounds the local sensitivity, since it is a worst case bound.

$$LS_f(\mathcal{D}) = \max_{\mathcal{D}' \in U^*, |\mathcal{D} \Delta \mathcal{D}'| = 1} |f(\mathcal{D}) - f(\mathcal{D}')|$$

Laplace mechanism. Dwork et al. (2006b) gave a simple ϵ-differentially private algorithm for computing a function $f : U^* \rightarrow \mathbb{R}$ on a given data set $\mathcal{D}$, based on the global sensitivity of f. The algorithm is to output $f(\mathcal{D}) + \text{Lap}\left(\frac{GS_f}{\epsilon}\right)$, where $\text{Lap}(\lambda)$ refers to a sample drawn from the Laplace distribution with scaling parameter λ. Recall that the density function for a Laplace random variable with scaling parameter λ is given by $\frac{1}{2\lambda} e^{-\frac{|x|}{\lambda}}$.

Appendix B. Discussion on Typical and Strongly-typical Assumptions

In this section we discuss the semantics of the *typical* and *strongly-typical* assumption (Assumptions 1 and 2) we use in this paper. The first comment we want to make is that the strongly-typical assumption differs from the typical assumption only up to additive terms depending on k (the stability parameter), s (the sparsity parameter) and constants. In terms of understanding the semantics of the assumptions, in this section we will restrict ourselves to the typical assumption only.

The *column normalization* assumption is trivially satisfied when each entry of the design matrix X is bounded. However, when the entries in X are bounded only in expectation (say, if the entries are Gaussian), then the assumption is strictly weaker than assuming that each entry of X is bounded.

In the *bounded parameter vector* assumption we need a lower bound on the nonzero entries of the true parameter vector $\boldsymbol{\theta}^*$ so that while recovering the true support, the algorithm should be able to distinguish between zero coordinates and nonzero coordinates in $\boldsymbol{\theta}^*$. An upper bound on the nonzero entries of $\boldsymbol{\theta}^*$ is a technical condition which ensures that the minimizer of a particular convex program lies away from the boundary of the convex set that constrains it. We conjecture that this condition can be removed via a more careful analysis.

Our *incoherence* assumption is weaker than the the one existent in the literature (Wainwright, 2006). Previous work made a much stronger assumption, namely that the condition holds when the sign of $\boldsymbol{\theta}^*$ is replaced by *any* vector with L_∞ norm of one. We observe that it is sufficient to have the condition hold for one vector, $sign(\boldsymbol{\theta}^*)$, and this simplifies the application of the definition. Now a random Gaussian matrix with appropriate parameters satisfies our definition with high probability,

but we do not know if it satisfies the stronger assumption. Wainwright (2006) actually had a separate argument for random Gaussian data (that is, the argument of Wainwright (2006) for Gaussian data did not proceed by showing that the deterministic conditions were satisfied with high probability).

The *bounded noise* assumption ensures that the noise vector $\boldsymbol{w}$ is not strongly correlated with any of the columns of the design matrix X (outside the columns of the support set Γ), after projection on to the directions orthogonal to the columns of X in Γ.

Appendix C. Stability and Privacy (Proofs)

C.1. Proof of Proposition 3

Proof [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 Dwork et al. (2006b), the variable $\tilde{d}$ (in Algorithm $\mathcal{A}_{dist}$) satisfies $(\epsilon, 0)$-differential privacy.

Theorem 15 (Laplace Mechanism Dwork et al. (2006b)) *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 [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. ■

C.2. Proof of Theorem 6 (Privacy and Utility Guarantee for Algorithm $\mathcal{A}_{samp}$)

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

Lemma 16 *Fix $q \in (0,1)$. Given $f : U^* \to \mathcal{R}$, let $\hat{f} : U^* \to \mathcal{R}$ 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 $\hat{d}(\mathcal{D}) = (count_{(1)} - count_{(2)})/(4mq) - 1$. 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 function $\hat{d}$ is a proxy for the distance to instability of $\hat{f}$.

(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 simultaneously 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.

Proof [Proof of Lemma 16] 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. ■

Appendix D. Consistency and Stability of Sparse Linear Regression via LASSO (Proofs)

D.1. Consistency of LASSO Estimator

Theorem 17 (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 the noise vector $\boldsymbol{w}$ be drawn from $\mathcal{N}(0, \sigma^2\mathbb{I}_n)$, 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 1, $(\boldsymbol{y}, X, \boldsymbol{\theta}^*)$ satisfies (s, Ψ, σ, Φ)-*Typical* with $\Phi = \frac{16\sigma}{\Psi}\sqrt{\frac{s \log p}{n}}$.*

Proof [Proof of Theorem 17 (Stochastic Consistency)] In the following we show that each of the Conditions 1, 3, 4, and 5 in Assumption 1 are satisfied with probability at least $15/16$. By union bound over the failure probabilities of these events, this will straightaway imply Theorem 17.

- **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 Raskutti et al. (2011), 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 $j \in [p-s]$,

$$|\left(X^T{}_{\Gamma^c} X_\Gamma \boldsymbol{u}\right)_j| = \left|\sum_{i\in[s]} u_i \langle \boldsymbol{a_j}, \boldsymbol{b_i}\rangle\right| = \left|\langle \boldsymbol{a_j}, \sum_{i\in[s]} u_i \boldsymbol{b_i}\rangle\right| \tag{3}$$

  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 (3) 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 $\tilde{\boldsymbol{a}}_{\boldsymbol{i}}$ 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 $\|X_{\Gamma^c}^T V \boldsymbol{w}\|_\infty \leq 2\sigma\sqrt{n \log p}$.

■

D.2. Stability of LASSO Estimator in the Fixed Data Setting (Proofs)

D.2.1. PROOF OF THEOREM 8 (STABILITY OF UNMODIFIED LASSO)

Proof of Theorem 8 follows directly from the following two lemmas and a claim (Lemmas 18 and 19 and Claim 20). 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 18 *Under the assumptions of Theorem 8 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 19 *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 18, $\boldsymbol{z}' = \hat{\boldsymbol{\theta}}(\mathcal{D}')$, 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 Wainwright (2006) under the name of *primal-dual construction* and was used to argue consistency in non-private sparse linear regression.

Claim 20 *Under assumptions of Lemma 19, $\hat{\boldsymbol{\theta}}(\mathcal{D})$ 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 [Proof of Lemma 18] In order to prove this lemma, we first prove that the minimizer $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is unique. We use Theorem 7 (which is a modified version of Theorem 1 from Wainwright (2006)) to prove the above claim.

Since from Theorem 7 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 a 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 $\hat{\boldsymbol{\theta}}(\mathcal{D}) = \arg\min_{\boldsymbol{\theta}\in\mathcal{C}} \frac{1}{2n}\|\boldsymbol{y} - X\boldsymbol{\theta}\|_2^2 + \frac{\Lambda}{n}\|\boldsymbol{\theta}\|_1$. ■

Proof [Proof of Lemma 19] 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 7, $\hat{\Gamma}$ 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 21 $\|\boldsymbol{z} - \boldsymbol{z}'\|_2 \leq \frac{4ks^{3/2}}{\Psi n}$.

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 22 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 22 *If $\Lambda = 4\sigma\sqrt{\log p}$, then 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, $\|\boldsymbol{\theta}^*\|_\infty \leq 1 - \max\left\{\frac{16\sigma}{\Psi}\sqrt{\frac{s\log p}{n}}, \frac{8ks^{3/2}}{\Psi n}\right\}$. Also from Theorem 7, 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 21, 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,

$$\begin{aligned} (n+k) \bigtriangledown \hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D}')_i &= n \bigtriangledown \hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D})_i + n(\bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D})_i - \bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D})_i) + \sum_{j=1}^{k}\bigtriangledown\ell(\boldsymbol{z}';\alpha_j)_i \\ \Rightarrow |(n+k) \bigtriangledown \hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D}')_i| &\leq \underbrace{|n \bigtriangledown \hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D})_i|}_{A} + \underbrace{|n(\bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D})_i - \bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D}))_i|}_{B} + \underbrace{\left|\sum_{j=1}^{k}\bigtriangledown\ell(\boldsymbol{z}';\alpha_j)_i\right|}_{C} \end{aligned} \tag{4}$$

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 (4): Notice that term A is equal to $|X^T(\boldsymbol{y} - X\boldsymbol{z})|_i$. We have argued in the proof of Lemma 18, 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

$$\frac{1}{n}\begin{bmatrix} {X_{\Gamma^*}}^T X_{\Gamma^*} & {X_{\Gamma^*}}^T X_{\Gamma^{*c}} \\ {X_{\Gamma^{*c}}}^T X_{\Gamma^*} & {X_{\Gamma^{*c}}}^T X_{\Gamma^*} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_{|\Gamma^*} - \boldsymbol{\theta}^*_{|\Gamma^*} \\ 0 \end{bmatrix} + \frac{1}{n}\begin{bmatrix} {X_{\Gamma^*}}^T \\ X_{\Gamma^{*c}}^T \end{bmatrix} \boldsymbol{w} + \frac{\Lambda}{n}\begin{bmatrix} \boldsymbol{v}_{|\Gamma^*} \\ \boldsymbol{v}_{|\Gamma^{*c}} \end{bmatrix} = 0 \tag{5}$$

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 (5) we have the following.

$$({X_{\Gamma^*}}^T X_{\Gamma^*})(\boldsymbol{z}_{|\Gamma^*} - {\boldsymbol{\theta}^*}_{\Gamma^*}) + {X_{\Gamma^*}}^T \boldsymbol{w} + \Lambda \boldsymbol{v}_{|\Gamma^*} = 0 \tag{6}$$

$$\Leftrightarrow (\boldsymbol{z}_{|\Gamma^*} - {\boldsymbol{\theta}^*}_{|\Gamma^*}) = -({X_{\Gamma^*}}^T X_{\Gamma^*})^{-1} {X_{\Gamma^*}}^T \boldsymbol{w} - \Lambda({X_{\Gamma^*}}^T X_{\Gamma^*})^{-1} \boldsymbol{v}_{|\Gamma^*} \tag{7}$$

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 (5) and (7), we have the following.

$$\begin{aligned}
(X_{\Gamma^{*c}}^T X_{\Gamma^*})(\boldsymbol{z}_{\Gamma^*} - \theta^*_{\Gamma^*}) + {X_{\Gamma^{*c}}}^T \boldsymbol{w} + \Lambda \boldsymbol{v}_{\Gamma^{*c}} &= 0 \\
\Leftrightarrow \boldsymbol{v}_{\Gamma^{*c}} &= \frac{1}{\Lambda}\Big(({X_{\Gamma^{*c}}}^T X_{\Gamma^*})({X_{\Gamma^*}}^T X_{\Gamma^*})^{-1} {X_{\Gamma^*}}^T \boldsymbol{w} - X_{\Gamma^{*c}}^T \boldsymbol{w} \\
&\quad - \Lambda({X_{\Gamma^{*c}}}^T X_{\Gamma^*})({X_{\Gamma^*}}^T X_{\Gamma^*})^{-1} \boldsymbol{v}_{\Gamma^*}\Big) \\
&= -({X_{\Gamma^{*c}}}^T X_{\Gamma^*})({X_{\Gamma^*}}^T X_{\Gamma^*})^{-1} \boldsymbol{v}_{\Gamma^*} \\
&\quad - \frac{{X_{\Gamma^{*c}}}^T}{\Lambda}\left(\mathbb{I}_{n\times n} - X_{\Gamma^*}({X_{\Gamma^*}}^T X_{\Gamma^*})^{-1} X_{\Gamma^*}^T\right)\boldsymbol{w} \\
\Leftrightarrow \|\boldsymbol{v}_{\Gamma^{*c}}\|_\infty &\leq \|({X_{\Gamma^{*c}}}^T X_{\Gamma^*})({X_{\Gamma^*}}^T X_{\Gamma^*})^{-1} \boldsymbol{v}_{\Gamma^*}\|_\infty \\
&\quad + \frac{1}{\Lambda}\|{X_{\Gamma^{*c}}}^T \left(\mathbb{I}_{n\times n} - X_{\Gamma^*}({X_{\Gamma^*}}^T X_{\Gamma^*})^{-1} X_{\Gamma^*}^T\right)\boldsymbol{w}\|_\infty \\
&= \|({X_{\Gamma^{*c}}}^T X_{\Gamma^*})({X_{\Gamma^*}}^T X_{\Gamma^*})^{-1} \boldsymbol{v}_{\Gamma^*}\|_\infty + \\
&\quad + \frac{1}{\Lambda}\|{X_{\Gamma^{*c}}}^T V \boldsymbol{w}\|_\infty
\end{aligned} \tag{8}$$

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 1), 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 (4): 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 21 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 (4): By the definition of $\ell(\boldsymbol{z}; \alpha_j)$ (where $\alpha_j = (y, \boldsymbol{x})$ is as defined in (4)), 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 22, 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 19. ■

To complete the proof of Theorem 8 (utility guarantee), all that is left is to prove Claim 20.

Proof [Proof of Claim 20] We need to show that the supports of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\hat{\boldsymbol{\theta}}(\mathcal{D}')$ are the same. From Lemma 19 it directly follows that $\text{supp}(\hat{\boldsymbol{\theta}}(\mathcal{D}')) \subseteq \text{supp}(\hat{\boldsymbol{\theta}}(\mathcal{D}))$. To prove equality, we provide the following argument.

From Theorem 7 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 21 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. ■

D.2.2. CONSISTENCY AND STABILITY OF HUBERIZED LASSO

In this section, we modify the LASSO program of (2) 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).)

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

$$\tilde{\boldsymbol{\theta}}(\mathcal{D}) = \arg\min_{\boldsymbol{\theta} \in \mathcal{C}} \frac{1}{n} \sum_{i=1}^{n} \hat{\ell}(\boldsymbol{\theta}; y_i, \boldsymbol{x}_i) + \frac{\Lambda}{n} \|\boldsymbol{\theta}\|_1 \tag{9}$$

In this section we show the correctness (Theorem 23) and stability property (Theorem 25) of $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ under Assumption Typical (Assumption 1).

Theorem 23 (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 $\|\tilde{\boldsymbol{\theta}}(\mathcal{D}) - \boldsymbol{\theta}^*\|_\infty \leq \frac{8\sigma}{\Psi}\sqrt{\frac{s \log p}{n}}$.*

Proof [Proof of Theorem 23 (Correctness Theorem)] We first show that the support of $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ in (9) will be the same as the output of LASSO in (2), i.e., the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ in (2) 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 24 *$\tilde{\boldsymbol{\theta}}(\mathcal{D})$ equals $\hat{\boldsymbol{\theta}}(\mathcal{D})$.*

Proof In order to prove this claim, we invoke Theorem 1 from Wainwright (2006) (see Theorem 7). 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 7). 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 (9). 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 (9). 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 18 since $\hat{\boldsymbol{\theta}}(\mathcal{D})$ is the unique minimizer of (2), it follows that $\tilde{\boldsymbol{\theta}}(\mathcal{D}) = \hat{\boldsymbol{\theta}}(\mathcal{D})$. ■

To conclude the proof of Theorem 23, we invoke Theorem 1 from Wainwright (2006). For completeness purposes we provide it in Theorem 7. ■

In the proof of Theorem 23 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 25 we show that as long as the data set size $n = \omega(s \log p, \frac{s^3 k^2 \log n}{\log p}, ks\sqrt{\log n})$, 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 25 is same as the proof structure of Theorem 8 for the unmodified LASSO.

Theorem 25 (Stability of huberized LASSO) *Fix $k > 1$. Under assumptions of Theorem 23 and $n = \omega(s \log p, \frac{s^3 k^2 \log n}{\log p})$, $(\boldsymbol{y}, X, \boldsymbol{\theta}^*)$ 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.*

Proof [Proof of Theorem 25 (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 8), we prove the stability guarantee via Lemma 26 and 27, and Claim 28.

Lemma 26 *Under assumptions of Theorem 23, if $\hat{\Gamma}$ is the support of $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ and $\tilde{\boldsymbol{\theta}}(\mathcal{D})_{\hat{\Gamma}} = \arg\min_{\boldsymbol{\theta} \in \mathcal{C}_{\hat{\Gamma}}} \frac{1}{n}\sum_{i=1}^n \hat{\ell}(\boldsymbol{\theta}; (y_i, X_i)) + \frac{\Lambda}{n}\|\boldsymbol{\theta}\|_1$, 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 27 *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 $\boldsymbol{z}' = \arg\min_{\boldsymbol{\theta} \in \mathcal{C}_{\hat{\Gamma}}} \frac{1}{|\mathcal{D}'|}\sum_{i=1}^{|\mathcal{D}'|} \hat{\ell}(\boldsymbol{\theta}; (y_i', X_i')) + \frac{\Lambda}{|\mathcal{D}'|}\|\boldsymbol{\theta}\|_1$. Under assumptions of Lemma 26, $\boldsymbol{z}' = \tilde{\boldsymbol{\theta}}(\mathcal{D}')$, where $\tilde{\boldsymbol{\theta}}(\mathcal{D}') = \arg\min_{\boldsymbol{\theta} \in \mathcal{C}} \frac{1}{|\mathcal{D}'|}\sum_{i=1}^{|\mathcal{D}'|} \hat{\ell}(\boldsymbol{\theta}; (y_i', X_i')) + \frac{\Lambda}{|\mathcal{D}'|}\|\boldsymbol{\theta}\|_1$.*

To prove the above lemma, we use a proof technique which was developed by Wainwright (2006) under the name of *primal-dual construction* and was used to argue consistency in non-private sparse linear regression.

Claim 28 *Under assumptions of Lemma 27, $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ 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 26 is exactly the same for Lemma 18 in Section D.2.1 and hence omitted here.

Proof [Proof of Lemma 27] 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 23, $\hat{\Gamma}$ 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 29 $\|\boldsymbol{z} - \boldsymbol{z}'\|_2 \leq \frac{10ks\sqrt{\log n}}{\Psi n}$.

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 30 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 30 *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 7, 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 29, 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,

$$\begin{aligned} (n+k) \bigtriangledown \hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D}')_i &= n \bigtriangledown \hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D})_i + n(\bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D})_i - \bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D})_i) + \sum_{j=1}^{k}\bigtriangledown\ell(\boldsymbol{z}';\alpha_j)_i \\ \Rightarrow |(n+k) \bigtriangledown \hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D}')_i| &\leq \underbrace{|n \bigtriangledown \hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D})_i|}_{A} + \underbrace{|n(\bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z}';\mathcal{D})_i - \bigtriangledown\hat{\mathcal{L}}(\boldsymbol{z};\mathcal{D}))_i|}_{B} + \underbrace{\left|\sum_{j=1}^{k}\bigtriangledown\ell(\boldsymbol{z}';\alpha_j)_i\right|}_{C} \end{aligned} \tag{10}$$

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 (10): Notice that term A is equal to $|X^T(\boldsymbol{y} - X\boldsymbol{z})|_i$. We have argued in the proof of Lemma 18, 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

$$\frac{1}{n}\begin{bmatrix} {X_{\Gamma^*}}^T X_{\Gamma^*} & {X_{\Gamma^*}}^T X_{\Gamma^{*c}} \\ {X_{\Gamma^{*c}}}^T X_{\Gamma^*} & {X_{\Gamma^{*c}}}^T X_{\Gamma^*} \end{bmatrix} \begin{bmatrix} \boldsymbol{z}_{|\Gamma^*} - \boldsymbol{\theta}^*_{|\Gamma^*} \\ 0 \end{bmatrix} + \frac{1}{n}\begin{bmatrix} {X_{\Gamma^*}}^T \\ X_{\Gamma^{*c}}^T \end{bmatrix} \boldsymbol{w} + \frac{\Lambda}{n}\begin{bmatrix} \boldsymbol{v}_{|\Gamma^*} \\ \boldsymbol{v}_{|\Gamma^{*c}} \end{bmatrix} = 0 \tag{11}$$

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 (11) we have the following.

$$({X_{\Gamma^*}}^T X_{\Gamma^*})(\boldsymbol{z}_{|\Gamma^*} - {\boldsymbol{\theta}^*}_{\Gamma^*}) + {X_{\Gamma^*}}^T \boldsymbol{w} + \Lambda \boldsymbol{v}_{|\Gamma^*} = 0 \tag{12}$$

$$\Leftrightarrow (\boldsymbol{z}_{|\Gamma^*} - {\boldsymbol{\theta}^*}_{|\Gamma^*}) = -({X_{\Gamma^*}}^T X_{\Gamma^*})^{-1} {X_{\Gamma^*}}^T \boldsymbol{w} - \Lambda({X_{\Gamma^*}}^T X_{\Gamma^*})^{-1} \boldsymbol{v}_{|\Gamma^*} \tag{13}$$

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 11 and 13, we have the following.

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

In the above expression $V = \left(\mathbb{I}_{n\times n} - X_{\Gamma^*}({X_{\Gamma^*}}^T X_{\Gamma^*})^{-1} X^T_{\Gamma^*}\right)$ is a projection matrix. Applying the bounds from Bullets 3 and 5 from Assumption Typical (Assumption 1), 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 (10): The term B is upper bounded by $\|X^T X(\boldsymbol{z}' - \boldsymbol{z})\|_\infty$. 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 $X^T X$ 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 29 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 (10): By the definition of $\ell(\boldsymbol{z}; \alpha_j)$ (where $\alpha_j = (y, \boldsymbol{x})$ is as defined in (10)), 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 30, 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 27. ■

To complete the proof of Theorem 25 (utility guarantee), all is left is to provide the proof for Claim 28.

Proof [Proof of Claim 28] We need to show that the supports of $\tilde{\boldsymbol{\theta}}(\mathcal{D})$ and $\tilde{\boldsymbol{\theta}}(\mathcal{D}')$ are the same. From Lemma 19 it directly follows that $\text{supp}(\tilde{\boldsymbol{\theta}}(\mathcal{D}')) \subseteq \text{supp}(\tilde{\boldsymbol{\theta}}(\mathcal{D}))$. To prove equality, we provide the following argument.

From Theorem 7 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 29 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, $\text{supp}(\tilde{\boldsymbol{\theta}}(\mathcal{D}')) = \text{supp}(\tilde{\boldsymbol{\theta}}(\mathcal{D}))$. This concludes the proof. ■

■

D.2.3. PROOFS OF THEOREMS 9 (k-STABILITY (PROXY VERSION)) AND 10 (Strongly-Typical $\Rightarrow$ k-STABILITY (PROXY VERSION))

Proof [Proof of Theorem 9] The proof of this theorem directly follows from Lemma 31 and Claims 32, 33, and 34 below. We prove these statements after stating them.

Lemma 31 *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 32 *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 $\mathcal{D}$),*

$$\left|g_1(\mathcal{D}) - g_1(\mathcal{D}')\right| \leq \frac{12s^2}{\Psi} = \Delta_1$$

Claim 33 *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 34 *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 [Proof of Lemma 31] We prove the lemma via the following three claims (Claims 35, 36 and 37).

Claim 35 *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 36 *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 37 *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 18, 19 and Claim 20 respectively.

■

Proof [Proof of Claim 32] 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 31 the support of $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\hat{\boldsymbol{\theta}}(\mathcal{D}')$ is the same. We now need to bound the following.

$$(n+1)\bigtriangledown\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}');\mathcal{D}') = n\bigtriangledown\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D});\mathcal{D}) + n(\bigtriangledown\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}');\mathcal{D}) - \bigtriangledown\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D});\mathcal{D})) + \bigtriangledown\ell(\hat{\boldsymbol{\theta}}(\mathcal{D}');d_{new}) \tag{14}$$

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.

$$\left|(n+1)\bigtriangledown\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}');\mathcal{D}')_i - n\bigtriangledown\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D});\mathcal{D})_i\right| \leq \underbrace{n\left|(\bigtriangledown\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D}');\mathcal{D})_i - \bigtriangledown\hat{\mathcal{L}}(\hat{\boldsymbol{\theta}}(\mathcal{D});\mathcal{D})_i)\right|}_{B} + \underbrace{\left|\bigtriangledown\ell(\hat{\boldsymbol{\theta}}(\mathcal{D}');d_{new})_i\right|}_{C} \tag{15}$$

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

Bound on term $B \leq \frac{4s^2}{\Psi}$ in (15): 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 21 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 (15): By the definition of $\ell(\hat{\boldsymbol{\theta}}(\mathcal{D}); \alpha_j)$ (where $\alpha_j = (y, \boldsymbol{x})$ is as defined in (4)), we have $\bigtriangledown \ell(\hat{\boldsymbol{\theta}}(\mathcal{D}); \alpha_j) = -\boldsymbol{x}(y - \langle \boldsymbol{x}, \hat{\boldsymbol{\theta}}(\mathcal{D})\rangle)$. From the assumed bounds on y and $\|\boldsymbol{x}\|_2$, we bound $|\bigtriangledown \ell(\hat{\boldsymbol{\theta}}(\mathcal{D}); \alpha_j)_i|$ by $2s^{3/2}$. Now, it directly follows that the term C is bounded by $2s^{3/2}$. ■

Proof [Proof of Claim 33] From Lemma 31 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 33, 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 [Proof of Claim 34] From Lemma 31 we know that the unique minimizers $\hat{\boldsymbol{\theta}}(\mathcal{D})$ and $\hat{\boldsymbol{\theta}}(\mathcal{D}')$ share the same support.

Now, from Claim 21, 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 [Proof of Theorem 10] 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 7 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 22 (more precisely (8)) 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 $g_1(\mathcal{D}) > t_1 + (k-1)\Delta_1$. ■

D.2.4. STABILITY ANALYSIS OF UNMODIFIED LASSO IN STOCHASTIC SETTING

In order to make sure that Theorem 8 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 8 is from the domain U^*, and ii) $(\hat{\boldsymbol{y}}, \hat{X}, \boldsymbol{\theta}^*)$ satisfy $(s, \Psi, \sigma, \Phi, k)$-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 and $\boldsymbol{y}$ 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 2 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 17, 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 8. 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 $\boldsymbol{w}$.

D.2.5. THEOREM AND PROOF OF LEMMA 13 [DISTANCE PROXY GUARANTEE FOR $\hat{d}$]

Proof In Theorem 9 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 32, 33, and 34 in Section D.2.3.) Now since $\hat{d}$ cannot be negative, global sensitivity of $\hat{d}$ is at most one. ■

Appendix E. Private Support Selection for Sparse Linear Regression (Proofs)

E.1. Support Selection via Sampling Stability

In order to argue that Algorithm 1 outputs the correct support, we make the following assumption (Assumption 4) about the data set $\mathcal{D}$ and the parameter vector $\boldsymbol{\theta}^*$. Under this assumption, we obtain the following utility guarantee for the support selection algorithm as a corollary to Theorem 6.

Assumption 4 *[(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.

Theorem 38 (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 4 (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* $1 - 3\delta$, *Algorithm 1 outputs the correct support of* $\boldsymbol{\theta}^*$.

E.2. Support Selection via Stability of LASSO

Theorem 39 *Let* $\mathcal{D} = (\boldsymbol{y}, X)$ *be a data set from* U^*, *let* $\Lambda = 4\sigma\sqrt{n\log p}$ *and suppose* $s \leq \sqrt{\frac{\sigma n^{1/2}\log^{1/2} p}{2k(1/\Psi+1)}}$. *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* $1 - \delta$.

The proof of this above theorem follows by combining Lemma D.2.5, Proposition 3 and Theorem 2.