PRIVATE LINEAR REGRESSION VIA A DOWN-SENSITIVITY TO PRIVACY REDUCTION

Ittai Rubinstein
Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
ittair@mit.edu

Chris Ge
Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
cge7@mit.edu

Samuel B. Hopkins
Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
samhop@mit.edu

ABSTRACT

We present a sample- and time-efficient (ε, δ)-differentially private (DP) algorithm for d-dimensional linear regression with a sample complexity of

$$n_{\text{STAR}} \approx \frac{d}{\alpha^2} + \frac{d\log(1/\delta)}{\alpha\varepsilon} + \frac{d\log(1/\delta)}{\varepsilon} + o(d)\,.$$

This improves upon prior polynomial-time algorithms whose sample complexity either depends on the condition number of the design matrix κ (for DP-SGD with gradient clipping), scales quadratically with the dimension (for Sum-of-Squares algorithms) or with the privacy parameter (for outlier removal algorithms such as insufficient statistics perturbation or ISSP),

$$n_{\text{DP-SGD}} = \widetilde{\Omega}\left(\frac{d\sqrt{\kappa}}{\varepsilon}\right), \quad n_{\text{SoS}} = \widetilde{\Omega}\left(\frac{d^2}{\alpha^2}\right), \quad n_{\text{ISSP}} = \widetilde{\Omega}\left(\frac{d}{\varepsilon^2}\right).$$

Our algorithm is based on a novel *subsample-test-aggregate* (STA) approach for ensuring privacy given only bounded *down-sensitivity* – robustness to removal, but not addition, of a small number of samples. The intuition that down-sensitivity should be related to privacy is not new, but STA formalizes this by providing an *efficient black-box reduction from down-sensitivity to privacy* which we expect to be applicable beyond the setting of linear regression.

Keywords: differential privacy; linear regression; OLS; high-dimensional

1 INTRODUCTION

As the fields of machine learning, medicine and econometrics become ever more data-driven, the problem of ensuring the privacy of individuals also grows in importance. *Differential privacy* (DP) is the gold standard mathematical definition for an algorithm to protect privacy of individuals represented in its input Dwork et al. (2006b). Despite two decades of intensive research, DP algorithms for foundational tasks in statistics and machine learning lag far behind their non-private counterparts in accuracy, running time, or both. Indeed, this is true even for basic tasks like estimation of mean and covariance and linear regression Alabi et al. (2020); Barrientos et al. (2024).

We study DP algorithms for high-dimensional linear regression, for two main reasons. First, linear regression is an important problem in its own right – it is widely used across the sciences, it forms the foundation for many more advanced statistical prediction methods, and it is a basic algorithmic building block for optimization, machine learning, and beyond. Second, it is a useful

lens on DP algorithm design for high-dimensional statistics more broadly, since improving on the privacy-accuracy tradeoffs of existing efficient algorithms for linear regression requires developing algorithmic techniques which go beyond the standard DP toolkit.

Our main contribution is STAR – an (ε, δ)-DP algorithm for linear regression, with runtime comparable to OLS (up to logarithmic factors) and a significantly better sample complexity than the previous state-of-the-art. This brings the dream of practical private linear regression one step closer to reality.

Prior polynomial-time algorithms require a sample complexity that depends on the condition number of the covariates κ, or scales quadratically either with the privacy parameter or with the dimension Brown et al. (2024); Anderson et al. (2025); Liu et al. (2023)

$$n_{\text{PRIOR}} \geq \Omega\left(\min\left\{\frac{d\sqrt{\kappa}}{\varepsilon}, \frac{d^2}{\alpha^2}, \frac{d}{\varepsilon^2}\right\}\right).$$

Our second contribution is a new framework for privately releasing the output of a non-Lipschitz function of the input dataset. We call this framework *subsample-test-aggregate* (STA), since it is partially inspired by the classic subsample-and-aggregate framework Nissim et al. (2007), and we instantiate it to give our main result on linear regression.

The ability to release highly sensitive functions privately is crucial for statistics, where many of the basic primitives – empirical means, covariances, regressors – are highly sensitive to the addition of a single outlier to a dataset. Standard tools like the Gaussian or Laplace mechanisms, or the Propose-Test-Release framework Dwork & Lei (2009) cannot be applied to these problems directly, since *every dataset has unbounded sensitivity to a single sample change*. While a range of techniques have appeared in recent years to address this challenge in the context of high-dimensional statistics Brown et al. (2024; 2023); Liu et al. (2022; 2023); Hopkins et al. (2022; 2023); Asi et al. (2023); Kuditipudi et al. (2023), when applied to linear regression all come at a significant cost to the sample or computational complexity of the resulting estimator, leading to the $\Omega\left(\min\left\{\frac{d\sqrt{\kappa}}{\varepsilon}, \frac{d^2}{\alpha^2}, \frac{d}{\varepsilon^2}\right\}\right)$ state-of-the-art. STA thus improves on the privacy-accuracy tradeoffs obtained by prior DP algorithm design methods in high-dimensional statistics, and hence we expect it to be of use more broadly.

1.1 Results

We now describe our results for linear regression more formally; we discuss the STA framework in Section 1.2 after describing more background. Recall the definition of (ϵ, δ)-differential privacy:

Definition 1.1 ((ε, δ)-Differential Privacy, Dwork et al. (2006a))**.** A (randomized) algorithm $\mathcal{A}$ with domain $\mathcal{X}^n$ satisfies (ε, δ)-differential privacy if for all pairs of neighboring datasets $X, X' \in \mathcal{X}^n$ that differ in at most one sample, and for all measurable subsets S of the output space of $\mathcal{A}$,

$$\Pr[\mathcal{A}(X) \in S] \;\leq\; e^{\varepsilon} \Pr[\mathcal{A}(X') \in S] + \delta.$$

In keeping with much recent work on DP statistics, e.g. Karwa & Vadhan (2017); Brown et al. (2024; 2023; 2021); Liu et al. (2022); Hopkins et al. (2023); Georgiev & Hopkins (2022); Narayanan et al. (2022); Asi et al. (2023); Kuditipudi et al. (2023); Cai et al. (2021); Kamath et al. (2019); Arbas et al. (2023); Biswas et al. (2020); Kamath et al. (2020); Ashtiani & Liaw (2022); Anderson et al. (2025); Dagan et al. (2024), we study algorithms which satisfy DP with respect to all inputs and have good prediction/estimation error when the data is drawn from a "nice" distribution such as Definition 1.2 (i.e., *worst-case* privacy guarantees, *average-case* utility guarantees).[1]

Definition 1.2. Let $\Sigma \succ 0$ with $\Sigma \in \mathbb{R}^{d\times d}$, $\sigma^2 > 0$, and $\beta^* \in \mathbb{R}^d$. The n-sample linear Gaussian model consists of n i.i.d. draws $(x_1, y_1), \ldots, (x_n, y_n)$ where

$$x_i \sim \mathcal{N}(0, \Sigma) \text{ and } y_i \mid x_i \sim \mathcal{N}(x_i^\top \beta^\star, \sigma^2).$$

[1] A distinct line of work, including *sufficient statistics perturbation* (SSP) and variants like AdaSSP Sheffet (2017); Wang (2018); Lev et al. (2026) focuses on DP linear regression under stronger assumptions on the covariates – typically ℓ_2 boundedness and small condition number. The state-of-the-art of these methods also has a sample complexity of $\frac{d^{3/2}}{\alpha\varepsilon} \geq \min\left\{\frac{d}{\varepsilon^2}, \frac{d^2}{\alpha^2}\right\}$, no better than the ISSP or SoS approaches Anderson et al. (2025); Brown et al. (2024).

We measure the performance of a linear regression estimator $\hat{\beta}$ by its (normalized) excess risk α, if

$$\alpha^2 \coloneqq \frac{1}{\sigma^2}\left(\mathop{\mathbb{E}}_{(x,y)}\left[(x^\top\hat{\beta}-y)^2\right] - \mathop{\mathbb{E}}_{(x,y)}\left[(x^\top\beta^\star-y)^2\right]\right),$$

where (x, y) is a fresh draw from the same linear Gaussian model.

We introduce a time and sample efficient (ε, δ)-DP regression algorithm, STAR (*Subsample-Test-Aggregate Regression*). We state its main guarantees here and describe the algorithm in Section 1.2.

Theorem 1.3 (Privacy and utility of STAR)**.** STAR *(Algorithm 6) is (ε, δ)-DP, runs in time $\widetilde{O}\left(nd^2(1+\lambda)\right)$, achieves excess risk α on well-behaved data (Definition 1.2) and has sample complexity*

$$n_{\text{STAR}} = \widetilde{O}\left(\frac{d}{\alpha^2} + \frac{d(1+\lambda)\log(1/\delta)}{\alpha\,\varepsilon} + \frac{d(1+\lambda)\log(1/\delta)}{\varepsilon}\right) \quad \text{for} \quad \lambda \coloneqq \frac{\log(1/\delta)}{\sqrt{d}}.$$

The computational complexity of STAR is dominated by $\widetilde{O}(n/d)$ linear algebra operations over $(1+\lambda)d \times d$ matrices. As such, it can be implemented with fast matrix multiplication achieving an asymptotic runtime of $\widetilde{O}\left(nd^{\omega-1}(1+\lambda)\right)$ – matching the runtime of OLS up to logarithmic factors.

From a statistical standpoint, even without privacy, estimating a d-dimensional regression parameter to error α requires on the order of d/α^2 samples, but privacy comes at a cost and information-theoretic lower bounds imply that any (ε, δ)-DP estimator achieving excess risk α with constant success probability must use

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

samples (see, e.g., Karwa & Vadhan (2017); Cai et al. (2023); Brown et al. (2024)).

These bounds are essentially tight for inefficient algorithms. Liu et al.'s high-dimensional propose-test-release (HPTR) framework gives an exponential-time estimator with sample complexity which nearly matches the lower bound up to logarithmic factors when α is a constant Liu et al. (2022). This highlights that the bottleneck for private regression algorithms is computational rather than statistical.

The most notable gap between n_{STAR} and the optimal sample complexity is the leading-order dependence on $\log(1/\delta)$, which is $\log(1/\delta)/\epsilon$ in the optimal sample complexity, but incurs an extra d factor in n_{STAR}. This is partially explained by statistical query lower bounds, which suggest that polynomial time DP regression algorithms using $o(d^2)$ samples must also use $\omega(\log(1/\delta)/\epsilon)$ samples Georgiev & Hopkins (2022); Anderson et al. (2025); thus, we do not expect to eliminate this d factor gap entirely. The other gaps between n_{STAR} and the optimal sample complexity are $o(d)$ additive terms, or logarithmic factors.

In Table 1, we summarize the state of DP linear regression algorithms. All existing polynomial-time algorithms fall short of the optimal sample complexity in (at least) one of three ways: requiring $\Omega\left(\frac{d^2}{\alpha^2}\right)$ samples Anderson et al. (2025), requiring $\Omega\left(\frac{d}{\varepsilon^2}\right)$ samples Brown et al. (2024), or requiring samples growing polynomially with the condition number of Σ.

1.2 Techniques

1.2.1 Setting and Challenges

The Gaussian Mechanism To get started, it will be useful to review the Gaussian mechanism, one of the foundational algorithmic tools in DP Dwork & Roth (2014). Suppose $f : \mathcal{X}^n \to \mathbb{R}^d$ is a function from n-element datasets to $\mathbb{R}^d$ which satisfies the following *bounded sensitivity* property: for all datasets $X, X' \in \mathcal{X}^n$ which differ on exactly one element, $\|f(X) - f(X')\| \leq 1$, where $\|\cdot\|$ is Euclidean/ℓ_2 norm. Then the following algorithm is (ϵ, δ)-DP: on input X, output $f(X) + Z$, where $Z \sim \mathcal{N}(0, O(\frac{\log(1/\delta)}{\epsilon^2}) \cdot I)$ is Gaussian.

At the highest level, our goal in what follows is to develop a variant of the Gaussian mechanism which allows us to privately release a (noisy version of) a function f which is a good estimator of

Paper	Sample Complexity	Method	Poly-time?
Liu et al. (2022)	$\frac{d}{\alpha^2} + \frac{d+\log(1/\delta)}{\alpha\varepsilon}$	HPTR	No
Varshney et al. (2022)	$\frac{d}{\alpha^2} + \frac{\kappa\sqrt{L}Rd\log(1/\delta)}{\alpha\varepsilon}$	DP-SGD	Yes
Liu et al. (2023)	$\frac{d}{\alpha^2} + \frac{\sqrt{\kappa}d\log(1/\delta)}{\alpha\varepsilon}$	DP-SGD	Yes
Asi et al. (2023)	$\frac{d}{\alpha^2} + \frac{d\log(R+\sqrt{\kappa})}{\alpha\varepsilon}$	exp. mech.	No
Brown et al. (2024)	$\frac{d}{\alpha^2} + \frac{d\sqrt{\log(1/\delta)}}{\alpha\varepsilon} + \frac{d\log^2(1/\delta)}{\varepsilon^2}$	outlier filtering	Yes
Anderson et al. (2025)	$\frac{d^2}{\alpha^2} + \frac{d}{\alpha\varepsilon} + \frac{d\log RL}{\varepsilon}$	SoS	Yes
Anderson et al. (2025)	$\frac{d^2}{\alpha^2} + \frac{d}{\alpha\varepsilon} + \frac{\log(1/\delta)}{\varepsilon}$	SoS	Yes
STAR (Theorem 1.3)	$\frac{d}{\alpha^2} + \frac{d\log(1/\delta)}{\alpha\varepsilon} + \frac{d\log(1/\delta)}{\varepsilon} + o(d)$	STA	Yes

Table 1: Sample complexities of private linear regression algorithms for well-behaved data (Definition 1.2). As in Anderson et al. (2025), we set d to be the dimension, ε, δ to be the privacy parameters, α to be the excess risk, $R \geq \|\beta^\star\|_2$ to be an upper bound on the norm of the ground truth model, κ an upper bound on the condition number of Σ and L an upper bound on the operator norm of Σ.

β^* when fed i.i.d. samples from a Gaussian linear model. The standard estimator of β^* is ordinary least squares (OLS). But the OLS estimator is very far from satisfying the stability property above.

Down-Sensitivity Even though they don't satisfy the strong stability conditions needed for the Gaussian mechanism, statistics like empirical means, covariances, and OLS regressors do satisfy a weaker stability notion: there are *efficiently verifiable* conditions that hold with high probability on "nice" input data (e.g., Definition 1.2) and suffice to guarantee robustness to *removing* one or a few samples. This property is called having bounded *down-sensitivity*, and it is at the heart of many modern algorithms for DP high-dimensional statistics.

For instance, it is a well known fact in the statistics community that an OLS is robust to removing any single sample if it has bounded leverages and residuals Cook & Weisberg (1982), and this is the core of the ISSP regression algorithm of Brown et al. (2024): starting from the entire regression, samples are carefully removed until this down-sensitivity condition is met. Similarly, the Sum of Squares (SoS) based regression algorithm of Anderson et al. (2025) utilizes a down-sensitivity condition introduced by Bakshi & Prasad (2021) that is particularly amenable to the SoS framework.

However, the route from bounded down-sensitivity to DP algorithms typically has many technical challenges, often overcome with problem-specific algorithmic strategies. For instance, the main challenge in the ISSP algorithm is deciding *which* samples to remove from the regression, leading to the $\frac{d\log^2(1/\delta)}{\varepsilon^2}$ term in their sample complexity, and the SoS approach of Anderson et al. (2025) comes at a significant cost to the runtime of their algorithm and is limited to down-sensitivity bounds that fit the SoS framework resulting in the $\frac{d^2}{\alpha^2}$ term of their sample complexity.

More generally, if we are given a new down-sensitivity bound, it might be possible to find a corresponding filtering method a-la ISSP, and it might be possible to convert it to a (slow) polynomial time SoS algorithm, but we are not guaranteed to get an efficient private algorithm from either framework.

Our STA technique charts a new course from down-sensitivity to DP estimation. Unlike previous approaches, STA is agnostic to the algorithm used to check down-sensitivity – instead, STA makes *black-box* use of an oracle which can certify down-sensitivity of a learning algorithm (in our case, OLS regression) on subsamples of an input dataset X. We turn next to a more detailed overview of STA.

Subsample-and-Aggregate Another important mechanism to consider in our comparison is *subsample-and-aggregate* Dwork & Roth (2014). In this framework, subsamples (i.e., small subsets

of the dataset) are drawn at random, the (potentially sensitive) learner is applied to each of the subsets separately, and the outputs of the learners are aggregated using a robust aggregation algorithm. Stability is obtained from the fact that changing any single entry in the dataset can only affect a small fraction of the subsamples allowing us to rely on the robustness guarantee of the aggregator. Finally, a standard mechanism (e.g., the Gaussian mechanism) is applied to this now-stable statistic to ensure privacy.

As a concrete example, a natural way to apply the subsample-and-aggregate framework to the problem of private linear regression, could be to draw a large number N of random subsamples from our original regression, to apply the standard OLS formula to each of them, and then aggregate using the BHS private mean estimation algorithm of Brown et al. (2023). In order to apply the OLS formula on each of these subsamples, we would most-likely need each subsample to contain at least $t \geq d/n$ fraction of the samples (otherwise the covariates might be ill-defined).

Therefore, an adversary changing a single sample in the regression might corrupt up to $\approx tN$ of the subsamples. Therefore, to ensure privacy of the overall subsample-and-aggregate algorithm to a single change in the dataset, we need to ensure privacy of the BHS aggregator to tN changes in its input. Composing the privacy guarantees of Brown et al. (2023) to the change in a single learner tN times yields a privacy-accuracy tradeoff of

$$N \geq \frac{d \log(1/\delta)}{\varepsilon} \cdot tN \Rightarrow n \geq \frac{d}{t} \geq \frac{d^2 \log(1/\delta)}{\varepsilon},$$

yielding a quadratic-in-d sample complexity.[2]

Subsample-Test-Aggregate Motivated by these shortcomings of existing privacy mechanisms, we introduce *Subsample-Test-Aggregate* (STA) – a privacy framework combining elements of subsample-and-aggregate and propose-test-release. At a high-level, the STA algorithm generates random subsamples of the dataset, uses a sensitivity test to reject sensitive subsamples of the dataset and then aggregates the results. STA differs from subsample-and-aggregate in that its stability now stems solely from the test (allowing us to use simple aggregators such as outputting the empirical mean plus Gaussian noise) and from propose-test-release in that it suffices to test only for down-sensitivity.

Concretely, STA has three main building blocks: a *learner* $A : \mathcal{X}^* \to \Omega$, which takes a dataset X in $\mathcal{X}^*$ and outputs a hypothesis in a hypothesis space Ω, a *sensitivity test* $\mathcal{T} : \mathcal{X}^* \to \{\text{ACCEPT}, \text{REJECT}\}$, which tests whether A has bounded down-sensitivity on a given dataset, and an *aggregator*, taking several hypotheses and aggregating them into a single one. Later, we will generalize the setting, to allow the tester $\mathcal{T}(X)$ to output a quantitative bound on the degree of down-sensitivity that A has on dataset X; we start with the simpler version for sake of exposition.

Algorithm 1: SUBSAMPLE-TEST-AGGREGATE

Input: Dataset $X_{1:n} \in \mathcal{X}^n$;
learner $A : \mathcal{X}^\star \to \Omega_A$;
stability test $\mathcal{T} : \mathcal{X}^\star \to \{\text{ACCEPT}, \text{REJECT}\}$.
Output: A model in Ω or REJECT.
for $j \leftarrow 1$ **to** N **do** `// draw N iid Poisson subsamples`
  Draw $S_j \subseteq [n]$ by including each $i \in [n]$ iid with probability t
  $T_j \leftarrow \mathcal{T}(X_{S_j})$
if $|\{j \in [N] \mid T_j = \text{ACCEPT}\}|/N$ *is not sufficiently high* **then**
  return REJECT `// private acceptance-rate check; see Thm. A.5`
$\mathcal{J} \leftarrow \{\, j \in [N] : T_j = \text{ACCEPT} \,\}$
return $\text{Aggregate}\big(\{\, A(X_{S_j}) : j \in \mathcal{J} \,\}\big)$

[2]This analysis is just one way to use subsample-and-aggregate for DP-OLS, and we do not claim to prove an impossibility result. However, the issues we raise seem to be structural and to the best of our knowledge, no known subsample-and-aggregate algorithm for high dimensional statistics like OLS or mean estimation avoids them.

The overall goal is to map an input dataset in $\mathcal{X}^n$ to a single good hypothesis for that dataset, in Ω, subject to DP. Crucially, the learner A *may have large or unbounded down-sensitivity on the input dataset* X – allowing such datasets is important to obtain a DP algorithm, since in our settings every dataset is adjacent to one where A has large down-sensitivity. The STA scheme leverages the existence of *subsets* of X on which A is down-stable.

The STA scheme does this as follows. Given an n-element dataset X, pick $X_{S_1}, \ldots, X_{S_N}$ as random subsamples of X. For each of the subsamples X_{S_i}, run the tester $\mathcal{T}(X_{S_i})$ to test whether A has bounded down-sensitivity on X_{S_i}. (We give the formal meaning of this below.) If sufficiently many of the tests pass, give the passing outputs of the learner $\{A(X_{S_i}) : \mathcal{T}(X_{S_i}) = \text{ACCEPT}\}$ to the aggregator, and output its aggregated hypothesis. Unlike subsample-and-aggregate, the aggregator does not need to satisfy any robustness or stability properties – we show that testing down-sensitivity of the subsamples is instead enough to obtain strong privacy guarantees.

1.2.2 STAR-LITE: A CONCRETE EXAMPLE OF STA

To gain better intuition for the STA framework, we begin with a simplified version of the STAR algorithm – STAR-LITE (Algorithm 2).

STAR-LITE utilizes only the down-sensitivity implied by bounded leverages and residuals Cook & Weisberg (1982), and has similar privacy and runtime guarantees to those of STAR, with only a slightly worse sample complexity than the full STAR algorithm (though it still obtains a strictly better sample complexity than ISSP).

$$n_{\text{STAR-LITE}} = \widetilde{O}\left(\frac{d}{\alpha^2} + \frac{d\log(1/\delta)}{\alpha\varepsilon} + \frac{d\log^2(1/\delta)}{\varepsilon}\right) < \widetilde{O}\left(\frac{d}{\alpha^2} + \frac{d\log^2(1/\delta)}{\varepsilon^2}\right).$$

Algorithm 2: STAR-LITE: A simplified version of STAR.

Input: Data $(X, y) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$
Output: Private linear model $\hat{\beta} \in \mathbb{R}^d$

Set $\mathcal{T}$ to be the *stability test* that accepts a subsample $S \subseteq [n]$ if the corresponding regression has bounded leverages $\lesssim L$ and residuals $\lesssim R$; `// where` $\lesssim$ `denotes a randomized soft threshold`

Define the Poisson-subsample process $S \subseteq [n]$ to have $i \in S \overset{i.i.d.}{\sim} \text{Bern}(t)$ with $t \approx \frac{\log(1/\delta)d}{n}$;

if $\Pr[\mathcal{T}(S)] \ll 1$ `// this can be checked privately - see Thm A.5.`
then
  return *REJECT*

Draw $k \approx \frac{\log(n)}{\alpha^2 \log(1/\delta)}$ fresh Poisson-subsamples S_j, conditioned on $\mathcal{T}$ accepting;
Compute $v_j := \text{OLS}(X_{S_j}, y_{S_j})$ and the down-sensitivity bounds $E_j := R^2 L(X_{S_j}^\intercal X_{S_j})^{-1}$;
return *a sample from*

$$\mathcal{N}\left(\frac{1}{k}\sum_{j=1}^{k} v_j, O\left(\frac{\log(1/\delta)}{k^2}\right)\sum_{j=1}^{k} E_j\right)$$

We will see that the test $\mathcal{T}(X_{S_j})$ ensures that the matrix E_j bounds the magnitude and "shape" of the change to v_j which can be achieved by removing any single element of S_j. Concretely, if Δ_j is this change, then we will show $\Delta_j \Delta_j^\top \preceq E_j$. Thus, the noise STAR-LITE adds to $\frac{1}{k}\sum_{j=1}^{k} v_j$ adapts to the down-sensitivity of the subsampled OLS instances.

1.2.3 SUBSAMPLE-TEST-AGGREGATE: A NEW DP MECHANISM

Down-Sensitivity of the Learning Algorithm The most important relationship between the tester and learner is that if the tester outputs ACCEPT, it constitutes a promise that the learner is not down-sensitive. For purposes of this overview, we adopt the following definition, which assumes that $\Omega = \mathbb{R}^d$.

Definition 1.4 (Down-Sensitivity Test (Simplified))**.** Let $E \succeq 0$. We say that $\mathcal{T}$ is an $E \in \mathbb{R}^{d\times d}$ down-sensitivity test for $A : \mathcal{X}^\star \to \mathbb{R}^d$, if

$$\forall S \subseteq [n] \forall i \in S \quad \Pr\left[\mathcal{T}(X_S) = \text{Accept}\right] > 0 \Rightarrow \left\| E^{-1/2}\left(A(X_S) - A(X_{S\setminus\{i\}})\right)\right\|_2 \leq 1\,.$$

The matrix E allows us to adapt down-sensitivity testing to a problem-appropriate norm – later we will return to the question of how to choose E for linear regression. Note that the definition of a down-sensitivity test is one-sided – it is possible for $\mathcal{T}$ to REJECT on some X where A actually does have bounded down-sensitivity. This is important, since natural two-sided versions of this testing problem are computationally intractable Moitra & Rohatgi (2022). But it also means that when we use STA, we need to avoid trivial tests $\mathcal{T}$ which always REJECT.

Down-Stability of the Test To ensure that STA is private, we will need a second down-stability property, this time *of the test itself* – the probability that the test accepts should not *decrease* by too much when a sample is removed. Notably, it could *grow* arbitrarily – this is crucial, since removing a single outlying sample can cause a dataset to go from highly down-sensitive to non-down-sensitive, meaning that the accept probability of the test should increase dramatically (perhaps from zero to nonzero).

Definition 1.5 (Down-Stable Sensitivity Test (Simplified))**.** We say that the test $\mathcal{T}$ is down-stable if for any dataset X and any index i, we have

$$\Pr\left[\mathcal{T}(X) = \text{Accept}\right] = O\left(\Pr\left[\mathcal{T}(X_{-i}) = \text{Accept}\right]\right)$$

where X_{-i} denotes X with the i-th sample removed. Here the probability is taken only with respect to internal randomness of the test $\mathcal{T}$.

Note that the existence of a down-stable sensitivity test $\mathcal{T}$ for a learning algorithm A implies bounds on the down-sensitivity of A to removal of more than one sample – if $\mathcal{T}(X)$ accepts with nonzero probability on X, then by down-stability it must accept with nonzero probability on all subsets X' of X, and hence A also has bounded down-sensitivity on all subsets X'. Later we will generalize the definition of down-stability of the test to allow for A's down-sensitivity to increase as samples are removed – this is necessary for OLS, which becomes increasingly down-sensitive when the number of samples removed is so great that the design matrix becomes (near) singular.

Stability and Privacy of STA The following theorem captures the first key intuition about STA: the *average* of $\mathcal{T}$-accepted subsets is stable, not just to sample removals but also to arbitrary changes to a single sample.

Theorem 1.6 (Stability of Mean Aggregation STA (Theorem A.8 with $\eta = 0$))**.** *Let $\mathcal{T}$ be a down-stable $E \in \mathbb{R}^{d\times d}$ down-sensitivity test for $A : \mathcal{X}^\star \to \mathbb{R}^d$. Let $S \subseteq [n]$ be drawn by including each index independently with sufficiently small probability t. Let $\mu(X) = \mathbb{E}\left[A(X_S)|\mathcal{T}(X_S) = \text{ACCEPT}\right]$ and $M(X) = E + \text{Cov}\left(A(X_S) \mid \mathcal{T}(X_S) = \text{ACCEPT}\right)$. Then, for any neighboring datasets X, X' that differ on exactly one point, we have*

$$\left\| M(X)^{-1/2}\left(\mu(X) - \mu(X')\right)\right\|_2 = O(t)\,,$$

and

$$(1 - O(t))M(X) \preceq M(X') \preceq (1 + O(t))M(X)\,.$$

The stability condition $\|M(X)^{-1/2}(\mu(X) - \mu(X'))\| \leq O(t)$ suggests that if we add noise with covariance $O(t^2) \cdot M(X)$ to $\mu(X)$, the result would be safe to release privately. However, this plan faces several complications. First, $M(X)$ itself depends on X, which means that the *shape* of the noise we add might leak private information – although since $M(X)$ itself is stable in spectral order we can hope to avoid too much privacy loss through this channel. Second, the privacy cost of adding Gaussian noise with covariance $O(t^2) \cdot M(X)$ is difficult to analyze.[3] Third, STA doesn't have access directly to $\mu(X)$; instead it can only draw samples from the distribution $A(X_S)$ for a random subset S.

[3]To the best of our knowledge, it may even violate privacy.

We address the challenges by using a different noise distribution with covariance proportional to $M(X)$. To sample $\mu(X) + (\text{noise})$, first, for an integer k, draw k samples from $A(X_S)$ conditioned on $\mathcal{T}(X_S) = \textsc{Accept}$. Then, average the samples and add $Z \sim \mathcal{N}(0, (\log(1/\delta)/k)E)$. The covariance of this distribution is similar to $M(X)$. Note that $\mu(X)$ never needs to be computed directly. To prove privacy of the resulting mechanism, we combine the stability arguments underlying Theorem 1.6 with a generalization of the "advanced composition" method from DP. This argument is captured in the following theorem, which we can use to analyze the "Aggregate" step of STA when the aggregator is the empirical average plus Gaussian noise.

Theorem 1.7 (Informal, see Theorem B.3)**.** *The mechanism*

$$X \to \text{empirical mean over } k \text{ samples from } \operatorname{Law}\left(\mathcal{N}\big(A(X_S), O(\log(1/\delta)E\big) \mid S \sim \mathsf{Acc}_X\right)$$

is $\left(O(\sqrt{k\log(1/\delta)}t, \delta\right)$*-private.*

The freedom to average over k draws from $A(X_S)$ is crucial. In our application to OLS, each subset S will be too small to get a good estimator of β^* using only X_S. By averaging over many draws of $A(X_S)$, we get accuracy comparable to (non-private) OLS.

To complete the privacy analysis of Algorithm 1, we also need to analyze the privacy of the "acceptance rate check" step, which makes sure that $\mathcal{T}(X_S)$ accepts with high-enough probability. Down-stability of $\mathcal{T}$ turns out to be enough to privately check the acceptance rate – see Theorem A.5.

Generalizations Needed for OLS We will need two key relaxations of the assumptions above to use STA for OLS regression. First, as we discussed above, we will need to allow for down-sensitivity of A to sometimes *increase* when a sample is removed. We can achieve this by modifying the definition of down-stability for $\mathcal{T}$ to read:

$$\Pr\left[\mathcal{T}(X_S) = \text{Accept}\right] = O\left(\Pr\left[\mathcal{T}(X_{S\setminus\{i\}}) = \text{Accept}\right]\right) + \eta\,,$$

where we will eventually use $\eta \approx \delta^{O(1)}$.

The second relaxation we make is to Definition 1.4, where instead of assuming that $\mathcal{T}(X) = \textsc{Accept}$ implies a bound on the down-sensitivity of A in some fixed E geometry, we allow $\mathcal{T}(X)$ to return a matrix $E(X)$, and an $\textsc{Accept}$ output guarantees bounded down-sensitivity for A in the $E(X)$-norm. We prove an analogue of Theorem 1.7 with varying $E(X)$, using a guarantee that $E(X)$ itself is monotonically non-increasing and changes smoothly

$$E(X) \preceq E(X_{-i}) \quad \text{and} \quad \Omega\left(|E(X_{-i})|\right) \le |E(X)| \le |E(X_{-i})|\,.$$

1.2.4 STAR: Subsample-Test-Aggregate Regression

We now turn to our instantiation of STA for linear regression. Our learner will be OLS regression itself, and our aggregator will be the mean-plus-Gaussian described in Theorem 1.7 (generalized to allow data-dependent $E(X_S)$ as described above). All that remains is to define our down-sensitivity test $\mathcal{T}$ and its corresponding leave-one-out bound $E(X_S)$. Motivated by the Algorithm for Certifying Robustness Efficiently (ACRE) of Rubinstein & Hopkins (2025), we propose the following test for robustness:

Definition 1.8 ((Simplified) ACRE Robustness)**.** Given a regression problem $X = (X_1, \ldots, X_n) \in \mathbb{R}^{n\times d}$, $y = (Y_1, \ldots, Y_n) \in \mathbb{R}^n$, define its hat matrix to be $H = X\left(X^\intercal X\right)^{-1} X^\intercal \in \mathbb{R}^{n\times n}$. We say that the regression is (L, R, ℓ)-ACRE if:

1. All samples have bounded leverages – i.e., the diagonal elements of the hat matrix satisfy $H_{i,i} \le L$.
2. All samples have bounded residuals. In terms of the hat matrix, this can be written as $\|y - Hy\|_\infty \le R$.
3. All sample pairs have bounded cross-leverages – i.e., the off-diagonal elements of the hat matrix are also bounded $|H_{i,j}| \le \ell$ for all $i \ne j$.

Using leverage and residual scores to test the robustness of a linear regression is a common technique Weisberg (2005), but by also testing for cross-leverages, we can effectively test the robustness of the regression to a larger number of sample removals. This is a very simplified version of the tests performed by the ACRE algorithm. We capture in a later Lemma (Lemma 1.11) that the ACRE properties imply bounded down-sensitivity, and move on now to ACRE's stability.

The key property of ACRE that we will use to ensure stability is that if a set of samples satisfies the ACRE conditions, any leave-one-out satisfies the ACRE condition with slightly worse parameters. This follows almost immediately from the Sherman-Morrison formula.

Lemma 1.9 (Stability of the ACRE Condition)**.** *If X, y is (L, R, ℓ)-ACRE for $\ell \leq L$, then any dataset obtained by dropping a single sample from X, y is $((1+c)L, (1+c)R, (1+c)\ell)$-ACRE, for $c = \frac{\ell}{1-L}$.*

A careful reader might note two remaining discrepancies in our definition of the ACRE test. First, the current test is deterministic, thus it cannot achieve our definition of down-stability. Second, the test has several hyperparameters that need to be set.

Soft-ACRE We use Lemma 1.9 to construct a randomized soft-threshold version of the ACRE test.

Algorithm 3: SOFT-ACRE

Input: Dataset $X, y \in \mathbb{R}^{n \times d} \times \mathbb{R}^n$ and target thresholds $(L, R, \ell) \in \mathbb{R}^3$
Output: ACCEPT or REJECT
Compute the hat matrix $H \leftarrow X(X^\intercal X)^{-1}X^\intercal$ `// if` $X^\intercal X$ `is singular, we reject.`

Find the worst violator of the ACRE conditions

$$\text{SCORE} = \max\left\{\max_{i\in[n]}\left\{\frac{H_{i,i}}{L}\right\}, \max_{i\neq j\in[n]}\left\{\frac{|H_{i,j}|}{\ell}\right\}, \frac{\|y - Hy\|_\infty}{R}\right\}$$

Draw $U \sim \text{Unif}[0, 1]$
if SCORE ≤ 2 ***and*** $U \leq \text{SCORE}^{-\frac{2L}{1-2\ell}}$ **then**
  return *ACCEPT*
else
  return *REJECT*

Lemma 1.10 (SOFT-ACRE is Down-Stable)**.** *The randomized test $\mathcal{T}$ =*SOFT-ACRE *is $(2, \eta)$ down-stable (see Definition A.1) for*

$$\eta = \exp\left(-\ln(2)\left(1 - \frac{2L}{1-2\ell}\right)\frac{1-2L}{2\ell}\right).$$

Finally, we need to show that SOFT-ACRE also implies a bound on the leave-one-out effect of a regression.

Lemma 1.11 (SOFT-ACRE implies bounded down-sensitivity)**.** *If X, y is a regression that is accepted by* SOFT-ACRE *with non-zero probability, then for $E(X) := 8R^2\frac{L}{(1-2L)^2}(X^\intercal X)^{-1}$ and any index $i \in [n]$, we have*

$$\left\|E(X)^{-1/2}\left(\text{OLS}(X, y) - \text{OLS}(X_{\neg i}, y_{\neg i})\right)\right\|_2 \leq 1.$$

Lemma 1.11 follows immediately from the fact that SOFT-ACRE never accepts regressions that are not $(2L, 2R, 2\ell)$-ACRE and the Sherman-Morrison formula.

Selecting the Hyperparameters of ACRE Given a subsample of size $|S|$, it is a well-known analysis that the average leverage scales like

$$\mathop{\mathbb{E}}_{i\in S}\left[X_i^\intercal\left(X_S^\intercal X_S\right)^{-1}X_i\right] = \frac{1}{|S|}\operatorname{tr}[H_S] = \frac{1}{|S|}\operatorname{tr}[I_d] = \frac{d}{|S|}.$$

Therefore, for the STAR algorithm we will set the leverage threshold to $L = \Theta\left(\frac{d}{tn}\right)$. To get the STAR-LITE algorithm, we fix $\ell = L$ (from the Cauchy-Schwarz inequality, this makes the cross-leverage test null, since we always have $|H_{i,j}| \leq \sqrt{H_{i,i}H_{j,j}} \leq L$). However, we can often gain a "blessing-of-dimensionality" effect for the cross-leverages, which allows us to use $\ell = \Theta\left(\frac{\sqrt{d\log(n)}}{tn}\right)$.

Finally, for the residual bound R, if we are told the variance of the underlying sample distribution, we may set $R = \Theta(\sigma\sqrt{\log(n)})$. To select R when σ is not known, we first estimate σ using Algorithm 6 of Brown et al. (2024) (a subsample-and-aggregate algorithm for estimating σ up to a constant factor using $\frac{d\log(1/\delta)}{\varepsilon}$ samples).

1.3 Related Work

Down-Sensitivity, Lipschitz Extensions, and Privacy Down-sensitivity (defined slightly differently than we do here) as a tool for DP emerged in Chen & Zhou (2013), and has seen use especially in private analysis of graphs – see Raskhodnikova & Smith (2016) for a survey. These ideas are extended somewhat beyond graphs, though still in settings where the output is one-dimensional, by Fang et al. (2022). And they are extended to a graph-theoretic statistical inference setting, graphon estimation, in Borgs et al. (2015; 2018). More recent works relating down-sensitivity and privacy in the context of privately releasing real-valued (one dimensional) functions include Linder et al. (2025); Cummings & Durfee (2020); Steinke & Steinke (2025). And in the statistical estimation setting, Ashtiani & Liaw (2022); Tsfadia et al. (2022) also use approaches which combine sensitivty considerations with subsampling.

Reductions from Private Algorithms to Robust Ones in High-Dimensional Statistics A number of recent works offer frameworks to convert learning algorithms satisfying robustness guarantees to differentially private ones. Hopkins et al. (2023); Asi et al. (2023) describe a black-box reduction based on the *inverse sensitivity mechanism* Asi & Duchi (2020a); Asi et al. (2023), which can obtain optimal privacy-accuracy tradeoffs, but the black-box reduction is not computationally efficient, even starting with a computationally efficient robust learning algorithm, and its computationally efficient variant in Hopkins et al. (2023) is non-black-box. These works were preceded by the foundational *high dimensional propose-test-release* framework of Liu et al. (2022), whose reduction is non-black-box and also not computationally efficient. Kothari et al. (2022) gives a framework for converting certain robust learning algorithms based on convex programming to private ones; their framework allows for computationally efficient private algorithms but is non-black-box. So far, none of these approaches offers a combination of computational efficiency and privacy-accuracy tradeoffs which achieves an algorithm with comparable guarantees to ours for linear regression.

Beyond Worst-Case Analysis in DP DP algorithms which add noise scaled to global sensitivity often do achieve optimal privacy-accuracy tradeoffs when accuracy is measured in a worst-case sense Dwork & Roth (2014), and yet the need for DP algorithms which adapt the amount of noise they add to the "shape" or "niceness" of each particular input has long been clear. How, then, do we measure the accuracy or quality of a DP algorithm in a way which differentiates algorithms which add less noise on "nice" inputs? The statistical estimation setting we study here is one way to formulate an average-case accuracy measure for a DP algorithm which distinguishes such algorithms. Other approaches include the *instance optimality* framework Asi & Duchi (2020a;b); Huang et al. (2021); McMillan et al. (2022); Feldman et al. (2024) and *subset-based instance optimality* Dick et al. (2023).

1.4 Organization

In Appendix A, we introduce the subsample-test-aggregate in more detail and prove some of its basic results. Appendix B is devoted to our privacy mechanism based on adding Gaussian noise to an empirical average over subsamples with bounded down-sensitivity. In Appendix C, we give more details on our STAR algorithm and prove Theorem 1.3. Finally, in Appendix D we introduce basic DP primitives used throughout our construction.

ACKNOWLEDGMENTS

We used large language models (LLMs) to assist with writing initial drafts of some of the proofs and for converting our submission to the ICLR DATA-FM format. All LLM-generated text was reviewed and edited by human authors, and the authors take full responsibility for the content and originality of this submission.

We thank Gautam Kamath and Stefan Tiegel for helpful comments and feedback on earlier versions of this paper.

This work was supported by NSF Award No. 2238080 and CSAIL Alliances. Ittai Rubinstein was additionally supported by the MIT EECS MathWorks Fellowship.

REFERENCES

Daniel Alabi, Audra McMillan, Jayshree Sarathy, Adam Smith, and Salil Vadhan. Differentially private simple linear regression. *arXiv preprint arXiv:2007.05157*, 2020.

Prashanti Anderson, Ainesh Bakshi, Mahbod Majid, and Stefan Tiegel. Sample-optimal private regression in polynomial time. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 2341–2349, 2025.

Jamil Arbas, Hassan Ashtiani, and Christopher Liaw. Polynomial time and private learning of unbounded gaussian mixture models. In *International Conference on Machine Learning*, pp. 1018–1040. PMLR, 2023.

Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning gaussians and beyond. In *Conference on Learning Theory*, pp. 1075–1076. PMLR, 2022.

Hilal Asi and John C Duchi. Instance-optimality in differential privacy via approximate inverse sensitivity mechanisms. *Advances in neural information processing systems*, 33:14106–14117, 2020a.

Hilal Asi and John C Duchi. Near instance-optimality in differential privacy. *arXiv preprint arXiv:2005.10630*, 2020b.

Hilal Asi, Jonathan Ullman, and Lydia Zakynthinou. From robustness to privacy and back. In *International Conference on Machine Learning*, pp. 1121–1146. PMLR, 2023.

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*, pp. 102–115, 2021.

Andrés F Barrientos, Aaron R Williams, Joshua Snoke, and Claire McKay Bowen. A feasibility study of differentially private summary statistics and regression analyses with evaluations on administrative and survey data. *Journal of the American Statistical Association*, 119(545):52–65, 2024.

Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. *Advances in Neural Information Processing Systems*, 33:14475–14485, 2020.

Christian Borgs, Jennifer Chayes, and Adam Smith. Private graphon estimation for sparse graphs. *Advances in Neural Information Processing Systems*, 28, 2015.

Christian Borgs, Jennifer Chayes, Adam Smith, and Ilias Zadik. Revealing network structure, confidentially: Improved rates for node-private graphon estimation. In *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, pp. 533–543. IEEE, 2018.

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 B. Hopkins, and Adam D. Smith. Fast, sample-efficient, affine-invariant private mean and covariance estimation for subgaussian distributions. *CoRR*, abs/2301.12250, 2023. doi: 10.48550/ARXIV.2301.12250.

Gavin Brown, Jonathan Hayase, Samuel Hopkins, Weihao Kong, Xiyang Liu, Sewoong Oh, Juan C Perdomo, and Adam Smith. Insufficient statistics perturbation: Stable estimators for private least squares. *arXiv preprint arXiv:2404.15409*, 2024.

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.

Shixi Chen and Shuigeng Zhou. Recursive mechanism: towards node differential privacy and unrestricted joins. In *Proceedings of the 2013 ACM SIGMOD International Conference on Management of Data*, pp. 653–664, 2013.

R Dennis Cook and Sanford Weisberg. Residuals and influence in regression. 1982.

Rachel Cummings and David Durfee. Individual sensitivity preprocessing for data privacy. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 528–547. SIAM, 2020.

Yuval Dagan, Michael Jordan, Xuelin Yang, Lydia Zakynthinou, and Nikita Zhivotovskiy. Dimension-free private mean estimation for anisotropic distributions. *Advances in Neural Information Processing Systems*, 37:120667–120698, 2024.

Travis Dick, Alex Kulesza, Ziteng Sun, and Ananda Theertha Suresh. Subset-based instance optimality in private estimation. In *International Conference on Machine Learning*, pp. 7992–8014. PMLR, 2023.

Cynthia Dwork and Jing Lei. Differential privacy and robust statistics. In *Proceedings of the forty-first annual ACM symposium on Theory of computing*, pp. 371–380, 2009.

Cynthia Dwork and Aaron Roth. The algorithmic foundations of differential privacy. *Foundations and trends® in theoretical computer science*, 9(3-4):211–487, 2014.

Cynthia Dwork, Krishnaram Kenthapadi, Frank McSherry, Ilya Mironov, and Moni Naor. Our data, ourselves: Privacy via distributed noise generation. In *Annual international conference on the theory and applications of cryptographic techniques*, pp. 486–503. Springer, 2006a.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Theory of cryptography conference*, pp. 265–284. Springer, 2006b.

Juanru Fang, Wei Dong, and Ke Yi. Shifted inverse: A general mechanism for monotonic functions under user differential privacy. In *Proceedings of the 2022 ACM SIGSAC Conference on Computer and Communications Security*, pp. 1009–1022, 2022.

Vitaly Feldman, Audra McMillan, Satchit Sivakumar, and Kunal Talwar. Instance-optimal private density estimation in the wasserstein distance. *Advances in Neural Information Processing Systems*, 37:90061–90131, 2024.

Kristian Georgiev and Samuel B. Hopkins. Privacy induces robustness: Information-computation gaps and sparse mean estimation. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/2d76b6a9f96181ab717c1a15ab9302e1-Abstract-Conference.html.

Samuel B. Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. In Stefano Leonardi and Anupam Gupta (eds.), *STOC '22: 54th Annual ACM SIGACT Symposium on Theory of Computing, Rome, Italy, June 20 - 24, 2022*, pp. 1406–1417. ACM, 2022. doi: 10.1145/3519935.3519947.

Samuel B Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pp. 497–506, 2023.

Ziyue Huang, Yuting Liang, and Ke Yi. Instance-optimal mean estimation under differential privacy. *Advances in Neural Information Processing Systems*, 34:25993–26004, 2021.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Conference on Learning Theory*, pp. 1853–1902. PMLR, 2019.

Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory*, pp. 2204–2235. PMLR, 2020.

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

Pravesh Kothari, Pasin Manurangsi, and Ameya Velingker. Private robust estimation by stabilizing convex relaxations. In *Conference on Learning Theory*, pp. 723–777. PMLR, 2022.

Rohith Kuditipudi, John Duchi, and Saminul Haque. A pretty fast algorithm for adaptive private mean estimation. In *The Thirty Sixth Annual Conference on Learning Theory*, pp. 2511–2551. PMLR, 2023.

Omri Lev, Moshe Shenfeld, Vishwak Srinivasan, Katrina Ligett, and Ashia C Wilson. Near-optimal private linear regression via iterative hessian mixing. *arXiv preprint arXiv:2601.07545*, 2026.

Ephraim Linder, Sofya Raskhodnikova, Adam Smith, and Thomas Steinke. Privately evaluating untrusted black-box functions. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pp. 2350–2361, 2025.

Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Conference on Learning Theory*, pp. 1167–1246. PMLR, 2022.

Xiyang Liu, Prateek Jain, Weihao Kong, Sewoong Oh, and Arun Sai Suggala. Near optimal private and robust linear regression. *arXiv preprint arXiv:2301.13273*, 2023.

Audra McMillan, Adam Smith, and Jon Ullman. Instance-optimal differentially private estimation. *arXiv preprint arXiv:2210.15819*, 2022.

Ankur Moitra and Dhruv Rohatgi. Provably auditing ordinary least squares in low dimensions. *arXiv preprint arXiv:2205.14284*, 2022.

Shyam Narayanan, Vahab Mirrokni, and Hossein Esfandiari. Tight and robust private mean estimation with few users. In *International Conference on Machine Learning*, pp. 16383–16412. PMLR, 2022.

Kobbi Nissim, Sofya Raskhodnikova, and Adam Smith. Smooth sensitivity and sampling in private data analysis. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pp. 75–84, 2007.

Sofya Raskhodnikova and Adam Smith. Differentially private analysis of graphs. In *Encyclopedia of Algorithms*, pp. 543–547. Springer, 2016.

Ittai Rubinstein and Samuel B. Hopkins. Robustness auditing for linear regression: To singularity and beyond. In *Proceedings of the Thirteenth International Conference on Learning Representations (ICLR 2025)*, 2025. URL https://openreview.net/forum?id=V5ns6uvRZ9.

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

Günter F Steinke and Thomas Steinke. Privately estimating black-box statistics. *arXiv preprint arXiv:2510.00322*, 2025.

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

Prateek Varshney, Abhradeep Thakurta, and Prateek Jain. (nearly) optimal private linear regression via adaptive clipping. *arXiv preprint arXiv:2207.04686*, 2022.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

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

Sanford Weisberg. *Applied Linear Regression*. Wiley Series in Probability and Statistics. John Wiley & Sons, Hoboken, NJ, 3 edition, 2005. ISBN 9780471663799.

A THE SUBSAMPLE-TEST-AGGREGATE PRIVACY MECHANISM

In this section we formalize the subsample-test-aggregate (STA) mechanism abstractly. We focus on the minimal properties we need from a Test procedure in order to obtain strong privacy guarantees by aggregating over many test-passing subsamples.

A.1 NOTATIONS

We fix notation for the *subsample–test* stage of the STA mechanism. Throughout, we will be careful about the two sources of randomness: the subsampling step and the randomized Test.

Subsampling randomness. Fix a dataset $X = (x_1, \ldots, x_n) \in \mathcal{X}^n$ and a subsampling rate $t \in (0, 1)$. As previously described, STA first draws a **random** subset $S \subseteq [n]$ by *Poisson subsampling* at rate t, which we write as $S \sim \mathsf{PoisSub}_t$. We write $X_S \stackrel{\text{def}}{=} (x_i)_{i \in S}$. When we need to represent a fixed subset, we will frequently use $T \subseteq [n]$.

Test randomness. The Test procedure $\mathcal{T}$ is randomized. Given a subsample S, we run the test on X_S and write its accept/reject bit as

$$b_S(X) \in \{\mathsf{ACCEPT}, \mathsf{REJECT}\}.$$

(When the test produces additional outputs—e.g., a certificate matrix—we will denote them explicitly later; for this notation section we only need the accept/reject bit.)

Acceptance event and pass probability. We define the acceptance event and its marginal probability by

$$\mathsf{Acc}_X \stackrel{\text{def}}{=} \{b_S(X) = \mathsf{ACCEPT}\}, \qquad p_X \stackrel{\text{def}}{=} \Pr[\mathsf{Acc}_X].$$

Unless otherwise specified, all probabilities and expectations involving Acc_X (e.g., $\Pr[\mathsf{Acc}_X]$, $\Pr[\cdot \mid \mathsf{Acc}_X]$, $\mathbb{E} \cdot \mathsf{Acc}_X$) are taken over the *joint* randomness of: (i) $S \sim \mathsf{PoisSub}_t$ and (ii) the internal randomness of $\mathcal{T}$.

When we condition on S (e.g., in $\Pr[\mathsf{Acc}_X \mid S]$), the probability is only over the internal randomness of $\mathcal{T}$.

The post-test subsample distribution. When $p_X > 0$, the *post-test subsample distribution* is simply the conditional distribution of S given acceptance, $S \mid \mathsf{Acc}_X$.

We will commonly switch between conditional and unconditional probabilities, as well as conditional and unconditional expectations, using the following identities from elementary probability.

$$\Pr[S = T \mid \mathsf{Acc}_X] = \frac{\Pr[S = T \wedge \mathsf{Acc}_X]}{p_X}.$$

and for any function $f : 2^{[n]} \to \mathbb{R}$,

$$\mathbb{E}[f(S) \mid \mathsf{Acc}_X] \;=\; \frac{\mathbb{E}\left[f(S) \cdot \mathbb{1}\{\mathsf{Acc}_X\}\right]}{p_X}.$$

General Notation We write $\mathbb{S}_+^d$ for the cone of $d \times d$ symmetric positive semidefinite matrices,

$$\mathbb{S}_+^d \;\stackrel{\text{def}}{=}\; \left\{M \in \mathbb{R}^{d\times d} : M = M^\top \text{ and } M \succeq 0\right\}.$$

A.2 Down-Stable Test

A.2.1 Definition

Definition A.1 (Down-Stable Test)**.** We say that the test $\mathcal{T} : \mathcal{X}^\star \to \{\mathsf{ACCEPT}, \mathsf{REJECT}\}$ is C_p, η*-down-stable*, if for any set of samples $X \in \mathcal{X}^*$, and any leave-one-out $X^{\neg i}$ of X, we have

$$\Pr\left[\mathcal{T}(X) = \mathsf{ACCEPT}\right] \le C_p \Pr\left[\mathcal{T}(X^{\neg i}) = \mathsf{ACCEPT}\right] + \eta\,.$$

Even before tying $\mathcal{T}$ to a learning algorithm, we can derive some properties just from the fact that it is down-stable.

A.2.2 Key Property: No single sample can be too over-represented by conditioning on a down-stable test

Theorem A.2 (No index is over-represented)**.** *If $\mathcal{T}$ is a (C_p, η)-down-stable test, $i \in [n]$ is any index and $T \subseteq [n]$ is any subset of the training set containing $i \in T$, then*

$$\Pr[S = T \mid \mathsf{Acc}_X] \;\le\; \frac{C_p\, t}{1-t} \cdot \left(\Pr[S = T \setminus \{i\} \mid \mathsf{Acc}_X] \;+\; \frac{\eta}{C_p p_X} \cdot \Pr[S = T \setminus \{i\}]\right),$$

where all probabilities are taken over the subsampling (and, for the event Acc_X, also over the randomness of $\mathcal{T}$).

Remark A.3 (Interpretation of Theorem A.2). Theorem A.2 says that conditioning on acceptance cannot dramatically increase the relative likelihood of subsamples that contain any fixed point i. Up to the expected inclusion rate (the factor $\frac{t}{1-t}$) and a constant factor C_p, sets containing i behave similarly to the corresponding sets without i, plus an additive slack controlled by η/p. In applications we will choose parameters so that $\eta \ll p$, making the slack term negligible.

Theorem A.2 is central to the STA framework. First, we can use it to certify that the acceptance rate of $\mathcal{T}$ on random subsamples is stable (Corollary A.4), which in turn allows us to privately check that this acceptance probability is not too low (Theorem A.5). Privately checking that the acceptance rate p_X is not too low is crucial, since otherwise our STA algorithm could have an unbounded runtime or condition on very low probability events.

This no over-representation property is also what powers our privacy guarantee for the single-draw aggregation mechanism (Lemma B.1), and we re-use the same proof technique to show that mean-aggregation STA is stable (Theorem A.8).

Proof of Theorem A.2

Proof of Theorem A.2. Fix $i \in [n]$ and $T \ni i$, and let $R \stackrel{\text{def}}{=} T \setminus \{i\}$. Write

$$q_T \stackrel{\text{def}}{=} \Pr\left[\mathcal{T}(X_T) = \mathsf{ACCEPT}\right] \qquad \text{and} \qquad q_R \stackrel{\text{def}}{=} \Pr\left[\mathcal{T}(X_R) = \mathsf{ACCEPT}\right],$$

where the probability is over the internal randomness of $\mathcal{T}$.

By our down-stability assumption (Definition A.1) applied to (X_T, X_R) which differ by one deletion, we have

$$q_T \;\le\; C_p\, q_R + \eta.$$

Multiplying by $\Pr[T]$ and using $\Pr[T \wedge \mathsf{Acc}_X] = \Pr[T] \cdot q_T$, we get

$$\Pr[T \wedge \mathsf{Acc}_X] \;\le\; C_p \, \Pr[T] \cdot q_R \;+\; \eta \, \Pr[T].$$

Under Poisson subsampling with rate t, inclusion of i is independent of the other indices, so for $T \ni i$ we have

$$\Pr[T] \;=\; \frac{t}{1-t} \, \Pr[R].$$

Substituting this identity and using $\Pr[R \wedge \mathsf{Acc}_X] = \Pr[R] \cdot q_R$, we obtain

$$\Pr[T \wedge \mathsf{Acc}_X] \;\le\; \frac{tC_p}{1-t} \, \Pr[R \wedge \mathsf{Acc}_X] \;+\; \frac{t\eta}{1-t} \, \Pr[R].$$

Finally, dividing by $p_X = \Pr[\mathsf{Acc}_X]$ gives

$$\Pr[T \mid \mathsf{Acc}_X] \;=\; \frac{\Pr[T \wedge \mathsf{Acc}_X]}{p_X} \;\le\; \frac{tC_p}{1-t} \, \Pr[R \mid \mathsf{Acc}_X] \;+\; \frac{t\eta}{(1-t)\, p_X} \, \Pr[R],$$

which is the desired inequality. □

Corollary: The acceptance rate of $\mathcal{T}$ is stable

Corollary A.4 (Stability of STA Acceptance Probability)**.** *Assume $\mathcal{T}$ is a C_p, η-down-stable test, run under independent Poisson subsampling at rate*

$$t \in \Big(0,\; \frac{1}{5(1+C_p)}\Big),$$

and let $X, X' \in \mathcal{X}^n$ be neighboring datasets that differ in exactly one coordinate.

Then, if

$$p_X \;\ge\; 10\eta,$$

we have

$$(1 - 3C_p t)\, p_{X'} \le p_X \;\le\; (1 + 3C_p t)\, p_{X'}. \tag{1}$$

A.2.3 Privately Testing p_X

The goal of this section is to develop a differentially private check that the probability of a subsample passing the test, p_X, is not too small. This is necessary because many of our stability results like Corollary A.4 require this passing probability to be much larger than the η error terms. Furthermore, having too small of a p_X would increase the number of subsamples we have to draw in the Monte Carlo estimates, increasing runtime.

We provide an algorithm for privately testing that p_X is sufficiently large, and prove a theorem about its guarantees. This algorithm calls PrivateBernoulliThreshold (Algorithm 7) as a subroutine.

Algorithm 4: PrivatePCheck$_{\varepsilon,\delta,\gamma,p_{\mathrm{acc}}}(X)$

Input: Dataset X; subsampling rate t (used by $\mathcal{T}$); test $\mathcal{T}$ with stability parameter C_p; privacy parameters $(\varepsilon, \delta) \in (0,1)^2$; slack $\gamma \in (0, 1/2)$; threshold $p_{\mathrm{acc}} \in (0,1)$; number of runs $N \in \mathbb{N}$.

Output: ACCEPT or REJECT.

$\tau \leftarrow -\log(1 - 3C_p t)$; `// well-defined since` $3C_p t < 1$ `under Cor. A.4`

$\varepsilon_0 \leftarrow \varepsilon/3$;

$\delta_0 \leftarrow \delta/12$;

$p_{\mathrm{th}} \leftarrow p_{\mathrm{acc}}\, e^{-\gamma}$;

for $j = 1$ **to** N **do**

  Draw S_j by independent Poisson subsampling at rate t;

  Run $\mathcal{T}$ on X_{S_j} and set $b_j \leftarrow \mathbb{1}\{\mathcal{T}(X_{S_j}) = \mathsf{ACCEPT}\}$;

$y \leftarrow \mathrm{PBT}_{\tau,\varepsilon_0,\delta_0}(b_1, \ldots, b_N;\; p_{\mathrm{th}})$; `// Algorithm 7`

return *ACCEPT if y = REJECT, else REJECT*;

Theorem A.5 (PRIVATEPCHECK: privacy, soundness, and completeness)**.** *Assume $\mathcal{T}$ is a C_p, η-down-stable test run under independent Poisson subsampling at rate t satisfying the conditions of Corollary A.4. For each dataset X, define the acceptance probability*

$$p_X \;:=\; \Pr[\mathcal{T}(X_S) = \textsf{ACCEPT}],$$

where the probability is over the Poisson subsample S and the internal randomness of $\mathcal{T}$.

Fix $\varepsilon \in (0,1)$, $\delta \in (0, 1/10)$, a slack parameter $\gamma \in (0, 1/2)$, and an acceptance threshold $p_{\mathrm{acc}} \in (0,1)$. Define

$$\tau \;:=\; -\log(1 - 3C_p t), \qquad \textit{and assume} \qquad \gamma \le \tau/2.$$

Define the low/high acceptance thresholds

$$p_{\mathrm{high}} \;:=\; p_{\mathrm{acc}}, \qquad p_{\mathrm{low}} \;:=\; p_{\mathrm{acc}} \cdot \exp\left(-\frac{6\tau}{\varepsilon}\log\frac{24}{\delta}\right) \cdot e^{-2\gamma}.$$

Assume also $\eta \le p_{\mathrm{low}}/20$ and

$$N \;\ge\; \frac{12}{\gamma^2\, p_{\mathrm{low}}} \cdot \log\frac{48}{\delta}. \tag{2}$$

Run Algorithm 4 with parameters $(\varepsilon, \delta, \gamma, p_{\mathrm{acc}}, N)$. Then it satisfies:

- ***Completeness.*** *If $p_X \ge p_{\mathrm{high}}$ then $\Pr[\textsc{PrivatePCheck}(X) = \textsf{ACCEPT}] \ge 1 - \delta$.*
- ***Soundness.*** *If $p_X \le p_{\mathrm{low}}$ then $\Pr[\textsc{PrivatePCheck}(X) = \textsf{REJECT}] \ge 1 - \delta$.*
- ***Privacy.*** PRIVATEPCHECK *is (ε, δ)-differentially private.*

Proof Sketch of Theorem A.5. From corollary A.4, we know that $z = \log(p_X)$ is stable. Moreover, standard concentration bounds can be used to show that when $p_X > p_{\mathrm{th}}$ is not too small, the empirical estimate p_X^* of p_X from sufficiently many samples ($N \gtrsim 1/p_{\mathrm{th}}$) suffices to estimate $z \approx z^* := \log(1/p_X^*)$.

Therefore, so long p_X is not too small, adding Laplace noise suffices to privatize z^*, and by post-processing, we may also reveal

$$\textsc{Decision} = \mathbb{1}\left\{z^* + \textsc{Laplace} > \textsc{Threshold}\right\}$$

for any fixed THRESHOLD.

Finally, to deal with cases where p_X is very small, we simply note that these are deep into the REJECT range for z, so even a bad approximation of p_X will suffice for them.

□

For a more detailed proof of Theorem A.5, see Section A.5.

A.3 MEAN-AGGREGATION STA

Once we have a framework for privately drawing test-passing subsamples of our training set, we want to start relating this test to the down-sensitivity of a learning algorithm, which will require some new definitions.

Definition A.6 (Down-Sensitivity Bound)**.** Let $A : \mathcal{X}^\star \to \mathbb{R}^d$ be a deterministic learning algorithm that maps a dataset $X \in \mathcal{X}^\star$ to a model $A(X) \in \mathbb{R}^d$. We say that the mapping $E : \mathcal{X}^\star \to \mathbb{S}_+^d$ is a *valid down-sensitivity bound* corresponding to the test $\mathcal{T} : \mathcal{X}^\star \to \{\textsf{ACCEPT}, \textsf{REJECT}\}$, if for any dataset $X = (X_1, \ldots, X_n)$ and for any $X^{\neg i}$ obtained from X by dropping a single sample i,

$$\Pr\left[\mathcal{T}(X) = \textsf{ACCEPT}\right] > 0 \Longrightarrow \forall i \in [n]\; \Delta_i \Delta_i^\intercal \preceq E(X) \quad \text{where} \quad \Delta_i = A(X) - A(X^{\neg i})\,.$$

We further say that E is C_E- stable, for $C_E > 0$, if

$$\Pr\left[\mathcal{T}(X) = \textsf{ACCEPT}\right] > 0 \;\Longrightarrow\; E(X) \;\preceq\; C_E\, E(X^{\neg i})$$

Definition A.7 (Envelope Condition)**.** Let $A : \mathcal{X}^\star \to \mathbb{R}^d$ be a deterministic learning algorithm, and let $\mathcal{T} : \mathcal{X}^\star \to \{\mathsf{ACCEPT}, \mathsf{REJECT}\}$ be a randomized test, with down-sensitivity bound $E : \mathcal{X}^\star \to \mathbb{S}^d_+$.

Fix PSD matrices $\mathcal{E}, \mathcal{V} \in \mathbb{S}^d_+$. We say that $\mathcal{T}$ implies a $\mathcal{E}, \mathcal{V}$ envelope on E, if

$$\Pr\left[\mathcal{T}(X) = \mathsf{ACCEPT}\right] > 0 \implies E(X) \preceq \mathcal{E}, \quad A(X)A(X)^\intercal \preceq \mathcal{V}$$

Theorem A.8 (Stability of Mean-Aggregation STA)**.** *Assume $\mathcal{T}$ is a C_p, η-down-stable test for A, run under independent Poisson subsampling at rate*

$$t \leq \frac{1}{c_0\big(1 + C_p(2 + 3C_E)\big)}$$

for a sufficiently large universal constant c_0. Assume $\mathcal{T}$ has corresponding C_E-stable down-sensitivity bound E, and that $\eta = 0$ or $\mathcal{T}$ implies a $\mathcal{E}, \mathcal{V}$ envelope on E.

Let $X, X' \in \mathcal{X}^n$ be datasets that differ in exactly one coordinate, such that $p_X \stackrel{\text{def}}{=} \Pr\left[\mathsf{Acc}_X\right] > 10\eta$. For a (deterministic) learning algorithm $A : \mathcal{X}^\star \to \mathbb{R}^d$, define the subsample output,

$$v_S(X) \stackrel{\text{def}}{=} A(X_S).$$

the conditional mean

$$\mu(X) \stackrel{\text{def}}{=} \mathbb{E}\left[v_S(X) | \mathsf{Acc}_X\right]$$

and the proxy matrix (using the returned certificate $E_S(X)$ from running the test $\mathcal{T}$ on subset S)

$$M(X) \stackrel{\text{def}}{=} \mathrm{Cov}\left(v_S(X)\right) + \mathbb{E}\left[E_S(X) | \mathsf{Acc}_X\right] \tag{3}$$

$$= \mathbb{E}\left[\left(v_S(X) - \mu(X)\right)\left(v_S(X) - \mu(X)\right)^\top + E_S(X) \middle| \mathsf{Acc}_X\right] \tag{4}$$

Define $\Delta \stackrel{\text{def}}{=} \mu(X) - \mu(X')$. Then there are universal constants $c_1, \ldots, c_6$ such that:

1. ***Conditional mean stability.***

$$\Delta\Delta^\top \preceq c_1\big(C_p(1 + C_E)t\big)^2 \cdot M(X) + c_2 \frac{C_p\, t^2}{p_X} \eta\, \mathcal{E} + c_3 \frac{C_p\, t^2}{p_X} \eta\, \mathcal{V}.$$

2. ***Proxy stability (multiplicative up to additive error).***

$$M(X) \preceq \left(1 + c_4\left((C_p + C_p C_E)t\right)\right) M(X') + \frac{c_5 t}{p_X}\eta\mathcal{E} + \frac{c_6 t}{p_X}\eta\mathcal{V}\,,$$

and

$$M(X) \succeq \left(1 - c_4\left((C_p + C_p C_E)t\right)\right) M(X') - \frac{c_5 t}{p_X}\eta\mathcal{E} - \frac{c_6 t}{p_X}\eta\mathcal{V}\,.$$

A.4 Proof of Corollary A.4

Proof of Corollary A.4. Let i be the (unique) index on which X and X' differ. If $i \notin S$ then $X_S = X'_S$, so $\mathcal{T}$ sees identical input and hence

$$\Pr[\mathsf{Acc}_X \wedge (i \notin S)] = \Pr[\mathsf{Acc}_{X'} \wedge (i \notin S)].$$

Therefore the entire difference comes from subsamples containing i:

$$|p_X - p_{X'}| = \Big|\Pr[\mathsf{Acc}_X \wedge (i \in S)] - \Pr[\mathsf{Acc}_{X'} \wedge (i \in S)]\Big| \leq \Pr[\mathsf{Acc}_X \wedge (i \in S)] + \Pr[\mathsf{Acc}_{X'} \wedge (i \in S)].$$

Writing

$$\delta_i(X) \stackrel{\text{def}}{=} \Pr[i \in S \mid \mathsf{Acc}_X], \qquad \delta_i(X') \stackrel{\text{def}}{=} \Pr[i \in S \mid \mathsf{Acc}_{X'}],$$

this becomes

$$|p_X - p_{X'}| \leq p_X\, \delta_i(X) + p_{X'}\, \delta_i(X').$$

We now bound $\delta_i(X)$ via Theorem A.2. Let

$$\alpha \stackrel{\text{def}}{=} \frac{C_p t}{1-t}.$$

For every $T \subseteq [n]$ with $i \in T$, Theorem A.2 gives

$$\Pr[S = T \mid \mathsf{Acc}_X] \le \alpha \left(\Pr[S = T \setminus \{i\} \mid \mathsf{Acc}_X] + \frac{\eta}{C_p p_X} \Pr[S = T \setminus \{i\}] \right).$$

Summing over all $T \ni i$ yields

$$\delta_i(X) = \sum_{T \ni i} \Pr[S = T \mid \mathsf{Acc}_X] \le \alpha \sum_{T \ni i} \Pr[S = T \setminus \{i\} \mid \mathsf{Acc}_X] + \alpha \cdot \frac{\eta}{C_p p_X} \sum_{T \ni i} \Pr[S = T \setminus \{i\}].$$

The change of variables $U = T \setminus \{i\}$ gives

$$\sum_{T \ni i} \Pr[S = T \setminus \{i\} \mid \mathsf{Acc}_X] = \Pr[i \notin S \mid \mathsf{Acc}_X] = 1 - \delta_i(X),$$

and under Poisson subsampling,

$$\sum_{T \ni i} \Pr[S = T \setminus \{i\}] = \Pr[i \notin S] = 1 - t.$$

Hence

$$\delta_i(X) \le \alpha \left(1 - \delta_i(X)\right) + t \cdot \frac{\eta}{p_X}.$$

Rearranging,

$$(1+\alpha)\, \delta_i(X) \le \alpha + t \cdot \frac{\eta}{p_X} \qquad \Longrightarrow \qquad \delta_i(X) \le \frac{\alpha + t \cdot \eta / p_X}{1+\alpha}.$$

The same argument gives

$$\delta_i(X') \le \frac{\alpha + t \cdot \eta / p_{X'}}{1+\alpha}.$$

Plugging into $|p_X - p_{X'}| \le p_X \delta_i(X) + p_{X'} \delta_i(X')$,

$$|p_X - p_{X'}| \le \frac{\alpha p_X + t\eta}{1+\alpha} + \frac{\alpha p_{X'} + t\eta}{1+\alpha} = \frac{\alpha(p_X + p_{X'}) + 2t\eta}{1+\alpha}.$$

Let $\beta \stackrel{\text{def}}{=} \frac{\alpha}{1+\alpha} = \frac{C_p t}{1+(C_p-1)t}$, so

$$|p_X - p_{X'}| \le \beta(p_X + p_{X'}) + \frac{2t\eta}{1+\alpha} \le \beta(p_X + p_{X'}) + 2t\eta.$$

Using the assumption $p_X \ge 10\eta$, we have $2t\eta \le \frac{t}{5} p_X$, and thus

$$|p_X - p_{X'}| \le \beta(p_X + p_{X'}) + \frac{t}{5} p_X.$$

This implies both

$$p_X \le p_{X'} + \beta(p_X + p_{X'}) + \frac{t}{5} p_X \qquad \Longrightarrow \qquad p_X \left(1 - \beta - \frac{t}{5}\right) \le (1+\beta) p_{X'},$$

and

$$p_X \ge p_{X'} - \beta(p_X + p_{X'}) - \frac{t}{5} p_X \qquad \Longrightarrow \qquad p_X \left(1 + \beta + \frac{t}{5}\right) \ge (1-\beta) p_{X'}.$$

Therefore

$$\frac{1-\beta}{1+\beta+\frac{t}{5}}\, p_{X'} \;\le\; p_X \;\le\; \frac{1+\beta}{1-\beta-\frac{t}{5}}\, p_{X'}.$$

Finally, since $t \le \frac{1}{5(1+C_p)}$, we have $C_p t \le \frac{1}{5}$ and also $\beta \le C_p t$ (and $t/5 \le (C_p t)/5$). Writing $u \stackrel{\text{def}}{=} C_p t$, we get

$$\frac{1+\beta}{1-\beta-\frac{t}{5}} = 1 + \frac{2\beta+\frac{t}{5}}{1-\beta-\frac{t}{5}} \le 1 + \frac{\frac{11}{5}u}{1-\frac{6}{5}u} \le 1 + \frac{\frac{11}{5}u}{\frac{19}{25}} < 1 + 3u = 1 + 3C_p t,$$

and similarly,

$$\frac{1-\beta}{1+\beta+\frac{t}{5}} = 1 - \frac{2\beta+\frac{t}{5}}{1+\beta+\frac{t}{5}} \ge 1 - \left(2\beta + \frac{t}{5}\right) \ge 1 - \frac{11}{5}u \ge 1 - 3u = 1 - 3C_p t.$$

Plugging these into the previous display yields (1). □

A.5 Proof of Theorem A.5

Proof. Fix a dataset X. In Algorithm 4, each run draws an independent Poisson(t) subsample S_j and runs $\mathcal{T}$ with fresh internal randomness. Therefore the bits

$$b_j \;:=\; \mathbf{1}\{\mathcal{T}(X_{S_j}) = \mathsf{ACCEPT}\} \qquad (j = 1, \dots, N)$$

are i.i.d. $\mathrm{Ber}(p_X)$, where $p_X = \Pr[\mathcal{T}(X_S) = \mathsf{ACCEPT}]$. Let y denote the output of the call to PBT inside Algorithm 4, and recall that PrivatePCheck(X) is the deterministic post-processing that flips y.

Internal parameters used by PBT. By inspection of Algorithm 4, the call to PBT uses

$$\kappa = \tau, \qquad \varepsilon_0 = \varepsilon/3, \qquad \delta_0 = \delta/12, \qquad p_{\mathrm{th}} = p_{\mathrm{acc}} e^{-\gamma}$$

Define

$$\beta_0 \;:=\; \delta/12, \qquad C_0 \;:=\; \frac{2\log(2/\delta_0)}{\varepsilon_0} \;=\; \frac{6}{\varepsilon}\log\frac{24}{\delta}.$$

Then the thresholds appearing in Theorem D.6 (applied to $\mathrm{Ber}(p_X)$) are

$$p_{\mathrm{high}}^{\mathrm{PBT}} = p_{\mathrm{th}} e^{\gamma} = p_{\mathrm{acc}} = p_{\mathrm{high}}, \qquad p_{\mathrm{low}}^{\mathrm{PBT}} = p_{\mathrm{th}}\, e^{-\kappa C_0} e^{-\gamma} = p_{\mathrm{acc}}\, e^{-\tau C_0} e^{-2\gamma} = p_{\mathrm{low}}.$$

Moreover, the sample size condition in Theorem D.6 with failure probability β_0 is exactly (2), since $\log(4/\beta_0) = \log(48/\delta)$.

Completeness and soundness. Apply Theorem D.6 to the i.i.d. bits $b_1, \dots, b_N \sim \mathrm{Ber}(p_X)$ with parameters $(\kappa, \varepsilon_0, \delta_0, p_{\mathrm{th}}, \gamma, \beta_0)$. If $p_X \ge p_{\mathrm{high}}$ then Theorem D.6(2) implies $y = \mathsf{REJECT}$ with probability at least $1-\beta_0$, hence PrivatePCheck(X) = ACCEPT with probability at least $1-\beta_0$. If $p_X \le p_{\mathrm{low}}$ then Theorem D.6(1) implies $y = \mathsf{ACCEPT}$ with probability at least $1-\beta_0$, hence PrivatePCheck(X) = REJECT with probability at least $1-\beta_0$. Since $\beta_0 = \delta/12 \le \delta$, both statements hold with probability at least $1-\delta$ as claimed.

Privacy. Fix neighboring datasets $X \sim X'$ and write $p := p_X$ and $p' := p_{X'}$. Let $\mathcal{D}_p$ and $\mathcal{D}_{p'}$ denote the output distributions of y under $\mathrm{Ber}(p)^{\otimes N}$ and $\mathrm{Ber}(p')^{\otimes N}$, respectively. Since PrivatePCheck is deterministic post-processing of y, it suffices to show $\mathcal{D}_p \approx_{\varepsilon,\delta} \mathcal{D}_{p'}$.

We consider two cases.

Case 1: $\min\{p, p'\} \ge p_{\mathrm{low}}/2$. By $\eta \le p_{\mathrm{low}}/20$, this implies $\min\{p, p'\} \ge 10\eta$, so Corollary A.4 applies (to both orientations) and yields

$$(1 - 3C_p t)\, p' \le p \le (1 + 3C_p t)\, p'.$$

Let $r := 1 - 3C_p t$, so $\tau = -\log r$ and $e^{-\tau} = r$. From the above equation,

$$p' \le \frac{p}{r} = e^{\tau} p, \qquad p' \ge \frac{p}{1+3C_p t} \ge (1 - 3C_p t)p = rp = e^{-\tau} p,$$

where we used $\frac{1}{1+a} \geq 1-a$ for $a \geq 0$. Hence $e^{-\kappa}p \leq p' \leq e^{\kappa}p$ with $\kappa = \tau$, and also $p, p' \geq p_{\text{low}}/2$. Therefore Theorem D.6(3) applies to $\mathcal{D}_p$ and $\mathcal{D}_{p'}$ and gives

$$\mathcal{D}_p \approx_{3\varepsilon_0,\, \delta_{\text{priv}}} \mathcal{D}_{p'}, \qquad \delta_{\text{priv}} = (1 + e^{\varepsilon_0} + e^{2\varepsilon_0})\delta_0 + (1 + e^{2\varepsilon_0})\beta_0.$$

Since $\varepsilon \in (0, 1)$ we have $e^{\varepsilon_0} = e^{\varepsilon/3} < 2$ and $e^{2\varepsilon_0} < 2$, so

$$\delta_{\text{priv}} \leq (1 + 2 + 2)\frac{\delta}{12} + (1 + 2)\frac{\delta}{12} = \frac{8}{12}\delta \leq \delta.$$

Also $3\varepsilon_0 = \varepsilon$. Thus $\mathcal{D}_p \approx_{\varepsilon,\delta} \mathcal{D}_{p'}$.

Case 2: $\min\{p, p'\} < p_{\text{low}}/2$. Without loss of generality assume $p < p_{\text{low}}/2$. We claim that then $p' \leq p_{\text{low}}$.

If $p' < 10\eta$, then $p' \leq 10\eta \leq p_{\text{low}}/2 < p_{\text{low}}$. Otherwise $p' \geq 10\eta$, so Corollary A.4 applies to X' (with roles swapped) and yields $p' \leq (1 + 3C_p t)\, p$. Since t satisfies the conditions of Corollary A.4, we have $3C_p t < 1$; thus using $p < p_{\text{low}}/2$ gives $p' < p_{\text{low}}$. Hence $p, p' \leq p_{\text{low}}$.

Now apply Theorem D.6(1): under both p and p', PBT outputs ACCEPT with probability at least $1 - \beta_0$. Equivalently,

$$\mathcal{D}_p(\text{REJECT}) \leq \beta_0, \qquad \mathcal{D}_{p'}(\text{REJECT}) \leq \beta_0.$$

Since the output space is $\{\text{ACCEPT}, \text{REJECT}\}$, for every event $\mathcal{S} \subseteq \{\text{ACCEPT}, \text{REJECT}\}$ we have

$$\left|\mathcal{D}_p(\mathcal{S}) - \mathcal{D}_{p'}(\mathcal{S})\right| \leq \beta_0 \leq \delta.$$

In particular, for every $\mathcal{S}$, $\mathcal{D}_p(\mathcal{S}) \leq \mathcal{D}_{p'}(\mathcal{S}) + \delta \leq e^{\varepsilon}\mathcal{D}_{p'}(\mathcal{S}) + \delta$, and symmetrically with p and p' swapped. Hence $\mathcal{D}_p \approx_{\varepsilon,\delta} \mathcal{D}_{p'}$ in this case as well.

Combining the two cases proves $\mathcal{D}_p \approx_{\varepsilon,\delta} \mathcal{D}_{p'}$ for all neighboring $X \sim X'$. Therefore the internal output y is (ε, δ)-DP, and since PRIVATEPCHECK is a deterministic post-processing of y, it is also (ε, δ)-DP. □

A.6 PROOF OF THEOREM A.8

Our privacy analysis studies distributions and expectations *conditioned on* Acc, and the following lemmas show that this conditioning does not allow any single example to have outsized influence.

Lemma A.9 (No single sample dominates the expected certificate)**.** *Assume $\mathcal{T}$ is a C_p, η-down-stable test for A, with corresponding C_E-stable down-sensitivity bound E, and that $\mathcal{T}$ implies a $\mathcal{E}, \mathcal{V}$ envelope on E. Then for every $i \in [n]$,*

$$\mathbb{E}\left[E_S \cdot \mathbb{1}\{i \in S\} | \text{Acc}_X\right] \;\preceq\; \frac{tC_pC_E}{1-t} \cdot \mathbb{E}\left[E_S | \text{Acc}_X\right] \;+\; t \cdot \frac{\eta\mathcal{E}}{p_X}.$$

Remark A.10 (Interpretation of Lemma A.9). Lemma A.9 formalizes the intuition that, provided the marginal acceptance probability p is not too small, conditioning on Acc cannot cause any single point to dominate the typical sensitivity certificate E_S. The second term is an additive error that becomes negligible when η/p is small (since $\mathcal{E}$ upper bounds E_S on acceptance).

Proof of Lemma A.9. Fix $i \in [n]$. Expanding the conditional expectation,

$$\mathbb{E}\left[E_S \cdot \mathbb{1}\{i \in S\} | \text{Acc}_X\right] \;=\; \sum_{S \ni i} \Pr[S \mid \text{Acc}_X]\, E_S.$$

Apply Theorem A.2 (with $R = S \setminus \{i\}$) to each summand:

$$\sum_{S \ni i} \Pr[S \mid \text{Acc}]\, E_S \;\preceq\; \frac{tC_p}{1-t} \sum_{S \ni i} \Pr[R \mid \text{Acc}_X]\, E_S \;+\; \frac{t\eta}{(1-t)p_X} \sum_{S \ni i} \Pr[R]\, E_S.$$

For the first term, note that under the conditioning on Acc, every set S in the sum satisfies $b(X_S) = \text{ACCEPT}$, and hence C_E stability (Definition A.6) gives

$$E_S \;\preceq\; C_E\, E_{S \setminus \{i\}} \;=\; C_E\, E_R.$$

Therefore,

$$\frac{tC_p}{1-t}\sum_{S\ni i}\Pr[R \mid \mathsf{Acc}_X]\,E_S \;\preceq\; \frac{tC_pC_E}{1-t}\sum_{S\ni i}\Pr[R \mid \mathsf{Acc}_X]\,E_R.$$

Changing variables from $S \ni i$ to $R = S \setminus \{i\}$ (which ranges over sets $R \subseteq [n] \setminus \{i\}$), the term on the right above becomes

$$\sum_{S\ni i}\Pr[R \mid \mathsf{Acc}_X]\,E_R \;=\; \sum_{R:\, i\notin R}\Pr[R \mid \mathsf{Acc}_X]\,E_R \;\preceq\; \sum_{R}\Pr[R \mid \mathsf{Acc}_X]\,E_R \;=\; \mathbb{E}\,[E_S|\mathsf{Acc}_X],$$

since all matrices are PSD.

For the second term, again every S in the sum satisfies $b(X_S) = \mathsf{ACCEPT}$, so the envelope condition (Definition A.7) implies $E_S \preceq \mathcal{E}$. Hence

$$\frac{t\eta}{(1-t)p_X}\sum_{S\ni i}\Pr[R]\,E_S \;\preceq\; \frac{t\eta}{(1-t)p_X}\sum_{S\ni i}\Pr[R]\cdot\mathcal{E} \;=\; \frac{t\eta}{(1-t)p_X}\Big(\sum_{S\ni i}\Pr[S\setminus\{i\}]\Big)\mathcal{E}.$$

Finally, $\sum_{S\ni i}\Pr[S\setminus\{i\}] = \Pr[i \notin S] = 1-t$ under Poisson subsampling, so this term equals $\frac{t\eta}{p_X}\mathcal{E}$.

Combining the bounds proves the lemma. □

A.6.1 Auxiliary Inequalities

Lemma A.11 (Bounded drop-set second moment implies bounded mean shift)**.** *Let* $\mathbf{v}$ *be an* $\mathbb{R}^d$*-valued random vector and let* $\Sigma \in \mathbb{S}^d_+$ *be any PSD matrix. Let* D *be an event with* $\Pr[\mathsf{D}] \geq 1-\delta$*. Suppose there is a parameter* $\gamma \geq 0$ *such that*

$$\mathbb{E}\left[\mathbb{1}\{\mathsf{D}^c\}\,\mathbf{v}\mathbf{v}^\top\right] \;\preceq\; \gamma\,\Sigma.$$

Define the (truncation) mean shift

$$\mathbf{m} \;\stackrel{\text{def}}{=}\; \mathbb{E}\,[\mathbf{v}] - \mathbb{E}\,[\mathbb{1}\{\mathsf{D}\}\mathbf{v}] \;=\; \mathbb{E}\,[\mathbb{1}\{\mathsf{D}^c\}\mathbf{v}].$$

Then

$$\mathbf{m}\mathbf{m}^\top \;\preceq\; \gamma\,\delta\,\Sigma.$$

In particular, if $\Sigma \succ 0$*, then* $\|\mathbf{m}\|_{\Sigma^{-1}} \leq \sqrt{\gamma\delta}$*.*

Proof. Fix any $\mathbf{u} \in \mathbb{R}^d$. Since $\mathbf{m} = \mathbb{E}[\mathbb{1}\{\mathsf{D}^c\}\mathbf{v}]$, we have

$$\langle \mathbf{u}, \mathbf{m}\rangle \;=\; \mathbb{E}\,[\mathbb{1}\{\mathsf{D}^c\}\,\langle \mathbf{u}, \mathbf{v}\rangle].$$

By Cauchy–Schwarz,

$$\langle \mathbf{u}, \mathbf{m}\rangle^2 \;\leq\; \mathbb{E}\,[\mathbb{1}\{\mathsf{D}^c\}]\cdot\mathbb{E}\left[\mathbb{1}\{\mathsf{D}^c\}\,\langle \mathbf{u}, \mathbf{v}\rangle^2\right] \;\leq\; \delta\cdot\mathbf{u}^\top\mathbb{E}\left[\mathbb{1}\{\mathsf{D}^c\}\mathbf{v}\mathbf{v}^\top\right]\mathbf{u} \;\leq\; \gamma\delta\cdot\mathbf{u}^\top\Sigma\mathbf{u}.$$

Equivalently,

$$\mathbf{u}^\top(\mathbf{m}\mathbf{m}^\top)\mathbf{u} \;\leq\; \gamma\delta\cdot\mathbf{u}^\top\Sigma\mathbf{u} \qquad \text{for all } \mathbf{u} \in \mathbb{R}^d.$$

This implies $\mathbf{m}\mathbf{m}^\top \preceq \gamma\delta\,\Sigma$.

If $\Sigma \succ 0$, taking the supremum of $\langle \mathbf{u}, \mathbf{m}\rangle^2 \leq \gamma\delta\;\|\mathbf{u}\|^2_\Sigma$ over $\mathbf{u} \neq 0$ yields $\|\mathbf{m}\|^2_{\Sigma^{-1}} \leq \gamma\delta$. □

Lemma A.12 (A useful PSD inequality)**.** *For any* $\mathbf{a}, \mathbf{b} \in \mathbb{R}^d$*,*

$$(\mathbf{a}+\mathbf{b})(\mathbf{a}+\mathbf{b})^\top \;\preceq\; 2\,\mathbf{a}\mathbf{a}^\top + 2\,\mathbf{b}\mathbf{b}^\top.$$

Proof. Rearranging,

$$2\,\mathbf{a}\mathbf{a}^\top + 2\,\mathbf{b}\mathbf{b}^\top - (\mathbf{a}+\mathbf{b})(\mathbf{a}+\mathbf{b})^\top \;=\; (\mathbf{a}-\mathbf{b})(\mathbf{a}-\mathbf{b})^\top \;\succeq\; 0,$$

which proves the claim. □

Corollary A.13 (Propagating a leave-one-out certificate to a second-moment bound)**.** *Let $\mathbf{o} \in \mathbb{R}^d$ be any fixed offset. Let $S \subseteq [n]$ and suppose $\mathcal{T}(X_S) = (\mathsf{ACCEPT}, E_S)$. Fix any $i \in S$, and define*

$$\mathbf{v} \stackrel{\text{def}}{=} A(X_S) + \mathbf{o}, \qquad \mathbf{u} \stackrel{\text{def}}{=} A(X_{S\setminus\{i\}}) + \mathbf{o}.$$

Then

$$\mathbf{v}\mathbf{v}^\top \preceq 2\big(\mathbf{u}\mathbf{u}^\top + E_S\big).$$

Proof. Let $\mathbf{\Delta} \stackrel{\text{def}}{=} \mathbf{v} - \mathbf{u} = A(X_S) - A(X_{S\setminus\{i\}})$. Since $\mathcal{T}(X_S)$ accepts, the down-sensitivity certificate condition (Definition A.6(1)) implies $\mathbf{\Delta}\mathbf{\Delta}^\top \preceq E_S$. Applying Lemma A.12 with $\mathbf{a} = \mathbf{u}$ and $\mathbf{b} = \mathbf{\Delta}$ gives

$$\mathbf{v}\mathbf{v}^\top = (\mathbf{u}+\mathbf{\Delta})(\mathbf{u}+\mathbf{\Delta})^\top \preceq 2\,\mathbf{u}\mathbf{u}^\top + 2\,\mathbf{\Delta}\mathbf{\Delta}^\top \preceq 2\,\mathbf{u}\mathbf{u}^\top + 2\,E_S,$$

which is equivalent to the stated bound. □

A.6.2 Stability Under Neighboring Datasets

We now formalize the key structural fact needed for privacy: under a stable down-sensitivity test, the *conditional* distribution over accepted subsamples does not change too much when we modify a single datapoint. Concretely, this implies that both (i) the conditional mean of the accepted estimator and (ii) the associated "proxy covariance" matrix are stable under neighboring datasets.

A.6.3 Conditioning on $j \notin S$ does not shift the mean too much

Lemma A.14 (Mean shift from excluding one index)**.** *Assume $\mathcal{T}$ is a C_p, η-down-stable test for A, with corresponding C_E-stable down-sensitivity bound E, and that $\mathcal{T}$ implies a $\mathcal{E}, \mathcal{V}$ envelope on E.*

Let

$$\mu^{\neg j}(X) \stackrel{\text{def}}{=} \mathbb{E}\left[v_S(X) | \mathsf{Acc}_X \wedge (j \notin S)\right] \qquad \textit{and} \qquad \Delta_j(X) \stackrel{\text{def}}{=} \mu(X) - \mu^{\neg j}(X).$$

Fix $X \in \mathcal{X}^n$ and $j \in [n]$. Let $\delta_j(X) \stackrel{\text{def}}{=} \Pr[j \in S \mid \mathrm{Acc}_X]$. Then

$$\Delta_j(X)\,\Delta_j(X)^\top \preceq \frac{\delta_j(X)}{(1-\delta_j(X))^2} \cdot \left(\frac{2tC_p\,(1+C_E)}{1-t}\,M(X) + \frac{2t}{p_X}\eta\,\mathcal{E} + \frac{8t}{(1-t)\,p_X}\eta\,\mathcal{V} \right). \quad (5)$$

Proof. Let $\tilde{v}_S(X) \stackrel{\text{def}}{=} v_S(X) - \mu(X)$. Then $\mathbb{E}\left[\tilde{v}_S(X) | \mathsf{Acc}_X\right] = 0$ and

$$\Delta_j(X) = \mu(X) - \mu^{\neg j}(X) = -\mathbb{E}\left[\tilde{v}_S(X) | \mathsf{Acc}_X \wedge (j \notin S)\right].$$

Let D be the event $\{j \notin S\}$, so $\Pr\left[D \mid \mathsf{Acc}_X\right] = 1 - \delta_j(X)$. Applying Lemma A.11 to the random vector $\tilde{v}_S(X)$ under the conditional distribution $\Pr\left[\,\cdot \;\middle|\mathsf{Acc}_X\right]$ and the event D gives (after dividing by $\Pr[D \mid \mathsf{Acc}_X]^2$ from both sides and using $\mathbb{E}\left[\tilde{v}_S(X) | \mathsf{Acc}_X\right] = 0$)

$$\left(\mathbb{E}\left[\tilde{v}_S(X) | \mathsf{Acc}_X \wedge D\right]\right)\left(\mathbb{E}\left[\tilde{v}_S(X) | \mathsf{Acc}_X \wedge D\right]\right)^\top \preceq \frac{\delta_j(X)}{(1-\delta_j(X))^2} \cdot \mathbb{E}\left[\tilde{v}_S(X)\tilde{v}_S(X)^\top \cdot \mathbb{1}\{j \in S\} \middle| \mathsf{Acc}_X\right].$$

Since $\Delta_j(X) = -\mathbb{E}\left[\tilde{v}_S(X) | \mathsf{Acc}_X \wedge D\right]$, it remains to upper bound the RHS.

On the event Acc_X and $j \in S$, the down-sensitivity bound (Definition A.6) with $X^{\neg j}$ obtained by dropping j implies

$$\big(v_S(X) - v_{S\setminus\{j\}}(X)\big)\,\big(v_S(X) - v_{S\setminus\{j\}}(X)\big)^\top \preceq E_S(X).$$

By the corollary of Lemma A.12, applied to the centered vectors

$$\tilde{v}_S(X) = \tilde{v}_{S\setminus\{j\}}(X) + \big(v_S(X) - v_{S\setminus\{j\}}(X)\big)\ ,$$

we get

$$\tilde{v}_S(X)\tilde{v}_S(X)^\top \preceq 2\tilde{v}_{S\setminus\{j\}}(X)\tilde{v}_{S\setminus\{j\}}(X)^\top + 2E_S(X).$$

Multiplying by $\mathbb{1}\{j \in S\}$ and taking $\mathbb{E}[\cdot|\mathsf{Acc}_X]$ yields

$$\mathbb{E}\left[\tilde{v}_S\tilde{v}_S^\top \cdot \mathbb{1}\{j \in S\}\middle|\mathsf{Acc}_X\right] \preceq 2\mathbb{E}\left[\tilde{v}_{S\setminus\{j\}}\tilde{v}_{S\setminus\{j\}}^\top \cdot \mathbb{1}\{j \in S\}\middle|\mathsf{Acc}_X\right] + 2\mathbb{E}\left[E_S \cdot \mathbb{1}\{j \in S\}|\mathsf{Acc}_X\right]. \tag{6}$$

We bound the second term using Lemma A.9, giving

$$\mathbb{E}\left[E_S \cdot \mathbb{1}\{j \in S\}|\mathsf{Acc}_X\right] \preceq \frac{tC_pC_E}{1-t} \cdot \mathbb{E}\left[E_S|\mathsf{Acc}_X\right] + \frac{t}{p_X}\eta\mathcal{E}. \tag{7}$$

For the first term in (6), apply Theorem A.2 to reweight from sets containing j to sets with j removed. Concretely, Theorem A.2 implies (after the $S \mapsto S \setminus \{j\}$ change of variables) that

$$\mathbb{E}\left[\tilde{v}_{S\setminus\{j\}}\tilde{v}_{S\setminus\{j\}}^\top \cdot \mathbb{1}\{j \in S\}\middle|\mathsf{Acc}_X\right] \preceq \frac{tC_p}{1-t} \cdot \mathbb{E}\left[\tilde{v}_S\tilde{v}_S^\top\middle|\mathsf{Acc}_X\right] + \frac{t\eta}{(1-t)\,p_X} \cdot \mathop{\mathbb{E}}_{S}\left[\tilde{v}_S\tilde{v}_S^\top\right].$$

By Definition A.7 and Lemma A.12, $\tilde{v}_S\tilde{v}_S^\top \preceq 2v_Sv_S^\top + 2\mu(X)\mu(X)^\top \preceq 4\mathcal{V}$, hence $\mathbb{E}_S\left[\tilde{v}_S\tilde{v}_S^\top\right] \preceq 4\mathcal{V}$. Therefore,

$$\mathbb{E}\left[\tilde{v}_{S\setminus\{j\}}\tilde{v}_{S\setminus\{j\}}^\top \cdot \mathbb{1}\{j \in S\}\middle|\mathsf{Acc}_X\right] \preceq \frac{tC_p}{1-t} \cdot \mathbb{E}\left[\tilde{v}_S\tilde{v}_S^\top\middle|\mathsf{Acc}_X\right] + \frac{4t\eta}{(1-t)\,p_X}\mathcal{V} \tag{8}$$

Plugging (7) and (8) into (6), and using the definition of $M(X)$ in (3) (so that $\mathbb{E}\left[\tilde{v}_S\tilde{v}_S^\top\middle|\mathsf{Acc}_X\right] \preceq M(X)$ and $\mathbb{E}\left[E_S|\mathsf{Acc}_X\right] \preceq M(X)$) yields

$$\mathbb{E}\left[\tilde{v}_S\tilde{v}_S^\top \cdot \mathbb{1}\{j \in S\}\middle|\mathsf{Acc}_X\right] \preceq \frac{2tC_p\,(1+C_E)}{1-t}\,M(X) + \frac{2t\eta}{p_X}\mathcal{E} + \frac{8t\eta}{(1-t)\,p_X}\mathcal{V}.$$

Substituting this bound into the initial application of Lemma A.11 completes the proof of (5). □

A.6.4 Stability of M under deletions

Let $X^{\neg j}$ denote the dataset obtained by dropping the jth sample from X. Under Poisson subsampling, sampling from $X^{\neg j}$ is equivalent to sampling S from X and conditioning on $j \notin S$; moreover, on this event the subsample itself is identical.

Lemma A.15 (Proxy stability under dropping one sample)**.** *Assume $\mathcal{T}$ is a C_p, η-down-stable test for A, with corresponding C_E-stable down-sensitivity bound E, and that $\mathcal{T}$ implies a $\mathcal{E}, \mathcal{V}$ envelope on E. Fix $X \in \mathcal{X}^n$ and $j \in [n]$.*

Let $\delta_j(X), \Delta_j(X)$ be as in Lemma A.14 and define

$$a \stackrel{\text{def}}{=} \frac{tC_p}{1-t}\,(2+3C_E)\,.$$

Then:

$$(1-\delta_j(X)) \cdot M\left(X^{\neg j}\right) \preceq M(X), \tag{9}$$

$$M(X) \preceq \frac{1}{1-a} \cdot \left(M\left(X^{\neg j}\right) + \Delta_j(X)\Delta_j(X)^\top + \frac{3t\eta}{p_X}\mathcal{E} + \frac{8t\eta}{(1-t)\,p_X}\mathcal{V}\right), \tag{10}$$

whenever $a < 1$.

Proof. Let $\tilde{v}_S(X) = v_S(X) - \mu(X)$ and define the PSD random matrix

$$W_S(X) \stackrel{\text{def}}{=} \tilde{v}_S(X)\tilde{v}_S(X)^\top + E_S(X).$$

By definition, $M(X) = \mathbb{E}\left[W_S(X)|\mathsf{Acc}_X\right]$.

Let D be the event $j \notin S$. As noted above, the conditional distribution of X_S given $\mathsf{Acc}_X \wedge D$ is the same as the accepted-subsample distribution for $X^{\neg j}$. A direct expansion gives

$$M\left(X^{\neg j}\right) = \mathbb{E}\left[W_S(X)|\mathsf{Acc}_X \wedge D\right] - \Delta_j(X)\Delta_j(X)^\top. \tag{11}$$

Next, observe that

$$\begin{aligned}\mathbb{E}\left[W_S(X)|\mathsf{Acc}_X \wedge D\right] &= \frac{1}{1-\delta_j(X)} \cdot \mathbb{E}\left[W_S(X) \cdot \mathbb{1}\{D\}|\mathsf{Acc}_X\right] \\ &= \frac{1}{1-\delta_j(X)} \cdot \left(M(X) - \mathbb{E}\left[W_S(X) \cdot \mathbb{1}\{j \in S\}|\mathsf{Acc}_X\right]\right).\end{aligned}$$

Plugging into (11) yields

$$M\left(X^{\neg j}\right) = \frac{1}{1-\delta_j(X)} \cdot \left(M(X) - \mathbb{E}\left[W_S(X) \cdot \mathbb{1}\{j \in S\}|\mathsf{Acc}_X\right]\right) \; - \; \Delta_j(X)\Delta_j(X)^\top. \quad (12)$$

Since $\mathbb{E}\left[W_S(X) \cdot \mathbb{1}\{j \in S\}|\mathsf{Acc}_X\right] \succeq 0$ and $\Delta_j(X)\Delta_j(X)^\top \succeq 0$, we immediately get $M(X^{\neg j}) \preceq \frac{1}{1-\delta_j(X)} M(X)$, which is equivalent to (9).

For the other direction, rearrange (12) as

$$M(X) = (1-\delta_j(X))\left(M\left(X^{\neg j}\right) + \Delta_j(X)\Delta_j(X)^\top\right) \; + \; \mathbb{E}\left[W_S(X) \cdot \mathbb{1}\{j \in S\}|\mathsf{Acc}_X\right]. \quad (13)$$

We upper bound the last term. By Lemma A.14's intermediate bounds (equations 6, 7, 8) , we have

$$\begin{aligned}\mathbb{E}\left[\tilde{v}_S\tilde{v}_S^\top \cdot \mathbb{1}\{j \in S\}\middle|\mathsf{Acc}_X\right] \; \preceq \; & \frac{2tC_p}{1-t} \cdot \mathbb{E}\left[\tilde{v}_S\tilde{v}_S^\top\middle|\mathsf{Acc}_X\right] \\ & + \; \frac{2tC_pC_E}{1-t} \cdot \mathbb{E}\left[E_S|\mathsf{Acc}_X\right] \\ & + \; \frac{2t\eta}{p_X}\mathcal{E} \; + \; \frac{8t\eta}{(1-t)\,p_X}\mathcal{V}.\end{aligned}$$

and by Lemma A.9,

$$\mathbb{E}\left[E_S \cdot \mathbb{1}\{j \in S\}|\mathsf{Acc}_X\right] \; \preceq \; \frac{tC_pC_E}{1-t} \cdot \mathbb{E}\left[E_S|\mathsf{Acc}_X\right] \; + \; \frac{t\eta}{p_X}\mathcal{E}.$$

Adding these gives

$$\begin{aligned}\mathbb{E}\left[W_S(X) \cdot \mathbb{1}\{j \in S\}|\mathsf{Acc}_X\right] \; &\preceq \; \frac{tC_p}{1-t}\left(2+3C_E\right) M(X) \; + \; \frac{3t\eta}{p_X}\mathcal{E} \; + \; \frac{8t\eta}{(1-t)\,p_X}\mathcal{V} \\ &= a\,M(X) + \frac{3t\eta}{p_X}\mathcal{E} + \frac{8t\eta}{(1-t)\,p_X}\mathcal{V}.\end{aligned}$$

Plugging into (13) and using $1-\delta_j(X) \leq 1$ yields

$$M(X) \; \preceq \; M\left(X^{\neg j}\right) + \Delta_j(X)\Delta_j(X)^\top \; + \; a\,M(X) \; + \; \frac{3t\eta}{p_X}\mathcal{E} \; + \; \frac{8t\eta}{(1-t)\,p_X}\mathcal{V}.$$

Rearranging completes the proof of (10). □

A.6.5 Concluding Theorem A.8

We now upgrade Lemmas A.14 and A.15 into the main theorem of this subsection for *substitution* neighbors. We'll use the fact from Theorem A.2 that $\delta_j(X) = O(t)$ when $p_X \gtrsim \eta$.

Proof. For (i), note that on the event $j \notin S$ we have $X_S = X'_S$, hence $v_S(X) = v_S(X')$ and the test output is identical; therefore $\mu^{\neg j}(X) = \mu^{\neg j}(X')$. Thus

$$\mu(X) - \mu(X') = \left(\mu(X) - \mu^{\neg j}(X)\right) - \left(\mu(X') - \mu^{\neg j}(X')\right).$$

Applying Lemma A.12 gives $\Delta\Delta^\top \preceq 2\Delta_j(X)\Delta_j(X)^\top + 2\Delta_j(X')\Delta_j(X')^\top$. Lemma A.14 bounds each term.

For (ii), let $Y = X^{\neg j} = (X')^{\neg j}$. By Lemma A.15 applied to X, we have $M(Y) \preceq \frac{1}{1-\delta_j(X)} M(X)$. Applying Lemma A.15 to X' and substituting this bound on $M(Y)$, along with Lemma A.14 on $\Delta_j(X')\Delta_j(X')$ yields the displayed inequality. The reverse direction follows symmetrically. □

Algorithm 5: AC-STA Aggregator

Input: $X \in \mathcal{X}^*$, $k \in \mathbb{N}$
Output: a model in $\mathbb{R}^d$ or REJECT

Draw $S_1, \ldots, S_k \overset{\text{iid}}{\sim} \mathsf{Acc}_X$;

for $j \leftarrow 1$ **to** k **do**
  Draw $v_j \leftarrow \mathcal{N}\left(A(X_{S_j}),\, 2\log(5/\delta)\, E(X_{S_j})\right)$

return $\frac{1}{k}\sum_{j=1}^{k} v_j$;

B Aggregation via Advanced Composition

B.1 Advanced Composition

In this section, we will analyze the privacy of our main STA mechanism – Algorithm 5. The key idea is that every draw from

$$\left(\mathcal{N}\big(A(X_S), O(\log(1/\delta))E(X_S)\big) \mid S \sim \mathsf{Acc}_X\right)$$

is $(O(t), \delta)$-private (Lemma B.1).

Using just a single draw from this distribution would not give us good utility, since it would only utilize a small fraction of our dataset. Therefore, we would want to average over $k \gg 1$ draws from the mechanism above. Utilizing the advanced composition theorem (Theorem B.2), we can achieve this for a very small cost to the privacy of our algorithm, since advanced composition yields a $(O(t\sqrt{k\log(1/\delta)} + kt^2), O(k\delta))$-privacy guarantee.

Lemma B.1 (Privacy of a Single Draw)**.** *Let $\mathcal{T}$ be a (C_p, η)-down-stable test for a down-stability bound $E(\cdot)$, and assume that*

$$E(X_S) \preceq E(X_{S\setminus\{i\}}) \quad \text{and} \quad C_E\left|E(X_{S\setminus\{i\}})\right| \le |E(X_S)| \le \left|E(X_{S\setminus\{i\}})\right|.$$

Assume $t \le \frac{1}{5(C_p+1)}$ and fix any $\delta \in (0, 1/10)$.

Let $X, X' \in \mathcal{X}^n$ be neighboring datasets such that $p_X \ \ge\ \frac{2\eta}{\delta_{\text{draw}}}$. Then

$$\left(\mathcal{N}\left(A(X_S),\, \sigma^2 E(X_S)\right) \mid S \sim \mathsf{Acc}_X\right) \approx_{\varepsilon_{\text{draw}},\, \delta_{\text{draw}}} \left(\mathcal{N}\left(A(X'_S),\, \sigma^2 E(X'_S)\right) \mid S \sim \mathsf{Acc}_{X'}\right),$$

for $\sigma^2 \ \overset{\text{def}}{=}\ 2\log(5/\delta_{\text{draw}})$ and $\varepsilon_{\text{draw}} = O(t)$.

Theorem B.2 (Advanced composition (Dwork & Roth, 2014, Thm. 3.20))**.** *Let $k \in \mathbb{N}$. For each $i \in [k]$, let $\mathcal{M}_i$ be an (interactive) mechanism that is (ε, δ)-differentially private. Then for every $\delta' \in (0, 1)$, the adaptive k-fold composition $\mathcal{M}_{1:k}$ is*

$$\left(\varepsilon\sqrt{2k\log(1/\delta')} \ +\ k\varepsilon(e^{\varepsilon} - 1),\ \ k\delta + \delta'\right)\text{-DP}.$$

Averaging over $k \approx \frac{\log(1/\delta)}{\alpha^2}$ draws from this mechanism suffice for the utility of our algorithm. Combining this with the advanced composition theorem allows us to ensure privacy so long as $t \le \frac{\varepsilon\alpha}{\log(1/\delta)}$.

Theorem B.3 (Privacy of AC-STA)**.** *Under the assumptions of Lemma B.1, the output of Algorithm 5 is private, in the sense that for any $\delta > 0$ and for any datasets X, X' as in Lemma B.1, it satisfies*

$$\text{AC-STA}(X) \approx_{\sqrt{k\log(1/\delta)}\varepsilon_{\mathsf{draw}} + 2k\varepsilon_{\mathsf{draw}}^2,\ k\delta_{\text{draw}} + \delta} \text{AC-STA}(X').$$

Theorem B.3 follows immediately from Lemma B.1 and Theorem B.2. The rest of this section will be devoted to proving Lemma B.1. Throughout this section, let X, X' be neighboring datasets and let $i \in [n]$ be the index on which they differ.

B.2 Proof of Lemma B.1 – Privacy of a Single Draw

We now proceed to prove that a single draw from $\left(\mathcal{N}(A(X_S), \log(1/\delta)E(X_S)) \mid S \sim \mathsf{Acc}_X\right)$ is private. The key idea we will use is that Theorem A.2 will allow us to show that $S \setminus \{i\}$ behaves the same for the datasets X and X' (Lemma B.4).

Lemma B.4 ($S \setminus \{i\}$ is Private)**.** *Under the assumptions of Lemma B.1,*

$$\left(S \setminus \{i\} \mid S \sim \mathsf{Acc}_X\right) \approx_{\varepsilon_S, \delta_S} \left(S \setminus \{i\} \mid S \sim \mathsf{Acc}_{X'}\right),$$

where

$$\varepsilon_S \;:=\; \log\left(\frac{1 + \frac{C_p t}{1-t}}{1 - 3C_p t}\right) = O(t) \qquad \textit{and} \qquad \delta_S \;:=\; \frac{t}{1-t}\,(1 + 3C_p t) \cdot \frac{\eta}{p_X} = O\left(\frac{\eta}{p_X}\right). \tag{14}$$

It is important to note that Lemma B.4 *does not* mean that we can reveal $S \setminus \{i\}$ while maintaining privacy, since the choice of i depended on X, X', while privacy of a statistic would require that it is private for any neighboring datasets.

We will then combine Theorem A.2 with our assumptions on the properties of E to show that the $\mathcal{N}(0, \log(1/\delta)E(X_S))$ normal noise is itself private (Lemma B.5). The idea is that if we are given $T = S \setminus \{i\}$, but not told if $S = T$ or $S = T \cup \{i\}$, then $S = T$ is much more likely (from Theorem A.2). If we are then also told that the output of the mechanism v is very far from 0, then the property $E(X_T) \succeq E(X_{T \cup \{i\}})$ tells us that v is more likely to be generated from $S = T$ than from $S = T \cup \{i\}$. Alternatively, if v is close to 0, then the second assumption that $\det(E(X_T)) = O(\det(E(X_{T \cup \{i\}})))$ tells us that the likelihood of drawing v from $T \cup \{i\}$ also cannot be too high. Combining the two tells us that regardless of the value of v, its likelihood of being drawn from T or a mixture of T and $T \cup \{i\}$ are similar.

Lemma B.5 ($\mathcal{N}(0, E(X_S))$ is Private)**.** *Under the assumptions of Lemma B.1,*

$$\left(\mathcal{N}\left(0, E(X_S)\right) \mid S \sim \mathsf{Acc}_X\right) \approx_{\varepsilon_E, \delta_E} \left(\mathcal{N}\left(0, E(X_{S \setminus \{i\}})\right) \mid S \sim \mathsf{Acc}_X\right),$$

where

$$\varepsilon_E \;:=\; \log\left(1 + \sqrt{C_E} \cdot \frac{C_p t}{1-t}\right) \qquad \textit{and} \qquad \delta_E \;:=\; \sqrt{C_E} \cdot \frac{t}{1-t} \cdot \frac{\eta}{p_X}. \tag{15}$$

We then utilize a similar argument, but this time combined with a standard argument for the privacy of the Gaussian mechanism to shifts to prove Lemma B.6.

Lemma B.6 (The Gaussian Mechanism for Spikes)**.** *Fix $\tau \in [0, 1]$ and $\delta \in (0, 1/2)$. Let $\Sigma \in \mathbb{S}^d_+$ be deterministic, and let $v \sim \mathcal{V} \in \mathbb{R}^d$ be random such that*

$$\Pr_{v \sim \mathcal{V}}[v = 0] \geq 1 - \tau \qquad \textit{and} \qquad \Sigma \succeq 2 \log\left(\frac{4}{\delta}\right) vv^\top \text{ almost surely.}$$

Interpreting $\mathcal{N}(v, \Sigma)$ as the mixture *distribution obtained by first drawing $v \sim \mathcal{V}$ and then drawing $z \sim \mathcal{N}(v, \Sigma)$, we have*

$$\mathcal{N}(0, \Sigma) \approx_{\varepsilon, \delta'} \mathcal{N}(v, \Sigma),$$

where one may take

$$\varepsilon \;:=\; \log\left((1-\tau) + \tau e\right) \;=\; \log\left(1 + \tau(e-1)\right) \qquad \textit{and} \qquad \delta' \;:=\; \tau \cdot \frac{\delta}{2}. \tag{16}$$

In particular, $\varepsilon \leq (e-1)\tau = O(\tau)$ and $\delta' \leq \delta$.

B.2.1 Proof of Lemma B.4

Proof of Lemma B.4. Fix any set $R \subseteq [n] \setminus \{i\}$. By a simple partition,

$$\Pr[S \setminus \{i\} = R \mid S \sim \mathsf{Acc}_X] = \Pr[S = R \mid S \sim \mathsf{Acc}_X] + \Pr[S = R \cup \{i\} \mid S \sim \mathsf{Acc}_X]. \tag{17}$$

Applying Theorem A.2 to the set $T := R \cup \{i\}$ (so $T \setminus \{i\} = R$), we get

$$\Pr[S = R \cup \{i\} \mid S \sim \mathsf{Acc}_X] \le \frac{C_p\, t}{1-t} \cdot \left(\Pr[S = R \mid S \sim \mathsf{Acc}_X] \;+\; \frac{\eta}{C_p p_X} \cdot \Pr[S = R] \right).$$

Plugging this into (17) yields the pointwise bound

$$\begin{aligned} \Pr[S \setminus \{i\} = R \mid S \sim \mathsf{Acc}_X] \;\le\; & \left(1 + \frac{C_p t}{1-t}\right) \Pr[S = R \mid S \sim \mathsf{Acc}_X] \\ & + \frac{t}{1-t} \cdot \frac{\eta}{p_X} \Pr[S = R]. \end{aligned} \tag{18}$$

Since $i \notin R$, we have $X_R = X'_R$, and therefore the conditional acceptance behavior is identical under X and X' given the event $S = R$:

$$\Pr[\mathsf{Acc}_X \mid S = R] = \Pr[\mathsf{Acc}_{X'} \mid S = R]\,.$$

Hence,

$$\Pr[S = R \mid S \sim \mathsf{Acc}_X] = \frac{\Pr[S = R]\,\Pr[\mathsf{Acc}_X \mid S = R]}{p_X} = \frac{p_{X'}}{p_X} \Pr[S = R \mid S \sim \mathsf{Acc}_{X'}]\,. \tag{19}$$

By Corollary A.4, we have $p_X \ge (1 - 3C_p t) p_{X'}$, i.e.

$$\frac{p_{X'}}{p_X} \le \frac{1}{1 - 3C_p t}\,. \tag{20}$$

Moreover, for every R,

$$\Pr[S = R \mid S \sim \mathsf{Acc}_{X'}] \le \Pr[S \setminus \{i\} = R \mid S \sim \mathsf{Acc}_{X'}]\,, \tag{21}$$

since the event $\{S = R\}$ implies $\{S \setminus \{i\} = R\}$.

Combining (18), (19), (20), and (21), we obtain

$$\begin{aligned} \Pr[S \setminus \{i\} = R \mid S \sim \mathsf{Acc}_X] \;\le\; & \frac{1 + \frac{C_p t}{1-t}}{1 - 3C_p t} \cdot \Pr[S \setminus \{i\} = R \mid S \sim \mathsf{Acc}_{X'}] \\ & + \frac{t}{1-t} \cdot \frac{\eta}{p_X} \Pr[S = R]. \end{aligned} \tag{22}$$

Let $\mathcal{E} \subseteq 2^{[n] \setminus \{i\}}$ be any event. Summing (22) over $R \in \mathcal{E}$ gives

$$\begin{aligned} \Pr[S \setminus \{i\} \in \mathcal{E} \mid S \sim \mathsf{Acc}_X] \le & \frac{1 + \frac{C_p t}{1-t}}{1 - 3C_p t} \cdot \Pr[S \setminus \{i\} \in \mathcal{E} \mid S \sim \mathsf{Acc}_{X'}] \\ & + \frac{t}{1-t} \cdot \frac{\eta}{p_X} \sum_{R \in \mathcal{E}} \Pr[S = R] \\ \le & \exp(\varepsilon_S) \cdot \Pr[S \setminus \{i\} \in \mathcal{E} \mid S \sim \mathsf{Acc}_{X'}] \\ & + \frac{t}{1-t} \cdot \frac{\eta}{p_X}\,, \end{aligned}$$

where ε_S is as in (14) and we used $\sum_{R \in \mathcal{E}} \Pr[S = R] \le 1$.

The reverse inequality (with X and X' swapped) is identical, except the additive term becomes $\frac{t}{1-t} \cdot \frac{\eta}{p_{X'}}$; using Corollary A.4 again, we have $p_{X'} \ge p_X/(1 + 3C_p t)$ and hence

$$\frac{t}{1-t} \cdot \frac{\eta}{p_{X'}} \le \frac{t}{1-t}\,(1 + 3C_p t) \cdot \frac{\eta}{p_X} = \delta_S\,.$$

Thus the two conditional laws are $(\varepsilon_S, \delta_S)$-close, proving the lemma.

Finally, to see the advertised scaling, note that for $t \in \left(0, \frac{1}{5(1+C_p)}\right)$ we have $t \le 1/5$ and $3C_p t \le 3/5$, so

$$\varepsilon_S = \log\left(1 + \frac{C_p t}{1-t}\right) + \log\left(\frac{1}{1 - 3C_p t}\right) \le \frac{C_p t}{1-t} + \frac{3C_p t}{1 - 3C_p t} \le \frac{5}{4} C_p t + \frac{15}{2} C_p t \le 9 C_p t\,.$$

□

B.2.2 Proof of Lemma B.5

Proof of Lemma B.5. Fix an arbitrary measurable event $\mathcal{A} \subseteq \mathbb{R}^d$ (where d is the ambient dimension of the output). For each $T \subseteq [n] \setminus \{i\}$, abbreviate

$$\Sigma_T := E(X_T) \qquad \text{and} \qquad \Sigma_{T\cup i} := E(X_{T\cup\{i\}})\,,$$

and let

$$P_T(\mathcal{A}) := \Pr_{z\sim\mathcal{N}(0,\Sigma_T)}[z \in \mathcal{A}] \qquad \text{and} \qquad Q_T(\mathcal{A}) := \Pr_{z\sim\mathcal{N}(0,\Sigma_{T\cup i})}[z \in \mathcal{A}]\,.$$

Comparing the two Gaussians for a fixed T. Let $p_T(\cdot)$ and $q_T(\cdot)$ denote the PDFs of $\mathcal{N}(0,\Sigma_T)$ and $\mathcal{N}(0,\Sigma_{T\cup i})$, respectively. Since $\Sigma_T \succeq \Sigma_{T\cup i}$, we have $\Sigma_T^{-1} \preceq \Sigma_{T\cup i}^{-1}$, and hence for every $v \in \mathbb{R}^d$,

$$\frac{q_T(v)}{p_T(v)} = \sqrt{\frac{\det(\Sigma_T)}{\det(\Sigma_{T\cup i})}} \cdot \exp\left(-\frac{1}{2} v^\top \left(\Sigma_{T\cup i}^{-1} - \Sigma_T^{-1}\right) v\right) \;\le\; \sqrt{\frac{\det(\Sigma_T)}{\det(\Sigma_{T\cup i})}} \;\le\; \sqrt{C_E}\,.$$

Integrating over $\mathcal{A}$ gives the uniform comparison

$$Q_T(\mathcal{A}) \le \sqrt{C_E}\, P_T(\mathcal{A})\,. \tag{23}$$

Expanding the two mixture distributions. Write $S \sim \mathsf{Acc}_X$ for Poisson subsampling (and test randomness) conditioned on Acc_X. Grouping by $T := S \setminus \{i\}$, we may write

$$\Pr_{S\sim\mathsf{Acc}_X}[\mathcal{N}(0, E(X_S)) \in \mathcal{A}] = \sum_{T\subseteq[n]\setminus\{i\}} \Pr_{S\sim\mathsf{Acc}_X}[S = T] \cdot P_T(\mathcal{A}) \tag{24}$$

$$+ \Pr_{S\sim\mathsf{Acc}_X}[S = T\cup\{i\}] \cdot Q_T(\mathcal{A}), \tag{25}$$

$$\Pr_{S\sim\mathsf{Acc}_X}\left[\mathcal{N}\left(0, E(X_{S\setminus\{i\}}\right)) \in \mathcal{A}\right] = \sum_{T\subseteq[n]\setminus\{i\}} \Pr_{S\sim\mathsf{Acc}_X}[S = T] \cdot P_T(\mathcal{A}) \tag{26}$$

$$+ \Pr_{S\sim\mathsf{Acc}_X}[S = T\cup\{i\}] \cdot P_T(\mathcal{A}). \tag{27}$$

Applying Theorem A.2 to the set $T \cup \{i\}$ yields, for every $T \subseteq [n] \setminus \{i\}$,

$$\Pr_{S\sim\mathsf{Acc}_X}[S = T\cup\{i\}] \le \frac{C_p t}{1-t} \cdot \Pr_{S\sim\mathsf{Acc}_X}[S = T] \;+\; \frac{t}{1-t} \cdot \frac{\eta}{p_X} \cdot \Pr_{S\sim\mathsf{Acc}_X}[S = T]\,. \tag{28}$$

Define $\rho := \frac{C_p t}{1-t}$ and $\xi := \frac{t}{1-t} \cdot \frac{\eta}{p_X}$ for brevity.

The first DP inequality. Using (23) and then (28) in (24) gives

$$\begin{aligned}
\Pr_{S\sim\mathsf{Acc}_X}[\mathcal{N}(0, E(X_S)) \in \mathcal{A}] &\le \sum_T \Pr_{S\sim\mathsf{Acc}_X}[S = T] \cdot P_T(\mathcal{A}) \\
&\quad + \Pr_{S\sim\mathsf{Acc}_X}[S = T\cup\{i\}] \cdot \sqrt{C_E}\, P_T(\mathcal{A}) \\
&\le \sum_T \left(1 + \sqrt{C_E}\rho\right) \Pr_{S\sim\mathsf{Acc}_X}[S = T] \cdot P_T(\mathcal{A}) \\
&\quad + \sqrt{C_E}\,\xi \sum_T \Pr_{S\sim\mathsf{Acc}_X}[S = T] \cdot P_T(\mathcal{A})\,.
\end{aligned}$$

Since $\sum_T \Pr_{S\sim\mathsf{Acc}_X}[S = T] \cdot P_T(\mathcal{A}) \le \sum_T \Pr_{S\sim\mathsf{Acc}_X}[S = T] \le 1$, and since

$$\begin{aligned}
\sum_T \Pr_{S\sim\mathsf{Acc}_X}[S = T] \cdot P_T(\mathcal{A}) &\le \sum_T \left(\Pr_{S\sim\mathsf{Acc}_X}[S = T] + \Pr_{S\sim\mathsf{Acc}_X}[S = T\cup\{i\}]\right) \cdot P_T(\mathcal{A}) \\
&= \Pr_{S\sim\mathsf{Acc}_X}\left[\mathcal{N}\left(0, E(X_{S\setminus\{i\}}\right)) \in \mathcal{A}\right],
\end{aligned}$$

we conclude

$$\Pr_{S\sim\mathsf{Acc}_X}[\mathcal{N}(0, E(X_S)) \in \mathcal{A}] \le \left(1 + \sqrt{C_E}\rho\right) \cdot \Pr_{S\sim\mathsf{Acc}_X}\left[\mathcal{N}\left(0, E(X_{S\setminus\{i\}}\right)) \in \mathcal{A}\right] \;+\; \sqrt{C_E}\,\xi\,.$$

This is exactly the first $(\varepsilon_E, \delta_E)$ inequality with parameters (15).

The second DP inequality. Starting from (26) and applying (28) directly (without needing (23)), we get

$$\begin{aligned}\Pr_{S\sim\mathsf{Acc}_X}\left[\mathcal{N}\left(0, E(X_{S\setminus\{i\}})\right)\in\mathcal{A}\right] &= \sum_T \Pr_{S\sim\mathsf{Acc}_X}[S=T]\cdot P_T(\mathcal{A}) \\ &\quad + \Pr_{S\sim\mathsf{Acc}_X}[S=T\cup\{i\}]\cdot P_T(\mathcal{A}) \\ &\le \sum_T (1+\rho)\Pr_{S\sim\mathsf{Acc}_X}[S=T]\cdot P_T(\mathcal{A}) \\ &\quad + \xi\sum_T \Pr_{S\sim\mathsf{Acc}_X}[S=T]\cdot P_T(\mathcal{A}) \\ &\le (1+\rho)\cdot\sum_T \Pr_{S\sim\mathsf{Acc}_X}[S=T]\cdot P_T(\mathcal{A}) \;+\; \xi\,.\end{aligned}$$

Finally, by (24) we have

$$\sum_T \Pr_{S\sim\mathsf{Acc}_X}[S=T]\cdot P_T(\mathcal{A}) \le \Pr_{S\sim\mathsf{Acc}_X}[\mathcal{N}(0, E(X_S))\in\mathcal{A}],$$

and since $1+\rho \le 1+\sqrt{C_E}\rho$ and $\xi \le \sqrt{C_E}\xi$, this yields

$$\Pr_{S\sim\mathsf{Acc}_X}\left[\mathcal{N}\left(0, E(X_{S\setminus\{i\}})\right)\in\mathcal{A}\right] \le \left(1+\sqrt{C_E}\rho\right)\cdot \Pr_{S\sim\mathsf{Acc}_X}[\mathcal{N}(0, E(X_S))\in\mathcal{A}] \;+\; \sqrt{C_E}\,\xi\,,$$

which is the second $(\varepsilon_E, \delta_E)$ inequality with the same parameters (15). This completes the proof. □

B.2.3 Proof of Lemma B.6

Proof of Lemma B.6. Fix any measurable event $\mathcal{A}\subseteq\mathbb{R}^d$.

Warm-up: analysis of the standard Gaussian mechanism with $\varepsilon = 1$. Fix any deterministic vector $v\in\mathbb{R}^d$ satisfying $\Sigma \succeq 2\log\left(\frac{4}{\delta}\right) vv^\top$. Let p_0 and p_v denote the PDFs of $\mathcal{N}(0,\Sigma)$ and $\mathcal{N}(v,\Sigma)$ (with respect to Lebesgue measure on the support subspace; equivalently one may interpret Σ^{-1} below as the Moore–Penrose pseudoinverse $\Sigma^\dagger$). A direct calculation gives the privacy-loss random variable

$$L(u) \;:=\; \log\left(\frac{p_0(u)}{p_v(u)}\right) \;=\; \frac{1}{2}\,v^\top\Sigma^\dagger v \;-\; u^\top\Sigma^\dagger v\,.$$

Let $\alpha := v^\top\Sigma^\dagger v$. The matrix domination assumption implies (e.g. by the rank-one PSD ordering characterization) that

$$\alpha \;\le\; \frac{1}{2\log\left(\frac{4}{\delta}\right)}\,. \tag{29}$$

Now, if $u\sim\mathcal{N}(0,\Sigma)$ then $u^\top\Sigma^\dagger v$ is a one-dimensional Gaussian with mean 0 and variance α, and therefore $L(u)\sim\mathcal{N}(\alpha/2,\alpha)$. Hence, using the standard Gaussian tail bound $\Pr_{Z\sim\mathcal{N}(0,1)}[Z\ge t]\le e^{-t^2/2}$,

$$\begin{aligned}\Pr_{u\sim\mathcal{N}(0,\Sigma)}[L(u)>1] &= \Pr_{Z\sim\mathcal{N}(0,1)}\left[Z > \frac{1-\alpha/2}{\sqrt{\alpha}}\right] \\ &\le \exp\left(-\frac{(1-\alpha/2)^2}{2\alpha}\right) \\ &= \exp\left(-\left(\frac{1}{2\alpha}-\frac{1}{2}+\frac{\alpha}{8}\right)\right) \\ &\le \exp\left(-\left(\log\left(\frac{4}{\delta}\right)-\frac{1}{2}\right)\right) \\ &= e^{1/2}\cdot\frac{\delta}{4} \;\le\; \frac{\delta}{2},\end{aligned}$$

where in the last line we used (29) to lower bound $\frac{1}{2\alpha} \geq \log(4/\delta)$.

Likewise, if $u \sim \mathcal{N}(v, \Sigma)$ then $u = v + g$ with $g \sim \mathcal{N}(0, \Sigma)$, so $u^\top \Sigma^\dagger v = \alpha + g^\top \Sigma^\dagger v$ and thus $L(u) \sim \mathcal{N}(-\alpha/2, \alpha)$. By symmetry, the same tail bound yields

$$\Pr_{u \sim \mathcal{N}(v,\Sigma)} [L(u) < -1] \leq \frac{\delta}{2}.$$

Define the good set $\mathcal{G} := \{u : |L(u)| \leq 1\}$. On $\mathcal{G}$ we have both $p_0(u) \leq e \cdot p_v(u)$ and $p_v(u) \leq e \cdot p_0(u)$. Therefore, for every event $\mathcal{A}$,

$$\Pr_{u \sim \mathcal{N}(0,\Sigma)} [u \in \mathcal{A}] \leq e \Pr_{u \sim \mathcal{N}(v,\Sigma)} [u \in \mathcal{A}] \; + \; \Pr_{u \sim \mathcal{N}(0,\Sigma)} [u \notin \mathcal{G}] \leq e \Pr_{u \sim \mathcal{N}(v,\Sigma)} [u \in \mathcal{A}] \; + \; \frac{\delta}{2},$$

and similarly

$$\Pr_{u \sim \mathcal{N}(v,\Sigma)} [u \in \mathcal{A}] \leq e \Pr_{u \sim \mathcal{N}(0,\Sigma)} [u \in \mathcal{A}] \; + \; \frac{\delta}{2}.$$

Thus, for every fixed v satisfying the matrix domination assumption,

$$\mathcal{N}(0, \Sigma) \approx_{1,\, \delta/2} \mathcal{N}(v, \Sigma). \tag{30}$$

Mixing over a spike at $v = 0$. Let $P := \mathcal{N}(0, \Sigma)$. Write $\mathcal{V}_{\neq 0}$ for the conditional law of v given $v \neq 0$, and define the conditional mixture

$$Q \; := \; \mathop{\mathbb{E}}_{v \sim \mathcal{V}_{\neq 0}} [\mathcal{N}(v, \Sigma)].$$

By (30) and linearity of expectation, we have $Q \approx_{1,\delta/2} P$. Moreover, using $\Pr[v = 0] \geq 1 - \tau$, the full mixture distribution is

$$\mathcal{N}(v, \Sigma) \; = \; (1 - \tau)\, P \; + \; \tau\, Q.$$

We now compare P and $(1 - \tau)P + \tau Q$. First,

$$\begin{aligned}
\Pr_{\mathcal{N}(v,\Sigma)} [u \in \mathcal{A}] &= (1 - \tau) \Pr_P [u \in \mathcal{A}] + \tau \Pr_Q [u \in \mathcal{A}] \\
&\leq (1 - \tau) \Pr_P [u \in \mathcal{A}] + \tau \left(e \Pr_P [u \in \mathcal{A}] + \delta/2 \right) \\
&= ((1 - \tau) + \tau e) \Pr_P [u \in \mathcal{A}] + \tau \cdot \frac{\delta}{2} = e^{\varepsilon} \Pr_P [u \in \mathcal{A}] + \delta',
\end{aligned}$$

with ε, δ' as in (16).

For the reverse inequality, using $P \approx_{1,\delta/2} Q$ we have $\Pr_P [u \in \mathcal{A}] \leq e \Pr_Q [u \in \mathcal{A}] + \delta/2$. Substituting $\Pr_Q [u \in \mathcal{A}] = \frac{1}{\tau} \Big(\Pr_{\mathcal{N}(v,\Sigma)} [u \in \mathcal{A}] - (1 - \tau) \Pr_P [u \in \mathcal{A}] \Big)$ and rearranging gives

$$\Pr_P [u \in \mathcal{A}] \leq \frac{e}{\tau + e(1 - \tau)} \Pr_{\mathcal{N}(v,\Sigma)} [u \in \mathcal{A}] \; + \; \frac{\tau}{\tau + e(1 - \tau)} \cdot \frac{\delta}{2}.$$

Since $\frac{e}{\tau + e(1-\tau)} = \frac{1}{(1-\tau) + \tau/e} \leq (1 - \tau) + \tau e = e^{\varepsilon}$ and $\frac{\tau}{\tau + e(1-\tau)} \cdot \frac{\delta}{2} \leq \tau \cdot \frac{\delta}{2} = \delta'$, we obtain

$$\Pr_P [u \in \mathcal{A}] \leq e^{\varepsilon} \Pr_{\mathcal{N}(v,\Sigma)} [u \in \mathcal{A}] \; + \; \delta'.$$

Thus $\mathcal{N}(0, \Sigma) \approx_{\varepsilon,\delta'} \mathcal{N}(v, \Sigma)$, completing the proof. □

B.2.4 Concluding Lemma B.1

Proof of Lemma B.1. Let i be the unique coordinate on which X and X' differ, and set

$$\sigma^2 \stackrel{\text{def}}{=} 2 \log \left(\frac{5}{\delta} \right), \qquad \alpha \stackrel{\text{def}}{=} \frac{C_p t}{1 - t}.$$

Write

$$\tau_X \stackrel{\text{def}}{=} \Pr[i \in S \mid \mathsf{Acc}_X], \qquad \tau_{X'} \stackrel{\text{def}}{=} \Pr[i \in S \mid \mathsf{Acc}_{X'}].$$

Summing Theorem A.2 over all $T \ni i$ gives

$$\tau_X \le \alpha(1-\tau_X) + t \cdot \frac{\eta}{p_X} \qquad \Longrightarrow \qquad \tau_X \le \frac{\alpha + t \cdot \eta/p_X}{1+\alpha}.$$

The same argument yields

$$\tau_{X'} \le \frac{\alpha + t \cdot \eta/p_{X'}}{1+\alpha}.$$

Since $\delta < 1/5$, the assumption $p_X \ge 2\eta/\delta$ implies $p_X \ge 10\eta$; hence Corollary A.4 applies and gives $p_{X'} \ge p_X/(1+3C_p t)$, so

$$\frac{\eta}{p_{X'}} \le (1+3C_p t) \cdot \frac{\eta}{p_X}.$$

Therefore,

$$\tau_X, \tau_{X'} \;\le\; \tau \;\stackrel{\text{def}}{=}\; \frac{\alpha + t\,(1+3C_p t) \cdot \eta/p_X}{1+\alpha}. \tag{31}$$

Applying Lemma B.6 to remove i from the mean. Define the spike

$$v \stackrel{\text{def}}{=} A(X_S) - A(X_{S\setminus\{i\}}).$$

By definition, $v = 0$ whenever $i \notin S$, and from the definition of E,

$$E(X_S) \succeq vv^\top \qquad \text{almost surely.}$$

Condition on S (so $\Sigma \stackrel{\text{def}}{=} \sigma^2 E(X_S)$ is fixed). Since $\sigma^2 = 2\log(5/\delta) \ge 2\log(4/\delta)$, we have

$$\Sigma \succeq 2\log\left(\frac{4}{\delta}\right) vv^\top.$$

Applying Lemma B.6 with τ from (31) and then shifting by the common mean $A(X_{S\setminus\{i\}})$ gives

$$\Big(\mathcal{N}\left(A(X_S),\, \sigma^2 E(X_S)\right) \mid S \sim \mathsf{Acc}_X\Big) \approx_{\varepsilon_G,\, \tau\delta/2} \Big(\mathcal{N}\left(A(X_{S\setminus\{i\}}),\, \sigma^2 E(X_S)\right) \mid S \sim \mathsf{Acc}_X\Big),$$

where $\varepsilon_G \stackrel{\text{def}}{=} \log\left(1+\tau(e-1)\right)$.

Applying Lemma B.5 to remove i from the covariance. By translation invariance and post-processing, Lemma B.5 applies with the same ε_E and δ_E to Gaussians whose mean is $A(X_{S\setminus\{i\}})$, and also with both covariances scaled by the common factor σ^2. Hence

$$\Big(\mathcal{N}\left(A(X_{S\setminus\{i\}}),\, \sigma^2 E(X_S)\right) \mid S \sim \mathsf{Acc}_X\Big) \approx_{\varepsilon_E,\, \delta_E} \Big(\mathcal{N}\left(A(X_{S\setminus\{i\}}),\, \sigma^2 E(X_{S\setminus\{i\}})\right) \mid S \sim \mathsf{Acc}_X\Big),$$

where ε_E is as in (15) and

$$\delta_E \;=\; \sqrt{C_E} \cdot \frac{t}{1-t} \cdot \frac{\eta}{p_X}.$$

Applying Lemma B.4 to replace $S \sim \mathsf{Acc}_X$ with $S \sim \mathsf{Acc}_{X'}$. Since X and X' differ only at coordinate i, for every set $U \subseteq [n]$ we have

$$X_U = X'_U \qquad \text{whenever } i \notin U.$$

In particular, $X_{S\setminus\{i\}} = X'_{S\setminus\{i\}}$ and hence also $A(X_{S\setminus\{i\}}) = A(X'_{S\setminus\{i\}})$ and $E(X_{S\setminus\{i\}}) = E(X'_{S\setminus\{i\}})$. Thus, Lemma B.4 and post-processing imply

$$\Big(\mathcal{N}\left(A(X_{S\setminus\{i\}}),\, \sigma^2 E(X_{S\setminus\{i\}})\right) \mid S \sim \mathsf{Acc}_X\Big) \approx_{\varepsilon_S,\, \delta_S} \Big(\mathcal{N}\left(A(X'_{S\setminus\{i\}}),\, \sigma^2 E(X'_{S\setminus\{i\}})\right) \mid S \sim \mathsf{Acc}_{X'}\Big),$$

where ε_S and δ_S are as in (14).

Re-introducing i into the covariance using Lemma B.5 (again). Applying Lemma B.5 to X' (and using $\eta/p_{X'} \le (1+3C_p t)\eta/p_X$ from above) yields

$$\left(\mathcal{N}\left(A(X'_{S\setminus\{i\}}),\, \sigma^2 E(X'_{S\setminus\{i\}})\right) \mid S \sim \mathsf{Acc}_{X'}\right) \approx_{\varepsilon_E,\, \delta'_E} \left(\mathcal{N}\left(A(X'_{S\setminus\{i\}}),\, \sigma^2 E(X'_S)\right) \mid S \sim \mathsf{Acc}_{X'}\right),$$

where

$$\delta'_E \;=\; \sqrt{C_E}\cdot\frac{t}{1-t}\cdot\frac{\eta}{p_{X'}} \;\le\; \sqrt{C_E}\cdot\frac{t}{1-t}\,(1+3C_p t)\cdot\frac{\eta}{p_X}.$$

Re-introducing i into the mean using Lemma B.6 (again). Exactly as in Step 1 (now for X'), we obtain

$$\left(\mathcal{N}\left(A(X'_{S\setminus\{i\}}),\, \sigma^2 E(X'_S)\right) \mid S \sim \mathsf{Acc}_{X'}\right) \approx_{\varepsilon_G,\, \tau\delta/2} \left(\mathcal{N}\left(A(X'_S),\, \sigma^2 E(X'_S)\right) \mid S \sim \mathsf{Acc}_{X'}\right).$$

Composition and the final δ bookkeeping. Composing the five steps, the total ε is

$$\varepsilon_{\mathsf{draw}} = \varepsilon_S + 2\varepsilon_E + 2\varepsilon_G = O(t),$$

as claimed, while the total δ is at most

$$\delta_{\mathsf{tot}} \;\le\; \delta_S + \delta_E + \delta'_E + \tau\delta.$$

Using $\eta/p_X \le \delta/2$, $t \le 1/5$, and $\alpha \le 1/4$ (from $t \le 1/(5(C_p+1))$), one checks that $\delta_S \le \delta/5$, $\delta_E \le \delta/8$, $\delta'_E \le \delta/5$, and $\tau \le 3/10$, hence $\delta_{\mathsf{tot}} \le \delta$. This completes the proof. □

C STAR: Subsample-Test-Aggregate for Linear Regression

In this section, we will describe instantiate our subsample-test-aggregate framework for linear regressions.

C.1 The STAR Algorithm

Algorithm 6: Subsample-Test-Aggregate Regression (STAR)

Input: Data $(X, y) \in \mathbb{R}^{n\times d} \times \mathbb{R}^n$
Output: Private linear model $\hat{\beta} \in \mathbb{R}^d$

Define the Poisson-subsample process $i \in S \overset{i.i.d.}{\sim} \mathrm{Bern}(t)$ with $t = \Theta\left(\frac{d+\sqrt{d\log(n)}}{n}\right)$.

Set $\mathcal{T}$ to be the SOFT-ACRE stability test with parameters as defined in Section 1.2.4.

Use the private p-check algorithm given in Theorem A.5 to privately check if $\Pr[\mathcal{T}(S)] \ll 1$.
if $\Pr[\mathcal{T}(S)] \ll 1$ **then**
  return *REJECT*

Draw $k = \frac{\log(n)\log(1/\delta)}{\alpha^2}$ fresh Poisson-subsamples S_j, conditioned on $\mathcal{T}$ accepting.
For each S_j, draw a sample

$$v_j \leftarrow \mathcal{N}\left(\mathrm{OLS}(X_{S_j}, y_{S_j}),\, \frac{8R^2L}{(1-2L)^2}(X_{S_j}^\intercal X_{S_j})^{-1}\right).$$

return *the empirical mean* $\frac{1}{k}\sum_{j=1}^{k} v_j$

C.2 Utility Analysis

In this section we assemble our utility analysis for STAR.

C.2.1 Accuracy Bounds for Subsampled OLS

For the duration of this section, let $(X_1, y_1), \ldots, (X_n, y_n)$ be iid draws from the Gaussian linear model where $X \sim \mathcal{N}(0, I)$ and $y = \langle X, \beta_0 \rangle + \epsilon$, where $\epsilon \sim \mathcal{N}(0, \sigma^2)$ and $\beta_0 \in \mathbb{R}^d$. Let $t \in [0, 1]$ be a subsampling rate. For $S \subseteq [n]$, let $\hat{\beta}_S = (X_S X_S^\top)^{-1} X_S y_S$ be the (minimum norm) OLS regressor using only the samples contained in S. Suppose we draw random subsets $S_1, \ldots, S_k$, where each subset is defined by including each sample i independently with probability t, and let $\overline{\beta} = \frac{1}{k} \sum_{i \le k} \hat{\beta}_{S_i}$. Our main result in this section is:

Theorem C.1. *Let E be event that $|S_i| \ge tn/4$ for all $i \le k$, and assume $tn \ge Cd$ for a big-enough constant C. Then*

$$\mathbb{E}[\|\overline{\beta} - \beta_0\|^2 \mid E] \le O\left(\frac{d}{tnk} + \frac{d}{n}\right) \cdot \sigma^2 ,$$

where the expectation is taken with respect to $(X_1, y_1), \ldots, (X_n, y_n)$ and $S_1, \ldots, S_k$.

To prove the theorem we assemble a few lemmas.

Lemma C.2. *Let $S, T \subseteq [n]$ be two independent Bernoulli(t) subsamples and let $I = S \cap T$. Let $W_S \in \mathbb{R}^{n \times n}$ be the diagonal indicator matrix for membership in S and similarly for W_T. Let X have rows $X_1, \ldots, X_n$. Write*

$$M_S := X^\top W_S X = \sum_{i \in S} x_i x_i^\top , \qquad M_T := X^\top W_T X = \sum_{i \in T} x_i x_i^\top .$$

Then

$$\mathbb{E}_\epsilon \left[\left\langle \widehat{\beta}_S - \beta_0, \widehat{\beta}_T - \beta_0 \right\rangle \,\middle|\, X, S, T \right] = \sigma^2 \sum_{i \in I} x_i^\top M_S^{-1} M_T^{-1} x_i ,$$

where the expectation is taken only with respect to $\epsilon_1, \ldots, \epsilon_n$.

Proof. Expanding the formula for OLS,

$$\widehat{\beta}_S - \beta_0 = M_S^{-1} \sum_{i \in S} x_i \epsilon_i, \qquad \widehat{\beta}_T - \beta_0 = M_T^{-1} \sum_{i \in T} x_i \epsilon_i.$$

Now condition on (X, S, T) and take expectation over the Gaussian noise ϵ: only the indices in the overlap $I := S \cap T$ contribute, because ϵ_i's are independent. A short calculation gives

$$\mathbb{E}_\epsilon \left[\left\langle \widehat{\beta}_S - \beta_0, \widehat{\beta}_T - \beta_0 \right\rangle \,\middle|\, X, S, T \right] = \sigma^2 \sum_{i \in I} x_i^\top M_S^{-1} M_T^{-1} x_i$$

which proves the lemma. □

Lemma C.3. *Let $M_S = X_S X_S^T$ and $M_T = X_T X_T^\top$ be Wishart matrices, where $X_S \in \mathbb{R}^{d \times m_1}$ has independent $\mathcal{N}(0, 1)$ entries and similarly for $X_T \in \mathbb{R}^{d \times m_2}$, with $m_1, m_2 \ge Cd$ for a sufficiently large constant C. Let $x \in \mathbb{R}^d$ be $\mathcal{N}(0, I)$-distributed. (X_S, X_T, and x may not be independent of each other.) Then*

$$\mathbb{E}\, x^\top M_S^{-1} M_T^{-1} x \le O\left(\frac{d}{m_1 m_2}\right)$$

Proof. Using Cauchy-Schwarz, we have

$$\mathbb{E}\, x^\top M_S^{-1} M_T^{-1} x \le (\mathbb{E}\, \|x\|^4)^{1/2} (\mathbb{E}\, \|M_S^{-1}\|^4)^{1/4} (\mathbb{E}\, \|M_T^{-1}\|^4)^{1/4}$$

By Vershynin (2010)(Corollary 5.35), $\mathbb{E}\, \|M_S^{-1}\|^4 \le O(m_1^{-4})$. The lemma follows. □

Now we can prove Theorem C.1.

Proof of Theorem C.1. We expand the error of the subsampled estimator, eliding the conditioning on E for simplicity:

$$\mathbb{E}\,\|\overline{\beta}-\beta_0\|^2 = \frac{1}{k^2}\sum_{i\le k}\mathbb{E}\,\|\hat{\beta}_{S_i}-\beta_0\|^2 + \frac{1}{k^2}\sum_{i\ne j\le k}\mathbb{E}\langle\hat{\beta}_{S_i}-\beta_0,\hat{\beta}_{S_j}-\beta_0\rangle$$

In the first term, $\mathbb{E}\,\|\hat{\beta}_{S_i}-\beta_0\|^2$ is the squared error of an OLS estimator using $|S_i|$ samples, hence is at most $O(d/|S_i|)$. Since $|S_i|\ge tn/4$, this is at most $O(d/(tn))$, and so the entire first term is at most $O(\sigma^2 d/(tnk))$. For the second term, combining Lemmas C.2 and C.3, we get $O(d/(tn)^2)\cdot\sigma^2\cdot\mathbb{E}\,|S\cap T|$, where S and T are independent subsamples. This is $O(d/(tn)^2)\cdot\sigma^2\cdot t^2 n = O(\sigma^2 d/n)$. □

C.2.2 Analysis of ACRE test on Gaussian data

Our goal in this section is to prove Theorem C.4. For this section, let $(x_1,y_1),\ldots,(x_n,y_n)$ be draws from a linear Gaussian model with $n\ge Cd$ for a big-enough constant C.

Theorem C.4 (ACRE passes with high probability on well-behaved data)**.** *If X,y is a regression drawn according to Definition 1.2, and $S\subseteq[n]$ is drawn from a Poisson-t subsampling procedure with $tn = C(d+\sqrt{d\log(n)}\log(1/\delta))$ for a sufficiently large constant C, then X_S,y_S is (L,ℓ,R)-ACRE (for L,ℓ,R as defined in Section 1.2.4) with probability $1-O(n^{-100})-\exp(-\Omega(tn))$.*

Theorem C.4 follows from the following lemmas:

Conditioning on the subsample size. Let $S\subseteq[n]$ be drawn by Poisson-t subsampling and write $m:=|S|$. (Equivalently, $m\sim\mathrm{Binomial}(n,t)$ in the usual Poisson-subsampling sense.) A standard Chernoff bound gives

$$\Pr\big[m\notin\big[\tfrac{1}{2}tn,2tn\big]\big]\le 2\exp(-\Omega(tn)). \tag{32}$$

In the rest of this subsection we write $X:=X_S$ and $y:=y_S$, and we work on the event $m\in[\frac{1}{2}tn,2tn]$.

Lemma C.5. *With probability at least $1-n^{-100}-\exp(-\Omega(tn))$, every leverage score satisfies*

$$\max_{i\in S}\; x_i^\top(X^\top X)^{-1}x_i \;\le\; \frac{C\,(d+\sqrt{d\log n}+\log n)}{tn}\,,$$

for a sufficiently large universal constant C.

Proof. Condition on $m=|S|$ and define the whitened design

$$Z:=X\Sigma^{-1/2}\in\mathbb{R}^{m\times d}.$$

The rows of Z are i.i.d. $N(0,I)$. By Vershynin (2010)[Corollary 5.35], for $m\ge C_0 d$,

$$\left\|Z^\top Z-mI\right\|\le\frac{1}{2}m$$

with probability at least $1-2\exp(-\Omega(m))$. Equivalently,

$$\frac{1}{2}mI\preceq Z^\top Z\preceq\frac{3}{2}mI.$$

Multiplying on both sides by $\Sigma^{1/2}$ gives

$$\frac{1}{2}m\,\Sigma\preceq X^\top X=\Sigma^{1/2}Z^\top Z\Sigma^{1/2}\preceq\frac{3}{2}m\,\Sigma,$$

which implies the displayed constant-factor sandwich (after loosening $\frac{3}{2}$ to 2).

On this event,

$$(X^\top X)^{-1}\preceq\frac{2}{m}\,\Sigma^{-1}.$$

Hence for any $i\in S$,

$$x_i^\top(X^\top X)^{-1}x_i\;\le\;\frac{2}{m}\,x_i^\top\Sigma^{-1}x_i\;=\;\frac{2}{m}\left\|\Sigma^{-1/2}x_i\right\|^2.$$

Since $\Sigma^{-1/2}x_i \sim N(0, I)$, $\|\Sigma^{-1/2}x_i\|^2 \sim \chi^2_d$. By the Laurent–Massart tail bound, for any $u > 0$,

$$\Pr\Big[\chi^2_d \geq d + 2\sqrt{du} + 2u\Big] \leq e^{-u}.$$

Taking $u = 110 \log n$ and union bounding over $i \in [n]$ gives, with probability at least $1 - n^{-100}$,

$$\max_{i\in[n]} x_i^\top \Sigma^{-1} x_i \;\leq\; d + 2\sqrt{110\, d \log n} + 220 \log n.$$

Finally, on the size event $m \geq \frac{1}{2}tn$ from (32), we obtain

$$\max_{i\in S} x_i^\top (X^\top X)^{-1} x_i \;\leq\; \frac{4}{tn}\left(d + 2\sqrt{110\, d \log n} + 220 \log n\right) \;=\; O\!\left(\frac{d + \sqrt{d \log n} + \log n}{tn}\right),$$

as claimed. Combining failure probabilities yields $n^{-100} + \exp(-\Omega(tn))$. □

Lemma C.6. *With probability at least $1 - n^{-100} - \exp(-\Omega(tn))$, every cross-leverage satisfies*

$$\max_{\substack{i,j\in S\\ i\neq j}} \left|x_i^\top (X^\top X)^{-1} x_j\right| \;\leq\; \frac{C\,(\sqrt{d \log n} + \log n)}{tn},$$

for a sufficiently large universal constant C.

Proof. Fix distinct $i, j \in S$ and let $M_{-i} := X_{-i}^\top X_{-i}$ denote the Gram matrix with row i removed. By Sherman–Morrison for adding $x_i x_i^\top$,

$$(X^\top X)^{-1} = M_{-i}^{-1} - \frac{M_{-i}^{-1} x_i x_i^\top M_{-i}^{-1}}{1 + x_i^\top M_{-i}^{-1} x_i},$$

hence

$$x_i^\top (X^\top X)^{-1} x_j = \frac{x_i^\top M_{-i}^{-1} x_j}{1 + x_i^\top M_{-i}^{-1} x_i}, \qquad \text{so} \qquad \left|x_i^\top (X^\top X)^{-1} x_j\right| \leq \left|x_i^\top M_{-i}^{-1} x_j\right|.$$

Now condition on X_{-i} (equivalently on M_{-i}) and on x_j. Let $v := M_{-i}^{-1} x_j$. Since x_i is independent of (X_{-i}, x_j) and $x_i \sim N(0, \Sigma)$, we have

$$x_i^\top v \,\big|\, (X_{-i}, x_j) \sim N\big(0,\; v^\top \Sigma v\big).$$

We next upper bound $v^\top \Sigma v$. For each fixed i, M_{-i} is the sum of $(m-1)$ independent $N(0, \Sigma)$ outer products. By Vershynin (2010)[Corollary 5.35] (applied to $(m-1)$ samples),

$$\Pr\big[M_{-i} \not\succeq \tfrac{1}{2}(m-1)\Sigma\big] \leq 2\exp(-\Omega(m)).$$

Union bounding over $i \in S$ (at most n choices) keeps this at $\exp(-\Omega(m))$. On the event $M_{-i} \succeq \frac{1}{2}(m-1)\Sigma$, we have

$$\Sigma^{1/2} M_{-i}^{-1} \Sigma^{1/2} \preceq \frac{2}{m-1} I \preceq \frac{4}{m} I,$$

and thus

$$v^\top \Sigma v = \left\|\Sigma^{1/2} M_{-i}^{-1} x_j\right\|^2 = \left\|\Sigma^{1/2} M_{-i}^{-1} \Sigma^{1/2}\, \Sigma^{-1/2} x_j\right\|^2 \leq \left(\frac{4}{m}\right)^2 \left\|\Sigma^{-1/2} x_j\right\|^2 = \frac{16}{m^2}\, x_j^\top \Sigma^{-1} x_j.$$

From the same χ^2 union bound as in Lemma C.5, with probability at least $1 - n^{-100}$ we have simultaneously for all $j \in [n]$ that

$$x_j^\top \Sigma^{-1} x_j \leq B \quad \text{where} \quad B := d + 2\sqrt{110\, d \log n} + 220 \log n.$$

Therefore, on the intersection of these events,

$$\mathrm{Var}\big(x_i^\top M_{-i}^{-1} x_j \mid X_{-i}, x_j\big) \leq \frac{16B}{m^2}.$$

A Gaussian tail bound then gives

$$\Pr\left[\left|x_i^\top M_{-i}^{-1} x_j\right| \geq \frac{4\sqrt{B}}{m}\sqrt{220\log n} \;\middle|\; X_{-i}, x_j\right] \leq 2n^{-110}.$$

Union bounding this tail over all ordered pairs (i,j) with $i \neq j$ (at most n^2 pairs) gives overall failure probability $\leq n^{-100}$ for the cross-term bound.

Finally, on the size event $m \geq \frac{1}{2}tn$ from (32), the right-hand side is

$$O\left(\frac{\sqrt{B\log n}}{m}\right) = O\left(\frac{\sqrt{d\log n}+\log n}{m}\right) \leq O\left(\frac{\sqrt{d\log n}+\log n}{tn}\right),$$

proving the lemma. Combining failure probabilities yields $n^{-100} + \exp(-\Omega(tn))$. □

Lemma C.7. *With probability at least* $1 - n^{-100} - \exp(-\Omega(tn))$*, every residual satisfies*

$$\max_{i\in S} \left|y_i - (Hy)_i\right| \;\leq\; C\,\sigma\sqrt{\log n},$$

where $H := X(X^\top X)^{-1}X^\top$ *is the hat matrix for* $(X,y) = (X_S, y_S)$*.*

Proof. Work on the event from (32) that $m \in [\frac{1}{2}tn, 2tn]$, and on the conditioning event from Lemma C.5 that $X^\top X \succeq \frac{1}{2}m\,\Sigma$. This intersection fails with probability $\exp(-\Omega(tn))$, and on it $X^\top X$ is invertible so H is well-defined.

Write $y = X\beta^\star + \epsilon$ with $\epsilon \sim \mathcal{N}(0, \sigma^2 I_m)$ independent of X. Since $HX = X$, the residual vector is

$$r := y - Hy = (I - H)\epsilon.$$

Condition on X so H is fixed. Then r is mean-zero Gaussian with covariance $\sigma^2(I-H)$. For each coordinate $k \in [m]$,

$$\mathrm{Var}(r_k \mid X) = \sigma^2(I-H)_{kk} = \sigma^2(1 - H_{kk}) \leq \sigma^2,$$

because H is an orthogonal projector and thus $0 \leq H_{kk} \leq 1$. Therefore, for any $u > 0$,

$$\Pr\left[|r_k| \geq \sigma\sqrt{2u} \;\middle|\; X\right] \leq 2e^{-u}.$$

Taking $u = 110\log n$ and union bounding over $k \in [m]$ (and using $m \leq n$) gives

$$\Pr\left[\|r\|_\infty \geq \sigma\sqrt{220\log n} \;\middle|\; X\right] \leq 2me^{-110\log n} \leq 2n \cdot n^{-110} \leq n^{-100}.$$

Combining with the $\exp(-\Omega(tn))$ failure probability of the size/conditioning event completes the proof. □

C.2.3 Proofs of Lemmas from Section 1.2.4

Proof of Lemma 1.9. Fix $i \in [n]$ and let $X_{\neg i}, y_{\neg i}$ be the dataset with sample i removed. The new covariance matrix $X_{\neg i}^\intercal X_{\neg i} = X^\intercal X - X_i X_i^\intercal$ is a rank-1-update of the original covariance matrix. Therefore, updating the hat matrix using the Sherman-Morrison formula, we have

$$\begin{aligned}
\forall j,k \neq i\ (H_{\neg i})_{j,k} &= X_j^\intercal \left(X^\intercal X - X_i X_i^\intercal\right)^{-1} X_k = \\
&= X_j^\intercal \left((X^\intercal X)^{-1} + \frac{1}{1-H_{i,i}}\,(X^\intercal X)^{-1} X_i X_i^\intercal (X^\intercal X)^{-1}\right) X_k = \\
&= H_{j,k} + \frac{H_{i,j}H_{i,k}}{1-H_{i,i}}\,.
\end{aligned}$$

Immediately, we see that the right-hand-side is at most $\ell + \frac{\ell^2}{1-L}$ for $j \neq k$ and $L + \frac{\ell^2}{1-L} \leq L + \frac{\ell L}{1-L}$ for $j = k$, yielding the stability of the leverages and cross-leverages.

To see the stability of the residuals, fix $j \neq i$ and write

$$(H_{\neg i}y)_j = X_j^\intercal (X^\intercal X - X_i X_i^\intercal)^{-1}(X^\intercal y - X_i y_i).$$

Applying Sherman–Morrison and rearranging gives the standard leave-one-out identity

$$(H_{\neg i}y)_j = (Hy)_j - \frac{H_{j,i}}{1 - H_{i,i}}\,(y_i - (Hy)_i).$$

Therefore,

$$|y_j - (H_{\neg i}y)_j| \leq |y_j - (Hy)_j| + \frac{|H_{j,i}|}{1 - H_{i,i}}|y_i - (Hy)_i| \leq R + \frac{\ell}{1-L}R = (1+c)R\,.$$

□

Proof of Lemma 1.10. Lemma 1.10 follows from Lemma 1.9.

If X, y is a regression that has probability $> \eta$ of being accepted by soft-ACRE, then it must have $\textsc{score}(X, y) \leq 2\left(1 - \frac{2L}{1-2\ell}\right) \leq 2$, so X, y must be $(2L, 2R, 2\ell)$-ACRE. Therefore, from Lemma 1.9, for any leave-one-out $X_{\neg i}, y_{\neg i}$, we know that

$$\textsc{score}(X_{\neg i}, y_{\neg i}) \leq \left(1 + \frac{2L}{1-2\ell}\right)\textsc{score}(X, y)\,,$$

so it must be accepted by SOFT-ACRE with probability at least $\frac{1}{2}$ that of X, y. □

C.3 Proof of Theorem 1.3

Proof of Theorem 1.3. To prove Theorem 1.3, we need to show that the STAR algorithm is private and efficient and that it has high utility for well-behaved data.

Privacy The STAR algorithm utilizes the AC-STA framework. It relies on a down-sensitivity test $\mathcal{T}$, which we proved to be $(O(1), \exp(-\Omega(1/\ell)))$-down-stable (Lemma 1.10). We set the subsampling rate t to be of order $\frac{d+\sqrt{d\log(n)\log(1/\delta)}}{n}$, and this allows us to set

$$L = \Theta\left(\frac{d}{tn}\right) < \frac{1}{10} \quad \text{and} \quad \ell = \Theta\left(\frac{\sqrt{d\log(n)}}{tn}\right) = O\left(\frac{1}{\log(1/\delta)}\right)\,.$$

We select the constants to ensure that $\eta = \exp\left(-\Omega(1/\ell)\right) < \delta^2$.

From our assumption that $n \geq \frac{(1+\lambda)d\log(1/\delta)}{\varepsilon}$, we have $t = O\left(\frac{\varepsilon}{\log(1/\delta)}\right)$. From Theorem A.5, this allows us to privately verify that $\Pr\left[\mathcal{T}(S) = \mathsf{ACCEPT}\right] > p_0$, but because of our regime of t, we have $p_0 = \Omega(1)$, and in particular $p_0/\delta \gg \delta^2 \geq \eta$.

From Lemma 1.11, we know that passing SOFT-ACRE implies bounded down-sensitivity in the geometry of

$$E(X_S) = \frac{8LR^2}{(1-L)^2}(X_S^\intercal X_S)^{-1}\,.$$

By its definition, removing a sample can only cause $(X_S^\intercal X_S)^{-1}$ to grow in the PSD sense

$$E(X_S) \preceq E(X_{S\setminus\{i\}})\,,$$

and because passing SOFT-ACRE with non-zero probability implies that X_S, y_S is $(2L, 2R, 2\ell)$-ACRE, and in particular, that implies that X_S has bounded leverages $\leq 2L$, a standard identity (see e.g., Brown et al. (2023)) gives us

$$(1-2L)\left|E(X_{S\setminus\{i\}})\right| \leq |E(X_S)|\,.$$

Therefore, the assumptions of Lemma B.1 hold with $C_E = 1-2L = \Theta(1)$, implying that the output of the STAR algorithm is $(O(\sqrt{k\log(1/\delta)}t + kt^2), O\left(k\frac{n}{p_0} + \delta\right))$-DP. From our assumption that

$$n \geq \frac{(1+\lambda)d\log(1/\delta)\sqrt{\log(n)}}{\alpha\varepsilon_{\text{target}}}\ ,\ t = \Theta\left(\frac{d + \sqrt{d\log(n)}\log(1/\delta)}{n}\right)\ ,\ k = \frac{\log(n)\log(1/\delta)}{\alpha^2},$$

we get that the desired privacy of the algorithm

$$\varepsilon = \Theta(\sqrt{k\log(1/\delta)}t + kt^2) = \Theta\left(\frac{\log(1/\delta)\sqrt{\log(n)}}{\alpha} \times \frac{\alpha\varepsilon_{\text{target}}}{\log(1/\delta)\sqrt{\log(n)}}\right) = \Theta\left(\varepsilon_{\text{target}}\right)\ .$$

Utility The utility of the STAR algorithm for well-behaved data follows directly from a combination of Theorem C.1 (which says that the empirical mean over the OLS fits of k subsamples not conditioned on being ACRE is more than needed to get the desired excess risk α), and Theorem C.4 (which tells us that SOFT-ACRE will accept all $O(k/p_0) \ll n^{100}$ tested subsamples with high probability). From a union bound on the two failure events (that even a single ACRE test failed or that the mean over the empirical k subsamples was far from $\beta^\star$).

Finally, note that the STAR algorithm also added Gaussian noise. The covariance of this noise was

$$\text{Covariance} = \frac{\log(1/\delta)}{k^2}\sum_{j=1}^{k} E(X_{S_j}) = \frac{8LR^2\log(1/\delta)}{(1-L)^2k^2}\sum_{j=1}^{k}\left(X_{S_j}^\intercal X_{S_j}\right)^{-1}\ .$$

From Lemma C.5 and a union bound on the k subsamples, we know that

$$\text{Covariance} \preceq \Theta\left(\frac{d\log(n)\log(1/\delta)\sigma^2}{tn}\frac{\alpha^2}{\log(d)\log(1/\delta)}\right)\Sigma^{-1} = \Theta\left(\alpha^2\sigma^2\right)\Sigma^{-1}\ ,$$

so with high probability

$$\|\text{Added Noise}\|_\Sigma = \Theta(\alpha\sigma)\ .$$

Therefore, from triangle inequality, with high probability, the Mahalanobis-norm distance between the output of our algorithm β^{STAR}, and the ground-truth answer $\beta^\star$, is at most $O(\alpha\sigma)$, as desired.

Time Complexity Finally, we consider the time complexity of STAR. Because $p_0 = \Omega(1)$ from our assumption that $t = O\left(\frac{\varepsilon}{\log(1/\delta)}\right)$, the private p-check could be run for the cost of $\log(1/\delta) \ll n/d$ calls to the SOFT-ACRE oracle, and each such call was dominated by the cost of computing the hat matrix which was in turn dominated by a matrix multiplication between two $d \times tn = d \times \widetilde{O}\left(d + \lambda d\right)$ matrices.

Similarly, to get our $k = \widetilde{O}\left(n/d\right)$ fresh draws of a $\mathcal{T}$-passing subsample S_j, it suffices to run $\mathcal{T}$ on $O(k/p_0) = O(k)$ hypothesis subsamples. □

D Basic Mechanisms for Privacy

In this appendix, we will explore some basic primitives we use to ensure the privacy of certain intermediate values throughout the paper.

D.1 Soft-Truncation Primitives for Enforcing Boundedness

D.1.1 Differentially-Private soft-Threshold Test

Lemma D.1 (Differentially-Private soft-Threshold Test)**.** *For any $\varepsilon, \delta \in (0,1)$, there exists a randomized mechanism $\varphi_{\varepsilon,\delta}$ that takes as input a real number z and returns* ACCEPT *or REJECT such that:*

1. ***Privacy:*** *for any $|z - z'| < 1$,*

$$\varphi_{\varepsilon,\delta}(z) \approx_{\varepsilon,\delta} \varphi_{\varepsilon,\delta}(z')$$

2. ***Correctness:*** *almost surely:*

$$\forall z \leq 0 \; \varphi_{\varepsilon,\delta}(z) = \text{ACCEPT} \quad \text{and} \quad \forall z \geq \frac{2\log(2/\delta)}{\varepsilon} \; \varphi_{\varepsilon,\delta}(z) = \text{REJECT}$$

Proof of Lemma D.1. It is a classic result in DP that for a 1-sensitive variable z, adding truncated Laplace noise suffices to ensure its privacy. In particular, for $z_{\text{private}} = z + L$, where L is drawn from the truncated Laplace law

$$\Pr[L = \lambda] \propto \mathbb{1}_{|\lambda| \leq \tau} \exp(-\varepsilon|\lambda|) \; ,$$

is ε, δ approx-DP when $\tau = \frac{\log(2/\delta)}{\varepsilon}$.

Therefore, from the post-processing theorem, we know that

$$\varphi_{\varepsilon,\delta}(z) \coloneqq \mathbb{1}_{z_{\text{private}} \leq \tau}$$

is also private.

Finally, to ensure the correctness of φ, we note that $|L| < \tau$ almost surely, so if $z \leq 0$, then $z_{\text{private}} \leq \tau$ almost surely, and if $z \geq 2\tau$, then $z_{\text{private}} > \tau$ almost surely, completing our proof. □

D.1.2 Soft-Truncation Kernels

We use a soft-threshold acceptance function $\mathsf{Acc}(\cdot) \in [0, 1]$. Formally, each $\mathsf{Acc}(s)$ is the acceptance probability of an underlying randomized accept/reject rule; when writing $\mathsf{Acc}(s) \approx_{\varepsilon,\delta} \mathsf{Acc}(s')$ we mean that the induced Bernoulli laws on $\{\text{ACCEPT}, \text{REJECT}\}$ are (ε, δ)-approximately indistinguishable.

Definition D.2 (Power-decay and exponential-decay kernels)**.** Fix threshold $\tau > 0$ and floor $\zeta \in (0, 1)$.

1. (Power) For exponent $\alpha > 0$ and $s > 0$,

$$\mathsf{Acc}^{\text{pow}}_{\tau,\alpha,\zeta}(s) \coloneqq \Pr[\varphi_{\alpha,\zeta}(\log(s/\tau)) = \text{ACCEPT}] \, .$$

2. (Exponential) For slope $\gamma > 0$ and $s \geq 0$,

$$\mathsf{Acc}^{\text{exp}}_{\tau,\gamma,\zeta}(s) \coloneqq \Pr[\varphi_{\gamma,\zeta}(s - \tau) = \text{ACCEPT}] \, .$$

Lemma D.3 (Truncation implies global bounds)**.** *Fix $\tau > 0$ and $\zeta \in (0, 1)$.*

1. *(Power) For any $\alpha > 0$ and any $s > 0$,*

$$s \leq \tau \implies \mathsf{Acc}^{\text{pow}}_{\tau,\alpha,\zeta}(s) = 1, \qquad s \geq \tau \exp\left(2\frac{\log(2/\zeta)}{\alpha}\right) \implies \mathsf{Acc}^{\text{pow}}_{\tau,\alpha,\zeta}(s) = 0.$$

2. *(Exponential) For any $\gamma > 0$ and any $s \geq 0$,*

$$s \leq \tau \implies \mathsf{Acc}^{\text{exp}}_{\tau,\gamma,\zeta}(s) = 1, \qquad s \geq \tau + 2\frac{\log(2/\zeta)}{\gamma} \implies \mathsf{Acc}^{\text{exp}}_{\tau,\gamma,\zeta}(s) = 0.$$

Proof. For the power kernel, set $z \coloneqq \log(s/\tau)$. If $s \leq \tau$ then $z \leq 0$ and by Lemma D.1 we have $\varphi_{\alpha,\zeta}(z) = \text{ACCEPT}$ almost surely, hence $\mathsf{Acc}^{\text{pow}}_{\tau,\alpha,\zeta}(s) = 1$. If instead $s \geq \tau \exp\left(2\frac{\log(2/\zeta)}{\alpha}\right)$, then $z \geq 2\frac{\log(2/\zeta)}{\alpha}$ and Lemma D.1 gives $\varphi_{\alpha,\zeta}(z) = \text{REJECT}$ almost surely, hence $\mathsf{Acc}^{\text{pow}}_{\tau,\alpha,\zeta}(s) = 0$.

For the exponential kernel, set $z \coloneqq s - \tau$ and apply Lemma D.1 identically: if $s \leq \tau$ then $z \leq 0$ so acceptance is almost sure, while if $s \geq \tau + 2\frac{\log(2/\zeta)}{\gamma}$ then $z \geq 2\frac{\log(2/\zeta)}{\gamma}$ so rejection is almost sure. □

Lemma D.4 (Smoothness of Soft-Truncation)**.** *Fix $\tau > 0$ and $\zeta \in (0, 1)$, and define $k \coloneqq \lceil L \rceil$.*

1. *(Power; multiplicative smoothness) For any $s, s' > 0$ such that $e^{-L}s' \le s \le e^{L}s'$,*

$$\mathsf{Acc}^{\mathrm{pow}}_{\tau,\alpha,\zeta}(s) \approx_{\varepsilon_L,\delta_L} \mathsf{Acc}^{\mathrm{pow}}_{\tau,\alpha,\zeta}(s') \quad \textit{with} \quad \varepsilon_L := k\alpha, \;\; \delta_L := \zeta \sum_{i=0}^{k-1} e^{i\alpha} \le k e^{(k-1)\alpha}\zeta.$$

2. *(Exponential; additive smoothness) For any $s, s' \ge 0$ such that $s' - L \le s \le s' + L$,*

$$\mathsf{Acc}^{\mathrm{exp}}_{\tau,\gamma,\zeta}(s) \approx_{\varepsilon'_L,\delta'_L} \mathsf{Acc}^{\mathrm{exp}}_{\tau,\gamma,\zeta}(s') \quad \textit{with} \quad \varepsilon'_L := k\gamma, \;\; \delta'_L := \zeta \sum_{i=0}^{k-1} e^{i\gamma} \le k e^{(k-1)\gamma}\zeta.$$

Proof. We first note the following chaining bound for $\varphi_{\varepsilon,\delta}$. Let $z_0, \ldots, z_k$ satisfy $|z_{i+1} - z_i| < 1$ for all i. By Lemma D.1(1), for every measurable event S and each i,

$$\Pr\left[\varphi_{\varepsilon,\delta}(z_i) \in S\right] \le e^{\varepsilon} \Pr\left[\varphi_{\varepsilon,\delta}(z_{i+1}) \in S\right] + \delta.$$

Iterating this inequality yields

$$\Pr\left[\varphi_{\varepsilon,\delta}(z_0) \in S\right] \le e^{k\varepsilon} \Pr\left[\varphi_{\varepsilon,\delta}(z_k) \in S\right] + \delta \sum_{i=0}^{k-1} e^{i\varepsilon},$$

and the reverse direction follows by swapping the roles of z_0 and z_k. Hence $\varphi_{\varepsilon,\delta}(z_0) \approx_{k\varepsilon,\, \delta \sum_{i=0}^{k-1} e^{i\varepsilon}} \varphi_{\varepsilon,\delta}(z_k)$.

For the power kernel, let $z := \log(s/\tau)$ and $z' := \log(s'/\tau)$. The assumption $e^{-L}s' \le s \le e^{L}s'$ implies $|z - z'| = |\log(s/s')| \le L$. Define $k := \lceil L \rceil + 1$ and interpolate $z_i := z + \frac{i}{k}(z' - z)$ so that $|z_{i+1} - z_i| = |z' - z|/k < 1$. Applying the chaining bound with $(\varepsilon, \delta) = (\alpha, \zeta)$ gives

$$\varphi_{\alpha,\zeta}(z) \approx_{k\alpha,\, \zeta \sum_{i=0}^{k-1} e^{i\alpha}} \varphi_{\alpha,\zeta}(z').$$

Finally, by definition, $\mathsf{Acc}^{\mathrm{pow}}_{\tau,\alpha,\zeta}(s)$ and $\mathsf{Acc}^{\mathrm{pow}}_{\tau,\alpha,\zeta}(s')$ are the acceptance probabilities of these two accept/reject distributions, so the same $(\varepsilon_L, \delta_L)$-approximate indistinguishability holds for the induced Bernoulli laws.

The exponential case is identical with $z := s - \tau$ and $z' := s' - \tau$: the condition $|s - s'| \le L$ implies $|z - z'| \le L$, and chaining with $(\varepsilon, \delta) = (\gamma, \zeta)$ yields the stated parameters. □

D.2 PRIVATEBERNOULLITHRESHOLD

Let $\varphi_{\varepsilon,\delta}$ denote the mechanism from Lemma D.1.

Algorithm 7: PRIVATEBERNOULLITHRESHOLD$_{\kappa,\varepsilon,\delta}$ (PBT)

Input: Bits $b_1, \ldots, b_N \in \{0, 1\}$; sensitivity parameter $\kappa > 0$; privacy parameters $(\varepsilon, \delta) \in (0, 1)^2$; threshold $p_{\mathrm{th}} \in (0, 1)$.
Output: ACCEPT or REJECT.

$\hat{p} \leftarrow \dfrac{1 + \sum_{j=1}^{N} b_j}{N + 1}$; `// Laplace smoothing;` $\hat{p} \in (0, 1]$ `always`

$\hat{z} \leftarrow \dfrac{1}{\kappa} \log\left(\dfrac{\hat{p}}{p_{\mathrm{th}}}\right) + (2\log(2/\delta)/\varepsilon)$;

return $\varphi_{\varepsilon,\delta}(\hat{z})$;

Deterministic accept/reject regions. By Lemma D.1 applied to $\hat{z}$:

- If $\hat{z} \le 0$ then PBT outputs ACCEPT almost surely.
- If $\hat{z} \ge (2\log(2/\delta)/\varepsilon)$ then PBT outputs REJECT almost surely.

Equivalently,

$$\hat{p} \;\le\; p_{\mathrm{th}}\, e^{-\kappa(2\log(2/\delta)/\varepsilon)} \implies \text{ACCEPT a.s.}, \qquad \hat{p} \;\ge\; p_{\mathrm{th}} \implies \text{REJECT a.s.}$$

This is the "gap" that permits privacy: the mechanism can be randomized only when $\hat{p} \in \left(p_{\text{th}} e^{-\kappa(2\log(2/\delta)/\varepsilon)},\ p_{\text{th}}\right)$.

First, we will prove a useful monotonicity lemma that the accepted probability of PBT increases monotonically with p, when the bits are sampled from $\mathrm{Ber}(p)$

Lemma D.5 (PBT is monotone)**.** *Fix parameters $N \in \mathbb{N}$, $\kappa > 0$, $(\varepsilon, \delta) \in (0,1)^2$, and $p_{\text{th}} \in (0,1)$, and let $\mathrm{PBT}_{\kappa,\varepsilon,\delta}$ be Algorithm 7. Define, for $z \in \mathbb{R}$,*

$$f(z) \ := \ \Pr\left[\varphi_{\varepsilon,\delta}(z) = \textit{REJECT}\right],$$

where the probability is over the internal randomness of $\varphi_{\varepsilon,\delta}$.

Assume that f is nondecreasing, i.e.,

$$z \le z' \quad \Longrightarrow \quad f(z) \le f(z'). \tag{$\heartsuit$}$$

(For the threshold-type construction in Lemma D.1, this monotonicity holds.)

Then the following monotonicity properties hold.

1. ***Monotonicity in the dataset.*** *For any two bit-vectors $b, b' \in \{0,1\}^N$ with $\sum_{j=1}^N b_j \le \sum_{j=1}^N b'_j$ (in particular, if $b_j \le b'_j$ for all j),*

   $$\Pr\left[\mathrm{PBT}(b) = \textit{REJECT}\right] \ \le \ \Pr\left[\mathrm{PBT}(b') = \textit{REJECT}\right],$$

   and hence $\Pr[\mathrm{PBT}(b) = \textit{ACCEPT}] \ge \Pr[\mathrm{PBT}(b') = \textit{ACCEPT}]$.

2. ***Monotonicity in the Bernoulli mean.*** *Let $\mathcal{D}_p$ denote the output distribution of PBT when $b_1, \ldots, b_N \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p)$. If $p \le p'$, then*

   $$\mathcal{D}_p(\textit{REJECT}) \ \le \ \mathcal{D}_{p'}(\textit{REJECT}), \qquad \textit{equivalently} \qquad \mathcal{D}_p(\textit{ACCEPT}) \ \ge \ \mathcal{D}_{p'}(\textit{ACCEPT}).$$

Proof. Let $X(b) \coloneqq \sum_{j=1}^N b_j$. Algorithm 7 forms

$$\hat{p}(b) \ = \ \frac{1 + X(b)}{N+1}, \qquad \hat{z}(b) \ = \ \frac{1}{\kappa} \log\left(\frac{\hat{p}(b)}{p_{\text{th}}}\right) \ + \ (2\log(2/\delta)/\varepsilon)\,.$$

By inspecting the expressions, we have:

- $X(b) \le X(b') \implies \hat{p}(b) \le \hat{p}(b')$ (because $\hat{p}$ is affine in X with positive slope);
- $\hat{p}(b) \le \hat{p}(b') \implies \hat{z}(b) \le \hat{z}(b')$ (because $\log(\cdot)$ is increasing and $\kappa > 0$).

Hence, whenever $X(b) \le X(b')$, we have $\hat{z}(b) \le \hat{z}(b')$.

Now condition on the (fixed) dataset b. The only remaining randomness in $\mathrm{PBT}(b)$ comes from the call to $\varphi_{\varepsilon,\delta}$, so

$$\Pr\left[\mathrm{PBT}(b) = \mathsf{REJECT}\right] = \Pr\left[\varphi_{\varepsilon,\delta}(\hat{z}(b)) = \mathsf{REJECT}\right] = f(\hat{z}(b)).$$

Therefore, if $X(b) \le X(b')$, then $\hat{z}(b) \le \hat{z}(b')$, and by the monotonicity assumption ($\heartsuit$),

$$\Pr\left[\mathrm{PBT}(b) = \mathsf{REJECT}\right] = f(\hat{z}(b)) \ \le \ f(\hat{z}(b')) = \Pr\left[\mathrm{PBT}(b') = \mathsf{REJECT}\right],$$

which proves part (1). Since the output is binary, the ACCEPT inequality follows immediately.

For part (2), fix $p \le p'$ and couple the two product measures $\mathrm{Ber}(p)^{\otimes N}$ and $\mathrm{Ber}(p')^{\otimes N}$ as follows: let $U_1, \ldots, U_N \overset{\text{i.i.d.}}{\sim} \mathrm{Unif}[0,1]$ and set

$$b_j \coloneqq \mathbf{1}\{U_j \le p\}, \qquad b'_j \coloneqq \mathbf{1}\{U_j \le p'\}.$$

Then $b_j \le b'_j$ for all j, hence $X(b) \le X(b')$ almost surely. Applying part (1) pointwise (for each realization of $(U_1, \ldots, U_N)$) gives

$$\Pr\left[\mathrm{PBT}(b) = \mathsf{REJECT} \,\middle|\, U_1, \ldots, U_N\right] \ \le \ \Pr\left[\mathrm{PBT}(b') = \mathsf{REJECT} \,\middle|\, U_1, \ldots, U_N\right].$$

Taking expectations over the uniforms yields

$$\Pr_{\mathrm{Ber}(p)^{\otimes N}}\big[\mathrm{PBT} = \mathsf{REJECT}\big] \;\le\; \Pr_{\mathrm{Ber}(p')^{\otimes N}}\big[\mathrm{PBT} = \mathsf{REJECT}\big],$$

i.e. $\mathcal{D}_p(\mathsf{REJECT}) \le \mathcal{D}_{p'}(\mathsf{REJECT})$. Again, the ACCEPT inequality follows since the output is binary. □

Theorem D.6 (PBT: privacy, soundness, and completeness)**.** *Let $b_1, \ldots, b_N \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p)$ and let $\mathrm{PBT}_{\kappa,\varepsilon,\delta}$ be Algorithm 7 with threshold p_{th}. Fix parameters $\beta \in (0,1)$ and $\gamma \in (0, 1/2)$ with $\gamma \le \frac{\kappa}{2}$, and define*

$$p_{\mathrm{low}} \;:=\; p_{\mathrm{th}}\, e^{-\kappa(2\log(2/\delta)/\varepsilon)}\, e^{-\gamma}, \qquad p_{\mathrm{high}} \;:=\; p_{\mathrm{th}}\, e^{\gamma}.$$

Assume

$$N \;\ge\; \frac{12}{\gamma^2\, p_{\mathrm{low}}} \cdot \log \frac{4}{\beta}. \tag{$\star$}$$

1. ***Completeness (low mean).*** *If $p \le p_{\mathrm{low}}$, then* PBT *outputs* ACCEPT *with probability at least $1-\beta$.*
2. ***Soundness (high mean).*** *If $p \ge p_{\mathrm{high}}$, then* PBT *outputs* REJECT *with probability at least $1-\beta$.*
3. ***Privacy-style stability (log-sensitivity κ).*** *Let p, p' satisfy $e^{-\kappa}p \le p' \le e^{\kappa}p$ and also $p, p' \ge p_{\mathrm{low}}/2$. Let $\mathcal{D}_p$ denote the output distribution of* PBT *under $\mathrm{Ber}(p)^{\otimes N}$. Then*

$$\mathcal{D}_p \;\approx_{3\varepsilon,\, \delta_{\mathrm{priv}}}\; \mathcal{D}_{p'}, \qquad \textit{where} \qquad \delta_{\mathrm{priv}} \;:=\; (1 + e^{\varepsilon} + e^{2\varepsilon})\,\delta \;+\; (1 + e^{2\varepsilon})\,\beta.$$

Moreover, the sample size requirement $(\star)$ can be rewritten as

$$N \;=\; O\left(\frac{\exp\big(2\kappa\log(2/\delta)/\varepsilon\big)}{\gamma^2\, p_{\mathrm{th}}} \cdot \log\frac{1}{\beta}\right).$$

Proof. Write $\bar{p} := \frac{1}{N}\sum_{j=1}^{N} b_j$ and recall $\hat{p} = \frac{1+N\bar{p}}{N+1}$.

Step 1: Useful facts about $\hat{p}, \bar{p}$, and p For $p \in (0,1)$, write $\Pr_p[\cdot]$ for the probability over $b_1, \ldots, b_N \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p)$ and the internal randomness of $\varphi_{\varepsilon,\delta}$. Let

$$C \;:=\; (2\log(2/\delta)/\varepsilon)\,, \qquad t_{\mathrm{acc}} \;:=\; p_{\mathrm{th}} e^{-\kappa C}.$$

By the deterministic regions of Lemma D.1 applied to $\hat{z}$,

$$\hat{p} \le t_{\mathrm{acc}} \implies \text{PBT OUTPUTS } \mathsf{ACCEPT} \text{ a.s.}, \qquad \hat{p} \ge p_{\mathrm{th}} \implies \text{PBT OUTPUTS } \mathsf{REJECT} \text{ a.s.} \tag{33}$$

Moreover,

$$\hat{p} = \frac{1+N\bar{p}}{N+1} = \bar{p} + \frac{1-\bar{p}}{N+1} \qquad \implies \qquad \bar{p} \le \hat{p} \le \bar{p} + \frac{1}{N+1}. \tag{34}$$

In particular, for any threshold $t \in (0,1)$,

$$\Pr_p[\hat{p} > t] \le \Pr_p\left[\bar{p} > t - \frac{1}{N+1}\right], \qquad \Pr_p[\hat{p} < t] \le \Pr_p[\bar{p} < t]. \tag{35}$$

Finally, by Lemma D.5,

$$p \le p_0 \implies \Pr_p[\mathrm{PBT} = \mathsf{ACCEPT}] \ge \Pr_{p_0}[\mathrm{PBT} = \mathsf{ACCEPT}], \tag{36}$$

$$p \ge p_0 \implies \Pr_p[\mathrm{PBT} = \mathsf{REJECT}] \ge \Pr_{p_0}[\mathrm{PBT} = \mathsf{REJECT}] \tag{37}$$

Step 2: completeness. Fix any $p \le p_{\text{low}}$. By (36) with $p_0 = p_{\text{low}}$, it suffices to analyze $p = p_{\text{low}}$. Using (33) and (35) (with $t = t_{\text{acc}}$),

$$\begin{aligned}\Pr_{p_{\text{low}}}[\text{PBT OUTPUTS } \mathsf{ACCEPT}] &\ge \Pr_{p_{\text{low}}}[\hat{p} \le t_{\text{acc}}] = 1 - \Pr_{p_{\text{low}}}[\hat{p} > t_{\text{acc}}] \\ &\ge 1 - \Pr_{p_{\text{low}}}\left[\bar{p} > t_{\text{acc}} - \frac{1}{N+1}\right]. \end{aligned} \tag{38}$$

Since $p_{\text{low}} = t_{\text{acc}} e^{-\gamma}$, we have $t_{\text{acc}} = e^{\gamma} p_{\text{low}} \ge (1+\gamma) p_{\text{low}}$. Also, $(\star)$ implies $N p_{\text{low}} \ge \frac{12}{\gamma^2} \log \frac{4}{\beta} \ge \frac{2}{\gamma}$, hence $\frac{1}{N+1} \le \frac{1}{N} \le \frac{\gamma}{2} p_{\text{low}}$. Therefore

$$t_{\text{acc}} - \frac{1}{N+1} \;\ge\; \Big(1 + \frac{\gamma}{2}\Big) p_{\text{low}}.$$

Applying the multiplicative Chernoff upper-tail bound (with $\gamma/2 \in (0,1)$) gives

$$\Pr_{p_{\text{low}}}\left[\bar{p} > \Big(1 + \frac{\gamma}{2}\Big) p_{\text{low}}\right] \le \exp\left(-\frac{(\gamma/2)^2}{3}\, p_{\text{low}} N\right) = \exp\left(-\frac{\gamma^2}{12}\, p_{\text{low}} N\right) \le \frac{\beta}{4} \le \beta,$$

where the penultimate inequality uses $(\star)$. Plugging this into (38) yields $\Pr_{p_{\text{low}}}[\text{PBT OUTPUTS } \mathsf{ACCEPT}] \ge 1 - \beta$, and (36) implies the same for all $p \le p_{\text{low}}$.

Step 3: soundness. Fix any $p \ge p_{\text{high}}$. By (37) with $p_0 = p_{\text{high}}$, it suffices to analyze $p = p_{\text{high}}$. Using (33) and (35) (with $t = p_{\text{th}}$),

$$\begin{aligned}\Pr_{p_{\text{high}}}[\text{PBT OUTPUTS } \mathsf{REJECT}] &\ge \Pr_{p_{\text{high}}}[\hat{p} \ge p_{\text{th}}] = 1 - \Pr_{p_{\text{high}}}[\hat{p} < p_{\text{th}}] \\ &\ge 1 - \Pr_{p_{\text{high}}}[\bar{p} < p_{\text{th}}]. \end{aligned} \tag{39}$$

Since $p_{\text{high}} = p_{\text{th}} e^{\gamma}$, we have $p_{\text{th}} = e^{-\gamma} p_{\text{high}}$. Thus the multiplicative Chernoff lower-tail bound implies

$$\Pr_{p_{\text{high}}}[\bar{p} < p_{\text{th}}] = \Pr_{p_{\text{high}}}[\bar{p} < e^{-\gamma} p_{\text{high}}] \le \exp\left(-\frac{(1-e^{-\gamma})^2}{2}\, p_{\text{high}} N\right) \le \exp\left(-\frac{\gamma^2}{8}\, p_{\text{high}} N\right),$$

where we used $1 - e^{-\gamma} \ge \gamma/2$ for $\gamma \in (0, 1/2)$. Since $p_{\text{high}} \ge p_{\text{low}}$, assumption $(\star)$ yields

$$\exp\left(-\frac{\gamma^2}{8}\, p_{\text{high}} N\right) \le \exp\left(-\frac{\gamma^2}{8}\, p_{\text{low}} N\right) \le \exp\left(-\frac{12}{8} \log \frac{4}{\beta}\right) = \left(\frac{\beta}{4}\right)^{3/2} \le \beta.$$

Plugging into (39) yields $\Pr_{p_{\text{high}}}[\text{PBT OUTPUTS } \mathsf{REJECT}] \ge 1 - \beta$, and (37) implies the same for all $p \ge p_{\text{high}}$.

Step 4: privacy-style stability. Recall our assumptions that $e^{-\kappa} p \le p' \le e^{\kappa} p$ and $p, p' \ge p_{\text{low}}/2$. Let

$$\hat{z} \;=\; \frac{1}{\kappa} \log\left(\frac{\hat{p}}{p_{\text{th}}}\right) + C, \qquad z(p) \;:=\; \frac{1}{\kappa} \log\left(\frac{p}{p_{\text{th}}}\right) + C, \qquad C \;:=\; (2\log(2/\delta)/\varepsilon)\,.$$

Then

$$|z(p) - z(p')| = \frac{1}{\kappa}\, |\log(p/p')| \le 1. \tag{3}$$

Step 4(a): multiplicative concentration implies log concentration. Define the good event

$$\mathcal{G}(p) \;:=\; \left\{e^{-\gamma} p \le \bar{p} \le e^{\gamma} p\right\}.$$

For $\gamma \in (0, 1/2)$ we have $e^{\gamma} - 1 \ge \gamma$ and also $1 - e^{-\gamma} \;\ge\; \frac{2}{3}\gamma$.

By standard multiplicative Chernoff bounds for $\bar{p}$:

$$\Pr_{p}[\bar{p} > e^{\gamma} p] \le \exp\left(-\frac{(e^{\gamma}-1)^2}{3}\, pN\right) \le \exp\left(-\frac{\gamma^2}{3}\, pN\right),$$

$$\Pr_{p}\left[\bar{p} < e^{-\gamma} p\right] \le \exp\left(-\frac{(1-e^{-\gamma})^2}{2}\, pN\right) \le \exp\left(-\frac{2\gamma^2}{9}\, pN\right),$$

Therefore, for all p,

$$\Pr_p[\mathcal{G}(p)^c] \;\le\; \exp\left(-\frac{\gamma^2}{3}\,pN\right) + \exp\left(-\frac{2\gamma^2}{9}\,pN\right) \;\le\; 2\exp\left(-\frac{2\gamma^2}{9}\,pN\right) \tag{40}$$

On $\mathcal{G}(p)$ we have $|\log(\bar{p}/p)| \le \gamma$.

Step 4(b): a good event implies $|\hat{z} - z(p)| \le 1$. Fix p with $p \ge p_{\text{low}}/2$ and suppose we are in the good event $\mathcal{G}(p)$ from the previous step. We always have $\bar{p} \le \hat{p} \le \bar{p} + \frac{1}{N+1}$ from Step 1. Under $(\star)$ we showed in Step 2 that

$$\frac{1}{N+1} \le \frac{\gamma}{2} p_{\text{low}} \le \gamma p. \tag{41}$$

On $\mathcal{G}(p)$ we thus have

$$\hat{p} \;\le\; \bar{p} + \frac{1}{N+1} \;\le\; e^{\gamma}p + \gamma p \;=\; p\,(e^{\gamma} + \gamma) \;\le\; p\,e^{2\gamma},$$

where the last inequality holds for all $\gamma \in (0, 1/2)$ since $e^{2\gamma} - e^{\gamma} - \gamma \ge 0$. Also $\hat{p} \ge \bar{p} \ge e^{-\gamma}p$ on $\mathcal{G}(p)$. Therefore, on $\mathcal{G}(p)$,

$$e^{-\gamma} \;\le\; \frac{\hat{p}}{p} \;\le\; e^{2\gamma} \qquad \Longrightarrow \qquad \left|\log\left(\frac{\hat{p}}{p}\right)\right| \le 2\gamma.$$

Equivalently,

$$|\hat{z} - z(p)| = \frac{1}{\kappa}\left|\log\left(\frac{\hat{p}}{p}\right)\right| \le \frac{2\gamma}{\kappa} \le 1, \tag{42}$$

where we used the assumption $\gamma \le \kappa/2$.

By (40) and $p \ge p_{\text{low}}/2$,

$$\Pr_p[\mathcal{G}(p)^c] \le 2\exp\left(-\frac{2\gamma^2}{9}\,pN\right) \le 2\exp\left(-\frac{\gamma^2}{9}\,p_{\text{low}}N\right) \le 2\exp\left(-\frac{4}{3}\log\frac{4}{\beta}\right) = 2\left(\frac{\beta}{4}\right)^{4/3} \le \beta,$$

so we may summarize:

$$\Pr_p\left[\,|\hat{z} - z(p)| \le 1\,\right] \;\ge\; 1 - \beta, \qquad \Pr_{p'}\left[\,|\hat{z}' - z(p')| \le 1\,\right] \;\ge\; 1 - \beta. \tag{43}$$

Step 4(c): a "DP under a good event" claim. For $z \in \mathbb{R}$ and an event $\mathcal{S} \subseteq \{\mathsf{ACCEPT}, \mathsf{REJECT}\}$, write

$$q(z) \;:=\; \Pr\left[\varphi_{\varepsilon,\delta}(z) \in \mathcal{S}\right].$$

By Lemma D.1, if $|z - z'| \le 1$ then

$$q(z) \;\le\; e^{\varepsilon}q(z') + \delta \qquad \text{and} \qquad q(z') \;\le\; e^{\varepsilon}q(z) + \delta. \tag{44}$$

Claim. Let Z be any real-valued random variable and $z_0 \in \mathbb{R}$. If $\Pr[|Z - z_0| \le 1] \ge 1 - \beta$, then for every event $\mathcal{S}$,

$$\Pr[\varphi_{\varepsilon,\delta}(Z) \in \mathcal{S}] \;\le\; e^{\varepsilon}\Pr[\varphi_{\varepsilon,\delta}(z_0) \in \mathcal{S}] + \delta + \beta, \tag{45}$$

and also

$$\Pr[\varphi_{\varepsilon,\delta}(z_0) \in \mathcal{S}] \;\le\; e^{\varepsilon}\Pr[\varphi_{\varepsilon,\delta}(Z) \in \mathcal{S}] + \delta + \beta. \tag{46}$$

Proof of claim. Let $E = \{|Z - z_0| \le 1\}$. Then, using (44) pointwise on E and the trivial bound $q(Z) \le 1$,

$$\Pr[\varphi(Z) \in \mathcal{S}] = \mathbb{E}[q(Z)\mathbf{1}_E] + \mathbb{E}[q(Z)\mathbf{1}_{E^c}] \le \mathbb{E}[(e^{\varepsilon}q(z_0) + \delta)\mathbf{1}_E] + \Pr[E^c] \le e^{\varepsilon}q(z_0) + \delta + \beta,$$

which is (45). For (46), apply the reverse inequality in (44) pointwise on E: $q(z_0) \le e^{\varepsilon}q(Z) + \delta$, multiply by $\mathbf{1}_E$, take expectations, and add the term $q(z_0)\Pr[E^c] \le \Pr[E^c] \le \beta$.

Step 4(d): chaining the three comparisons. Let $\mathcal{D}_p$ denote the output distribution of PBT under $\mathrm{Ber}(p)^{\otimes N}$. For an event $\mathcal{S} \subseteq \{\mathsf{ACCEPT}, \mathsf{REJECT}\}$, we have

$$\mathcal{D}_p(\mathcal{S}) = \Pr_p[\varphi_{\varepsilon,\delta}(\hat{z}) \in \mathcal{S}].$$

Applying the claim with $Z = \hat{z}$ and $z_0 = z(p)$, and using (43), gives

$$\mathcal{D}_p(\mathcal{S}) \;\le\; e^{\varepsilon}\,\Pr[\varphi_{\varepsilon,\delta}(z(p)) \in \mathcal{S}] + \delta + \beta. \tag{47}$$

Next, since $|z(p) - z(p')| \le 1$ by (3), Lemma D.1 implies

$$\Pr[\varphi_{\varepsilon,\delta}(z(p)) \in \mathcal{S}] \;\le\; e^{\varepsilon}\,\Pr[\varphi_{\varepsilon,\delta}(z(p')) \in \mathcal{S}] + \delta. \tag{48}$$

Finally, applying the claim again (now in the reverse direction (46)) with $Z = \hat{z}'$ and $z_0 = z(p')$ yields

$$\Pr[\varphi_{\varepsilon,\delta}(z(p')) \in \mathcal{S}] \;\le\; e^{\varepsilon}\mathcal{D}_{p'}(\mathcal{S}) + \delta + \beta. \tag{49}$$

Combining (47), (48), and (49) gives

$$\mathcal{D}_p(\mathcal{S}) \;\le\; e^{3\varepsilon}\,\mathcal{D}_{p'}(\mathcal{S}) \;+\; (1 + e^{\varepsilon} + e^{2\varepsilon})\,\delta \;+\; (1 + e^{2\varepsilon})\,\beta.$$

Swapping the roles of p and p' yields the reverse inequality, hence $\mathcal{D}_p \approx_{3\varepsilon,\delta_{\text{priv}}} \mathcal{D}_{p'}$ with $\delta_{\text{priv}} = (1 + e^{\varepsilon} + e^{2\varepsilon})\delta + (1 + e^{2\varepsilon})\beta$. □

D.3 Privacy of Poisson

Lemma D.7 (Poisson rate change is (ε, δ)-close)**.** *Let $\delta \in (0, 1/4]$, $\varepsilon \in (0, 1]$, and $\lambda \ge 8\log(4/\delta)$. Fix $\alpha \in (0, 1/4]$ and $\lambda' := (1 - 2\alpha)\lambda$. If*

$$\alpha \;\le\; \frac{\varepsilon}{16\sqrt{3\lambda\log(4/\delta)}},$$

then

$$\text{Pois}(\lambda) \;\approx_{\varepsilon,\delta}\; \text{Pois}(\lambda').$$

Proof of Lemma D.7. Set $q := 1 - 2\alpha$ and write P for $\text{Pois}(\lambda)$ and Q for $\text{Pois}(\lambda')$. For $k \in \mathbb{N}$,

$$\log\frac{P(k)}{Q(k)} = (\lambda' - \lambda) + k\log(\lambda/\lambda') = -2\alpha\lambda + k\log(1/q).$$

Let $r := \sqrt{3\lambda\log(4/\delta)}$ and define the good sets

$$G := \{k : |k - \lambda| \le r\}, \qquad G' := \{k : |k - \lambda'| \le r\}.$$

A standard two-sided Poisson tail bound implies

$$P(G^c) \le \delta/2 \qquad \text{and} \qquad Q((G')^c) \le \delta/2,$$

using $\lambda \ge 8\log(4/\delta)$ and $\lambda' \ge \lambda/2$ (since $\alpha \le 1/4$).

Now take any event $T \subseteq \mathbb{N}$. Then

$$P(T) \le P(T \cap G) + \delta/2 \le \exp\Big(\sup_{k\in G}\log\frac{P(k)}{Q(k)}\Big)\,Q(T) + \delta/2.$$

For $k \in G$, write $k = \lambda + u$ with $|u| \le r$. Then

$$\log\frac{P(k)}{Q(k)} = \lambda\big(\log(1/q) - 2\alpha\big) + u\log(1/q).$$

For $\alpha \le 1/4$ we have $\log(1/q) = \log\big(\frac{1}{1-2\alpha}\big) \le 4\alpha$ and $\log(1/q) - 2\alpha \le 4\alpha^2$, hence

$$\sup_{k\in G}\log\frac{P(k)}{Q(k)} \le 4\alpha^2\lambda + 4\alpha r.$$

Under the stated bound on α and $\varepsilon \le 1$ (so that $4\alpha^2\lambda \le 4\alpha r \le \varepsilon/2$), we get $\sup_{k\in G}\log\frac{P(k)}{Q(k)} \le \varepsilon$, and therefore

$$P(T) \le e^{\varepsilon}Q(T) + \delta/2.$$

The reverse inequality $Q(T) \le e^{\varepsilon}P(T) + \delta/2$ is identical by swapping (λ, P, G) with (λ', Q, G') and using the same bound on the likelihood ratio. Combining the two directions gives $P \approx_{\varepsilon,\delta} Q$. □