Insufficient Statistics Perturbation: Stable Estimators for Private Least Squares

Gavin Brown[*] GRBROWN@CS.WASHINGTON.EDU
Jonathan Hayase[*] JHAYASE@CS.WASHINGTON.EDU
Samuel Hopkins[†] SAMHOP@MIT.EDU
Weihao Kong[‡] KWEIHAO@GMAIL.COM
Xiyang Liu[*] XIYANGL@CS.WASHINGTON.EDU
Sewoong Oh[*] SEWOONG@CS.WASHINGTON.EDU
Juan C. Perdomo[§] JCPERDOMO@G.HARVARD.EDU
Adam Smith[¶] ADS22@BU.EDU

Abstract

We present a sample- and time-efficient differentially private algorithm for ordinary least squares, with error that depends linearly on the dimension and is independent of the condition number of $X^\top X$, where X is the design matrix. All prior private algorithms for this task require either $d^{3/2}$ examples, error growing polynomially with the condition number, or exponential time. Our near-optimal accuracy guarantee holds for any dataset with bounded statistical leverage and bounded residuals. Technically, we build on the approach of Brown et al. (2023) for private mean estimation, adding scaled noise to a carefully designed stable nonprivate estimator of the empirical regression vector.

1. Introduction

We present a sample- and time-efficient differentially private algorithm for ordinary least squares (OLS) regression. Central throughout the theory and practice of data science, OLS is used in numerous domains, ranging from causal inference, to control theory, to (of course) supervised learning.

Given covariates $X \in \mathbb{R}^{n\times d}$ and responses $y \in \mathbb{R}^n$, the OLS estimator is defined as

$$\beta_{\text{ols}} \;=\; \left(X^\top X\right)^{-1} X^\top y\,.$$

Among the many reasons for the popularity of OLS is the fact that it is a statistically and computationally efficient way of solving linear regression. Speaking informally, OLS has low excess error whenever the number of samples n is as large as the problem dimension d. Crucially, its statistical performance does not depend on the condition number $\kappa(X^\top X)$, the ratio between the maximum and minimum eigenvalues. Furthermore, it can be computed in closed-form using only basic linear-algebraic operations, with no need for the subtle hyperparameter tuning often inherent in first-order methods.

[*] Paul G. Allen School of Computer Science and Engineering, University of Washington. Part of this work was done while G.B. was at Boston University.
[†] Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology.
[‡] Google Research
[§] Harvard University
[¶] Department of Computer Science, Boston University.

Given its widespread use in the analysis of personal data, there is a long line of work giving differentially private algorithms to approximate OLS. However, designing practical and efficient algorithms for this problem has been a particularly challenging endeavor; so far there are no clear answers even in the $d = 1$ case when x and y are both scalars (Alabi et al., 2022). Existing algorithms for DP regression suffer from one of three limitations: they either have poor dimension dependence in their sample complexity, place unnaturally restrictive assumptions on the geometry of the data, or run in exponential time.

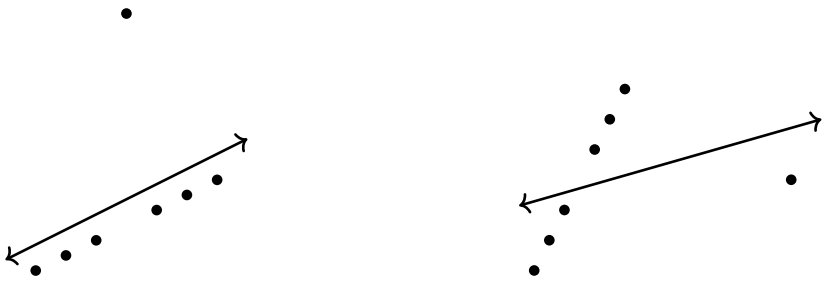
In terms of private algorithm design, one natural and well-established approach is *sufficient statistics perturbation*, which privately produces separate estimates of $X^\top X$ and $X^\top y$ and then combines them to produce a single parameter estimate. Such approaches are often efficient and some versions come with formal accuracy guarantees. An exemplar in this line is the *AdaSSP* algorithm of Wang (2018). The central drawback in all these algorithms, however, is the sample complexity as d grows: privately producing an accurate estimate of $X^\top X$ requires roughly $d^{3/2}$ samples (Dwork et al., 2014). Furthermore, many approaches within this class add noise proportional to the worst-case sensitivity of $X^\top X$ and $X^\top y$ (see, e.g., Sheffet, 2017). To deal with the fact that this sensitivity is *unbounded* in the case of real-valued data, these results assume uniform norm bounds on the covariates x and responses y (e.g., $\|x\| \leq B_y, |y| \leq B_y$). While conceptually simple, they fail to capture the intrinsic complexity of the problem and do not satisfy natural properties like scale invariance.

Another approach comes via private optimization, searching for a parameter estimate that approximately minimizes the sum of squared errors. Despite a wealth private convex optimization methods that can be applied directly to linear regression, off-the-shelf approaches again require $d^{3/2}$ samples for accurate estimates. A notable exception is the recent work from Varshney et al. (2022), whose algorithm based on private gradient descent succeeds with only roughly d samples. However, its error grows with the square of the condition number, a high price to pay for many problems. A polynomial dependence on $\kappa(X^\top X)$ is inherent in private first-order optimization for linear regression, as the smoothness of the optimization task is directly linked to the condition number.

The only known approach that avoids these two issues is the exponential-time algorithm of Liu et al. (2022), which comes from the framework they call *high-dimensional propose-test-release* (HPTR), after the *propose-test-release* (PTR) framework of Dwork and Lei (2009).

We see the mirror of this story in private mean estimation, where Kuditipudi et al. (2023) and Brown et al. (2023) recently gave the first sample- and time-efficient private algorithms with error guarantees that adapt to the covariance of the data. All prior private algorithms achieving this guarantee require $d^{3/2}$ examples, error growing polynomially with the condition number of the covariance, or exponential time.

In this work, we build on the work of Brown et al. (2023) and present the first computationally efficient (in fact, practically implementable) differentially private estimator for linear regression with sample complexity independent of $\kappa(X^\top X)$ and the optimal linear dependence on the dimension d. Furthermore, we make no use of norm bounds. We establish its utility under the "textbook" conditions one would typically require to run OLS in the non-private setting. More specifically, the algorithm is accurate as long no observation has high statistical leverage or a large residual, formalized in Definition 2.



(*a*) Large residual (*b*) High leverage

Figure 1: As famously illustrated by Anscombe (1973), a point may be influential because of its large residual (*a*) or its large leverage (*b*). Definition 6 controls both quantities.

1.1. Our Results

In this work, we introduce a new algorithm, `ISSP`, for differentially private linear regression. At a high level, `ISSP` works in two main phases. In the first, we search for a reweighting of the dataset such that running OLS on this reweighted version is roughly stable. Having successfully found this set of weights, we simply compute the OLS solution on this weighted version of the data and add appropriately shaped Gaussian noise to the solution. While the approach is conceptually very simple, establishing its correctness requires several significant technical advances.

Our estimator satisfies *differential privacy* (DP), the gold standard for privacy protection in statistical data analysis. DP requires that an algorithm provides approximately the same output on any datasets that differ in only one entry.

Definition 1 (Dwork et al. (2006)) *Let $\mathcal{X}$ and $\mathcal{Y}$ be sets. An algorithm $\mathcal{A} : \mathcal{X}^n \to \mathcal{Y}$ is (ε, δ)-differentially private if for every $x = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and $x' = (x'_1, \ldots, x'_n) \in \mathcal{X}^n$ such that x, x' agree on all but one coordinate and for all $Y \subseteq \mathcal{Y}$,*

$$\mathbb{P}\big[\mathcal{A}(x) \in Y\big] \;\leq\; e^{\varepsilon}\,\mathbb{P}\big[\mathcal{A}(x') \in Y\big] + \delta\,.$$

One of the core advances we make, in light of most previous results on DP regression, is that we do not require any norm bounds on the data. We only assume the types of conditions a circumspect statistician would always verify to ensure that OLS is a sensible procedure. In particular, we establish the utility of our estimator whenever the dataset is free of outliers, or "good."

Definition 2 ((L, R)-goodness) *Fix parameters $L, R > 0$. A dataset $(X, y) \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ is called* (L, R)-good *if $X^\top X$ is invertible and the following conditions hold for all $i \in [n]$.*

(1) Bounded leverage: $x_i^\top (X^\top X)^{-1} x_i \leq L$.

(2) Bounded residuals: $\big|\langle x_i, \beta_{\text{ols}}\rangle - y_i\big| \leq R$.

Note that both of these conditions hold in various natural, well-studied settings. For instance, when x, y are both subgaussian and drawn from a well-specified linear model with true parameter β^*, these conditions hold with high probability when $L \approx d/n$ and $R \approx \sigma$ where $\sigma^2 = \mathbb{E}[(y - \langle \beta^*, x\rangle)^2]$ (see Theorem 5). This idea, of outliers being observations with high leverage

or large residuals, is quite classical and found across standard texts. For instance, a standard rule-of-thumb identifies high-leverage points as those with leverage greater than $2d/n$ or $3d/n$ (Hoaglin and Welsch, 1978; Velleman and Welsch, 1981; Mendenhall et al., 2003). The precise forms of stability we need to ensure privacy, however, are far from classical. They require a carefully designed algorithm, which we elaborate in Section 1.3. Such a stable estimator, in turn, implies that we can achieve differential privacy with small amounts of noise.

In these well-behaved instances, our mechanism takes a most straightforward form: it returns the OLS solution plus a small amount of Gaussian noise.

Claim 3 *If (X, y) is (L, R)-good for parameters $R > 0$ and $L \leq c'\varepsilon^2 \log^{-2}(1/\delta)$, for some constant c', then* $\texttt{ISSP}(X, y; \varepsilon, \delta, L, R)$ *releases a sample drawn from $\mathcal{N}\big(\beta_{\text{ols}}, c^2(X^\top X)^{-1}\big)$, where $c^2 = \Theta\left(L\, R^2 \log(1/\delta)/\varepsilon^2\right)$.*

Attentive readers will detect a modest sleight-of-hand: (L, R)-goodness is a property of the data and a priori unknown, yet the algorithm gets L and R as inputs! Nevertheless, an analyst with beliefs about the data generation process can set these parameters appropriately. The maximum leverage score does not depend on the scale of the data, only its concentration properties. Since it lies within $[0, 1]$, one might pick the L hyperparameter adaptively by calling `ISSP` a handful of times. Similarly, if the analyst believes the labels are generated by a process such as $y_i \leftarrow \langle x_i, \beta^* \rangle + \mathcal{N}(0, \sigma^2)$, they can privately produce an accurate estimate σ using standard tools (see Appendix D). We believe alternative standardized or studentized definitions could remove this need for prior knowledge about σ. These alternatives would likely increase the complexity of our proofs.

The difficulty in our work lies in proving that `ISSP` is differentially private (Theorem 5); reasoning about utility is simple once we have Claim 3. More specifically, seeing how the output distribution on good data matches standard statistical practice (and classical CLT-like analyses of OLS), we can quickly derive error bounds. For instance, in the simplest case of *fixed design*, where we consider only the randomness of the labels generated from a well-specified linear model $y_i = \langle x_i, \beta^* \rangle + \mathcal{N}(0, \sigma^2)$, we have $\beta_{\text{ols}} \sim \mathcal{N}(\beta^*, \sigma^2 \cdot (X^\top X)^{-1})$. Hence, from Claim 3, we see that, relative to the empirical OLS solution, the private estimator is just a slightly noisier version of the true parameter (and has the same kind of error covariance).

More formally, we can analyze the mean squared error (MSE) of our algorithm on any good dataset.

Corollary 4 *Fix (X, y), $\varepsilon > 0$, and $\delta \in [0, 1]$. If (X, y) is (L, R)-good for $L \leq c'\varepsilon^2 \log^{-2}(1/\delta)$ for some constant c' and $R > 0$, then* $\texttt{ISSP}(X, y; \varepsilon, \delta, L, R)$*, releases $\tilde{\beta}$ such that, for some absolute constant c',*

$$\mathbb{E}\left[\frac{1}{n}\big\|y - X\tilde{\beta}\big\|^2\right] \;=\; \frac{1}{n}\big\|y - X\beta_{\text{ols}}\big\|^2 + c' L R^2 \,\frac{d}{n}\,\frac{\log 1/\delta}{\varepsilon^2}.$$

Proof By Claim 3, we have $\beta_{\text{ols}} - \tilde{\beta} = c \cdot (X^\top X)^{-1/2} u$ for $u \sim \mathcal{N}(0, \mathbb{I})$. We then expand:

$$\begin{aligned}
\mathbb{E}_u\big\|y - X\tilde{\beta}\big\|^2 &= \mathbb{E}_u\big\|y - X\beta_{\text{ols}} + X(\beta_{\text{ols}} - \tilde{\beta})\big\|^2 \\
&= \mathbb{E}_u\big\|y - X\beta_{\text{ols}} + X \cdot c(X^\top X)^{-1/2} u\big\|^2 \\
&= \|y - X\beta_{\text{ols}}\|^2 + c^2 \cdot \mathbb{E}_u\left[u^\top (X^\top X)^{-1/2} X^\top X (X^\top X)^{-1/2} u\right],
\end{aligned}$$

where the cross terms drop out as u is independent and mean-zero. The matrices cancel and we are left with $\mathbb{E}[u^\top u]$, which is exactly d. ■

We emphasize that this result holds without any assumption that the data arises from a specific family of distributions. It assumes (X, y) is fixed and (L, R)-good to bound the difference from the empirical OLS solution on (X, Y). However, if we do add such distributional assumptions, it is easy to show that our algorithm produces a private parameter estimate that closely approximates the true regression parameter. We state this fact as part of the following theorem, our main result.

Theorem 5 (Main Theorem) *Fix $\varepsilon, \eta \in (0, 1)$, $\delta \in (0, \varepsilon/10]$, and $n, d \in \mathbb{N}$.* `ISSP` *takes a dataset $(X, y) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$, privacy parameters ε, δ, and outlier thresholds L_0, R_0.*

(1) `ISSP` *is (ε, δ)-differentially private.*

(2) Let $X \in \mathbb{R}^{n\times d}$ be drawn i.i.d. from a d-dimensional subgaussian distribution $\mathcal{D}$ with mean 0 and covariance $\Sigma \succ 0$. Let $y_i = \beta^\top x_i + z_i$ where the z_i are drawn i.i.d. from a subgaussian distribution with mean 0 and variance σ^2 (see Definition 29 in Appendix B). [1] *If $L_0 = \widetilde{\Theta}(d/n)$, $R_0 = \widetilde{\Theta}(\sigma)$, and*

$$n \;=\; \widetilde{\Omega}\left(\frac{d}{\alpha^2} + \frac{d\sqrt{\log 1/\delta}}{\alpha\varepsilon} + \frac{d(\log 1/\delta)^2}{\varepsilon^2}\right),$$

with a large enough constant for some $\alpha > 0$, then `ISSP` *returns $\tilde{\beta}$ such that, with high probability,*

$$\left\|\tilde{\beta} - \beta\right\|_\Sigma \;\leq\; \sigma\alpha.$$

Here, $\widetilde{\Theta}$ and $\widetilde{\Omega}$ hide logarithmic factors in $1/\alpha$, $\log(1/\varepsilon)$, and $\log(1/\delta)$ as well as polynomial factors of the subgaussian parameters.

(3) Algorithm 1 can be implemented to require one product of the form $A^\top A$ for $A \in R^{n\times d}$, one product of the form AB for $A \in R^{n\times d}$ and $B \in R^{d\times d}$, one inversion of a positive definite matrix in $R^{d\times d}$; and further computational overhead of $\tilde{O}(nd/\varepsilon)$.

Informally, this running time corresponds to $\widetilde{O}(nd^{\omega-1} + nd/\varepsilon)$, where $\omega < 2.38$ is the matrix multiplication exponent. For modest privacy parameters, the running time of our algorithm is dominated by the time needed to compute the nonprivate OLS solution itself.

This is the first computationally efficient algorithm whose sample complexity is linear in d and has no dependence on the condition number $\kappa(X^\top X)$. This almost matches the best known sample complexity of an exponential-time algorithm from Liu et al. (2022); we have an additional $d(\log(1/\delta))^2/\varepsilon^2$ term, but this term does not depend on the final accuracy α.

We now briefly sketch the steps of the proof and discuss the paper's organization. We establish Theorem 5's subclaims in Lemmas 22, 25 and 26. As outlined above, the utility analysis is straightforward once we have Claim 3 in hand. The full analysis is presented in Section 5. It is easy to see that Algorithm 1 runs in polynomial time. In Section 6, we analyze a careful implementation.

1. We state the utility guarantees of our estimator for the case where data is drawn from a well-specified linear model to simplify the presentation and enable direct comparisons to previous work. However, as per Corollary 4, on good data our algorithm is always close to the OLS solution. Hence, we can prove closeness to the population quantity whenever the OLS solution concentrates.

Algorithm 1: InSufficient Statistics Perturbation (`ISSP`)

Input : dataset (X, y); privacy parameters ε, δ, outlier thresholds (L_0, R_0)

$k \leftarrow \left\lceil \frac{12 \log 3/\delta}{\varepsilon} \right\rceil + 8$; $\quad c^2 \leftarrow 56448 \exp\big(432k^2 L_0\big) L_0 R_0^2 \cdot \frac{\log(12/\delta)}{\varepsilon^2}$;

if $L_0 > 1/(96k)$ *or* $L_0 > 3\varepsilon/(56 \log 12/\delta)$ **then**
  return `FAIL`;
end

$\texttt{SCORE}_1, w \leftarrow \texttt{StableLeverageFiltering}(X, L_0, k)$; `// Algorithm 5`

$\texttt{SCORE}_2, v \leftarrow \texttt{StableResidualFiltering}(X, y, w, L_0, R_0, k)$; `// Algorithm 3`

if $\mathcal{M}_{\mathrm{PTR}}^{\varepsilon/3, \delta/3}(\max\{\texttt{SCORE}_1, \texttt{SCORE}_2\}) = \texttt{FAIL}$ **then**
  return `FAIL`;
else
  $S_v \leftarrow X^\top \operatorname{diag}(v) X$;
  $\hat{\beta} \leftarrow (S_v)^{-1} X^\top \operatorname{diag}(v) y$; `// OLS weighted by` v
  return $\tilde{\beta} \sim \mathcal{N}\big(\hat{\beta}, c^2 S_v^{-1}\big)$;
end

The bulk of the work comes in the privacy analysis. In Section 2, we analyze the greedy residual thresholding algorithm, with the main result about that algorithm being Claim 11, the "intertwining" property. Then, in Section 3, we establish our guarantees for `StableResidualFiltering`. The main results about `StableResidualFiltering` are Claim 13, which says that the score is low-sensitivity, and Claim 14, which says that the weights are stable. Section 4 pulls these together to establish the privacy of `ISSP`.

Appendix A covers additional related work. Appendix B provides necessary preliminaries. Appendix C contains proofs deferred from the main text. Appendix D, via standard tools, shows how to privately estimatie R. Appendix E contains details on the lower bound of Cai et al. (2023).

1.2. Optimality

For modest values of the privacy parameters, the error of our algorithm is dominated by the empirical error of OLS. Informally speaking, we obtain privacy "for free."

Formally, our error guarantees are close to tight for random-design regression with subgaussian covariates and subgaussian label noise. Suppressing constants and logarithmic factors other than $\log 1/\delta$, Theorem 5 says that we can achieve $\|\tilde{\beta} - \beta\|_\Sigma \le \sigma\alpha$ with high probability with

$$n \approx \frac{d}{\alpha^2} + \frac{d\sqrt{\log 1/\delta}}{\alpha\varepsilon} + \frac{d(\log 1/\delta)^2}{\varepsilon^2}.$$

Known lower bounds imply this task requires

$$n \gtrsim \frac{d}{\alpha^2} + \frac{d}{\alpha\varepsilon} + \frac{\log 1/\delta}{\varepsilon}. \tag{1}$$

The first term corresponds to the classical analysis of OLS. The second term was established by Cai et al. (2023) and holds even for parameter estimation in ℓ_2 norm; see Appendix E for a more

detailed discussion. The third term, the minimal number of samples required to produce any estimate, is from Karwa and Vadhan (2018) and holds even for one-dimensional mean estimation with known variance. The exponential-time algorithm of Liu et al. (2022) nearly matches all three terms. For constant ε and $\delta = 1/\mathrm{poly}(n)$, our algorithm's error guarantee in this setting is tight up to logarithmic factors. An exciting topic for future work is determining the existence or impossibility of efficient algorithms with error matching Eq. (1) up to constant factors.

1.3. Techniques

At a high level, our algorithm follows the blueprint for private mean estimation laid out by Brown et al. (2021) and made computationally efficient by Kuditipudi et al. (2023) and Brown et al. (2023). Our approach closely follows that of Brown et al. (2023), henceforth BHS. We now sketch our algorithm, discuss how our analysis differs from that of BHS, and investigate how the notions we use are, in a sense, "correct" for the task of private least squares.

Overview of `ISSP` Perhaps the most natural approach for private estimation of regression coefficients is to perturb the ordinary least square estimator, β_{ols}. However, without restrictions on the data, the sensitivity of β_{ols} is unbounded. Our key observation is that, on datasets with bounded leverage and bounded residuals, the OLS solution is actually quite stable. If we could restrict our inputs to only such outlier-free data sets, we might hope to release β_{ols} plus noise with shape $(X^\top X)^{-1}$.

While this would provide accuracy, it fails on privacy: we must accommodate worst-case data. We use the PTR framework of Dwork and Lei (2009) to test if our input contains a large good subset. We propose a greedy pruning algorithm which, in each iteration, removes the data point with the largest residual and recomputes OLS on the remaining data. Similar approaches abound in the literature on robust statistics, but we prove key new properties about how this algorithm behaves across adjacent data sets and different outlier thresholds.

Adaptively selecting outlier thresholds Our algorithm takes as input target bounds L and R for the leverage and residuals, respectively. This simplifies our analysis but is not strictly necessary. The maximum leverage can only lie within the interval $[0, 1]$, so one could imagine calling `ISSP` repeatedly within this space (via a well-chosen grid or binary search) to find an appropriate setting, perhaps via a small validation set. Independently, one could privately learn an appropriate value for R directly through standard techniques; we give a complete description in Appendix D.

Proof techniques While our work builds on a long line of research connecting robust statistics and differential privacy, it especially relies on the recent algorithmic approach of BHS, who gave improved algorithms for private mean and covariance estimation. At a bird's eye view, our recipe for private linear regression follows the main ideas behind the mean estimator of BHS. However, key parts of the implementation and analysis differ significantly in the more complicated linear regression setting.

We start by discussing the ways in which our main proof strategy is similar to BHS. As mentioned previously, we introduce a notion of "good" outlier-free datasets for linear regression. We repeatedly call a greedy algorithm to find a series of good weight vectors across a range of carefully chosen outlier thresholds. We use these vectors to privately test that our input data is sufficiently close to the good set and to finally produce a vector of weights over the input. Crucially, this weight-

finding procedure is stable: if run on any adjacent dataset, it would produce a vector that is close in ℓ_1 distance.

When adapting their analysis, however, we run into immediate issues. For both their definitions of good set, BHS prove a number of strong properties that are false in the context of regression. For mean and covariance estimation, the good sets are unique (i.e., for any dataset and outlier threshold, there exists a unique largest good set), are directly found by the natural greedy algorithm, and enjoy a form of monotonicity (e.g., introducing a new point to the dataset cannot alter the good set very much). For the definition we use, there is no unique "largest good set" which introduces significantly more complexity in the analysis. What's more, adding a single point may significantly affect the downstream choices made by the greedy algorithm which further complicate the relevant stability calculations.

In more detail, a key step in the BHS analysis establishes the following "intertwining" property in the context of mean and covariance estimation. Suppose we call the greedy algorithm on a dataset D with outlier threshold B and find a largest good subset $S \subseteq [n]$. If we then call the same algorithm on an adjacent dataset D' with a slightly larger outlier threshold B', the largest good subset T will satisfy the property that $S \subseteq T$ (ignoring the index that differs between D and D').

We establish an analogous statement (Claim 11) about the output of our greedy residual thresholding, Algorithm 2, even though we cannot prove the same uniqueness and monotonicity statements. More specifically, we develop a novel regression-specific argument that uses the closed-form expressions describing how the least squares solution changes when an observation is added or removed. The exact arguments are formalized in Claim 7 and Claim 8, but we sketch the ideas here.

For a dataset (X, y) and index $i \in [n]$, let $\hat{y}_i = x_i^\top \beta_{\text{ols}}$ be the fitted value. We denote by e_i the i-th residual: $e_i \stackrel{\text{def}}{=} \hat{y}_i - y_i$. Recall the *hat matrix*:

$$H \stackrel{\text{def}}{=} X(X^\top X)^{-1} X^\top,$$

so called because it maps the true labels to their "hat" values: $\hat{y} = Hy$. The *leverage scores* (also known as *sensitivities* or *self-influences*) form its diagonal entries, while its off-diagonal entries will be called (by us) the *cross-leverage scores*:

$$h_i = H_{i,i} = x_i^\top (X^\top X)^{-1} x_i \quad \text{and} \quad H_{i,j} = x_i^\top (X^\top X)^{-1} x_j.$$

Note that, by Cauchy–Schwarz, the cross-leverages are no larger in magnitude than the leverages.

What happens if we remove an observation, say, (x_j, y_j), from the dataset? This takes the form of a rank-one update. Applying the Sherman-Morrison formula we can derive closed-form expressions for the changes in the OLS solution as well as (for any $i \in [n]$) the leverage score and residual of point i after removing j. Using the subscript "$(-j)$" to denote the quantity after removal, we have

$$\begin{aligned}
h_i - h_{i(-j)} &= -\frac{H_{i,j}^2}{1 - h_j} \\
\beta_{\text{ols}} - \beta_{\text{ols}(-j)} &= \frac{(X^\top X)^{-1} x_j}{1 - h_j} \cdot \big(\langle x_j, \beta_{\text{ols}} \rangle - y_j\big) \\
e_i - e_{i(-j)} &= H_{i,j} \cdot \frac{e_j}{1 - h_j}.
\end{aligned}$$

These well-known formulas have elementary derivations; the second and third correspond to the DFBETA ("difference in β") and DFFITS ("difference in fits") regression diagnostics (see textbooks such as Mendenhall et al., 2003; Belsley et al., 2005; Huber, 2011). All three seamlessly generalize to the case where the points are weighted. We can reuse them to reason about what happens when we add points to the dataset.

Beyond their use in our formal arguments, these formulas show how our goodness definition in Definition 6 is essentially the "right" one to analyze stability. The leverage score and the magnitude of the residual *exactly* determine the sensitivity of the least-squares solution to adding or removing that data point. To see this in more detail, consider the effect of dropping a point from a typical dataset:

$$\begin{aligned}\left\|(X^\top X)^{1/2}(\beta_{\text{ols}} - \beta_{\text{ols}(-j)})\right\|^2 &= \left\|(X^\top X)^{1/2} \cdot \frac{(X^\top X)^{-1}x_j}{1-h_j} \cdot (y_j - \langle x_j, \beta_{\text{ols}}\rangle)\right\|^2 \\ &= \frac{h_j \cdot e_j^2}{(1-h_j)^2} = \Delta.\end{aligned}$$

Arguing heuristically for now, if removing a point changes the OLS solution by Δ, to ensure privacy one must ensure noise of magnitude *at least* Δ. It is impossible to do any better. Note that by working with (L, R)-good sets we can guarantee that the noise we add for privacy, $\mathcal{N}(0, c^2(X^\top X)^{-1})$ where $c^2 \geq LR^2 \cdot \frac{\log 1/\delta}{\varepsilon^2}$, has magnitude roughly Δ. This insight shows our accuracy guarantees are sharp.

1.4. Notation

We use $[n]$ to denote the set $\{1, \ldots, n\}$ and $\mathbb{N} = \{1, 2, \ldots\}$. For a vector $v \in \mathbb{R}^n$ its support is $\text{supp}(v) = \{i \in [n] \mid v_i \neq 0\}$. If we have a set $S \subseteq [n]$, then $\Pi_S(v) \in \mathbb{R}^n$ has $(\Pi_S(v))_i = v_i$ for $i \in S$ and $(\Pi_S(v))_i = 0$ otherwise. Also we define $\overline{S} = [n] \setminus S$. We use $\|v\| \stackrel{\text{def}}{=} \|v\|_2$ and $\|v\|_S \stackrel{\text{def}}{=} \|S^{1/2}v\|$. If $M \in \mathbb{R}^{n\times n}$ is a matrix, then $\|M\|_2$ denotes its spectral norm.

2. Analysis of Greedy Residual Thresholding

Algorithm 2: `ResidualThresholding`

input : dataset X, y; outlier threshold R; starting weights w

while `TRUE` **do**
  $\beta_w \leftarrow \texttt{WeightedOLS}(X, y, w)$;
  $i^* \leftarrow \text{argmax}_{i\in\text{supp}(w)} \left|y_i - x_i^\top \beta_w\right|$;
  if $|y_{i^*} - x_{i^*}^\top \beta_w| > R$ **then**
    $w_{i^*} \leftarrow 0$;
  else
    return w;
  end
end

In this section we establish the key properties of our greedy residual-thresholding algorithm. This analysis contains the bulk of the technical novelty in our work. The main result is Claim 11, the "intertwining" property that relates the outputs of `ResidualThresholding` on adjacent datasets.

Since we will be dealing extensively with weighted sets from now on, we expand the definition of good sets in Definition 2 to vectors of weights.

Definition 6 ((L, R)-goodness, weighted) *Fix a dataset $(X, y) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$ and parameters $L, R > 0$. A vector $w \in [0, 1]^n$ is (L, R)-*good for (X, y) *if, denoting $W = \operatorname{diag}(w)$, $X^\top WX$ is invertible and the following two conditions hold for all $i \in \operatorname{supp}(w)$.*

(1) Bounded leverage: $x_i^\top (X^\top WX)^{-1} x_i \le L$.

(2) Bounded residuals: $\left|\langle x_i, \beta_w\rangle - y_i\right| \le R$, where $\beta_w = (X^\top WX)^{-1} X^\top Wy$.

Furthermore, we will say that w is (L, ∞)-good for (X, y) if (1) holds, but not (2).

2.1. Stability and Goodness for Ordinary Least Squares

Our analyses rely on how goodness is affected when adding and or removing mass. As discussed in Section 1.3, closed-form expressions characterize the effects of removing a single point (Mendenhall et al., 2003; Belsley et al., 2005; Huber, 2011). The following claim generalizes these results to removing multiple weighted points, or adding weight to points already included in the regression. In addition, it shows how these results interact with goodness. We defer the proof to Appendix C.

Claim 7 (Changing Weight Within Support) *Let $w, w' \in [0, 1]^n$ satisfy $\operatorname{supp}(w') \subseteq \operatorname{supp}(w)$ and $\|w - w'\|_1 L \le \frac{1}{2}$. If w is (L, ∞)-good for (X, y), then for all $i \in \operatorname{supp}(w)$,*

$$x_i^\top (X^\top \operatorname{diag}(w')X)^{-1} x_i \le (1 + 2L\|w - w'\|_1)\, L. \tag{2}$$

If, in addition to the previous conditions, it also holds that w is (L, R)-good for (X, y), then

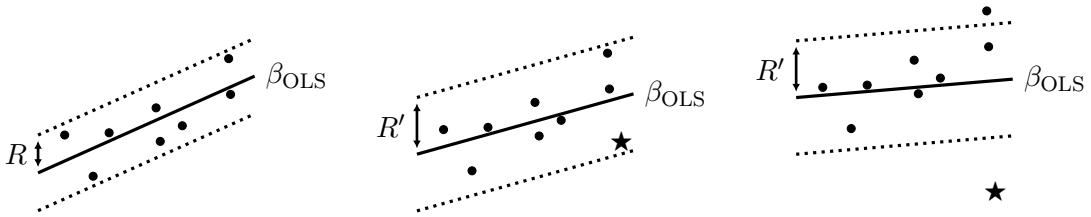
$$|x_i^\top \beta_w - x_i^\top \beta_{w'}| \le 2\|w - w'\|_1 LR. \tag{3}$$

In particular, since $\operatorname{supp}(w') \subseteq \operatorname{supp}(w)$, Equations (2) and (3) apply to all $i \in \operatorname{supp}(w')$. Consequently, w' is $(\eta L, \eta R)$-good for (X, y), where $\eta = 1 + 2L\|w - w'\|_1$.

We next present a claim about adding a point to existing good weights: either the expanded weights are good or the new point has a large residual (in which case our greedy algorithm, presented later, will identify it). We illustrate these cases in Fig. 2. Such a claim also holds when we add sets of points.

Mathematically, this proof contains little innovation beyond Claim 7. However, it provides a key conceptual bridge. We see that it connects directly to our greedy algorithm, which removes large residuals.

Claim 8 (Adding Weight Outside Support) *Let $w' \in [0, 1]^n$ be an (L, R)-good vector for a dataset (X, y) and let $v \in [0, 1]^n$ be a vector such that $\operatorname{supp}(w') \cap \operatorname{supp}(v) = \emptyset$. Define $w = w' + v$ and $\eta = 1 + 8\|v\|_1 L$. Assume the following two conditions hold:*



(*a*) A good dataset. (*b*) Add star with small residual. (*c*) Add star with large residual.

Figure 2: We illustrate our analysis of adding a new observation (star) to an (L, R)-good dataset (*a*). In (*b*), the added star is close to the original regression line. The new largest residual may be greater than R but is less than R'. In (*c*), we instead add a significant outlier. Multiple points may have residuals larger than R', but the largest belongs to the star. In this case, residual thresholding discards the star and recovers the original dataset.

(1) The matrix $X^\top \operatorname{diag}(w)X$ is invertible and for all $j \in \operatorname{supp}(w)$,

$$x_j^\top (X^\top \operatorname{diag}(w)X)^{-1} x_j \leq 2L.$$

(2) The weights v satisfy $\|v\|_1 \cdot L \leq \frac{1}{8}$

If $\max_{i \in \operatorname{supp}(w)} \left|y_i - x_i^\top \beta_w\right| > \eta R$, then $\operatorname{argmax}_{i \in \operatorname{supp}(w)} \left|y_i - x_i^\top \beta_w\right| \subseteq \operatorname{supp}(v)$.

Proof We prove the contrapositive: if there exists $j^* \in \operatorname{argmax}_{i \in \operatorname{supp}(w)} |y_i - x_i^\top \beta_w|$ with $j^* \notin \operatorname{supp}(v)$, then for all $i \in \operatorname{supp}(w)$, $|y_i - x_i^\top \beta_w| \leq \eta R$.

Note that since $\operatorname{supp}(w) = \operatorname{supp}(w') \cup \operatorname{supp}(v)$ and $\operatorname{supp}(w') \cap \operatorname{supp}(v) = \emptyset$, $j^* \notin \operatorname{supp}(v)$ implies $j^* \in \operatorname{supp}(w')$. We first produce a *lower bound* on the j^* residual under w'. By the triangle inequality,

$$\begin{aligned}
\left|e'_{j^*}\right| = \left|y_{j^*} - x_{j^*}^\top \beta_{w'}\right| &= \left|y_{j^*} - x_{j^*}^\top \beta_w + x_{j^*}^\top \beta - x_{j^*}^\top \beta_{w'}\right| \\
&\geq \left|y_{j^*} - x_{j^*}^\top \beta_w\right| - \left|x_{j^*}^\top \beta - x_{j^*}^\top \beta_{w'}\right| \\
&= \left|e_{j^*}\right| - \left|x_{j^*}^\top \beta_w - x_{j^*}^\top \beta_{w'}\right|.
\end{aligned}$$

Note that by assumption, w is $(2L, |e_{j^*}|)$-good for (X, y). Since $\operatorname{supp}(w') \subseteq \operatorname{supp}(w)$, and $\|w - w'\|_1 = \|v\|_1 \leq \frac{1}{8L}$, we can apply Claim 7 to get that $|x_{j^*}^\top \beta_w - x_{j^*}^\top \beta_{w'}| \leq 2\|w - w'\|_1 (2L)|e_{j^*}|$. Using this upper bound, we get that:

$$\left|e'_{j^*}\right| \geq \left|e_{j^*}\right| - 4\left\|w - w'\right\|_1 L \left|e_{j^*}\right|$$

To complete the proof, we use the upper bound $|e'_j| \leq R$, which holds by the assumption that $j \in \operatorname{supp}(w')$ (and that w' is (L, R)-good). Rearranging our previous inequality, we get that for all $i \in \operatorname{supp}(w)$,

$$|y_i - x_i^\top \beta_w| \leq |e_{j^*}| \leq \frac{|e'_{j^*}|}{1 - 4L\|w - w'\|_1} \leq \left(1 + 8L\|w - w'\|_1\right) \cdot R,$$

where we have used the inequality $(1 - z)^{-1} \leq 1 + 2z$, which holds for all $z \in (0, 1/2]$. ■

2.2. Guarantees for Leverage Filtering

`ResidualThresholding` receives as input a vector w, the "starting weights," and iteratively zeros out any weights corresponding to residual outliers, recomputing the weighted OLS solution as it goes. These starting weights w will come from the `StableLeverageFiltering` subroutine of BHS, which filters out high-leverage outliers. The exact algorithm we use differs superficially from the version in BHS, who use it for covariance estimation and call it "Stable Covariance." Our application only needs its properties as a leverage-filtering procedure. We give a complete description of our variant in Appendix B.2.

The filtering algorithm has "goodness" guarantees (when the score is modest, many points receive weight and no point with weight has high leverage), utility guarantees (on outlier-free data, all points receive full weight), and stability guarantees on adjacent datasets (the score is low-sensitivity and the weights are stable). We now give the formal statement.

Theorem 9 (Guarantees for `StableLeverageFiltering`, Brown et al. (2023)) *There is a deterministic algorithm* `StableLeverageFiltering` *receiving as input a list of vectors $X \in \mathbb{R}^{n\times d}$, a leverage threshold L, and a discretization parameter $k \in \mathbb{Z}$ and returning as output an integer* `SCORE` *and a vector $w \in [0,1]^n$. Let $W = \mathrm{diag}(w)$. Assume $kL \le 1$. If* `SCORE` $< k$ *the following hold.*

(1) $\|w\|_1 \ge n - k$. *As a consequence,* $|\mathrm{supp}(w)| \ge n - k$.

(2) For all $i \in \mathrm{supp}(w)$, *we have* $x_i^\top (X^\top W X)^{-1} x_i \le L$.

On "outlier-free" data as defined below, the algorithm's output is as follows.

(3) If $x_i^\top (X^\top X)^{-1} x_i \le L/2e^2$ *for all* $i \in [n]$ *then* `SCORE` $= 0$ *and* $w = \mathbf{1}$.

To present the stability guarantees, let X and X' be datasets that differ in one entry. For any values of k and L, consider

$$\texttt{SCORE}, w \leftarrow \texttt{StableLeverageFiltering}(X, L, k)$$
$$\texttt{SCORE}', w' \leftarrow \texttt{StableLeverageFiltering}(X', L, k).$$

We have the following sensitivity bounds.

(4) $|\texttt{SCORE} - \texttt{SCORE}'| \le 2$.

(5) If `SCORE`, `SCORE`$' < k$ *then* $\|w - w'\|_1 \le 2$.

2.3. Properties of `ResidualThresholding`

The first claim we prove says that, when we run `StableLeverageFiltering` followed immediately by `ResidualThresholding`, the returned weights are good.

Claim 10 *Let (X, y) be a dataset, $k \in \mathbb{N}$ be a discretization parameter, and $L, R > 0$ be outlier thresholds. Assume $kL \le 1/2$. Consider the outputs of the following calls, where the latter uses the output of the former:*

$$\texttt{SCORE}, w \leftarrow \texttt{StableLeverageFiltering}(X, L, k)\ ,$$
$$u \leftarrow \texttt{ResidualThresholding}(X, y, R, w).$$

If `SCORE` $< k$ *and* $\|u\|_1 \ge n - k$ *then u is $(2L, R)$-good for X, y.*

Proof By the guarantees of `StableLeverageFiltering`, Theorem 9, when $\texttt{SCORE} < k$ the weights w give us a bound of L on leverage. That is, for all $i \in \mathrm{supp}(w)$,

$$x_i^\top (X^\top W X)^{-1} x_i \le L.$$

Furthermore, since `ResidualThresholding` only alters w by setting some entries to 0, we have that $\|u\|_1 = \|w\|_1 - \|w - u\|_1$. Using the assumption that $\|u\|_1 \ge n - k$ and the trivial bound $\|w\| \le n$, we get that $\|w - u\|_1 \le k$. Thus, setting $U = \mathrm{diag}(u)$, by the first part of †Claim 7, we have for all $i \in \mathrm{supp}(u)$ that

$$x_i^\top (X^\top U X)^{-1} x_i \le \big(1 + 2L\|w - u\|_1\big) \cdot L \le 2L,$$

where we used the assumption that $Lk \le 1/2$. Since `ResidualThresholding` only returns a vector when the largest absolute residual is no greater than R, we are done. ■

Our next claim relates the runs of residual thresholding on adjacent datasets at nearby residual thresholds. This is the main result about our thresholding procedure.

Claim 11 (Intertwining) *Let (X, y) and (X', y') be adjacent datasets that differ on index i^*. Let $k \in \mathbb{N}$ be a discretization parameter. Let L, R, and $R' > 0$ be any outlier thresholds such that $kL \le \frac{1}{96}$ and $R' \ge \exp(108kL)R$. Consider the outputs of the following calls:*

$$w, \texttt{SCORE} \leftarrow \texttt{StableLeverageFiltering}(X, L, k)$$
$$w', \texttt{SCORE}' \leftarrow \texttt{StableLeverageFiltering}(X', L, k),$$

which we feed into:

$$u \leftarrow \texttt{ResidualThresholding}(X, y, R, w)$$
$$u' \leftarrow \texttt{ResidualThresholding}(X', y', R', w').$$

Define $I = \mathrm{supp}(u) \cap \mathrm{supp}(w') \setminus \{i^*\}$. If $\texttt{SCORE}, \texttt{SCORE}' < k$ and $\|u\|_1 \ge n - k$ then $I \subseteq \mathrm{supp}(u')$.

As we will see in Section 3, where `StableResidualFiltering` uses `ResidualThresholding` to obtain stable weights, many indices of the weights are easily accounted for. This includes i^*, which can be handled as a special case, as well as $\mathrm{supp}(w) \setminus \mathrm{supp}(w')$ and $\mathrm{supp}(w') \setminus \mathrm{supp}(w)$ whose stability is established by Theorem 9. Ignoring those cases for now, we wish to show that any point that is not filtered under (X, y) will also not be filtered under (X', y') provided that the threshold used to filter (X', y') is sufficiently large. We illustrate these cases in Fig. 3. Now we are ready to state the proof of Claim 11.

Proof Our first goal will be to show that $\Pi_I(w')$ is sufficiently good. (Recall our notation: $\Pi_I(w') \in [0, 1]^n$ takes the value w'_i for $i \in I$ and 0 elsewhere.) First, we see that u is $(2L, R)$-good for (X, y) by noting that $\texttt{SCORE} < k$, $\|u\|_1 \ge n - k$, and $kL \le 1/2$ and applying Claim 10. Next, we show that $\Pi_I(w')$ is close to u in ℓ_1 distance. In particular, by definition of `ResidualThresholding`, if $i \in \mathrm{supp}(u)$ then $u_i = w_i$. Hence,

$$\begin{aligned}\big\|u - \Pi_I(w')\big\|_1 &= \sum_{i=1}^{n} \big|u_i - \big(\Pi_I(w')\big)_i\big| \\ &= \sum_{i \in \mathrm{supp}(u)} \big|u_i - \big(\Pi_I(w')\big)_i\big| + \sum_{i \notin \mathrm{supp}(u)} \big|u_i - \big(\Pi_I(w')\big)_i\big|\end{aligned}$$

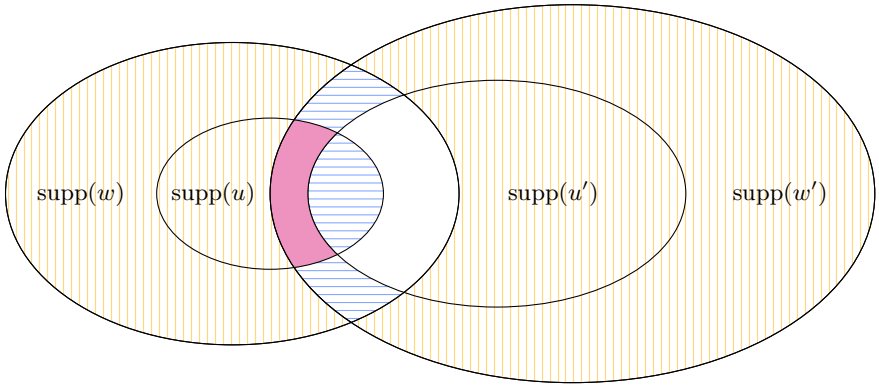


Figure 3: Graphical depiction of Claim 11's "intertwining." Here w represents the weights on dataset (X, y) after leverage filtering and u the weights on (X, y) after residual filtering. w' and u' represent the analogous weights on an adjacent dataset (X', y'). By Theorem 9's guarantees for leverage filtering, the yellow, vertically hatched regions represent a small amount of weight. The blue, horizontally hatched regions represent identical outcomes after residual filtering (either kept on both (X, y) and (X', y') or discarded on both). The claim's main consequence is that only *one* index can fall in the magenta, solid region, which receives weight under u and w' but not u'. This is i^*, the index that differs between (X, y) and (X', y').

If $i \notin \mathrm{supp}(u)$, then $u_i = 0$ and hence $\left(\Pi_I(w')\right)_i = 0$ since $\Pi_I(w')$ is by definition only nonzero outside the support of u. Hence, the second term in the last equation is 0. Moving on, by definition of the set I,

$$\begin{aligned}
\left\|u - \Pi_I(w')\right\|_1 &= \sum_{i \in \mathrm{supp}(u)} \left|w_i - \left(\Pi_I(w')\right)_i\right| \\
&= \sum_{i \in \mathrm{supp}(u)} \left|w_i - \left(\Pi_{\mathrm{supp}(w') \setminus \{i^*\}}(w')\right)_i\right| \\
&\le \left|w_{i^*} - \left(\Pi_{\mathrm{supp}(w') \setminus \{i^*\}}(w')\right)_{i^*}\right| + \sum_{i \in \mathrm{supp}(u)} \left|w_i - \left(\Pi_{\mathrm{supp}(w')}(w')\right)_i\right| \\
&= |w_{i^*} - 0| + \sum_{i \in \mathrm{supp}(u)} |w_i - w'_i| \\
&\le |w_{i^*}| + \|w - w'\|_1 \\
&\le 3 \,.
\end{aligned}$$

The the last inequality follows from the last part of Theorem 9.

Since $I \subseteq \mathrm{supp}(u)$, and $L \le kL \le 1/12$ by assumption, it holds that $\|u - \Pi_I(w')\|_1 \le 3 \le 1/(2L)$, and we can apply Claim 7 to show that $\Pi_I(w')$ is $(2\eta_1 L, \eta_1 R)$-good for (X, y) where $\eta_1 = 1 + 12L$. Furthermore, $\Pi_I(w')$ is $(2\eta_1 L, \eta_1 R)$-good for (X', y') because $(\Pi_I(w'))_{i^*} = 0$.

Now, we will show that during the execution of `ResidualThresholding`(X', y', R', w') we will never discard any $i \in I$. Let $w'^{(j)}$ denote the weights obtained in the j^{th} iteration of the while-

loop out of m total iterations, such that $w'^{(0)} = w'$ and $w'^{(m)} = u'$. We proceed to show the loop invariant $\Pi_I(w'^{(j)}) = \Pi_I(w')$ for $j \in \{0, \dots, m\}$. Since $w' = w'^{(0)}$, the invariant holds initially.

In each iteration, by the loop invariant we can decompose $w'^{(j-1)}$ as

$$w'^{(j-1)} = \Pi_I(w'^{(j-1)}) + \Pi_{\overline{I}}(w'^{(j-1)}) = \Pi_I(w') + \Pi_{\overline{I}}(w'^{(j-1)}).$$

Now, we note that we have $|\mathrm{supp}(u)| \geq \|u\|_1 \geq n - k$ by assumption and since $\texttt{SCORE}' < k$, we also have $|\mathrm{supp}(w')| \geq n - k$ by Theorem 9. Therefore by inclusion-exclusion, $\left|\overline{I}\right| \leq 2k + 1$ and so $\left\|\Pi_{\overline{I}}(w'^{(j-1)})\right\| \leq 2k + 1$. Now since $96kL \leq 1$ by assumption, we note that

$$2\left\|\Pi_{\overline{I}}(w'^{(j-1)})\right\|\eta_1 L \leq 2(2k+1)\eta_1 L \leq 2(2k+1)(1+12L)L \leq 12kL \leq \frac{1}{8}$$

and for $\eta_2 = 1 + 16(2k+1)\eta_1 L$,

$$\begin{aligned}
\eta_2\eta_1 R &= (1 + 16(2k+1)(1+12L)L)(1+12L)R \\
&\leq (1 + 96kL)(1+12L)R \\
&\leq \exp(96kL)\exp(12L)R \\
&\leq \exp(108kL)R \\
&\leq R'.
\end{aligned}$$

Thus by the $(2\eta_1 L, \eta_1 R)$-goodness of $\Pi_I(w')$ and Claim 8, if $\max|y_i - x_i^\top \beta_{w'^{(j-1)}}| \geq R'$ then $\mathrm{argmax}|y_i - x_i^\top \beta_{w'^{(j-1)}}| \notin I$ and so $\Pi_I(w'^{(j)}) = \Pi_I(w')$. Finally, it follows that $\Pi_I(u') = \Pi_I(w')$. Therefore, since $I \subseteq \mathrm{supp}(w')$, we have $I \subseteq \mathrm{supp}(u')$. ■

We now observe that our greedy residual thresholding subroutine only removes more points when run with smaller thresholds. We now state this simple fact for future reference.

Observation 12 *Fix a dataset (X, y) and starting weights w. For outlier thresholds $R \leq R'$, consider running*

$$\begin{aligned}
u &\leftarrow \texttt{ResidualThresholding}(X, y, R, w) \\
u' &\leftarrow \texttt{ResidualThresholding}(X, y, R', w).
\end{aligned}$$

For all $i \in [n]$, $u_i \leq u'_i$.

3. Analysis of `StableResidualFiltering`

In this section, we prove the stability guarantees for Algorithm 3, our new regression estimator. Algorithm 3 repeatedly calls Algorithm 2, `ResidualThresholding`, over a range of slowly increasing outlier thresholds. These thresholds are indexed by a number $j \in \{0, 1, \dots, 2k\}$, where k is a discretization parameter. (Later, we connect this discretization to the privacy parameters, setting $k \approx \log(1/\delta)/\varepsilon$.) The key lemma used in these proofs is Claim 11, which relates the weights found on a dataset (X, y) at level j to the weights found on an adjacent dataset (X', y') at level $j + 1$.

We start by showing that the `SCORE` value and the weight vector, v, returned by Algorithm 3 are low-sensitivity.

Algorithm 3: `StableResidualFiltering`

input : dataset X, y; base outlier thresholds L_0, R_0; weights w; discretization parameter k

$\forall j \in [2k], R_j \leftarrow (\exp(108kL))^j \cdot R_0$;
for $j \in \{0, \dots, 2k\}$ **do**
  $u^{(j)} \leftarrow \texttt{ResidualThresholding}(X, y, R_j, w)$;
  $\texttt{SCORE}^{(j)} \leftarrow \min\{k, n - \|u^{(j)}\|_1 + j\}$;
end
$\texttt{SCORE} \leftarrow \min_{j \in \{0,\dots,k\}} \texttt{SCORE}^{(j)}$;
$v \leftarrow \frac{1}{k} \sum_{j=k+1}^{2k} u^{(j)}$;
return $\texttt{SCORE}, v$

Claim 13 (Score is Low-Sensitivity) *Let (X, y) and (X', y') be adjacent datasets. Fix outlier thresholds L, R and discretization parameter k. Assume $kL \le \frac{1}{96}$. Let*

$$\texttt{SCORE}_1, w \leftarrow \texttt{StableLeverageFiltering}(X, L, k)$$
$$\texttt{SCORE}'_1, w' \leftarrow \texttt{StableLeverageFiltering}(X', L, k)$$

and

$$\texttt{SCORE}_2, v \leftarrow \texttt{StableResidualFiltering}(X, y, w, L, R, k)$$
$$\texttt{SCORE}'_2, v' \leftarrow \texttt{StableResidualFiltering}(X', y', w', L, R, k).$$

If $\texttt{SCORE}_1, \texttt{SCORE}'_1 < k$, then $|\texttt{SCORE}_2 - \texttt{SCORE}'_2| \le 4$.

Proof We observe that all `SCORE` variables are at most k by construction. Without loss of generality, assume $\texttt{SCORE}_2 \le \texttt{SCORE}'_2$. First, we consider the case when $\texttt{SCORE}_2 = k$. In this setting, since $\texttt{SCORE}'_2 \le k$ we must have $\texttt{SCORE}_2 = \texttt{SCORE}'_2$, hence $|\texttt{SCORE}_2 - \texttt{SCORE}'_2| = 0$ and we are done.

Now, consider the case when $\texttt{SCORE}_2 < k$. Then, by definition of the `StableResidualFiltering` algorithm, there must exist a $j^* \in \{0, \dots, k\}$ such that $n - \|u^{(j^*)}\|_1 + j^* = \texttt{SCORE}_2$, where $u^{(j^*)}$ are the weights returned by the `ResidualThresholding` subroutine (run within `StableResidualFiltering`) at outlier threshold R_{j^*}.

Let $u = u^{(j^*)}$ and $u' = (u')^{(j^*+1)}$ denote the weights returned by `ResidualThresholding` on dataset and outlier thresholds $(X, y), R_{j^*}$ and $(X', y'), R_{j^*+1}$ respectively. Defining I as in Claim 11, we note that

$$\|u'\|_1 \ge \left\|\Pi_I(u')\right\|_1 = \left\|u - (u - \Pi_I(u'))\right\|_1 \ge \|u\|_1 - \|u - \Pi_I(u')\|_1.$$

Now, seeking to bound the last term using Claim 11, we note that

(1) $R_{j^*+1}/R_{j^*} \ge \exp(108kL)$ by the definition in Algorithm 3,

(2) $kL \le \frac{1}{96}$ by assumption

(3) Since, $\|u^{(j^*)}\|_1 = \|u\|_1$ and $\texttt{SCORE}_2 = n - \|u^{(j^*)}\|_1 + j^* < k$, it holds that

$$\|u\|_1 > n - k + j^* \ge n - k.$$

Therefore, we can apply Claim 11, which implies that $\Pi_I(u') = \Pi_I(w')$. This gives

$$\begin{aligned}\|u - \Pi_I(u')\| &= \left\|\Pi_{\mathrm{supp}(u)}(w) - \Pi_I(w')\right\| \\ &= \left\|\Pi_{\mathrm{supp}(u)}(w) - \Pi_{\mathrm{supp}(u)\setminus\{i^*\}}(w')\right\| \\ &\le \left\|\Pi_{\mathrm{supp}(u)}(w - w')\right\| + 1 \\ &\le 3.\end{aligned}$$

Finally, combining the previous results gives

$$\begin{aligned}\texttt{SCORE}_2' &\le n - \left\|(u')^{(j^*+1)}\right\|_1 + (j^* + 1) \\ &\le n - (\|u^{(j^*)}\|_1 - 3) + j^* + 1 \\ &= (n - \|u^{(j^*)}\|_1 + j^*) + 4 \\ &= \texttt{SCORE}_2 + 4.\end{aligned}$$

The first inequality in the calculation about holds by definition of `StableResidualFiltering`. The second one uses our previous two calculations. ■

Claim 14 (Weights are Stable) *Let (X, y) and (X', y') be adjacent datasets. Fix outlier thresholds L, R and discretization parameter k. Assume $kL \le \frac{1}{96}$. Let*

$$\texttt{SCORE}_1, w \leftarrow \texttt{StableLeverageFiltering}(X, L, k)$$
$$\texttt{SCORE}_1', w' \leftarrow \texttt{StableLeverageFiltering}(X', L, k)$$

and

$$\texttt{SCORE}_2, v \leftarrow \texttt{StableResidualFiltering}(X, y, w, L, R, k)$$
$$\texttt{SCORE}_2', v' \leftarrow \texttt{StableResidualFiltering}(X', y', w', L, R, k).$$

If $\texttt{SCORE}_1, \texttt{SCORE}_1', \texttt{SCORE}_2, \texttt{SCORE}_2' < k$, then $\|v - v'\|_1 \le 5$.

Proof Consider the execution of `ResidualThresholding` resulting in a weight vector u. We observe that `ResidualThresholding` receives weight vector w as input and modifies it by setting a subset of the weights to zero. Thus we can write the weight $u_i = w_i \cdot 1\{u_i \neq 0\}$. This (rather trivial) modification allows us to write the output of `StableResidualFiltering` in terms of counts: letting $c_i = \sum_{j=k+1}^{2k} 1\{u_i^{(j)} \neq 0\}$, we have

$$v_i = \frac{1}{k}\sum_{j=k+1}^{2k} u_i^{(j)} = \frac{w_i}{k}\sum_{j=k+1}^{2k} 1\{u_i^{(j)} \neq 0\} = \frac{w_i c_i}{k}.$$

Now note that by Observation 12, for $j \in \{k+1, \ldots, 2k\}$ we can have $u_i^{(j-1)} \neq 0$ only if $u_i^{(j)} \neq 0$. This implies that $u_i^{(2k-c_i)} \neq 0$ and $u_i^{(2k-c_i-1)} = 0$.

Now, since $\texttt{SCORE}_2 < k$, we know that there exists some $j^* \in \{0, \ldots, k\}$ such that $n - \left\|u^{(j^*)}\right\|_1 + j^* < k$. Applying, Observation 12 again, we see that $\|u^{(j)}\|_1 \ge \|u^{(j^*)}\|_1 > n - k$ for all $j \ge j^*$. From this, we can conclude that $|\{i \mid c_i \neq k\}| < k$.

Now, define c' analogously as the counts under (X', y') and note that all of the previous observations apply under (X', y') as well. Consider some $i \in \operatorname{supp}(w) \cap \operatorname{supp}(w') \setminus \{i^*\}$ and suppose without loss of generality that $c_i \geq c'_i$. Our goal will be to show that $c'_i \leq c_i \leq c'_i + 1$. If we have $c_i = k$, then $c_i = c'_i$, so we turn our attention to the case where $c_i < k$. We know that $u_i^{(2k-c_i)} \neq 0$. Now, seeking to show that $(u'_i)^{(2k-c_i+1)} \neq 0$ using Claim 11, we note that

(1) $u_i^{(2k-c_i)}$ and $(u'_i)^{(2k-c_i+1)}$ were computed using outlier thresholds R_{2k-c_i} and R_{2k-c_i+1} which satisfy $R_{2k-c_i+1}/R_{2k-c_i} \geq \exp(108kL)$ by the definition in Algorithm 3,

(2) we have $kL \leq \frac{1}{96}$ by assumption, and

(3) we have $\left\|u_i^{(2k-c_i)}\right\|_1 > n - k$ as we observed previously.

Therefore we can apply Claim 11, which implies that $(u'_i)^{(2k-c_i+1)} \neq 0$. Recalling our previous observation, we obtain $c_i \leq c'_i + 1$ as desired. In summary, if $i \in \operatorname{supp}(w) \cap \operatorname{supp}(w') \setminus \{i^*\}$ then we can write $c'_i = c_i + \Delta_i$ where $|\Delta_i| \leq 1$.

Now, define $D = \{i \in \operatorname{supp}(w) \cap \operatorname{supp}(w') \setminus \{i^*\} \mid \Delta_i \neq 0\}$ and note that $|D| \leq 2k$ since c_i and c'_i both contain at most k elements not equal to k.

Now, we are ready to complete the proof by noting that we can decompose the quantity we wish to bound into four terms

$$k\|v - v'\|_1 = \left|c_{i^*} w_{i^*} - c'_{i^*} w'_{i^*}\right| + \sum_{\substack{i \in \operatorname{supp}(w) \\ i \notin \operatorname{supp}(w') \\ i \neq i^*}} |c_i w_i| + \sum_{\substack{i \notin \operatorname{supp}(w) \\ i \in \operatorname{supp}(w') \\ i \neq i^*}} |c'_i w'_i| + \sum_{\substack{i \in \operatorname{supp}(w) \\ i \in \operatorname{supp}(w') \\ i \neq i^*}} |c_i w_i - c'_i w'_i|$$

This is valid because $|c_i w_i - c'_i w'_i|$ appears exactly once on the right for each i. Now we will consider each term separately. The first term is at most k because c_i, c'_i are bounded by k and w_i, w'_i are bounded by 1. The summands of the second and third terms can be rewritten as $c_i|w_i - w'_i|$ and $c'_i|w_i - w'_i|$ and are thus bounded by $k|w_i - w'_i|$ and $k|w_i - w'_i|$ respectively. Now focusing on the last term, we have

$$\begin{aligned}
\sum_{\substack{i \in \operatorname{supp}(w) \\ i \in \operatorname{supp}(w') \\ i \neq i^*}} |c_i w_i - c'_i w'_i| &\leq \sum_{\substack{i \in \operatorname{supp}(w) \\ i \in \operatorname{supp}(w') \\ i \neq i^*}} \left(|c_i w_i - c_i w'_i| + |\Delta_i w'_i|\right) \\
&\leq \sum_{\substack{i \in \operatorname{supp}(w) \\ i \in \operatorname{supp}(w') \\ i \neq i^*}} |c_i w_i - c_i w'_i| + \sum_{\substack{i \in \operatorname{supp}(w) \\ i \in \operatorname{supp}(w') \\ i \neq i^* \\ i \in D}} |\Delta_i w'_i| \\
&\leq \sum_{\substack{i \in \operatorname{supp}(w) \\ i \in \operatorname{supp}(w') \\ i \neq i^*}} k|w_i - w'_i| + 2k.
\end{aligned}$$

Combining the bounds for each term, we have

$$\begin{aligned} k\|v - v'\|_1 &\le 3k + \sum_{\substack{i \in \mathrm{supp}(w) \\ i \notin \mathrm{supp}(w') \\ i \neq i^*}} k|w_i - w'_i| + \sum_{\substack{i \notin \mathrm{supp}(w) \\ i \in \mathrm{supp}(w') \\ i \neq i^*}} k|w_i - w'_i| + \sum_{\substack{i \in \mathrm{supp}(w) \\ i \in \mathrm{supp}(w') \\ i \neq i^*}} k|w_i - w'_i| \\ &\le 3k + k\sum_i |w_i - w'_i| \\ &= 3k + k\|w - w'\|_1 \\ &\le 5k \end{aligned}$$

where the last line is an application of Theorem 9. ■

The previous claim shows that the weights produced by `StableResidualFiltering` on adjacent datasets are close in ℓ_1 (when `SCORE` is less than k). In the next claim, we prove that the weights are good. Note that this does not follow immediately from Claim 10, which says that the weights returned by `ResidualThresholding` are good. Since `StableResidualFiltering` returns an average of the vectors returned by `ResidualThresholding`, we have to argue that the average of good sets is good.

Claim 15 (Weights are Good) *Fix a dataset (X, y), outlier thresholds L, R, and discretization parameter k. Assume $Lk \le \frac{1}{4}$. Consider the following calls:*

$$\begin{aligned} \texttt{SCORE}_1, w &\leftarrow \texttt{StableLeverageFiltering}(X, L, k) \\ \texttt{SCORE}_2, v &\leftarrow \texttt{StableResidualFiltering}(X, y, w, L, R, k). \end{aligned}$$

Then the vector v is $(4L, 2R_{2k})$-good for (X, y), where $R_{2k} = \exp(216k^2L) \cdot R$.

Proof `StableResidualFiltering` calls `ResidualThresholding` repeatedly, producing a vector $u^{(j)}$ for each residual threshold R_j. Recall from Observation 12 that for all $i \le j$, since $R_i \le R_j$ (within `StableResidualFiltering`) we have that $u^{(i)} \le u^{(j)}$ elementwise. This implies that the support of $u^{(2k)}$ contains all other supports, including that of the average v. Additionally, since $\texttt{SCORE}_2 < k$ (by construction within the algorithm), there exists some $j^* \in \{0, \dots, k\}$ with $\|u^{(j^*)}\|_1 \ge n - k$, so the same lower bound holds for all $j \ge k$. Together, these facts imply that $\|u^{(2k)} - u^{(j)}\|_1 \le k$ for all $j \ge k$. Since the ℓ_1 norm is convex and $v = \mathbb{E}_j[u^{(j)}]$, by Jensen's inequality we have $\|u^{(2k)} - v\|_1 \le k$ as well.

To finish the proof, we apply Claim 7: since $\mathrm{supp}(v) \subseteq \mathrm{supp}\big(u^{(2k)}\big)$, $\|u^{(2k)} - v\|_1 \le k$, and $u^{(2k)}$ is $(2L, R_{2k})$-good, we conclude that v is (L', R')-good for

$$\begin{aligned} L' &\le (1 + 4Lk) \cdot 2L \le 4L \\ R' &\le (1 + 4Lk) \cdot R_{2k} \le 2R_{2k}. \end{aligned}$$

Recalling that $R_{2k} = (\exp(108kL))^{2k} \cdot R_0$, we finish the proof. ■

4. Privacy Analysis of Algorithm 1

Our privacy analysis follows the blueprint established by Brown et al. (2021); Kuditipudi et al. (2023); Brown et al. (2023). We use the well-known propose-test-release (PTR) framework of Dwork and Lei (2009) and first privately check (via our low-sensitivity SCORE) if it is safe to proceed. If this check passes, we compute a vector of weights $v \in [0,1]^n$. We use this vector to compute a weighted covariance S_v and weighted least squares solution $\hat{\beta}_v$. The output is then drawn from $\mathcal{N}(\hat{\beta}_v, c^2 S_v^{-1})$ for some appropriate constant c.

On adjacent datasets, we may compute different weights v, v'. We know that, when the PTR checks pass, these vectors are close in ℓ_1. The main work in this section, then, lies in connecting this stability of weights to stability of parameters, which in turn implies $\mathcal{N}(\hat{\beta}_v, c^2 S_v^{-1}) \approx_{(\varepsilon,\delta)} \mathcal{N}(\hat{\beta}_{v'}, c^2 S_{v'}^{-1})$. Note that this is more complicated than the standard Gaussian mechanism, since both the shape and location of the noise change.

Before proving Lemma 22, our main privacy claim, we collect the necessary statements. First, we recall the privacy check of BHS, which (in place of the standard Laplace-noise-and-threshold) simplifies our analysis.

Claim 16 (PTR Mechanism) *Fix $0 < \varepsilon \le 1$, $0 < \delta \le \frac{\varepsilon}{10}$, and $0 < \Delta$. There is an algorithm $\mathcal{M}_{\mathrm{PTR}}^{\varepsilon,\delta} : \mathbb{R} \to \{\texttt{PASS}, \texttt{FAIL}\}$ that satisfies the following conditions:*

(1) Let $\mathcal{U}$ be a set and $g : \mathcal{U}^n \to \mathbb{R}_{\ge 0}$ a function. If, for all $x, x' \in \mathcal{U}^n$ that differ in one entry, $|g(x) - g(x')| \le \Delta$, then $\mathcal{M}_{\mathrm{PTR}}^{\varepsilon,\delta}(g(\cdot))$ is (ε,δ)-DP.

(2) $\mathcal{M}_{\mathrm{PTR}}^{\varepsilon,\delta}(0) = \texttt{PASS}$.

(3) For all $z \ge \frac{\Delta \log 1/\delta}{\varepsilon} + 2\Delta$, $\mathcal{M}_{\mathrm{PTR}}^{\varepsilon,\delta}(z) = \texttt{FAIL}$.

The next claim relates bounded leverage, ℓ_1 closeness, and covariance closeness. This statement comes directly from BHS, Lemma 23; similar claims were used in Brown et al. (2021); Kuditipudi et al. (2023). We use the notation $d_{\mathrm{PD}}(S_1, S_2)$ to denote the maximum of $\left\| S_1^{-1/2} S_2 S_1^{-1/2} - \mathbb{I} \right\|_{\mathrm{tr}}$ and $\left\| S_2^{-1/2} S_1 S_2^{-1/2} - \mathbb{I} \right\|_{\mathrm{tr}}$. Recall that $d_{\mathrm{PD}}(S_1, S_2) = d_{\mathrm{PD}}(S_1^{-1}, S_2^{-1})$ (Fact 27).

Claim 17 *Let $L \in (0,1)$ and let $X, X' \in \mathbb{R}^{n \times d}$ be adjacent (i.e., they differ in one out of n rows). For vectors $v, w \in [0,1]^n$, let $S_v = X^\top \operatorname{diag}(v) X$ and $S_w = (X')^\top \operatorname{diag}(w) X'$. Suppose v and w both have bounded leverage: for all $i \in \operatorname{supp}(v)$, $x_i^\top S_v^{-1} x_i \le L$ and for all $j \in \operatorname{supp}(w)$, $x_j^\top S_w^{-1} x_j \le L$. Then S_v and S_w are positive definite and, if $(1 + \|v - w\|_1)L \le \frac{1}{2}$, satisfy*

$$d_{\mathrm{PD}}(S_v, S_w) \le 2\big(2 + \|v - w\|_1\big) L.$$

An analogous claim says that, if we have two vectors v and v' that are (L,R)-good on adjacent datasets and are close in ℓ_1, then the regression parameters they induce are close. We defer the proof to Appendix C, as similar claims appear in the robust statistics literature (Klivans et al., 2018; Bakshi and Prasad, 2021).

Claim 18 *Let (X, y) and (X', y') be datasets differing in one entry. Let vector v be (L,R)-good for (X, y) and let vector w be (L,R)-good for (X', y'). Set $V = \operatorname{diag}(v)$ and likewise W. Let $S_v = X^\top V X$, $\beta_v = (X^\top V X)^{-1} X^\top V y$, and $\beta_w = ((X')^\top W X')^{-1} (X')^\top W y'$. Assume $(\|v - w\|_1 + 2)L \le \frac{1}{4}$. We have $\|S_v^{1/2}(\beta_v - \beta_w)\|^2 \le 4(\|v - w\|_1 + 2)^2 L R^2$.*

We use the following relationship between the closeness of covariance matrices and the indistinguishability of their induced Gaussians (as in Brown et al., 2021; Alabi et al., 2023; Kuditipudi et al., 2023; Brown et al., 2023).

Claim 19 *Fix $\varepsilon \in (0,1)$ and $\delta \in (0, 1/10]$ and let $S_1, S_2 \in \mathbb{R}^{d\times d}$ be positive definite matrices. If $d_{\mathrm{PD}}(S_1, S_2) \le \frac{\varepsilon}{3\log 2/\delta}$ then $\mathcal{N}(0, S_1) \approx_{(\varepsilon,\delta)} \mathcal{N}(0, S_2)$.*

We also need two standard privacy facts: privacy of the Gaussian mechanism and the DP almost-triangle inequality.

Fact 20 (Gaussian Mechanism) *Fix $\varepsilon, \delta \in (0,1)$ and let u, v be vectors. If $\|u - v\|_2 \le \Delta$, then for any $c^2 \ge \Delta^2 \cdot \frac{2\log 2/\delta}{\varepsilon^2}$ we have $\mathcal{N}(u, c^2\mathbb{I}) \approx_{(\varepsilon,\delta)} \mathcal{N}(v, c^2\mathbb{I})$.*

Fact 21 (See Vadhan (2017)) *Suppose for some ε and δ that distributions p_1, p_2, and p_3 satisfy $p_1 \approx_{(\varepsilon,\delta)} p_2$ and $p_2 \approx_{(\varepsilon,\delta)} p_3$. Then $p_1 \approx_{(2\varepsilon,(1+e^{\varepsilon})\delta)} p_3$.*

We are now ready to prove our main privacy claim.

Lemma 22 (Main privacy guarantee) *For $\varepsilon \in (0,1)$, $\delta \in (0, \varepsilon/10]$, and $L_0, R_0 > 0$, Algorithm 1 is (ε, δ)-differentially private.*

Proof Consider the execution of Algorithm 1 on two adjacent datasets (X, y) and (X', y'), yielding $\texttt{SCORE}_1, \texttt{SCORE}_2, v, \hat{\beta}$ and $\texttt{SCORE}'_1, \texttt{SCORE}'_2, v', \hat{\beta}'$ respectively. Note that in order to not immediately fail, we must have

$$L_0 \le \min\left\{\frac{1}{96k}, \frac{3\varepsilon}{56\log 12/\delta}\right\}$$

where $k = \lceil (12\log 3/\delta)/\varepsilon \rceil + 8$.

Privacy of the test First, we will show that

$$\left|\max\{\texttt{SCORE}_1, \texttt{SCORE}_2\} - \max\{\texttt{SCORE}'_1, \texttt{SCORE}'_2\}\right| \le 4.$$

By Theorem 9, we have $|\texttt{SCORE}_1 - \texttt{SCORE}'_1| \le 2$. Without loss of generality, assume that $\texttt{SCORE}_1 \ge \texttt{SCORE}'_1$.

Considering the case where $\texttt{SCORE}_1 = k$, we have $\max\{\texttt{SCORE}_1, \texttt{SCORE}_2\} = k$ and

$$\max\{\texttt{SCORE}'_1, \texttt{SCORE}'_2\} \ge \texttt{SCORE}'_1 \ge \texttt{SCORE}_1 - 2 \ge k - 2$$

so in this case, $|\max\{\texttt{SCORE}_1, \texttt{SCORE}_2\} - \max\{\texttt{SCORE}'_1, \texttt{SCORE}'_2\}| \le 2$.

Now if $\texttt{SCORE}_1 < k$ then $\texttt{SCORE}'_1 < k$ as well, so we can apply Claim 13 to get $|\texttt{SCORE}_2 - \texttt{SCORE}'_2| \le 4$. Then by noting that $\max$ is 1-Lipschitz in the ∞-norm, we have

$$\begin{aligned}
&\left|\max\{\texttt{SCORE}_1, \texttt{SCORE}_2\} - \max\{\texttt{SCORE}'_1, \texttt{SCORE}'_2\}\right| \\
&\quad\le \max\left\{|\texttt{SCORE}_1 - \texttt{SCORE}'_1|, |\texttt{SCORE}_2 - \texttt{SCORE}'_2|\right\} \\
&\quad\le \max\{2, 4\} \\
&\quad\le 4.
\end{aligned}$$

Finally we see that $\mathcal{M}_{\mathrm{PTR}}^{\varepsilon/3,\delta/3}(\max\{\texttt{SCORE}_1, \texttt{SCORE}_2\})$ is $(\varepsilon/3, \delta/3)$-DP by Claim 16.

Privacy of the parameter estimate Now we will proceed under the assumption that

$$\mathcal{M}_{\mathrm{PTR}}^{\varepsilon/3,\delta/3}(\max\{\texttt{SCORE}_1, \texttt{SCORE}_2\}) = \mathcal{M}_{\mathrm{PTR}}^{\varepsilon/3,\delta/3}(\max\{\texttt{SCORE}'_1, \texttt{SCORE}'_2\}) = \texttt{PASS},$$

with the goal of showing that $\mathcal{N}(\hat{\beta}, c^2 S_v^{-1}) \approx_{2\varepsilon/3, 2\delta/3} \mathcal{N}(\hat{\beta}', c^2 S_{v'}^{-1})$. Since the PTR checks passed, Claim 16 says that

$$\texttt{SCORE}_1, \texttt{SCORE}_2, \texttt{SCORE}'_1, \texttt{SCORE}'_2 < k$$

where $k = \lceil (12 \log 3/\delta)/\varepsilon \rceil + 8$, which matches the assignment in Algorithm 1. Now we can apply Claim 14 to obtain $\|v - v'\|_1 \leq 5$ and observe that v, v' are both $(4L_0, 2\exp(216k^2 L_0)R_0)$-good by Claim 15. We will use the stability and goodness of the weights to establish the stability of both $\hat{\beta}$ and S_v.

Claim 18 requires $28L_0 \leq \frac{7}{24k} \leq \frac{1}{4}$, which is true by assumption. The claim implies that $\|S_v^{1/2}(\hat{\beta} - \hat{\beta})\|^2 \leq \Delta^2$ where $\Delta^2 = 3136 \exp(432k^2 L_0) L_0 R_0^2$. Next, we see that transforming β, β' by $(S_v)^{-1/2}$ allows us to apply Fact 20, giving

$$\mathcal{N}(\hat{\beta}, c^2 S_v^{-1}) \approx_{\varepsilon/3, \delta/6} \mathcal{N}(\hat{\beta}', c^2 S_v^{-1}).$$

as long as $c^2 \geq \Delta^2 \cdot \frac{18 \log 12/\delta}{\varepsilon^2}$, which is satisfied by construction in Algorithm 1.

Then, since $24L_0 \leq 1/(4k) \leq 1/2$ by assumption, Claim 17 tells us that $d_{\mathrm{PD}}(S_v, S_{v'}) \leq 56L_0$. We apply Fact 27 and Claim 19 to obtain

$$\mathcal{N}(\hat{\beta}', c^2 S_v^{-1}) \approx_{\varepsilon/3, \delta/6} \mathcal{N}(\hat{\beta}', c^2 S_{v'}^{-1}),$$

since $56L_0 \leq 3\varepsilon/(\log 12/\delta)$, which we assumed to be true. Finally, we apply Fact 21 to combine the two results, observing that $e^\varepsilon < e < 3$, to complete the proof. ■

5. Utility Analysis of Algorithm 1

Given the privacy guarantee of Lemma 22, we analyze the utility of Algorithm 1 under the standard subgaussian linear model. The definition of subgaussian variables and necessary concentration inequalities are provided in Appendix B.1. We first note that data from the standard subgaussian linear model is good with high probability.

Lemma 23 (Subgaussian data is good) *Let $X \in \mathbb{R}^{n \times d}$ be drawn i.i.d. from a d-dimensional subgaussian distribution $\mathcal{D}$ with mean 0, (full-rank) covariance Σ, and subgaussian parameter $K_{\mathcal{D}}$. Let $y_i = \beta^\top x_i + z_i$ where the z_i are drawn i.i.d. from a subgaussian distribution with mean 0, variance σ^2, and subgaussian parameter K_σ. There exists constants $K_L, K_R, K_n > 0$ such that for any $\eta \in (0,1)$, if $n \geq K_n K_{\mathcal{D}}^4 (d + \log(3/\eta))$ then (X, y) is (L, R)-good, where*

$$L = K_L K_{\mathcal{D}}^2 \cdot \frac{d + \log(3n/\eta)}{n} \qquad \textit{and} \qquad R = K_R K_\sigma \sigma \sqrt{\log(3n/\eta)},$$

with probability at least $1 - \eta$.

Proof [Proof Sketch] Identical calculations about leverage appeared in Brown et al. (2021); Kuditipudi et al. (2023); Brown et al. (2023).

Recall that we can write the vector of residuals $e = \hat{y} - y = (H - \mathbb{I})z$, where H is the hat matrix and z the noise vector. Denote by r_i the i-th row of $H - \mathbb{I}$. Then $e_i = r_i^\top z$, which (for any fixed H) implies e_i has subgaussian norm $\|r_i\| K_\sigma$. We know that $\|r_i\| \leq 1$, since $\mathbb{I} - H$ is idempotent and symmetric: $1 \geq (\mathbb{I} - H)_{i,i} = r_i^\top r_i$. Thus all $|e_i|$ will be bounded with high probability. ■

As we noted in Claim 3, when the input data is good `ISSP` returns the OLS estimate plus noise. We use a bound on the error of the OLS estimate from prior work.

Lemma 24 (OLS error under random design, restatement of Theorem 1, Hsu et al. (2011)) *Under the distributional assumption of Lemma 23, there exists an absolute constant K_{OLS} such that, for any $\delta \in (0,1)$, if $n > K_{\mathrm{OLS}} K_{\mathcal{D}}(d + \log(1/\delta))$, then with probability $1 - \delta$, we have*

$$\left\|\hat{\beta}_{OLS} - \beta\right\|_\Sigma^2 \leq \frac{K_{\mathrm{OLS}} K_\sigma^2 \sigma^2 (d + \log(1/\delta))}{n}.$$

Now, we are ready to prove the main accuracy lemma by bounding the norm of the added noise.

Lemma 25 (Main accuracy guarantee) *Let $X \in \mathbb{R}^{n \times d}$ be drawn i.i.d. from a d-dimensional subgaussian distribution $\mathcal{D}$ with mean 0, (full-rank) covariance Σ, and subgaussian parameter $K_{\mathcal{D}}$. Let $y_i = \beta^\top x_i + z_i$ where the z_i are drawn i.i.d. from a subgaussian distribution with mean 0, variance σ^2, and subgaussian parameter K_σ. There exists constants $K_L, K_R > 0$ such that for any $\eta \in (0,1)$, if*

$$L_0 = K_L K_{\mathcal{D}}^2 \cdot \frac{d + \log(3n/\eta)}{n}, \qquad R_0 = K_R K_\sigma \sigma \sqrt{\log(3n/\eta)},$$

and

$$n = \widetilde{\Omega}\left(K_{\mathcal{D}}^4 \left(d + \log\left(\frac{1}{\varepsilon\eta}\right)\right) \frac{(\log 1/\delta)^2}{\varepsilon^2} \right),$$

then with probability at least $1 - \eta$ Algorithm 1 successfully returns $\tilde{\beta}$ such that

$$\left\|\tilde{\beta} - \beta\right\|_\Sigma \leq O\left(K_\sigma \sigma \sqrt{\frac{d + \log(1/\eta)}{n}} + K_{\mathcal{D}} K_\sigma \sigma \cdot \frac{(d + \log(n/\eta))\sqrt{\log(n/\eta)\log(1/\delta)}}{\varepsilon n} \right),$$

where $\widetilde{\Omega}$ hides log factors in K_{D} and $\log 1/\delta$.

Proof We begin by determining how many samples are needed to ensure that (i) the algorithm does not fail immediately and (ii) the data is (L_0, R_0)-good (for the specified values) with high probability.

In order to not fail, we require $L_0 = O(\varepsilon / \log(1/\delta))$. Meanwhile, in order to apply Claim 3, we require $L_0 = O((\varepsilon / \log(1/\delta)^2))$. It is clear that the second requirement implies the first. Thus, we can expand our choice of L_0 to get

$$L_0 = K_L K_{\mathcal{D}}^2 \cdot \frac{d + \log(3n/\eta)}{n} = O\left(\frac{\varepsilon^2}{(\log 1/\delta)^2} \right).$$

Using the fact that $a/\log a = \Omega(b)$ implies $a = \Omega(b \log b)$, this translates to

$$n = \Omega\left(\frac{K_{\mathcal{D}}^2(\log 1/\delta)^2}{\varepsilon^2}\left(d + \log\left(\frac{K_{\mathcal{D}}^2(\log 1/\delta)^2}{\varepsilon^2\eta}\right)\right)\right).$$

We note that this implies $n = \Omega\big(K_{\mathcal{D}}(d + \log(1/\delta))\big)$ as required by Lemma 24. Now, in order to apply Lemma 23, we additionally require $n = \Omega\big(K_{\mathcal{D}}^4(d + \log(1/\eta))\big)$. As in Lemma 23, this requirement gives us $\left\|\Sigma^{-1/2}\widehat{\Sigma}\Sigma^{-1/2} - \mathbb{I}\right\|_2 \le 1/2$ by Claim 31, which we will use later. Combining the two outstanding requirements and dropping lower order terms gives

$$n = \Omega\left(K_{\mathcal{D}}^4\left(d + \log\left(\frac{K_{\mathcal{D}}^2(\log 1/\delta)^2}{\varepsilon^2\eta}\right)\right)\frac{(\log 1/\delta)^2}{\varepsilon^2}\right).$$

This ensures that (X, y) is (L_0, R_0)-good with probability $1 - O(\eta)$. When this happens the PTR check passes deterministically. Thus, we now turn to evaluating the accuracy of our regression estimate. We apply the triangle inequality about β_{ols}:

$$\|\beta - \tilde{\beta}\|_\Sigma \le \|\beta - \beta_{\text{ols}}\|_\Sigma + \|\beta_{\text{ols}} - \tilde{\beta}\|_\Sigma. \tag{4}$$

We analyze these terms separately.

The first term in Eq. (4) is solely about the empirical quantity. By Lemma 24, with probability at least $1 - O(\eta)$ we have

$$\|\beta_{\text{ols}} - \beta\|_\Sigma = O\left(K_\sigma\sigma\sqrt{\frac{d + \log(1/\eta)}{n}}\right).$$

To bound the second term in Eq. (4), we apply Claim 3, which states that on good data $\tilde{\beta}$ is drawn from $\mathcal{N}(\beta_{\text{ols}}, c^2(X^\top X)^{-1})$ where $c^2 = \Theta(L_0\, R_0^2 \log(1/\delta)/\varepsilon^2)$. Equivalently, we draw $z \sim \mathcal{N}(0, \mathbb{I})$ and set $\tilde{\beta} \leftarrow \beta_{\text{ols}} + c(X^\top X)^{-1/2}z$. Plugging this in, we have

$$\begin{aligned}\|\beta_{\text{ols}} - \tilde{\beta}\|_\Sigma &= \|c(X^\top X)^{-1/2}z\|_\Sigma \\ &= c \cdot \|\Sigma^{1/2}(X^\top X)^{-1/2}z\|_2.\end{aligned}$$

We plug in $\hat{\Sigma} = \frac{1}{n}X^\top X$ the empirical covariance and apply Cauchy–Schwarz:

$$\|\beta_{\text{ols}} - \tilde{\beta}\|_\Sigma \le \frac{c}{\sqrt{n}} \cdot \|\Sigma^{1/2}\hat{\Sigma}^{-1/2}\|_2 \cdot \|z\|_2.$$

By Claim 31 the matrix norm is at most a constant, and by Claim 30 we can bound $\|z\|_2^2 = O(d + \log 1/\eta)$ with probability at least $1 - O(\eta)$. Plugging these in, along with our expressions for c, L_0, and R_0, we arrive at the expression in the lemma. Applying a union bound over the three failure cases finishes the proof. ∎

6. Running-Time Analysis of Algorithm 1

In this section we prove the following guarantee about the running time of `ISSP`, whose computational requirements are quite lightweight. The core ideas in this proof appeared in the analogous claim of BHS.

Lemma 26 (Running Time) *Algorithm 1 can be implemented to require*

(1) one product of the form $A^\top A$ for $A \in R^{n\times d}$,

(2) one product of the form AB for $A \in R^{n\times d}$ and $B \in R^{d\times d}$,

(3) one inversion of a positive definite matrix in $R^{d\times d}$; and

(4) further computational overhead of $\tilde{O}(nd/\varepsilon)$.

Ignoring bit complexity, this corresponds to time $\tilde{O}(nd^{\omega-1} + nd/\varepsilon)$, where $\omega < 2.38$ is the matrix multiplication exponent. For modest privacy parameters, the running time of our algorithm is dominated by the time needed to compute the nonprivate OLS solution itself.

To establish this claim, we provide a second version of `StableResidualFiltering`, Algorithm 4, which is more computationally efficient. We show that this alternative algorithm is functionally equivalent.

Proof [Proof of Lemma 26] From BHS, Lemma 20 in Section 2.3, we see that we can implement `StableLeverageFiltering` using one product $A^\top A$, one product AB, one matrix inversion, and at most $O(\log(1/\delta)/\varepsilon)$ additional operations, each of which requires $\tilde{O}(nd)$ time. We need two additional conclusions from their analysis: `StableLeverageFiltering` can be implemented to return the inverse weighted covariance $(X^\top W X)^{-1}$ in the same asymptotic running time and we can update all leverage scores in time $\tilde{O}(nd)$ when removing a single observation.

With the weights w and inverse covariance in hand, we call `StableResidualFiltering`. The initial regression parameter can be computed in $\tilde{O}(nd)$ time, as we compute the vector $X^\top W y$ with a matrix-vector product (since W is diagonal) and multiply it with the inverse covariance. Computing all residuals is linear-time.

Each outlier removal and associated set of updates can also be implemented in $\tilde{O}(nd)$ time. This is because the removal of a single point corresponds to a rank-one update, which can be done efficiently. Recall from Section 1.3 the equation for updating the least squares solution after removing a data point:

$$\beta_{\mathrm{ols}(-j)} = \beta_{\mathrm{ols}} + \frac{(X^\top X)^{-1} x_j}{1 - h_j} \cdot \big(y_j - \langle x_j, \beta_{\mathrm{ols}} \rangle\big).$$

(A nearly identical formula applies when the data are weighted.) Since we have the previous leverage scores and inverse covariance, this update can be performed in time $O(nd)$. As before, with the new regression parameter all the residuals can be recalculated in linear time.

Setting these details aside, we turn to the crux of the analysis: `StableResidualFiltering` is functionally equivalent to Algorithm 4, our efficient version.

Algorithm 4 iterates through the residual thresholds in decreasing order. This is identical to independently calling the greedy algorithm repeatedly from scratch, since the removal process is deterministic (we can break ties in a consistent manner, e.g., using the index of the points). Formally, for

Algorithm 4: `StableResidualFiltering`, More Efficient Implementation

input : dataset X, y; base outlier thresholds L_0, R_0; weights w; discretization parameter k

$\forall j \in [2k], R_j \leftarrow (\exp(108kL))^j \cdot R_0$;
$\mathrm{COUNT} \leftarrow 0$;
for $j \in \{2k, 2k-1, \ldots, 0\}$ **do**
  while `TRUE` **do**
    `// check for large residuals`
    $\beta_w \leftarrow \texttt{WeightedOLS}(X, y, w)$; `/* via rank-one update */`
    $i^* \leftarrow \mathrm{argmax}_{i \in \mathrm{supp}(w)} \left| y_i - x_i^\top \beta_w \right|$;
    if $|y_{i^*} - x_{i^*}^\top \beta_w| \le R_j$ or $\mathrm{COUNT} \ge k$ **then**
      break; `/* too many outliers or no large residuals */`
    end
    $w_{i^*} \leftarrow 0$; `/* otherwise, remove weight */`
    $\mathrm{COUNT} \leftarrow \mathrm{COUNT} + 1$;
  end
  if $\mathrm{COUNT} \ge k$ **then**
    `// too many outliers`
    $\forall i \le j, \texttt{SCORE}^{(i)} \leftarrow k$;
    $\forall i \le j, u^{(i)} \leftarrow 0^n$;
    break;
  end
  `// store result and move to next threshold`
  $u^{(j)} \leftarrow w$;
  $\texttt{SCORE}^{(j)} \leftarrow \min\{k, n - \|u^{(j)}\|_1 + j\}$;
end
$\texttt{SCORE} \leftarrow \min_{j \in \{0,\ldots,k\}} \texttt{SCORE}^{(j)}$;
$v \leftarrow \frac{1}{k} \sum_{j=k+1}^{2k} u^{(j)}$;
return SCORE, v

any $R > R'$ and any fixed X, y, w, the result of $\texttt{ResidualThresholding}(X, y, R', w)$ is identical to calling $u \leftarrow \texttt{ResidualThresholding}(X, y, R, w)$ and then $\texttt{ResidualThresholding}(X, y, R', u)$.

Algorithm 4 also tracks a count of observations it removes and halts if that number reaches k. If it halts at level ℓ, for all $j \leq \ell$ it sets $u^{(j)} = 0^n$ and $\texttt{SCORE}^{(j)} = k$. To show that this has no effect on the outcome of the algorithm, we suppose Algorithm 4's count reaches k and analyze two cases. Let $\ell \in \{0, \ldots, 2k\}$ be the residual threshold index at which the count k was reached. We know that $u^{(\ell)}$ returned by $\texttt{ResidualThresholding}(X, y, R_\ell, w)$ in Algorithm 3, the main version of `StableResidualFiltering`, satisfies $\|u^{(\ell)}\|_1 \leq n - k$, since after any removal the weight is zero. This also holds for all $u^{(j)}$ with $j \leq \ell$.

Case 1: Suppose $\ell > k$, i.e., ℓ falls among the indices used to compute the weights. Then for all $j \in \{0, \ldots, k\}$, the indices used to compute the scores, Algorithm 3 computes $u^{(j)}$ with $\|u^{(j)}\|_1 \leq n - k$. This means Algorithm 3 computes $\texttt{SCORE} = k$, as does Algorithm 4 (since it sets $\texttt{SCORE}_j = k$ for all $j \leq k$). (Recall that this causes `ISSP` to fail deterministically, so the weights do not impact the output.)

Case 2: If $\ell \leq k$, then ℓ falls among the indices used to compute the score. (Thus, Algorithms 3 and 4 return the same weights.) Algorithm 4 sets $\texttt{SCORE}^{(j)} = k$ for all $j \leq \ell$. We claim that Algorithm 3 also computes $\texttt{SCORE}^{(j)} = k$ for all $j \leq \ell$. To see this, recall that on these indices Algorithm 3 computes $u^{(j)}$ with $\|u^{(j)}\|_1 \leq n - k$.

To finish the proof, we note that the final β_v and S_v^{-1} computed by `ISSP` can be computed with at most k rank-one updates from their initial values. Since $k = O(\log(1/\delta)/\varepsilon)$, we are done. ■

Acknowledgements

This work is supported in part by the National Science Foundation under grant no. 2019844, 2112471, 2238080, and 2229876, NSF Graduate Research Fellowships Program, the Machine Learning Alliance at MIT CSAIL, and Microsoft Grant for Customer Experience Innovation. Part of this work was done while GB was at Boston University. AS and GB (while at Boston University) were supported in part by NSF awards CCF-1763786 and CNS-2120667 as well as Faculty Awards from Google and Apple. JCP was supported in part by the Harvard Center for Research on Computation and Society.

References

Ishaq Aden-Ali, Hassan Ashtiani, and Gautam Kamath. On the sample complexity of privately learning unbounded high-dimensional gaussians. In *Algorithmic Learning Theory*, pages 185–216. Proceedings of Machine Learning Research, 2021.

Daniel Alabi, Audra McMillan, Jayshree Sarathy, Adam Smith, and Salil Vadhan. Differentially private simple linear regression. *Proceedings on Privacy Enhancing Technologies*, 2022.

Daniel Alabi, Pravesh K Kothari, Pranay Tankala, Prayaag Venkat, and Fred Zhang. Privately estimating a gaussian: Efficient, robust, and optimal. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 483–496, 2023.

Francis J Anscombe. Graphs in statistical analysis. *The american statistician*, 27(1):17–21, 1973.

Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning gaussians and beyond. In *Conference on Learning Theory*, pages 1075–1076. Proceedings of Machine Learning Research, 2022.

Pranjal Awasthi, Abhimanyu Das, Weihao Kong, and Rajat Sen. Trimmed maximum likelihood estimation for robust generalized linear model. In *Advances in Neural Information Processing Systems*, volume 35, pages 862–873. Curran Associates, Inc., 2022.

Ainesh Bakshi and Adarsh Prasad. Robust linear regression: Optimal rates in polynomial time. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 102–115, 2021.

Raef Bassily, Adam Smith, and Abhradeep Thakurta. Private empirical risk minimization: Efficient algorithms and tight error bounds. In *2014 IEEE 55th Annual Symposium on Foundations of Computer Science*, pages 464–473. IEEE, 2014.

David A Belsley, Edwin Kuh, and Roy E Welsch. *Regression diagnostics: Identifying influential data and sources of collinearity*. John Wiley & Sons, 2005.

Kush Bhatia, Prateek Jain, and Purushottam Kar. Robust regression via hard thresholding. *Advances in neural information processing systems*, 28, 2015.

Kush Bhatia, Prateek Jain, Parameswaran Kamalaruban, and Purushottam Kar. Consistent robust regression. *Advances in Neural Information Processing Systems*, 30, 2017.

Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. *arXiv preprint arXiv:2006.06618*, 2020.

Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakynthinou. Covariance-aware private mean estimation without private covariance estimation. *Advances in Neural Information Processing Systems*, 34:7950–7964, 2021.

Gavin Brown, Samuel Hopkins, and Adam Smith. Fast, sample-efficient, affine-invariant private mean and covariance estimation for subgaussian distributions. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5578–5579. Proceedings of Machine Learning Research, 2023.

Gavin Brown, Krishnamurthy Dvijotham, Georgina Evans, Daogao Liu, Adam Smith, and Abhradeep Thakurta. Private gradient descent for linear regression: Tighter error bounds and instance-specific uncertainty estimation. *arXiv preprint arXiv:2402.13531*, 2024.

Mark Bun and Thomas Steinke. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. *arXiv preprint arXiv:1906.02830*, 2019.

Mark Bun, Gautam Kamath, Thomas Steinke, and Steven Z Wu. Private hypothesis selection. In *Advances in Neural Information Processing Systems*, pages 156–167, 2019.

T Tony Cai, Yichen Wang, and Linjun Zhang. The cost of privacy: Optimal rates of convergence for parameter estimation with differential privacy. *The Annals of Statistics*, 49(5):2825–2850, 2021.

T Tony Cai, Yichen Wang, and Linjun Zhang. Score attack: A lower bound technique for optimal differentially private learning. *arXiv preprint arXiv:2303.07152*, 2023.

Sitan Chen, Frederic Koehler, Ankur Moitra, and Morris Yau. Online and distribution-free robustness: Regression and contextual bandits with huber contamination. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 684–695. IEEE, 2022.

Christos Dimitrakakis, Blaine Nelson, Aikaterini Mitrokotsa, and Benjamin IP Rubinstein. Robust and private bayesian inference. In *International Conference on Algorithmic Learning Theory*, pages 291–305. Springer, 2014.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the 41st ACM Symposium on Theory of Computing*, STOC '09, pages 371–380. ACM, 2009.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam D. Smith. Calibrating noise to sensitivity in private data analysis. In Shai Halevi and Tal Rabin, editors, *Theory of Cryptography, Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006, Proceedings*, volume 3876 of *Lecture Notes in Computer Science*, pages 265–284. Springer, 2006. doi: 10.1007/11681878_14. URL https://doi.org/10.1007/11681878_14.

Cynthia Dwork, Kunal Talwar, Abhradeep Thakurta, and Li Zhang. Analyze gauss: optimal bounds for privacy-preserving principal component analysis. In *Proceedings of the forty-sixth annual ACM symposium on Theory of computing*, pages 11–20, 2014.

James Foulds, Joseph Geumlek, Max Welling, and Kamalika Chaudhuri. On the theory and practice of privacy-preserving bayesian data analysis. *arXiv preprint arXiv:1603.07294*, 2016.

David C Hoaglin and Roy E Welsch. The hat matrix in regression and anova. *The American Statistician*, 32(1):17–22, 1978.

Samuel B Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In *Proceedings of the 54th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1406–1417, 2022.

Daniel Hsu, Sham M Kakade, and Tong Zhang. An analysis of random design linear regression. *arXiv preprint arXiv:1106.2363*, 6, 2011.

Peter J Huber. Robust statistics. In *International encyclopedia of statistical science*, pages 1248–1251. Springer, 2011.

Arun Jambulapati, Jerry Li, and Kevin Tian. Robust sub-gaussian principal component analysis and width-independent schatten packing. *Advances in Neural Information Processing Systems*, 33: 15689–15701, 2020.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pages 1853–1902. Proceedings of Machine Learning Research, 2019.

Gautam Kamath, Argyris Mouzakis, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. A private and computationally-efficient estimator for unbounded gaussians. In *Conference on Learning Theory*, pages 544–572. Proceedings of Machine Learning Research, 2022.

Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. *arXiv preprint arXiv:1711.03908*, 2017.

Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *9th Innovations in Theoretical Computer Science Conference (ITCS 2018)*, volume 94, page 44. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2018.

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

Adam Klivans, Pravesh K Kothari, and Raghu Meka. Efficient algorithms for outlier-robust regression. In *Conference On Learning Theory*, pages 1420–1430. PMLR, 2018.

Rohith Kuditipudi, John Duchi, and Saminul Haque. A pretty fast algorithm for adaptive private mean estimation. In *Conference on Learning Theory*, pages 2511–2551. Proceedings of Machine Learning Research, 2023.

Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. *Advances in Neural Information Processing Systems*, 34:3887–3901, 2021.

Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, pages 1167–1246. Proceedings of Machine Learning Research, 2022.

Xiyang Liu, Prateek Jain, Weihao Kong, Sewoong Oh, and Arun Suggala. Label robust and differentially private linear regression: Computational and statistical efficiency. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

William Mendenhall, Terry Sincich, and Nancy S Boudreau. *A second course in statistics: regression analysis*, volume 6. Prentice Hall Upper Saddle River, NJ, 2003.

Kentaro Minami, HItomi Arai, Issei Sato, and Hiroshi Nakagawa. Differential privacy without sensitivity. In *Advances in Neural Information Processing Systems*, pages 956–964, 2016.

Darakhshan J Mir. *Differential privacy: an exploration of the privacy-utility landscape*. Rutgers The State University of New Jersey-New Brunswick, 2013.

Shyam Narayanan. Better and simpler lower bounds for differentially private statistical estimation. *arXiv preprint arXiv:2310.06289*, 2023.

Ankit Pensia, Varun Jog, and Po-Ling Loh. Robust regression with covariate filtering: Heavy tails and adversarial contamination. *arXiv preprint arXiv:2009.12976*, 2020.

Or Sheffet. Differentially private ordinary least squares. In *International Conference on Machine Learning*, pages 3105–3114. PMLR, 2017.

Or Sheffet. Old techniques in differentially private linear regression. In *Algorithmic Learning Theory*, pages 789–827. PMLR, 2019.

Yanyao Shen and Sujay Sanghavi. Iterative least trimmed squares for mixed linear regression. *Advances in Neural Information Processing Systems*, 32, 2019a.

Yanyao Shen and Sujay Sanghavi. Learning with bad training data via iterative trimmed loss minimization. In *International Conference on Machine Learning*, pages 5739–5748. PMLR, 2019b.

Arun Sai Suggala, Kush Bhatia, Pradeep Ravikumar, and Prateek Jain. Adaptive hard thresholding for near-optimal consistent robust regression. In *Conference on Learning Theory*, pages 2892–2897. PMLR, 2019.

Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, and Uri Stemmer. Friendlycore: Practical differentially private aggregation. In *International Conference on Machine Learning*, pages 21828–21863. Proceedings of Machine Learning Research, 2022.

Salil Vadhan. The complexity of differential privacy. In *Tutorials on the Foundations of Cryptography*, pages 347–450. Springer, 2017.

Prateek Varshney, Abhradeep Thakurta, and Prateek Jain. (nearly) optimal private linear regression for sub-gaussian data via adaptive clipping. In *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 1126–1166. PMLR, 02–05 Jul 2022.

Paul F Velleman and Roy E Welsch. Efficient computing of regression diagnostics. *The American Statistician*, 35(4):234–242, 1981.

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

Duy Vu and Aleksandra Slavkovic. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, pages 138–143. IEEE, 2009.

Yu-Xiang Wang. Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. *arXiv preprint arXiv:1803.02596*, 2018.

Yu-Xiang Wang, Stephen Fienberg, and Alex Smola. Privacy for free: Posterior sampling and stochastic gradient monte carlo. In *International Conference on Machine Learning*, pages 2493–2502. PMLR, 2015.

Appendix A. Additional Related Work

Private Mean Estimation Many of the developments in private linear regression have analogs in private mean estimation, albeit rearranged chronologically. Consider the canonical mean estimation problem with "covariance-adaptive" error guarantees, which respect the shape of the dataset: the error is measured in Mahalanobis norm with respect to the covariance matrix Σ of the data X, $\|\Sigma^{-1/2}(\hat{\mu} - \mu)\|$. This scales each direction according to the directional variance, providing a more relevant measure of utility. This is closely related to how linear regression error corresponds to the Σ-norm, $\|\hat{\beta} - \beta\|_\Sigma$.

For non-private mean estimation, the distinction between Euclidean and Mahalanobis norm error are minor. For example, the empirical mean achieves such a geometry-aware guarantee and, like the OLS estimator, is accurate with roughly d samples with no dependence on the condition number of the covariance of the data.

For private estimation, geometry-aware estimation is significantly more challenging, since the privately learning the geometry, i.e., the covariance matrix, is more sample expensive than the primary task of mean estimation. At the same time, it seemed like it was necessary to design the privacy noise that matches the shape of the data covariance. The standard Gaussian mechanism privately estimates the mean with only d samples but its error depends polynomially on the condition number, a high price to pay when the estimator does not respect the geometry. The work of Kamath et al. (2019) allows us to privately learn the covariance and apply the Gaussian mechanism on whitened data, but, as with the SSP approaches for linear regression, this requires $d^{3/2}$ samples. A long line of work that follows either makes more strict assumptions on the geometry or pays a price in the sample complexity (Karwa and Vadhan, 2017; Biswas et al., 2020; Cai et al., 2021; Aden-Ali et al., 2021; Bun et al., 2019; Bun and Steinke, 2019; Liu et al., 2021; Kamath et al., 2022; Hopkins et al., 2022; Alabi et al., 2023). Of particular relevance for our work are frameworks introduced by Tsfadia et al. (2022) and Ashtiani and Liaw (2022), which remove outliers in a way that depends on the rest of the dataset (e.g., asking that inliers be close to a large number of other examples). These frameworks bear some similarity to our techniques and to those of Brown et al. (2023), especially their "Stable Mean" estimator. Informally, our approach improves over theirs in the ability to adapt the definition of *outlier* and the resulting geometry as points are removed.

Brown et al. (2021) was the first to address this geometry-aware challenge in private mean estimation. They termed it the *covariance estimation bottleneck* and gave two exponential-time approaches for avoiding it, achieving accurate estimation with $\tilde{O}(d)$ samples and no dependence on the condition number. The first, which combined the exponential mechanism with PTR (Propose-Test-Release), served as a direct inspiration to the HPTR (High-dimensional PTR) framework of Liu et al. (2022). The concurrent works of Kuditipudi et al. (2023) and Brown et al. (2023) built on the second algorithm of Brown et al. (2021), giving time-efficient algorithms matching the guarantees of the exponential-time approaches. The sample complexity has linear dependence on the dimension d and no dependence on the condition number $\kappa(X^\top X)$. As in Kuditipudi et al. (2023) and Brown et al. (2023), our goal is to achieve the same for linear regression.

Private Linear Regression. Commensurate with its centrality in statistical theory and practice, significant effort has gone into producing differentially private algorithms for least squares (Vu and Slavkovic, 2009; Kifer et al., 2012; Mir, 2013; Dimitrakakis et al., 2014; Bassily et al., 2014; Wang et al., 2015; Foulds et al., 2016; Minami et al., 2016).

One standard theme in many of these works is the class of assumptions that directly enable global sensitivity analysis. Prime examples include assuming that the covariates satisfy an ℓ_2 norm bound or that the true parameter lies in some ball about the origin. Such guarantees are incomparable with our definition of goodness (for example, our definition allows arbitrarily large covariates, but covariates with bounded norms may still have high leverage). Under some collections of these assumptions, state-of-the-art guarantees are achieved in (Wang, 2018; Sheffet, 2019), which in our setting translates into a sample complexity of $n = \Omega(d^{1.5}/(\alpha\varepsilon))$ to achieve $(1/\sigma)\|\hat{\beta} - \beta\|_\Sigma \leq \alpha$. Both these prior algorithms and ours analyze accuracy under the assumption that the input data is "outlier-free." The prior work uses conditions on the norm of the covariates or the magnitude of the

labels. These assumptions lend themselves handily to global sensitivity calculations. In contrast, our work uses a notion of outlier-freeness which is more in line with standard statistical practice: we ask that the dataset have no high-leverage or high-residual points.

When applied to data from the standard sub-Gaussian linear models, `ISSP` is the first computationally efficient algorithm to achieve linear dependence in the dimension d and no dependence on the condition number $\kappa(X^\top X)$ (see Theorem 5). This nearly matches the best known sample complexity of Liu et al. (2022) that relies on an exponential time approach of HPTR: $n = \tilde{O}(d/\alpha^2 + (d + \log(1/\delta))/(\alpha\varepsilon))$ samples suffice to achieve an error of $(1/\sigma)\|\hat{\beta} - \beta\|_\Sigma \le \alpha$. Existing computationally efficient approaches based on gradient descent either assume the covariance matrix is close to identity (Cai et al., 2021; Brown et al., 2024) or have polynomial dependence on the condition number $\kappa(X^\top X)$ (Varshney et al., 2022; Liu et al., 2023). The best known sample complexity of an efficient algorithm is by Liu et al. (2023): $n = \tilde{O}(d/\alpha^2 + (\kappa^{1/2} d \log(1/\delta))/(\alpha\varepsilon))$.

Iterative Thresholding. Our `ResidualThresholding` algorithm is a special case of the family of iterative thresholding algorithms, a longstanding heuristic for robust linear regression that dates back to Legendre. Its theoretical properties in the non-asymptotic regime have been extensively studied recently in Bhatia et al. (2015, 2017); Suggala et al. (2019); Pensia et al. (2020); Chen et al. (2022). Shen and Sanghavi (2019b) and Awasthi et al. (2022) studied the iterative trimmed estimator under generalized linear models and Shen and Sanghavi (2019a) studied the mixed linear regression setting. It is worth noting that most iterative thresholding algorithm in the robust linear regression setting will alternate between finding the OLS solution of the current set and finding the set with the smallest residual under the current regression coefficient, and no data point is permanently removed in each iteration. In contrary, our algorithm will permanently remove one data point in each iteration before recomputing the OLS solution.

Appendix B. Preliminaries

We collect here known preliminary results that we use in our analyses.

Fact 27 *Let S_1, S_2 be positive-definite matrices and define*

$$d_{\mathrm{PD}}(S_1, S_2) = \max\Big\{ \Big\| S_1^{-1/2} S_2 S_1^{-1/2} - \mathbb{I} \Big\|_{\mathrm{tr}}, \Big\| S_2^{-1/2} S_1 S_2^{-1/2} - \mathbb{I} \Big\|_{\mathrm{tr}} \Big\}.$$

Then $d_{\mathrm{PD}}(S_1, S_2) = d_{\mathrm{PD}}(S_1^{-1}, S_2^{-1})$.

Proof Note that $S_1^{-1/2} S_2 S_1^{-1/2}$ and $S_2^{1/2} S_1^{-1} S_2^{1/2}$ are similar and likewise, $S_2^{-1/2} S_1 S_2^{-1/2}$ and $S_1^{1/2} S_2^{-1} S_1^{1/2}$ are similar. Thus,

$$\begin{aligned}
d_{\mathrm{PD}}(S_1, S_2) &= \max\Big\{ \Big\| S_1^{-1/2} S_2 S_1^{-1/2} - \mathbb{I} \Big\|_{\mathrm{tr}}, \Big\| S_2^{-1/2} S_1 S_2^{-1/2} - \mathbb{I} \Big\|_{\mathrm{tr}} \Big\} \\
&= \max\Big\{ \Big\| S_2^{1/2} S_1^{-1} S_2^{1/2} - \mathbb{I} \Big\|_{\mathrm{tr}}, \Big\| S_1^{1/2} S_2^{-1} S_1^{1/2} - \mathbb{I} \Big\|_{\mathrm{tr}} \Big\} \\
&= d_{\mathrm{PD}}(S_2^{-1}, S_1^{-1}) \\
&= d_{\mathrm{PD}}(S_1^{-1}, S_2^{-1})
\end{aligned}$$

as desired. ■

B.1. Subgaussian Random Variables and Concentration Inequalities

For formal proofs of claims and further discussion we refer to Vershynin (2018).

Definition 28 (Subgaussian Norm) *Let $y \in \mathbb{R}$ be a random variable. The* subgaussian norm of y, *denoted $\|y\|_{\psi_2}$, is $\|y\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\exp\left(y^2/t^2\right) \le 2\right\}$.*

Definition 29 (Subgaussian Random Variable) *Let $y \in \mathbb{R}^d$ be a random variable with mean μ and covariance Σ. Call y* subgaussian with parameter K *if there exists $K \ge 1$ such that for all $v \in \mathbb{R}^d$ we have*

$$\|\langle y - \mu, v\rangle\|_{\psi_2} \quad \le \quad K\sqrt{v^\top \Sigma v}.$$

For example, the Gaussian distribution $\mathcal{N}(\mu, \Sigma)$ is subgaussian with parameter $K = O(1)$.

Claim 30 (Concentration of Norm) *Let $y_1, \dots, y_n$ be drawn i.i.d. from a d-dimensional subgaussian distribution with parameter $K_y > 0$, mean μ, and (full-rank) covariance Σ. There exists a constant $K_1 > 0$ such that, with probability at least $1 - \beta$, we have both*

$$\left\|\Sigma^{-1/2}(y_1 - \mu)\right\|^2 \le K_1 K_y^2 (d + \log 1/\beta) \quad \textit{and} \quad \left\|\Sigma^{-1/2}\left(\frac{1}{n}\sum_{i=1}^n y_i - \mu\right)\right\|^2 \le K_1 K_y^2 \cdot \frac{d + \log 1/\beta}{n}.$$

Claim 31 (Concentration of Covariance) *Let $y_1, \dots, y_n$ be drawn i.i.d. from a d-dimensional subgaussian distribution with parameter $K_y > 0$, mean $\mu = 0$, and (full-rank) covariance Σ. Let $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n y_i y_i^\top$ be the empirical covariance. There exist positive absolute constants K_1 and K_2 such that, for any $\beta \in (0, 1)$, if $n \ge K_2(d + \log 1/\beta)$, then with probability at least $1 - \beta$ we have*

$$\|\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2} - \mathbb{I}\|_2 \quad \le \quad K_1 K_y^2 \sqrt{\frac{d + \log 1/\beta}{n}}.$$

B.2. Details on `StableLeverageFiltering`

As a preprocessing step, `ISSP` performs a leverage-score filtering routine introduced by BHS. The algorithm we use differs only superficially from their version. (For instance, they compute a set of weights $w \in [0, 1]^n$ and a weighted covariance estimate, while we only care about the weights themselves.) For completeness, we now state the version we use here. Recall that Theorem 9 contains the relevant guarantees proved by BHS.

Appendix C. Deferred Proofs

We now give the proof for Claim 7, which characterizes the effect of removing weighted points from a least squares model. This is a natural generalization of standard results (Mendenhall et al., 2003; Belsley et al., 2005; Huber, 2011).

Proof [Proof of Claim 7] We start by setting up notation and bounding a term useful in proving both (2) and (3).

Algorithm 5: Stable Leverage Filtering (`StableLeverageFiltering`), BHS

Input: dataset $X \in \mathbb{R}^{n \times d}$; outlier threshold L_0; discretization parameter $k \in \mathbb{N}$

$A \leftarrow [n]$;
for $j = 2k, 2k-1, \ldots, 0$ **do**
  $L_j = \exp\{j/k\} \cdot L_0$;
  repeat
    $S_A \leftarrow \sum_{i \in A} x_i x_i^\top$;
    $\texttt{OUT} \leftarrow \left\{ i \in A : x_i^\top (S_A)^{-1} x_i > L_j \right\}$;
    $A \leftarrow A \setminus \texttt{OUT}$
  until $\texttt{OUT} = \emptyset$;
  $A_j \leftarrow A$;
end
$\texttt{SCORE} \leftarrow \min\{k, \min_{0 \le j \le k}\{n - |A_j| + j\}\}$;
for $i = 1, \ldots, n$ **do**
  $w_i \leftarrow \frac{1}{k} \sum_{j=k+1}^{2k} \mathbb{1}\{i \in A_j\}$;
end
return $\texttt{SCORE}, w$;

Setup Assume without loss of generality that $\mathrm{supp}(w) = [n]$, as any points outside the support of w are irrelevant. Let $v = w' - w$, $\|v\|_1 = \rho$, $W = \mathrm{diag}(w)$ (likewise W' and V), and $C = X^\top W X$. Decompose $V = PN$ (for "positive" and "negative") where P, N are diagonal matrices with $P_{i,i} = \sqrt{|v_i|}$ and $N_{i,i} = \mathrm{sign}(v_i) \cdot \sqrt{|v_i|}$.

Let $\Delta = I + Y$ where $Y = NXC^{-1}X^\top P$. If $\|Y\|_2 \le \varepsilon < 1$ then $I + Y \succeq (1-\varepsilon)I$. Consequently, Δ is invertible and $\|\Delta^{-1}\|_2 \le 1/(1-\varepsilon)$. To prove that $\|Y\|_2 \le \varepsilon$, we use the fact that $\|Y\|_2 \le \|Y\|_F$ and compute:

$$\begin{aligned}
\|NXC^{-1}X^\top P\|_F^2 &= \sum_{i,j} \left(N_{i,i} P_{j,j} \cdot x_i^\top C^{-1} x_j \right)^2 \\
&\le \sum_{i,j} N_{i,i}^2 P_{j,j}^2 L^2 \\
&= L^2 \sum_{i,j} |v_i||v_j| \\
&= L^2 \sum_i |v_i| \sum_j |v_j| \\
&= L^2 \cdot \|w - w'\|_1^2.
\end{aligned}$$

Ultimately, we arrive at $\|\Delta^{-1}\|_2 \le (1 - \|w - w'\|_1 L)^{-1}$. In the first line of the calculation above, we used the fact that, by our initial assumption on the goodness of (X, w),

$$\|x_i^\top C^{-1} x_j\|_2 \le \|C^{-1/2} x_i\|_2 \|C^{-1/2} x_j\|_2 \le L^2.$$

Since, $\|w - w'\|_1 L \le 1/2$, we conclude that $\|\Delta^{-1}\|_2 \le 2$.

Bounding the Leverage Scores Consider an index $i \in \mathrm{supp}(w)$ Write out its leverage under w' and plug in our notation:

$$\begin{aligned} h_i' &= x_i^\top \Big(X^\top W' X\Big)^{-1} x_i \\ &= x_i^\top \Big(X^\top W X + X^\top V X\Big)^{-1} x_i \\ &= x_i^\top \Big(C + (PX)^\top N X\Big)^{-1} x_i. \end{aligned}$$

Recalling the fact that, $x_i^\top C^{-1} x_i = h_i$ and $\Delta = I + NXC^{-1}X^\top P$, we can apply the Woodbury matrix identity to arrive at the following relationship between h_i' and h_i:

$$\begin{aligned} h_i' &= x_i^\top \left[C^{-1} - C^{-1}(PX)^\top \Big(I + NXC^{-1}(PX)^\top\Big)^{-1} NXC^{-1}\right] x_i \\ &= h_i - x_i^\top C^{-1} X^\top P \Delta^{-1} NXC^{-1} x_i. \end{aligned}$$

We want to upper bound this leverage, h_i', so we take the absolute value of the right-hand term and apply Cauchy–Schwarz:

$$|x_i^\top C^{-1} X^\top P \Delta^{-1} NXC^{-1} x_i| \le \|\Delta^{-1}\|_2 \cdot \|PXC^{-1}x_i\|_2 \cdot \|NXC^{-1}x_i\|_2.$$

We already argued that $\|\Delta^{-1}\|_2 \le 2$, so we turn to the second term in the product:

$$\begin{aligned} \|PXC^{-1}x_i\|_2^2 &= \sum_{j=1}^n P_{j,j}^2 (x_j^\top C^{-1} x_i)^2 \\ &\le \sum_{j=1}^n |v_i| \cdot L^2 = \|w - w'\|_1 \cdot L^2. \end{aligned}$$

Note that an identical bound also holds for $\|NXC^{-1}x_i\|_2$, hence we have

$$\begin{aligned} h_i' &\le h_i + 2\|w - w'\|_1 L^2 \\ &\le L + 2\|w - w'\|_1 L^2 = \big(1 + 2\|w - w'\|_1 L\big) \cdot L. \end{aligned}$$

Bounding the Residuals As above, we use the Woodbury matrix identity to derive an expression for the regression line $\beta_{w'}$.

$$\begin{aligned} \beta_{w'} &= \Big(X^\top W' X\Big)^{-1} X^\top W' y \\ &= \Big(X^\top W X + X^\top V X\Big)^{-1} \Big(X^\top W y + X^\top V y\Big) \\ &= \Big(C + (PX)^\top N X\Big)^{-1} \Big(X^\top W y + X^\top V y\Big) \\ &= \Big(C^{-1} - C^{-1}(PX)^\top \Delta^{-1} NXC^{-1}\Big) \Big(X^\top W y + X^\top V y\Big). \end{aligned}$$

This expression expands into four terms, which simplify nicely:

$$\begin{aligned}
\beta_{w'} &= \beta_w \\
&\quad + C^{-1}X^\top Vy \\
&\quad - C^{-1}X^\top P\Delta^{-1}NXC^{-1}X^\top Wy \\
&\quad - C^{-1}X^\top P\Delta^{-1}NXC^{-1}X^\top Vy \\
&= \beta_w \\
&\quad + C^{-1}X^\top P\Delta^{-1}N(P\Delta^{-1}N)^{-1}Vy \\
&\quad - C^{-1}X^\top P\Delta^{-1}NX\beta_w \\
&\quad - C^{-1}X^\top P\Delta^{-1}NXC^{-1}X^\top Vy \\
&= \beta_w + C^{-1}X^\top P\Delta^{-1}N\Big(N^{-1}\Delta P^{-1}Vy - X\beta_w - XC^{-1}X^\top Vy\Big).
\end{aligned}$$

We can further simplify the expression in the parentheses. In particular, by plugging in the definition of Δ, we have that $N^{-1}\Delta P^{-1}Vy - X\beta_w - XC^{-1}X^\top Vy$ can be rewritten as:

$$\begin{aligned}
&N^{-1}\Big(I + NXC^{-1}X^\top P\Big)P^{-1}Vy - X\beta_w - XC^{-1}X^\top Vy \\
&= N^{-1}P^{-1}Vy + N^{-1}NXC^{-1}X^\top PP^{-1}Vy - X\beta_w - XC^{-1}X^\top Vy \\
&= V^{-1}Vy + XC^{-1}X^\top Vy - X\beta_w - XC^{-1}X^\top Vy \\
&= y - X\beta_w.
\end{aligned}$$

Plugging back in, we arrive at

$$\beta_{w'} = \beta_w + C^{-1}X^\top P\Delta^{-1}N(y - X\beta_w).$$

To finish the proof, we consider the residual on a point $x_i \in \operatorname{supp}(w)$:

$$\begin{aligned}
\left|y_i - x_i^\top \beta_{w'}\right| &= \left|y_i - x_i^\top \beta_w + x_i^\top \beta_w - x_i^\top \beta_{w'}\right| \\
&\le \left|y_i - x_i^\top \beta_w\right| + \left|x_i^\top \beta_w - x_i^\top \beta_{w'}\right| \\
&\le R + \left|x_i^\top C^{-1}X^\top P\Delta^{-1}N(y - X\beta_w)\right|,
\end{aligned}$$

The bound $\left|y_i - x_i^\top \beta_w\right|$ follows from our initial assumption that w is (L, R)-good for (X, y). We can bound the second term in an analogous way to how we bounded the leverage scores. In particular, using the fact that $\|N\|_2 \le \sqrt{\|w - w'\|_1}$,

$$\begin{aligned}
\left|x_i^\top C^{-1}X^\top P\Delta^{-1}N(y - X\beta_w)\right| &\le \left\|PXC^{-1}x_i\right\|_2 \cdot \left\|\Delta^{-1}\right\|_2 \cdot \|N(y - X\beta_w)\|_2 \\
&\le 2LR\|w - w'\|_1.
\end{aligned}$$

This completes the proof. ■

We now prove Claim 18, which says that, if two weight vectors on adjacent datasets are both good and close in total variation distance, then their least-squares solutions are close as well. We start by considering the setting where the vectors correspond to the same dataset.

Claim 32 *Assume v, w are both (L, R)-good for dataset (X, y). Define $S_v = X^\top \mathrm{diag}(v) X$, $\beta_v = S_v^{-1} X^\top V y$, and likewise S_w, β_w. If $\|v - w\|_1 L \le \frac{1}{4}$, then $\|S_v^{1/2}(\beta_v - \beta_w)\|^2 \le 4\|v - w\|_1^2 L R^2$.*

From this claim, Claim 18 is an easy corollary.

Proof [Proof of Claim 18] Vectors $v, w \in [0,1]^n$ on adjacent datasets (X, y) and (X', y'), respectively, correspond to vectors $v', w' \in [0,1]^{n+1}$ over the datasets' union. We have $\|v' - w'\|_1 \le \|v - w\|_1 + 2$. ■

Proof [Proof of Claim 32] We start by expanding out the squared norm and substituting in the definition of S_v:

$$\begin{aligned}\big\|S_v^{1/2}(\beta_v - \beta_w)\big\|^2 &= \big\langle \beta_v - \beta_w, S_v(\beta_v - \beta_w)\big\rangle \\ &= \left\langle \beta_v - \beta_w, \left(\sum_{i\in[n]} v_i \cdot x_i x_i^\top\right)(\beta_v - \beta_w)\right\rangle.\end{aligned}$$

Next we expand the sum across $\beta_v - \beta_w$ and add and subtract $\sum_i v_i \cdot x_i y_i$, which makes the right-hand side of the inner product look like a pair of gradients.

$$\begin{aligned}\big\|S_v^{1/2}(\beta_v - \beta_w)\big\|^2 &= \left\langle \beta_v - \beta_w, \sum_i v_i \cdot x_i \langle x_i, \beta_v\rangle - \sum_i v_i \cdot x_i \langle x_i, \beta_w\rangle \right\rangle \\ &= \left\langle \beta_v - \beta_w, \sum_i v_i \cdot x_i\big(\langle x_i, \beta_v\rangle - y_i\big) - \sum_i v_i \cdot x_i\big(\langle x_i, \beta_w\rangle - y_i\big) \right\rangle.\end{aligned}$$

By definition, β_v is the vector that sets the first gradient to zero, so we have

$$\big\|S_v^{1/2}(\beta_v - \beta_w)\big\|^2 = \left\langle \beta_v - \beta_w, 0 - \sum_i v_i \cdot x_i\big(\langle x_i, \beta_w\rangle - y_i\big) \right\rangle.$$

We now add and subtract the gradient at β_w weighted by w, which leaves a gradient term (also zero by definition) and the differences $v_i - w_i$:

$$\begin{aligned}\big\|S_v^{1/2}(\beta_v - \beta_w)\big\|^2 &= \left\langle \beta_v - \beta_w, -\sum_i w_i \cdot x_i\big(\langle x_i, \beta_w\rangle - y_i\big) + \sum_i (w_i - v_i) \cdot x_i\big(\langle x_i, \beta_w\rangle - y_i\big) \right\rangle \\ &= \left\langle \beta_v - \beta_w, 0 - \sum_i (w_i - v_i) \cdot x_i\big(\langle x_i, \beta_w\rangle - y_i\big) \right\rangle \\ &= \sum_i \Big\langle \beta_v - \beta_w, (w_i - v_i) \cdot x_i\big(\langle x_i, \beta_w\rangle - y_i\big) \Big\rangle.\end{aligned}$$

We now insert $S_v^{1/2} S_v^{-1/2}$ in the middle of each inner product. We apply Cauchy–Schwarz to each term and pull out the scalars (recall that w_i and v_i are scalars, x_i is a vector):

$$\begin{aligned}\big\|S_v^{1/2}(\beta_v - \beta_w)\big\|^2 &= \sum_i \Big\langle S_v^{1/2}(\beta_v - \beta_w), (w_i - v_i) \cdot S_v^{-1/2} x_i\big(\langle x_i, \beta_w\rangle - y_i\big) \Big\rangle \\ &\le \sum_i \big\|S_v^{1/2}(\beta_v - \beta_w)\big\| \cdot \big\|(w_i - v_i) \cdot S_v^{-1/2} x_i\big(\langle x_i, \beta_w\rangle - y_i\big)\big\| \\ &= \sum_i |v_i - w_i| \cdot \big\|S_v^{1/2}(\beta_v - \beta_w)\big\| \cdot \big\|S_v^{-1/2} x_i\big\| \cdot |\langle x_i, \beta_w\rangle - y_i|.\end{aligned}$$

Both sides of the equation have a $\left\|S_v^{1/2}(\beta_v - \beta_w)\right\|$ term; these cancel. We apply ℓ_1/ℓ_∞ on the weight differences, leverage scores, and residuals. There is some subtlety here: define set $U = \operatorname{supp}(v) \cup \operatorname{supp}(w)$. Then we have

$$\begin{aligned}\left\|S_v^{1/2}(\beta_v - \beta_w)\right\| &\le \sum_i |v_i - w_i| \cdot \|S_v^{-1/2} x_i\| \cdot \left|\langle x_i, \beta_w\rangle - y_i\right| \\ &\le \|v - w\|_1 \cdot \left(\max_{i \in U} \|S_v^{-1/2} x_i\| \cdot \left|\langle x_i, \beta_w\rangle - y_i\right|\right).\end{aligned}$$

By the definition of goodness, if $i \in \operatorname{supp}(v)$ then we have $\|S_v^{-1/2} x_i\| \le \sqrt{L}$. Similarly, if $i \in \operatorname{supp}(v)$, we have $|y_i - x_i^\top \beta_w| \le R$.

However, these bounds may not hold for points outside the relevant support. Claim 17 allows us to bound the leverage: for all $i \in \operatorname{supp}(w) \cup \operatorname{supp}(v)$ we have $\|S_v^{-1/2} x_i\|_2^2 \le 2L$, since we assumed $(1 + \|v - w\|_1)L \le \frac{1}{2}$.

Similarly, bounding the residual involves a simple trick alongside Claim 7. Let $\check{w}$ be the entry-wise minimum of $\{w, v\}$, so $\check{w}_i = \min\{w_i, v_i\}$. We have $\|\check{w} - w\|_1, \|\check{w} - v\|_1 \le \|w - v\|_1$ and, furthermore, the support of $\check{w}$ is contained in both the support of w and that of v. Thus, we can apply Claim 7: assuming $i \in \operatorname{supp}(w)$ (since otherwise $i \in \operatorname{supp}(v)$ and we have a bound on the residual)

$$\begin{aligned}|y_i - x_i^\top \beta_v| &= |y_i - x_i^\top \beta_w + x_i^\top \beta_w - x_i^\top \beta_{\check{w}} + x_i^\top \beta_{\check{w}} - x_i^\top \beta_v| \\ &\le |y_i - x_i^\top \beta_w| + |x_i^\top \beta_w - x_i^\top \beta_{\check{w}}| + |x_i^\top \beta_{\check{w}} - x_i^\top \beta_v| \\ &\le R + 2\|\check{w} - w\|_1 LR + 2\|\check{w} - v\|_1 LR \\ &\le \big(1 + 4\|v - w\|_1 L\big)R.\end{aligned}$$

Since $\|v - w\|_1 L \le \frac{1}{4}$, this is at most $2R$. ■

Appendix D. Estimation of σ^2

Algorithm 6: Private σ^2 Estimator

Input: $S = \{(x_i, y_i)\}_{i=1}^n$, target privacy $(\varepsilon_0, \delta_0)$, target failure probability ζ.
Partition S into $k = \lfloor C_1 \log(1/(\delta_0\zeta))/\varepsilon \rfloor$ subsets of equal size and let G_j be the j-th partition, where each dataset is of size $b = |G_j| = \lfloor n/k \rfloor$.
For each $j \in [k]$, denote $\psi_j = \min_\beta (1/|G_j|) \sum_{i \in G_j} (y_i - \beta^\top x_i)^2$.
Partition $[0, \infty)$ into bins of geometrically increasing intervals
$\Omega := \left\{\ldots, \left[2^{-2/4}, 2^{-1/4}\right), \left[2^{-1/4}, 1\right), \left[1, 2^{1/4}\right), \left[2^{1/4}, 2^{2/4}\right), \ldots\right\} \cup \{[0, 0]\}$
Run $(\varepsilon_0, \delta_0)$-DP histogram learner of Lemma 34 on $\{\psi_j\}_{j=1}^k$ over Ω
if all the bins are empty **then** Return $\perp$
Let $[\ell, r]$ be a non-empty bin that contains the maximum number of points in the DP histogram
Return ℓ

Lemma 33 *Algorithm 6 is $(\varepsilon_0, \delta_0)$-DP. Let $S = \{(x_i, y_i)\}_{i=1}^n$ be a dataset of i.i.d. samples with $x_i \sim \mathcal{N}(0, \Sigma)$, $y_i = x_i^\top \beta^* + z_i$ and $z_i \sim \mathcal{N}(0, \sigma^2)$ for some unknown true parameter $\beta^* =$*

$\Sigma^{-1}\mathbb{E}[y_i x_i] \in \mathbb{R}^d$ *and unknown* Σ *and* σ^2*. Suppose*

$$n = \tilde{O}\left(\frac{d\log(1/(\delta_0\zeta))}{\varepsilon_0}\right) ,$$

with a large enough constant then Algorithm 6 returns ℓ *such that, with probability* $1-\zeta$*,*

$$\frac{1}{\sqrt{2}}\sigma^2 \le \ell \le \sqrt{2}\sigma^2 ,$$

where $\tilde{O}(\cdot)$ *hides logarithmic factors in* $\log(1/\varepsilon_0), \log(1/\delta_0)$*.*

We provide a proof in Appendix. D.1.

D.1. Proof of Lemma 33 on the private σ^2 estimation

The privacy proof follows from the DP-histogram from Lemma 34. We provide proof for utility.

For each of the partition G_j, we show $a_i := \frac{1}{b}\sum_{i\in G_j}(y_i - \hat{\beta}_j^\top x_i)^2$ concentrates around the true parameter β^* where $\hat{\beta}_j := \mathrm{argmin}_\beta(1/|G_j|)\sum_{i\in G_j}(y_i - \beta^\top x_i)^2$. Let $f(\beta) = \frac{1}{b}\sum_{i\in G_j}(y_i - \beta^\top x_i)^2$. We know $f(\hat{\beta}_j) = \min_\beta \frac{1}{b}\sum_{i\in G_j}(y_i - \beta^\top x_i)^2 \le \frac{1}{b}\sum_{i\in G_j}(y_i - \beta^{*\top} x_i)^2 = \frac{1}{b}\sum_{i\in G_j} z_i^2$

Since z_i^2 are sub-exponential, from Bernstein bound, we know there exists constant $c_1 > 0$ such that with proability $1-\zeta$, $\frac{1}{b}\sum_{i=1}^b z_i^2 \le \sigma^2(1 + c\sqrt{\frac{\log(1/\zeta)}{b}} + c\frac{\log(1/\zeta)}{b})$.

Now we show lower bound of $f(\hat{\beta}_j)$. For any β, we also have

$$f(\beta) = \frac{1}{b}\sum_{i\in G_j}(y_i - w^\top x_i)^2 = \frac{1}{b}\sum_{i\in G_j}(z_i + x_i^\top(w^* - w))^2 .$$

Let $\tilde{\beta} := \left(\Sigma^{1/2}(\beta^* - \beta), \sigma\right) \in \mathbb{R}^{d+1}$ and $\tilde{x}_i := \left(\Sigma^{-1/2}x_i, z_i/\sigma\right) \in \mathbb{R}^{d+1}$ for $i \in [n]$. By definition, we can see that $\tilde{x}_i$ is zero-mean sub-Gaussian with covariance $\mathbf{I}_{d+1}$.

$$f(\beta) = \frac{1}{b}\sum_{i\in G_j}(\tilde{\beta}^\top \tilde{x}_i)^2.$$

Following Lemma 9 from Jambulapati et al. (2020), we know for any vector $\tilde{\beta}$, there exists $c_2 > 0$ such that with probability $1-\zeta$,

$$\left|\frac{1}{b}\sum_{i\in G_j}(\tilde{\beta}^\top \tilde{x}_i)^2 - \|\tilde{\beta}\|^2\right| = \left|\tilde{\beta}^\top\left(\frac{1}{b}\sum_{i\in G_j}\tilde{x}_i\tilde{x}_i^\top - \mathbf{I}_d\right)\tilde{\beta}\right| \le c_2\sqrt{\frac{d+1+\log(1/\zeta)}{b}} + c_2\frac{d+1+\log(1/\zeta)}{b}$$

This means for any w, we have

$$\begin{aligned} f(\beta) = \frac{1}{b}\sum_{i\in G_j}^b(\tilde{\beta}^\top \tilde{x}_i)^2 &\ge (1 - c_2\sqrt{\frac{d+1+\log(1/\zeta)}{b}} - c_2\frac{d+1+\log(1/\zeta)}{b})(\|\Sigma^{1/2}(\beta-\beta^*)\|^2 + \sigma^2) \\ &\ge (1 - c_1\sqrt{\frac{d+1+\log(1/\zeta)}{b}} - c_2\frac{d+1+\log(1/\zeta)}{b})\sigma^2 . \end{aligned}$$

Together with the upper bound, this implies that there exists constant $c_3 > 0$, such that with probability $1 - \zeta$,

$$\left| \frac{1}{b} \sum_{i \in G_j} a_i - \sigma^2 \right| = \left| f(\hat{\beta}_j) - \sigma^2 \right| \leq c_3 \left(\sqrt{\frac{d + 1 + \log(1/\zeta)}{b}} + \frac{d + 1 + \log(1/\zeta)}{b} \right) \sigma^2 \, .$$

By union bound, there exists a constant $c_4 > 0$ such that if $b \geq c_4(d + \log(k/\zeta))$, then for all $j \in [k]$,

$$|\psi_j - \sigma^2| = |\frac{1}{b} \sum_{i \in G_j} a_i - \sigma^2| \leq 2^{1/8} \sigma^2 \, .$$

With probability $1 - \zeta$, $\{\psi_j\}_{j=1}^k$ lie in interval of size $2^{1/4}\sigma^2$. Thus, at most two consecutive bins are filled with $\{\psi_j\}_{j=1}^k$. Denote them as $I = I_1 \cup I_2$.

Our analysis indicates that $\mathbb{P}(\psi_i \in I) \geq 0.99$. By private histogram in Lemma 34, if $k \geq c_5 \log(1/(\delta_0 \zeta))/\varepsilon_0$, $|\hat{p}_I - \tilde{p}_I| \leq 0.01$ where $\hat{p}_I$ is the empirical count on I and $\tilde{p}_I$ is the noisy count on I. Under this condition, one of these two intervals are released. This results in multiplicative error of $\sqrt{2}$.

Lemma 34 (Stability-based histogram (Karwa and Vadhan, 2018, Lemma 2.3)) *For every $K \in \mathbb{N} \cup \{\infty\}$, domain Ω, for every collection of disjoint bins $B_1, \ldots, B_K$ defined on Ω, $n \in \mathbb{N}$, $\varepsilon \geq 0, \delta \in (0, 1/n)$, $\beta > 0$ and $\alpha \in (0, 1)$ there exists an (ε, δ)-differentially private algorithm $M : \Omega^n \to \mathbb{R}^K$ such that for any set of data $X_1, \ldots, X_n \in \Omega^n$*

(1) $\hat{p}_k = \frac{1}{n} \sum_{X_i \in B_k} 1$

(2) $(\tilde{p}_1, \ldots, \tilde{p}_K) \leftarrow M(X_1, \ldots, X_n)$, *and*

(3)

$$n \geq \min \left\{ \frac{8}{\varepsilon\beta} \log(2K/\alpha), \frac{8}{\varepsilon\beta} \log(4/(\alpha\delta)) \right\}$$

then,

$$\mathbb{P}(|\tilde{p}_k - \hat{p}_k| \leq \beta) \geq 1 - \alpha$$

Appendix E. Lower Bound

Series of advances have been made in designing tools for lower bounds in statistical estimation.

Fingerprinting Narayanan (2023).

Our lower bound is a direct corollary of a similar lower bound on linear regression from Cai et al. (2023).

Theorem 35 *Let $\mathcal{P}_{\Sigma,\sigma^2}$ be a class of distributions over $(x_i, z_i) \in \mathbb{R}^d \times \mathbb{R}$, where x_i are i.i.d. samples from a d-dimensional subgaussian distribution with mean 0 and covariance $\Sigma \succ 0$, and z_i are i.i.d. samples from a subgaussian distribution with mean 0 and variance σ^2 (see Definition 29*

in Appendix B). We observe labelled examples from linear model: $y_i = \beta^\top x_i + z_i$ *with* $\mathbb{E}[x_i z_i] = 0$. *Let* $\mathcal{M}_{\varepsilon,\delta}$ *be a class of* (ε, δ)*-DP estimators that are functions over the datasets* $S = \{(x_i, y_i)\}_{i=1}^n$. *Then if* $0 < \varepsilon < 1$, $d \lesssim n\varepsilon$, $\delta \lesssim n^{-(1+\gamma)}$ *for some* $\gamma > 0$, *there exists constant* $C > 0$ *such that*

$$\inf_{M \in \mathcal{M}_{\varepsilon,\delta}} \sup_{\mathcal{P}_{\Sigma,\sigma^2},\beta} \mathbb{E}\|M(y,x) - \beta\|_\Sigma^2 \geq C\sigma^2 \left(\frac{d}{n} + \frac{d^2}{n^2\varepsilon^2}\right).$$

Proof We will apply the lower bound below from Cai et al. (2023).

Theorem 36 ((Cai et al., 2023, Theorem 3.1)) *Consider i.i.d. observations* $\{(y_1, x_1), \cdots, (y_n, x_n)\}$ *drawn from the Gaussian linear model:*

$$f_\beta(y \mid x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-\left(y - x^\top\beta\right)^2}{2\sigma^2}\right); x \sim f_x \, .$$

Suppose $\mathbb{E}[xx^\top]$ *is diagonal, and* $\lambda_{\max}(\mathbb{E}[xx^\top]) < C' < \infty$, $\|X\|_2 \lesssim \sqrt{d}$ *almost surely. If* $d \lesssim n\varepsilon, 0 < \varepsilon < 1$ *and* $\delta \lesssim n^{-(1+\gamma)}$ *for some* $\gamma > 0$, *then*

$$\inf_{M \in \mathcal{M}_{\varepsilon,\delta}} \sup_{\beta \in \mathbb{R}^d} \mathbb{E}\|M(y,x) - \beta\|_2^2 \gtrsim \sigma^2 \left(\frac{d}{n} + \frac{d^2}{n^2\varepsilon^2}\right) \, .$$

Note that this lower bound holds for every construction of x_i that satisfies the assumption. We construct one instance of joint distribution $P \in \mathcal{P}_{\Sigma,\sigma^2}$ such that it also satisfies the assumptions in Theorem 36. Let $\{x_i\}_{i=1}^n$ be i.i.d. samples from $\mathcal{N}(0, \mathbf{I}_d)$. And let $\tilde{x}_i = x_i \cdot \mathbf{I}[x_i \leq \sqrt{d}]$. Clearly, $\{\tilde{x}_i\}_{i=1}^n$ satisfies that $\mathbb{E}[\tilde{x}\tilde{x}^\top]$ is diagonal, $\lambda_{\max}(\mathbb{E}[\tilde{x}\tilde{x}^\top]) < 1$ and $\tilde{x}_i$ are bounded by $\sqrt{d}$. Let z_i be independent Gaussian distribution with variance σ^2. By Theorem 36, we know there exists constant C such that

$$\inf_{M \in \mathcal{M}_{\varepsilon,\delta}} \sup_{\mathcal{P}_{\Sigma,\sigma^2},\beta} \mathbb{E}\|M(y,x) - \beta\|_\Sigma^2 \geq C\sigma^2 \left(\frac{d}{n} + \frac{d^2}{n^2\varepsilon^2}\right).$$

■