Unbiased Statistical Estimation and Valid Confidence Intervals Under Differential Privacy

Christian Covington, Xi He, James Honaker, Gautam Kamath

Harvard University, University of Waterloo,

Anonym & Harvard University, University of Waterloo

Note that, throughout the supplement, we refer to theorems, definitions, algorithms, etc. via their corresponding number in the main text. Those which are hyperlinked and preceded by the letter S refer to items within the supplement.

S1 Definitions

S1.1 Differential Privacy

We begin with an introduction to the core definitions of DP.

Definition 1 (Neighboring data sets)**.** Let $\mathcal{X}$ be a data universe and $D, D' \in \mathcal{X}^n$. We say that D, D' are neighboring if

$$\max\left(|D \setminus D'|, |D' \setminus D|\right) = 1.$$

We also define the set of all neighboring data sets as

$$\mathcal{D}_n = \{(D, D') \in \mathcal{X}^n \times \mathcal{X}^n : D, D' \text{ are neighbors}\}.$$

Definition 2 (Rényi divergence (Rényi, 1961))**.** Let P, Q be probability measures over a measurable space (Ω, Σ). Then we define the α-Rényi divergence between P, Q as

$$H_\alpha(P\|Q) = \frac{1}{\alpha - 1} \ln \int_\Omega P(x)^\alpha Q(x)^{1-\alpha} dx.$$

Definition 3 (Global Function Sensitivity)**.** Let $\mathcal{X}$ be a data domain, $\gamma : \mathcal{X}^n \to \mathbb{R}^d$, and $\mathcal{D}_n$ be the set of all neighboring data sets as in Definition 1. Then we write the global sensitivity of γ with respect to a distance metric d as

$$GS_d(\mathcal{X}^n, \gamma) = \max_{D, D' \in \mathcal{D}_n} d\left(\gamma(D), \gamma(D')\right).$$

Algorithms can be made to respect DP in a variety of ways, but the most common way (as well as the approach we use in this work) is via an *additive noise mechanism*. This just entails running the algorithm as one would normally, and then adding random noise scaled relative to the algorithm's sensitivity.

Throughout this work, we use a popular additive noise mechanism called the *Gaussian mechanism*.

Lemma 1 (Gaussian Mechanism)**.** *Let $f : \mathcal{X}^n \to \mathbb{R}^d$ have global ℓ_2 sensitivity $GS_{\ell_2}(\mathcal{X}^n, f)$. Then the Gaussian mechanism*

$$\mathcal{M}_f(D) = f(D) + N\left(0, \left(\frac{GS_{\ell_2}(\mathcal{X}^n, f)}{\sqrt{2\rho}}\right)^2 I_d\right)$$

satisfies ρ-zCDP.

Note that it is often necessary to bound the data domain $\mathcal{X}$ to ensure that $GS_{\ell_2}(\mathcal{X}^n, f) < \infty$. For example, let $\mathcal{X} = \mathbb{R}^d$, $D = (D_1, \ldots, D_n)$ with $D_i \in \mathcal{X}$, and $f : \mathbb{R}^{n \times d} \to \mathbb{R}^d$ be such that $f(D) = n^{-1} \sum_{i=1}^n D_i$. If we let $D' = (\infty, D_2, \ldots, D_n)$, then D, D' are neighbors (they differ only in the first element), but $\|f(D) - f(D')\|_2 = \infty$. If instead $\mathcal{X} = [0, 1]^d$, then the D, D' that induce the largest difference in f are $D = (\vec{1}, D_2, \ldots, D_n)$ and $D' = (\vec{0}, D_2, \ldots, D_n)$. In this scenario, $\|f(D) - f(D')\|_2 = \|n^{-1}(\vec{1} - \vec{0})\|_2 = n^{-1}\sqrt{d}$, and thus $GS_{\ell_2}(\mathcal{X}^n, f) = n^{-1}\sqrt{d}$.

These bounds must be set without looking at the particular D_i, and are generally chosen by a data analyst based on public metadata and/or their beliefs about the data-generating process.

S1.2 Statistical Inference

This need to bound $\mathcal{X}$ introduces complications for doing statistical inference under DP, while maintaining the types of guarantees we often want from non-private estimators. We focus specifically on unbiased estimators

and valid confidence sets.

Definition 4 (Unbiased Estimator)**.** Let $\theta \in \mathbb{R}^d$ be a model parameter we wish to estimate. We collect data $D \sim \mathcal{D}$ and estimate θ with a random variable $\hat{\theta} : \mathcal{D} \to \mathbb{R}^m$. We say that $\hat{\theta}$ is an unbiased estimator of θ if $\mathbb{E}\left[\hat{\theta}(D)\right] = \theta$, with randomness taken over the sampling of $D \sim \mathcal{D}$, as well as any other randomness in $\hat{\theta}$.

Many applied statisticians, particularly those interested in estimating causal effects using linear models, prize unbiased parameter estimation and are willing to sacrifice on other fronts to achieve it. For example, the standard OLS estimator (which is the minimum-variance unbiased estimator under the assumptions of the Gauss-Markov theorem) is used for estimating parameters of a linear regression model in favor of other biased estimators, such as the James-Stein estimator (Stein, 1956; Stein and James, 1961), which dominate it in terms of ℓ_2 error of the parameter estimates.

Definition 5 (Confidence Set)**.** Let $\theta \in \mathbb{R}^d$ be a model parameter we wish to estimate using data $D \sim \mathcal{D}$. For arbitrary $\alpha \in [0, 1]$, a $(1 - \alpha)$-level confidence set for θ is a random set $S \subseteq \mathbb{R}^d$ such that

$$\mathbb{P}\left(\theta \in S\right) = 1 - \alpha,$$

with randomness taken from the sampling of D and any other randomness

in the construction of S.

Ideally, we would be able to find a perfectly-calibrated confidence set, where the coverage probability (i.e. $\mathbb{P}(\theta \in S)$) is exactly $1 - \alpha$. However, this is often impossible to compute exactly and so practitioners tend to default to being overly conservative instead. In this setting, we require $\mathbb{P}(\theta \in S) \geq 1 - \alpha$ and call such an S a *valid confidence set*. In this work, we focus on *confidence regions*, which are contiguous confidence sets, and occasionally *confidence intervals*, which are univariate confidence regions.

We can simplify the general problem of constructing confidence sets by restricting our attention to estimators whose sampling distribution belongs to a symmetric multivariate location-scale family.

Definition 6 (Location-Scale Family)**.** A set of probability distributions is a location-scale family if any density $f(x; \mu, \Sigma)$ in the set is written as $f(x; \mu, \Sigma) = c|\Sigma|^{-1/2} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$. for some normalization constant c.

Our restriction to location-scale families ensures that estimating the mean and (co)variance of the estimator is sufficient to characterize its distribution.

S2 Step 1: Bag of Little Bootstraps

This algorithm statement is adapted and simplified for our purposes; readers interested in the original version should consult Kleiner et al. (2014). We say that $\mathcal{X}$ is our data universe, $\mathcal{D}$ is a distribution over $\mathcal{X}$, and our realized data $X \in \mathbb{R}^{n \times m}$ are drawn from $\mathcal{D}^n$. For an arbitrary estimator $\hat{\theta} : \mathcal{X}^n \to \mathbb{R}^d$, we define $\hat{\theta}(\mathcal{D}) = \mathbb{E}_{X \sim \mathcal{D}^n}\left[\hat{\theta}(X)\right]$.

Algorithm 1 Bag of little bootstraps (BLB)

Input: data set $X \in \mathbb{R}^{n \times m}$, estimator $\hat{\theta} : \mathcal{X}^n \to \mathbb{R}^d$, estimator quality assessment ξ, k number of subsets of partition, r number of bootstrap simulations

Output: k estimates of $\hat{\theta}(\mathcal{D})$

procedure BLB($X, \hat{\theta}, k, r$)
  Randomly partition X into k subsets $\{X_i\}_{i \in [k]}$
  for $i \in [k]$ **do**
    $b = |X_i|$
    $\{\hat{\theta}_{i,c}\}_{c \in [r]} = \varnothing$
    for $c \in [r]$ **do**
      sample $(n_1, \ldots, n_b) \sim \text{Multinomial}(n, \mathbf{1}_b/b)$
      create $X_i^U \in \mathbb{R}^{n \times m}$ by including the j^{th} element of X_i n_j times
      $\hat{\theta}_{i,c} = \hat{\theta}(X_i^U)$
    $\hat{\theta}_i = \xi\left(\{\hat{\theta}_{i,c}\}_{c \in [r]}\right)$
  return $\{\hat{\theta}_i\}_{i \in [k]}$

S3 Step 2: Differentially Private Estimation

Definition 7. Let $\mathcal{B}(\mu, \Sigma)$ and $\mathcal{C}(\mu, \Sigma)$ be families of distributions and B, C be random variables drawn from each such that $\mathbb{E}(B) = \mathbb{E}(C) = \mu$ and $\text{Cov}(B) = \text{Cov}(C) = \Sigma$. Let PSD_d be the set of all $d \times d$ PSD matrices. We say that $\mathcal{B}$ is *heavier-tailed* than $\mathcal{C}$ if for all $\mu \in \mathbb{R}^d, \Sigma \in PSD_d,$ and $v \in \mathbb{R}^d$ such that $\|v\|_2 = 1$, then $\mathbb{P}\left[v^T(B - \mu) \leq z\right] \leq \mathbb{P}\left[v^T(C - \mu) \leq z\right]$ for all $z > 0$.

S3.1 Modified CoinPress Algorithm

Algorithm 2 Modified CoinPress

Input: $X = (x_1, \ldots, x_k)$ from a distribution D with mean μ and covariance Σ, $\tilde{\Sigma}$ such that $\Sigma \preceq \tilde{\Sigma}$, $B_2(\tilde{\mu}_0, r_0)$ containing μ, family of distributions $Q_X(\cdot, \Sigma_\mu)$ with heavier tails than D, number of iterations $t \in \mathbb{N}^+$, zCDP privacy loss parameter $\rho > 0$, failure probability $\beta > 0$

Output: t estimates of μ that jointly respect ρ-zCDP

procedure MVMRec$(X, \tilde{\mu}_0, r_0, \tilde{\Sigma}, Q, t, \rho, \beta)$

  $S = \tilde{\Sigma}^{1/2}$

  $\tilde{\mu}_0 = S^{-1}\tilde{\mu}_0$

  $r_0 = \max\left(\text{diag}\left(S^{-1}\right)\right) \cdot r_0$

  Define $\bar{X} \in \mathbb{R}^{k \times d}$ such that $\forall j \in [d], \forall m \in [k] : \bar{X}_{m,j} = \frac{1}{k}\sum_{m'=1}^{k} x_{m',j}$. Note that each row $\bar{X}_{m,:}$ is equal to the d-dimensional empirical mean of X

  $X' = \left(X - \bar{X}\right) S^{-1}$

  for $m \in [t-1]$ **do**

    $(\tilde{\mu}_m, r_m, \sigma_m) = \text{MVM}(X', \tilde{\mu}_{m-1}, r_{m-1}, Q_X(0, I_d), \frac{\rho}{2(t-1)}, \frac{\beta}{t})$ ▷ Algorithm 3

  $(\tilde{\mu}_t, r_t, \sigma_t) = \text{MVM}(X', \tilde{\mu}_{t-1}, r_{t-1}, Q_X(0, I_d), \frac{\rho}{2}, \frac{\beta}{t})$

  $\forall m \in [t] : \tilde{\mu}_m = (S\tilde{\mu}_m) + \bar{\mu}_{1,:}$ ▷ convert mean estimates to proper scale

  $\forall m \in [t] : \vec{\sigma}_m^2 = \text{diag}\left(S\sigma_m\right)^2$ ▷ convert private noise variances to proper scale

  return $\{(\tilde{\mu}_m, \vec{\sigma}_m^2)\}_{m \in [t]}$

S3.2 Modified CoinPress Algorithm - One Step Improvement

Algorithm 3 One Step Private Improvement of Mean Ball

Input: $x = (x_1, \ldots, x_k)$ from a distribution with mean 0 and covariance with smaller Löwner order than I_d, $B_2(\tilde{\mu}, r)$ containing 0, family of distributions $Q_X(\cdot, I_d)$, zCDP privacy loss parameter $\rho_m > 0$, failure probability $\beta_m > 0$

Output: A ρ_s-zCDP ball $B_2(\tilde{\mu}', r')$ and scale of the privatizing noise σ

procedure MVM($\hat{M}, \tilde{\mu}, r, Q_X, \rho_m, \beta_m$)

$\beta_s = \beta_m/2$

Let $R \sim Q_X(0, I_d)$

Set γ_1 such that $\mathbb{P}(\|R\|_2 > \gamma_1) \leq \frac{\beta_s}{k}$

Set γ_2 such that $\mathbb{P}(\|R\|_2 > \gamma_2) \leq \beta_s$

Project each x_i into $B_2(\tilde{\mu}, r + \gamma_1)$.

$\Delta = 2(r + \gamma_1)/k$.

$\sigma = \frac{\Delta}{\sqrt{2\rho_s}}$

Compute $\tilde{\mu}' = \frac{1}{k}\sum_i x_i + Y$, where $Y \sim \mathrm{N}\left(0, \sigma^2 I_d\right)$.

$r' = \gamma_2 \sqrt{\frac{1}{k} + \frac{2(r+\gamma_1)^2}{k^2 \rho_s}}$

return $(\tilde{\mu}', r', \sigma)$.

S3.3 Privacy Analysis of Algorithm 2

Theorem 1 (Modified CoinPress Privacy Statement)**.** *Algorithm 2 produces t estimates of μ that jointly respect ρ-zCDP.*

Proof. Algorithm 2 begins and ends by scaling the data to have empirical mean 0 and covariance which is Löwner upper bounded by I_d. The covariance scaling parameter is chosen independently of the data and the rest of the steps in the algorithm are invariant under location shift. So, our

privacy analysis rests on the application of Algorithm 3 in lines 8 and 9 of Algorithm 2.

Algorithm 3 interacts with the raw data only in line 9, so satisfying DP reduces to correct specification of Δ (the ℓ_2 sensitivity of the mean) and application of the Gaussian mechanism. The data are projected into $B_2(\tilde{\theta}, r+\gamma_1)$, and so the most a single data point can be changed in ℓ_2 norm is $2(r+\gamma_1)$. Because neighboring data sets X, Y differ in only one point (call it z), the ℓ_2 norm of the $k-1$ other points remains the same and so

$$\left\| \frac{1}{k} \sum_{x \in X} x - \frac{1}{k} \sum_{y \in Y} y \right\|_2 = \left\| \frac{1}{k} z \right\|_2 = \frac{1}{k} \|z\|_2 \leq \frac{2(r+\gamma_1)}{k}$$

as desired. Thus, each step of CoinPress satisfies zCDP at the stated level of its privacy parameter ρ. For each step $m \in [t-1]$, we see in line 8 that the privacy parameter is $\frac{\rho}{2(t-1)}$. For step t, we see in line 9 that the privacy parameter is $\frac{\rho}{2}$. Because zCDP parameters compose additively, the zCDP parameter for the entire CoinPress algorithm is $(t-1)\frac{\rho}{2(t-1)} + \frac{\rho}{2} = \rho$. □

S3.4 Proof of Theorem 5

Proof. We start with Assumption 3 so we have $\mu \in B_2(\tilde{\mu}_0, r_0)$. Note that the clipping bounds, parameterized by γ_1, in line 4 of Algorithm 3 are set such that, with probability $1-\beta_s$, no points are affected by the bounding; this follows because any given point is affected only if it falls outside the

clipping ball, which occurs with probability $\leq \frac{\beta_s}{k}$ and so, by the union bound, every point is unaffected with probability $\geq 1 - \beta_s$. Thus, with probability $\geq 1 - \beta_s$:

$$\begin{aligned}
\mu' &\sim \frac{1}{k}\sum_{i}^{k} \hat{\mu}_i + Y \\
&\sim \hat{\mu} + Y && \text{(definition of } \hat{\mu}\text{)} \\
&\sim \mathrm{N}\left(\hat{\mu}, \sigma^2 I_d\right).
\end{aligned}$$

We now consider γ_2, which is set as a $1-\beta_s$ probability upper bound on the ℓ_2 norm of the privatized mean of k draws from $Q(0, \tilde{\Sigma})$. Conditional on no points being clipped so that $\tilde{\mu}' = \sum_{i=1}^{k} \hat{\mu}_i + Y$, we have

$$1 - \beta_s \leq \mathbb{P}\left(\left\|\frac{1}{k}\sum_{i=1}^{k}\hat{\mu}_i - \mu + Y\right\|_2 \leq \gamma_2\right) \tag{S3.1}$$

$$= \mathbb{P}\left(\|\tilde{\mu}' - \mu\|_2 \leq \gamma_2\right). \tag{S3.2}$$

So, having $\mu \in B_2(\tilde{\mu}_0, r_0)$ implies that $\mathbb{P}\left(\mu \in B_2(\tilde{\mu}', r')\right) \geq 1 - 2\beta_s = 1 - \beta_m$. Using the fact that $\sum_m^t \beta_m = \beta^\mu$ and a union bound, we proceed by induction over the t steps of the algorithm and see that with probability $1 - \beta^\mu$ we have

$$\forall m \in [t] : \mu \in B_2(\tilde{\mu}_m, r_m)$$

and

$$\forall m \in [t] : \mu'_m \sim \mathrm{N}\left(\hat{\mu}, \sigma_m^2 I_d\right).$$

Scaling μ'_m, σ^2_m back up as in Lines 10 and 11 give the desired result. □

S3.5 Setting γ_1, γ_2

This section is concerned with how to set γ_1, γ_2 in lines 4, 5 of Algorithm 3 for various $Q_{\tilde{\mu}}$. We start with a general statement that works for arbitrary $Q_{\tilde{\mu}}$.

Fact 2 (Chebyshev's Inequality)**.** If X is a d-dimensional random vector with expected value $\mu = \mathbb{E}(X)$ and covariance $\Sigma = \mathbb{E}\left((X-\mu)(X-\mu)^T\right)$, then

$$\mathbb{P}\left(\sqrt{(X-\mu)^T\Sigma^{-1}(X-\mu)} > t\right) \leq \frac{d}{t^2},$$

provided that Σ is positive definite.

Corollary 1. *For any R in Algorithm 3,* $\mathbb{P}\left(\|R\|_2 > \sqrt{d/\beta}\right) \leq \beta$.

Proof. By construction of R, we know that $\mu = 0$ and $\Sigma = I_d$. Let R_j be the j^{th} element of R. Then we can write

$$\sqrt{(R-\mu)^T\Sigma^{-1}(R-\mu)} = \sqrt{R^TR} = \left(\sum_{j=1}^{d} R_j^2\right)^{1/2} = \|R\|_2$$

We can set $t = \sqrt{d/\beta}$ and rewrite Chebyshev's Inequality as

$$\mathbb{P}\left(\|R\|_2 > \sqrt{d/\beta}\right) \leq \beta.$$

□

In practice, it is beneficial to set tighter bounds based on the specified $Q_{\tilde{\mu}}$. This can hypothetically be done via Monte Carlo sampling and empirical CDF inequalities. However, this can be computationally expensive for γ_1 in particular, as you need at least k/β draws (and often far more) from the random variable to get a proper upper bound.

Some $Q_{\tilde{\mu}}$ also admit analytical bounds, which avoid the need for the costly computation. If $Q_{\tilde{\mu}}$ is multivariate Gaussian, we can use the following:

Fact 3 (Lemma 1 of Laurent and Massart (2000))**.** Let $Q_{\tilde{\mu}}$ be multivariate Gaussian such that $Q_{\tilde{\mu}}(\mu, \Sigma) = N(\mu, \Sigma)$. Then if $R \sim Q_{\tilde{\mu}}(0, I_d)$, we know that

$$\forall \beta \in (0, 1] : \mathbb{P}\left(\|R\|_2 > \sqrt{d + \sqrt{d \log(1/\beta)} + 2\log(1/\beta)} \right) \leq \beta.$$

We present a similar bound for when $Q_{\tilde{\mu}}$ is multivariate Laplace, based heavily on a result from Corollary 3.1 from Vladimirova et al. (2020).

Theorem 4. *Let $Q_{\tilde{\mu}}$ be multivariate Laplace with mean $\mu = 0$ and covariance $\Sigma = I_d$. Then if $R \sim Q_{\tilde{\mu}}(0, I_d)$, we know that*

$$\forall \beta \in (0, 1] : \mathbb{P}\left(\|R\|_2 > \sqrt{e \cdot d \log^2(\beta)} \right) \leq \beta,$$

where $e \approx 2.718$ is Euler's number.

Proof. We start by noting that $\|R\|_2 = \left(\sum_{j=1}^{d} R_j^2 \right)^{1/2}$. We know that the

R_j are Laplace with mean 0 and variance 1, and thus $\forall j \in [d] : R_j^2 \sim Weibull(\lambda = 1/2, k = 1/2)$. For ease of notation, we'll call $X_j = R_j^2$.

We now define sub-Weibull random variables, as is done in Vladimirova et al. (2020). We call a random variable X_j *sub-Weibull* with tail parameter θ if there exists $\theta, a, b > 0$ such that $\forall x > 0 : \mathbb{P}(|X_j| \geq x) \leq a \exp\left(-bx^{1/\theta}\right)$. For context, sub-Gaussian random variables are sub-Weibull with $\theta = 1/2$, sub-Exponentials are sub-Weibull with $\theta = 1$, and Weibull random variables themselves are sub-Weibull with $\theta = 2$.

We can state an alternative condition, also from Vladimirova et al. (2020), that X_j is sub-Weibull with tail parameter θ if $\exists c > 0$ s.t. $\forall t \geq 1 : \|X_j\|_t \leq ct^\theta$. Our goal is to find the smallest c that holds for Weibull random variables in particular. We recall that $X_j \sim Weibull(\lambda = 1/2, k = 1/2)$ and $\theta = 2$. Thus, for all $t \geq 1$:

$$\begin{aligned}
&\|X_j\|_t \leq ct^\theta \\
\iff &\left(\mathbb{E}\left(|X_j|^t\right)\right)^{1/t} \leq ct^\theta \\
\iff &\lambda\Gamma\left(\frac{t}{k}+1\right)^{1/t} \leq ct^\theta \qquad \text{(S3.3)} \\
\iff &\frac{1}{2t^2}\Gamma(2t+1)^{1/t} \leq c. \qquad \text{(S3.4)}
\end{aligned}$$

Line S3.3 follows by using the MGF of a Weibull random variable, and line S3.4 follows by plugging in the parameter values.

Our goal is to find the smallest c such that $\|X\|_t \leq ct^\theta$ for all $t \geq 1$. The lefthand side of line S3.4 is decreasing in t for $t \geq 1$, so finding the smallest possible c for $t = 1$ will be sufficient for all $t \geq 1$. Plugging in $t = 1$, we get $c = 1$.

We can finally appeal to Corollary 3.1 from Vladimirova et al. (2020), which states that if $X_1, \ldots, X_d$ are i.i.d. Weibull random variables with tail parameter θ, then for all $x \geq dK_\theta$ we have

$$\mathbb{P}\left(\left|\sum_{j=1}^{d} X_j\right| \geq x\right) \leq \exp\left(-\left(\frac{x}{K_\theta d}\right)^{1/\theta}\right)$$

for $K_\theta = ec$. Plugging in the $c = 1$ we found for Weibull random variables yields

$$\mathbb{P}\left(\left|\sum_{j=1}^{d} X_j\right| \geq x\right) \leq \exp\left(-\left(\frac{x}{e \cdot d}\right)^{1/\theta}\right).$$

We want the probability to be less than β, so we sub this in and get

$$\mathbb{P}\left(\left|\sum_{j=1}^{d} X_j\right| \geq e \cdot d \log^2(\beta)\right) \leq \beta.$$

We note that $\|X\|_2 = \sqrt{\left|\sum_{j=1}^{d} X_j\right|}$, so setting the bound at $\sqrt{e \cdot d \log^2(\beta)}$ gives our desired result. □

S3.6 Trick for setting $\tilde{\Sigma}$ for $\hat{\theta}^{BLB}$ estimation

In our GVDP algorithm, we independently estimate the means of both the $\{\hat{\theta}_i^{BLB}\}_{i\in[k]}$ and $\{\hat{\Sigma}_i^{BLB}\}_{i\in[k]}$, each requiring (among other things) that the

analyst specify $\tilde{\Sigma}$, a Löwner upper bound on the sample covariance of the BLB samples. If we estimate the mean of $\{\hat{\Sigma}_i^{BLB}\}_{i\in[k]}$ and do our post-processing to find a private covariance estimate $\tilde{\Sigma}$ prior to estimating the mean of $\{\hat{\theta}_i^{BLB}\}_{i\in[k]}$, we can actually leverage some extra information that will generally improve our estimates with a small cost to the theoretical guarantee. All the experimental results in the paper use this trick.

Although we are scaling up our subsets to the original data size within the BLB to get correct overall covariance estimates, this does not imply that the covariance of the $\{\hat{\theta}_i^{BLB}\}_{i\in[k]}$ match this correct scaling. In fact, this covariance will often be roughly the same as if $\hat{\theta}$ were simply run on subsets of size $\frac{n}{k}$. So, the covariance of the $\{\hat{\theta}_i^{BLB}\}_{i\in[k]}$ should be roughly $\frac{r(n/k)}{r(n)}\hat{\Sigma}$, where r is the convergence rate of the estimator in question. For example, the covariance of OLS coefficients decays with $\frac{1}{n}$, so if $\hat{\theta}$ represents OLS estimation we would say the covariance is $\frac{1/(n/k)}{1/n}\hat{\Sigma} = k\hat{\Sigma}$. We upper bound this with $k\tilde{\Sigma}$.

Under Assumption 1, this strategy gives us a $1-\beta^{\tilde{\Sigma}}$ probability guarantee that $k\tilde{\Sigma}$ will Löwner upper bound the empirical covariance of the $\{\hat{\theta}_i^{BLB}\}_{i\in[k]}$. So, by using $k\tilde{\Sigma}$ as the upper covariance bound for our mean estimation for $\{\hat{\theta}_i^{BLB}\}_{i\in[k]}$, we generally start with a pretty tight bound and can dramatically improve the accuracy of our estimates. This does lose a

bit of theoretical strength in the results; we generally assume that the analyst's upper bound is an actual upper bound on the empirical covariance with probability 1, whereas this trick provides a guarantee with probability $1 - \beta^{\tilde{\Sigma}}$.

S3.7 Generalizing CoinPress beyond multivariate sub-Gaussians

In Figure 1 we provide evidence that our generalization of CoinPress beyond sub-Gaussian distributions delivers on its promises. We pretend as if the estimator and data were such that the distributions induced by the BLB were are dominated by the multivariate Laplace (i.e. they are sub-Exponential), and the analyst overestimated the relevant parameters by a factor of 100. We show results corresponding to two different methods for calculating the clipping parameters at each step of CoinPress. The *analytic* solution calculates the bound using a theoretical bound given in Theorem 4, while the *approximate* solution calculates an approximate upper bound using Monte Carlo sampling.

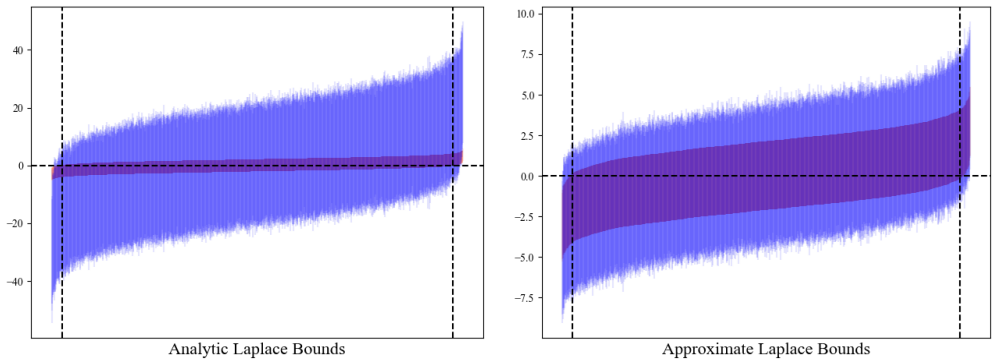


Figure 1: Distribution of coefficient estimates and 95% confidence intervals for $k = 5{,}000, d = 10, \rho = 0.1$ for multivariate Laplace distribution

S4 Step 3: Postprocessing

S4.1 Proof of Theorem 6

Proof. Our goal is to find weights $\{A_m\}_{m\in[t]}$ with $A_m \in \mathbb{R}^{d\times d}$ such that the Löwner order of $\mathrm{Cov}\left(\sum_{m=1}^{t} A_m\hat{\tau}_m\right)$ is minimized. Because we want our weighted estimator to remain unbiased, we restrict ourselves to sets of A_m such that $\sum_{m=1}^{t} A_m = I_d$.

We note that the A_m are constants and $\hat{\tau}_m$ are independent, so

$$\begin{aligned}\mathrm{Cov}\left(\sum_{m=1}^{t} A_m\hat{\tau}_m\right) &= \sum_{m=1}^{t} \mathrm{Cov}\left(A_m\hat{\tau}_m\right)\\ &= \sum_{m=1}^{t} A_m^T\mathrm{Cov}\left(\hat{\tau}_m\right)A_m.\end{aligned}$$

Assume $\hat{\tau}_m \in \mathbb{R}^d$ and let $\{B_m\}_{m\in[t]}$ with $B_m \in \mathbb{R}^{d\times d}$ be an arbitrary

weighting. Then we can write

$$\begin{aligned}
&\operatorname{Cov}\left(\sum_{m=1}^{t} A_m \hat{\tau}_m\right) \preceq \operatorname{Cov}\left(\sum_{m=1}^{t} B_m \hat{\tau}_m\right) \\
\iff &\forall v \in \mathbb{R}^d \setminus \{0\} : v^T \operatorname{Cov}\left(\sum_{m=1}^{t} A_m \hat{\tau}_m\right) v \leq v^T \operatorname{Cov}\left(\sum_{m=1}^{t} B_m \hat{\tau}_m\right) v.
\end{aligned}$$

Note that the quantities on the righthand side of the statement above are scalars, so we have translated the problem of finding a minimal Löwner bound into minimizing a one-dimensional quantity.

Let $v \in \mathbb{R}^d \setminus \{0\}$ be arbitrary. We now have a one-dimensional constrained optimization problem; we want to find $\{A_m\}_{m\in[t]}$ which minimizes $v^T \operatorname{Cov}\left(\sum_{m=1}^{t} A_m \hat{\tau}_m\right) v$ subject to $\sum_{m=1}^{t} A_m = I_d$. We can solve this using a Lagrange multiplier.

We write

$$\mathcal{L}\left(\{A_m\}_{m\in[t]}, \lambda\right) = v^T \operatorname{Cov}\left(\sum_{m=1}^{t} A_m \hat{\tau}_m\right) v - \lambda v^T \left(\sum_{m=1}^{t} A_m - I_d\right) v$$

and differentiate with respect to A_m. Recall that $\operatorname{Cov}(\hat{\tau}_m) = S_m$. Then we have

$$\begin{aligned}
\frac{\partial \mathcal{L}\left(\{A_m\}_{m\in[t]}, \lambda\right)}{\partial A_m} &= \frac{\partial v^T \operatorname{Cov}\left(\sum_{m=1}^{t} A_m \hat{\tau}_m\right) v - \lambda v^T \left(\sum_{m=1}^{t} A_m - I_d\right) v}{\partial A_m} \\
&= \frac{\partial \left(\sum_{m=1}^{t} v^T A_m^T \operatorname{Cov}(\hat{\tau}_m) A_m v\right) - \lambda v^T \left(\sum_{m=1}^{t} A_m - I_d\right) v}{\partial A_m} \\
&= S_m A_m v v^T + S_m^T A_m v v^T - \lambda v v^T \qquad \text{(S4.5)} \\
&= 2\left(S_m A_m - \lambda I_d\right) v v^T.
\end{aligned}$$

(S4.5) comes from a matrix calculus identity that for vectors a, b and matrix C all independent of X, $\frac{\partial (Xa)^T C (Xb)}{\partial X} = CXba^T + C^T Xab^T$ and noting that the partial with respect to A_m influences the sum only in the m^{th} term.

We set this to 0 to find a stationary point.

$$\begin{aligned} 0 &= 2\left(S_m A_m - \lambda I_d\right) vv^t \\ \lambda I_d vv^T &= S_m A_m vv^t \\ A_m &= \lambda S_m^{-1} I_d vv^T (vv^T)^{-1} \\ &= \lambda S_m^{-1}. \end{aligned}$$

We know from our constraint that $\sum_{m=1}^{t} A_m = I_d$, so

$$\begin{aligned} \sum_{m=1}^{t} \lambda S_m^{-1} &= I_d \\ \lambda &= \left(\sum_{m=1}^{t} S_m^{-1} \right)^{-1}, \end{aligned}$$

and thus our stationary point is achieved at $A_m = \left(\sum_{m=1}^{t} S_m^{-1}\right)^{-1} S_m^{-1}$.

We have shown that choosing A_m in this way achieves a stationary point, but we want to show that it is a global minimum. For that, we need to check the second partial derivative test, which states that our stationary point is a global minumum if $\frac{\partial^2 \mathcal{L}\left(\{A_m\}_{m\in[t]}, \lambda\right)}{\partial^2 A_m}$ is PD.

We first note that

$$\frac{\partial^2 \mathcal{L}\left(\{A_m\}_{m\in[t]}, \lambda\right)}{\partial^2 A_m} = \frac{\partial}{\partial A_m} 2\left(S_m A_m - \lambda I_d\right) vv^T$$
$$= 2(vv^T) \otimes S_m,$$

where $\otimes$ is the Kronecker product.

We know vv^T is PD, because $\forall z \in \mathbb{R}^d \backslash \{0\}$ we get $z^T vv^T z = (z^T v)(v^T z) = (v^T z)^T (v^T z) > 0$. The strict inequality comes because we know that both v and z are non-zero. We know S_m is PD by assumption and that, in general, if a matrix Y is PD then so is $2Y$. Finally, the Kronecker product of PD matrices is also PD, so $2(vv^T) \otimes S_m$ is PD and our second partial derivative condition is met. So $\mathcal{L}$ is convex and our local minimum is also a global minimum. Thus, our choice of A_m achieves the $\text{Cov}\left(\sum_{m=1}^t A_m \hat{\tau}_m\right)$ with minimal Löwner order. □

S4.2 Proof of Theorem 7

Proof. From Theorem 5, we know that, with probability $\geq 1 - \beta^{\tilde{\Sigma}}$:

$$\forall m \in [t] : \tilde{S}_m \sim \text{N}\left(\hat{S}^{BLB}, \vec{\sigma}^2_{\tilde{\Sigma},m} I_{d'}\right),$$

where $\hat{S}^{BLB}$ is the flattened form of $\hat{\Sigma}^{BLB}$. Assumption 1 then let's us substitute in $\hat{S}$ for $\hat{S}^{BLB}$. For the rest of the proof, we assume that this condition is met.

Thus, by Theorem 6, we know that a precision-weighted $\tilde{S}$ will have mean $\mathbb{E}[\tilde{S}] = \hat{S}$ and covariance $\text{Cov}\left(\tilde{S}\right) = \left(\Sigma_{m=1}^{t} \vec{\sigma}_{\tilde{\Sigma},m}^{-2} I_{d'}\right)^{-1}$ Moreover, this $\tilde{S}$ is itself multivariate Gaussian because it is a linear combination of multivariate Gaussians. That is, we can write

$$\tilde{S} \sim \text{N}\left(\hat{S}, \left(\sum_{m=1}^{t} \vec{\sigma}_{\tilde{\Sigma},m}^{2} I_{d'}\right)^{-1}\right) =: \text{N}\left(\hat{S}, \vec{\sigma}_{\tilde{S}}^{2} I_{d'}\right).$$

Let $\tilde{\Sigma}'$ be the unflattened matrix constructed from $\tilde{S}$. Then we can write $\tilde{\Sigma}'_{i,j} \sim N\left(\hat{\Sigma}_{i,j}, b_{i,j}^2\right)$, where $b_{i,j} = \text{unflatten}\left(\vec{\sigma}_{\tilde{S}}\right)_{i,j}$. Then, by Theorem 1.1 from Bandeira and Van Handel (2016), we know that

$$\mathbb{E}\|\tilde{\Sigma}' - \hat{\Sigma}\|_2 \leq (1+\epsilon)\left(2\max_{i\in[d]}\|b_{i,\cdot}\|_2 + \frac{6\sqrt{\log d}}{\log(1+\epsilon)} \max_{i,j\in[d]\times[d]} |b_{i,j}|\right),$$

for arbitrary $\epsilon \in (0, 1/2]$, where $\|\cdot\|_2$ is the spectral norm. Moreover, by Corollary 3.9 from Bandeira and Van Handel (2016) we have that, for any $t \geq 0$:

$$\|\tilde{\Sigma}' - \hat{\Sigma}\|_2 \leq (1+\epsilon)\left(2\max_{i\in[d]}\|b_{i,\cdot}\|_2 + \frac{6\sqrt{\log d}}{\log(1+\epsilon)} \max_{i,j\in[d]\times[d]} |b_{i,j}|\right) + t$$

with probability $\geq 1 - \exp\left(\frac{-t^2}{4\max_{i,j} b_{i,j}^2}\right)$. Setting $t = \sqrt{\frac{\ln(1/\beta^{ub})}{4\max_{i,j} b_{i,j}^2}}$ yields a $1 - \beta^{ub}$ probability bound.

Now define

$$c = \min_{\epsilon\in(0,1/2]} (1+\epsilon)\left(2\max_{i\in[d]}\|b_{i,\cdot}\|_2 + \frac{6\sqrt{\log d}}{\log(1+\epsilon)} \max_{i,j\in[d]\times[d]} |b_{i,j}|\right) + \sqrt{\frac{\ln(1/\beta^{ub})}{4\max_{i,j} b_{i,j}^2}}.$$

The spectral norm $\|\cdot\|_2$ of a matrix is its largest singular value (or equivalently, the square root of the absolute value of its largest magnitude eigenvalue). So, if $\|\tilde{\Sigma}' - \hat{\Sigma}\|_2 \leq c$, we know that the smallest eigenvalue of $\tilde{\Sigma}' - \hat{\Sigma}$ is necessarily at least $-c$. Therefore, the smallest eigenvalue of $\tilde{\Sigma}' + cI_d - \hat{\Sigma}$ is at least 0 or, equivalently, $\tilde{\Sigma} + cI_d \succeq \hat{\Sigma}$. This statement holds with probability $1 - \beta^{ub}$. Combining this with the initial $1 - \beta^{\tilde{\Sigma}}$ probability guarantee on the form of our estimator completes the proof. □

The statement for c simplifies significantly in the case where the $\{\tilde{\Sigma}_m\}_{m\in[t]}$ are diagonal matrices (which occurs if we care only about confidence intervals for each parameter rather than a joint confidence region). Let $q(p, \mu, \sigma^2) := \sqrt{2}\sigma \text{erf}^{-1}(2p-1) + \mu$ be the quantile function for a $N(\mu, \sigma^2)$ distribution where $\text{erf}^{-1}(\cdot)$ is the inverse error function.

Corollary 2. *Given diagonal covariance estimates and privacy variances* $\{\tilde{\Sigma}_m, \vec{\sigma}^2_{\Sigma,m}\}_{m\in[t]}$*, let* $\tilde{S}_m \in \mathbb{R}^{d'}$ *be the flattened version of* $\tilde{\Sigma}_m$*. We can construct a precision-weighted estimator* $\tilde{S}$*:* $\tilde{S} := \frac{\sum_{m=1}^t \tilde{S}_m/\vec{\sigma}^2_{\Sigma,m}}{\sum_{m=1}^t 1/\vec{\sigma}^2_{\Sigma,m}}$.

Let $\tilde{\Sigma}'$ *be the diagonal* $d \times d$ *matrix created by unflattening* $\tilde{S}$ *and* b *be the unflattened* $d \times d$ *diagonal matrix where* $b^2_{i,i} = Var\left(\tilde{\Sigma}'_{i,i}\right)$ *(i.e. the diagonal values of the covariance matrix of the flattened precision-weighted*

estimator). For $\beta^{ub} \in (0,1)$*, define* $\vec{c} = \{c_j\}_{j\in[d]}$ *where*

$$c_j = q\left(1 - \frac{\beta^{ub}}{d}, 0, b_{j,j}^2\right).$$

Then, for $\tilde{\Sigma} = \tilde{\Sigma}' + \vec{c}I_d$ *we have* $\mathbb{P}\left(\forall j \in [d] : \hat{\Sigma}_{j,j} \leq \tilde{\Sigma}_{j,j}\right) \geq 1 - \beta^{\tilde{\Sigma}} - \beta^{ub}$.

Proof. We start as in the proof in Section S4.2, but we know additionally that $\tilde{\Sigma}'_{i,j} = 0$ for $i \neq j$. We know that if our assumptions hold, which happens with probability $\geq 1 - \beta^{\tilde{\Sigma}}$, we can write $\tilde{\Sigma}'_{j,j} \sim N\left(\hat{\Sigma}_{j,j}, b_{j,j}^2\right)$ where $b_{j,j} = (\vec{\sigma}_{\tilde{S}})_j$.

By definition of the quantile function, we know then that, for arbitrary $j \in [d]$:

$$\begin{aligned}
1 - \frac{\beta^{ub}}{d} &= \mathbb{P}\left(\tilde{\Sigma}'_{j,j} \leq q\left(1 - \frac{\beta^{ub}}{d}, \hat{\Sigma}_{j,j}, b_{j,j}^2\right)\right) \\
&= \mathbb{P}\left(\tilde{\Sigma}'_{j,j} \leq \hat{\Sigma}_{j,j} + q\left(1 - \frac{\beta^{ub}}{d}, 0, b_{j,j}^2\right)\right) \\
&= \mathbb{P}\left(\tilde{\Sigma}'_{j,j} - c_j \leq \hat{\Sigma}_{j,j}\right) && \text{(definition of } c_j\text{)} \\
&= \mathbb{P}\left(\tilde{\Sigma}'_{j,j} + c_j \geq \hat{\Sigma}_{j,j}\right) && \text{(symmetry of the Gaussian)}.
\end{aligned}$$

Applying a union bound over the d failure probabilities and combining with the $1 - \beta^{\tilde{\Sigma}}$ probability that our required assumptions hold yields the desired result. □

S4.3 Proof of Theorem 8

Proof. From Theorem 5, we know that, with probability $\geq 1 - \beta^{\tilde{\theta}}$:

$$\forall m \in [t] : \tilde{\theta}_m \sim \mathrm{N}\left(\hat{\theta}^{BLB}, \vec{\theta}^2_{\tilde{\theta},m} I_{d'}\right),$$

where $\hat{S}$ is the flattened form of $\hat{\Sigma}$. Assumption 1 then let's us substitute in $\hat{\theta}$ for $\hat{\theta}^{BLB}$.

However, we opt to use the trick from Section S3.6 to avoid having to make Assumption 4. Theorem 8 gives us a covariance bound that, after appropriate scaling, satisfies Assumption 4 with probability $1 - \beta^{\tilde{\Sigma}} - \beta^{ub}$, so we fold this into our failure probability. Our result then follows directly from the precision-weighting procedure in Theorem 6. □

S5 Confidence Region/Intervals

S5.1 Proof of Theorem 9

Proof. We assume that our estimation of $\tilde{\theta}$ and $\tilde{\Sigma}$ worked as described at the top of Section 2.4, which comes with a $1 - \beta^{\tilde{\Sigma}} - \beta^{ub} - \beta^{\tilde{\theta}}$ probability guarantee. This means that $\mathbb{E}(\tilde{\theta}) = \mathbb{E}(\hat{\theta}) = \theta$ and $\tilde{\Sigma} \succeq \hat{\Sigma} \succeq \Sigma$ where our non-private estimator $\hat{\theta} \sim G(\theta, \Sigma)$. Furthermore, we assume that $Q_{\hat{\theta}}$ is heavier-tailed than G.

Summarizing this, we have a high-probability guarantee that three con-

ditions hold:

(1) $\mathbb{E}(Z) = \mathbb{E}(\hat{\theta}) = \theta$

(2) $\Sigma \preceq \tilde{\Sigma}$

(3) $Q_{\hat{\theta}}$ is heavier-tailed than G.

Under these conditions, Z and $\hat{\theta}$ have the same mean and $\hat{\theta}$ is more concentrated than Z, so a confidence region that is valid for Z is valid for $\hat{\theta}$. In other words,

$$\forall \alpha \in (0,1) : \mathbb{P}(Z \in C) \geq 1 - \alpha \implies \mathbb{P}(\theta \in C) \geq 1 - \alpha.$$

□

The result for confidence intervals, rather than a single region, follows trivially by noting that a confidence interval is a 1-dimensional confidence region.

Corollary 3 (Confidence Intervals (valid with high probability))**.** *Let Z be a d-dimensional random variable such that $Z \sim Q_{\hat{\theta}}\left(\hat{\theta} + N(0, \Sigma_{\tilde{\theta}}), \tilde{\Sigma}\right)$. Suppose $\{(ci_j^l, ci_j^u)\}_{j \in [d]}$ is a set of intervals such that*

$$\forall j \in [d] : \mathbb{P}\left[Z_j \in (ci_j^l, ci_j^u)\right] \geq 1 - \alpha_j,$$

for some $\{\alpha_j\}_{j \in [d]}$ with $\alpha_j \in (0,1)$. Then, with probability $1 - \beta^{\tilde{\Sigma}} - \beta^{ub} - \beta^{\tilde{\theta}}$,

$$\forall j \in [d] : \mathbb{P}\left[\hat{\theta}_j^{BLB} \in \left(ci_j^l, ci_j^u\right)\right] \geq 1 - \alpha_j.$$

Likewise, with probability $1 - \beta^{\tilde{\Sigma}} - \beta^{ub} - \beta^{\tilde{\theta}}$,

$$\mathbb{P}\left[\forall j \in [d] : \hat{\theta}_j^{BLB} \in \left(ci_j^l, ci_j^u\right)\right] \geq 1 - \sum_j^d \alpha_j.$$

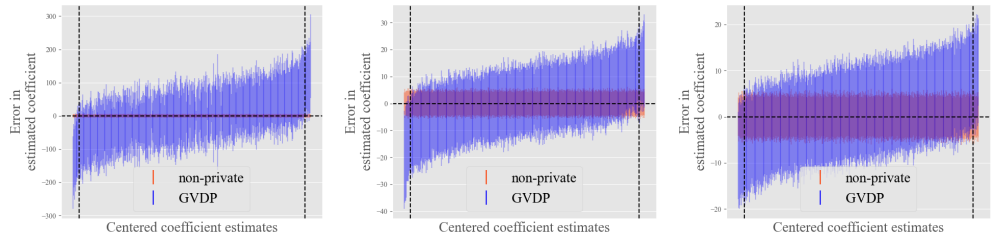
Proof. All but the last statement is a trivial application of Theorem 9 to the 1-dimensional case. The last statement follows via a union bound. □

S6 Empirical Results

S6.1 Logistic regression with imbalanced class output

In Figure 2 we show qualitatively similar results for a more challenging setting; logistic regression with imbalanced classes.

We generate data as we did for Figure 2 but run the outcome variable y through a scaled logistic function to get a new outcome variable $y' \in \{0,1\}^n$. Specifically, for $p_i = \frac{1}{1+\exp(-X_i\beta)}$ and $\bar{p} = \frac{1}{n}\sum_{i=1}^n p_i$, we have $\mathbb{P}(y_i = 1) = \frac{p_i}{\bar{p}} \cdot 0.05$. This induces a minority class that occurs with probability ≈ 0.05. Having such imbalanced classes introduces a practical problem in choosing a good k for GVDP. If $\frac{n}{k}$ is small, it becomes likely that we will see only the majority class in any given subset and the model will not be able to be fit. Thus, our experiments here have larger n than we used for the OLS experiments.



(a) $n = 10{,}000, k = 250$ (b) $n = 100{,}000, k = 1{,}000$ (c) $n = 100{,}000, k = 2{,}000$

Figure 2: Logistic Regression: Distribution of coefficient estimates and 95% confidence intervals

S6.2 Logistic regression with fully sparse data

The requisite bootstrap assumptions do not hold for all estimators and data distributions. We make our setting more difficult again for Figure 3, by making both the outcome and covariates a sparse binary vector/matrix respectively. It is unlikely that an analyst would want to use GVDP in this setting, because knowing that the data are binary (which we assume an analyst would know) immediately provides tight clipping bounds for the data. Nevertheless, we include it as an example because it's the most natural scenario we found where our method fails because of poor performance of the BLB.

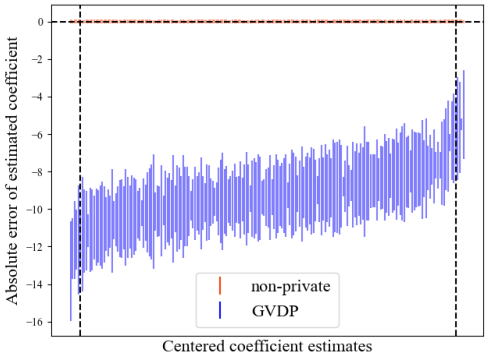
We create a new set of covariates X' such that $\forall j \in [d] : X'_{i,j} = \mathbb{1}(X_{i,j} \geq z_j)$ where $z_j = \min_{r \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_{i,j} > r) \leq 0.05$. That is, X' is itself now a binary matrix with highly imbalanced classes. The rightmost plot shows distributions of the d coefficient estimates induced by BLB, with a black

dotted line at the value of the non-private coefficient. Note that at $n =$ 100,000, the BLB distributions are essentially point masses at two extreme points, and our algorithm yields biased estimates and confidence intervals with insufficient coverage. For $n =$ 1,000,000, the BLB distributions are much closer to being a symmetric distribution about the true coefficient value, and our algorithm yields the promised guarantees.

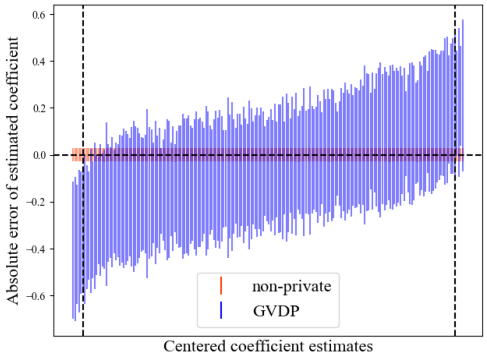
The left plots show poor confidence interval coverage because the BLB distribution is not a good approximation of the non-private sampling distribution. The right plots show better confidence interval coverage because the BLB approximation is successful.

S6.3 Explanation of Table 1

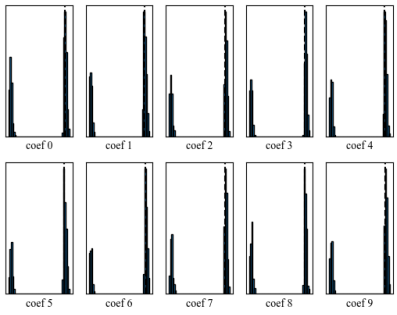
The GVDP and AdaSSP algorithms differ in a few key ways. First, AdaSSP does not attempt to do unbiased parameter estimation or give valid confidence intervals; instead, it is trying to estimate OLS coefficients with minimal ℓ_2 error. For purposes of comparison, we will ignore confidence intervals altogether and focus only on the parameter estimates. Second, AdaSSP assumes only bounds on the data, assuming that we can specify data domains $\mathcal{X}, \mathcal{Y}$ for our covariates and outcome, respectively, such that $\|\mathcal{X}\| = \sup_{x \in \mathcal{X}} \|x\|_2$ where $x \in \mathbb{R}^m$ and $\|\mathcal{Y}\| = \sup_{y \in \mathcal{Y}} |y|$. Ignoring the



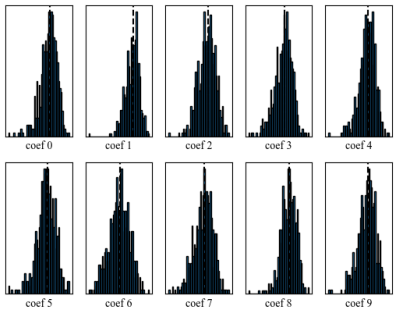
(a) $n = 100{,}000, k = 500, d = 10$



(b) $n = 1{,}000{,}000, k = 500, d = 10$



(c) BLB coefficient distributions for $n = 100{,}000, k = 500$



(d) BLB coefficient distributions for $n = 1{,}000{,}000, k = 500$

Figure 3: Logistic Regression with unbalanced binary fatures: Distribution of coefficient estimates and 95% confidence intervals

assumptions needed for confidence intervals for now, GVDP requires Assumptions 3, and 4 on the distribution of covariances induced by the bag of little bootstraps, as well as Assumption 3 on the distribution of means. It's worth noting the qualitative difference between these methods; AdaSSP requires the user to bound the data, GVDP requires the user to bound

moments of the parameter distribution. For most analyses, we expect the AdaSSP bounds to be easier to specify tightly than those of GVDP. However, GVDP is designed to scale more gracefully under overly conservative bounds.

We generate data just as we did for our OLS demonstration and compare AdaSSP and GVDP across a number of what we call "overestimation factors". Say we have realized data $X \in \mathbb{R}^{n \times d}, y \in \mathbb{R}^d$. For an overestimation factor of c, we set the bounds for AdaSSP to $c \cdot \sup_{x \in X} \|x\|_2$ and $c \cdot \sup_{y \in Y} \|y\|$. For GVDP, we perform the BLB step to get our $\{\hat{\theta}_i^{BLB}\}_{i \in [k]}$, which we'll say has empirical mean $\hat{\mu} \in \mathbb{R}^d$ and empirical covariance $\hat{\Sigma} \in \mathbb{R}^{d \times d}$. We set our ℓ_2 bounding ball for the mean of the distribution as $B_2\left(\hat{\mu}, c\left(\max_{j \in [d]} \hat{\mu}_j\right)\right)$ and our Löwner upper bound on the covariance as $c\left(\text{diag}(\hat{\Sigma}) I_d\right)$. Runs of non-private OLS are included for comparison, but the overestimation factor does not affect them.

The experiment in the body of the paper was run with $n = 500{,}000, k = 2{,}500, d = 10$, and $\rho = 0.1$. We run each method over 100 simulations, estimating d coefficients at each iteration, so each method produces 1,000 coefficient estimates overall.

Comparison with AdaSSP We again consider OLS, but now compare GVDP's performance to that of the Adaptive Sufficient Statistic Perturbation (AdaSSP)

OF	1	1.5	4	5	10	100	1000	10000
non-private	39.23	39.23	39.23	39.23	39.23	39.23	39.23	39.23
AdaSSP	40.31	49.66	1801.04	23395.30	71790.88	46382.71	82803.01	76377.38
GVDP	58.23	59.01	58.64	57.74	58.61	61.57	70.24	102.05

Table 1: Average ℓ_2 estimation error for each algorithm by overestimation factor (OF)

algorithm from Wang (2018), one of the best-performing algorithms for DP OLS. AdaSSP assumes bounds on the underlying data and attempts to estimate the OLS coefficients with minimal ℓ_2 error. We consider the performance of AdaSSP vs. GVDP as a function of the "overestimation factor" (OF), which is a multiplicative factor by which we overestimate the bounds of the data (for AdaSSP) or parameters (for GVDP).

Despite our presentation above, AdaSSP and GVDP can't really be directly compared because the OFs have qualitatively different meanings. However, the general comparison is still useful; AdaSSP performs well with slightly overestimated bounds but scales poorly with overly conservative bounds, while GVDP performs a bit less well at low overestimation factors but scales much better when the bounds are poorly set. More information can be found in Section S6.3.

S6.4 Replication of Card (1999)

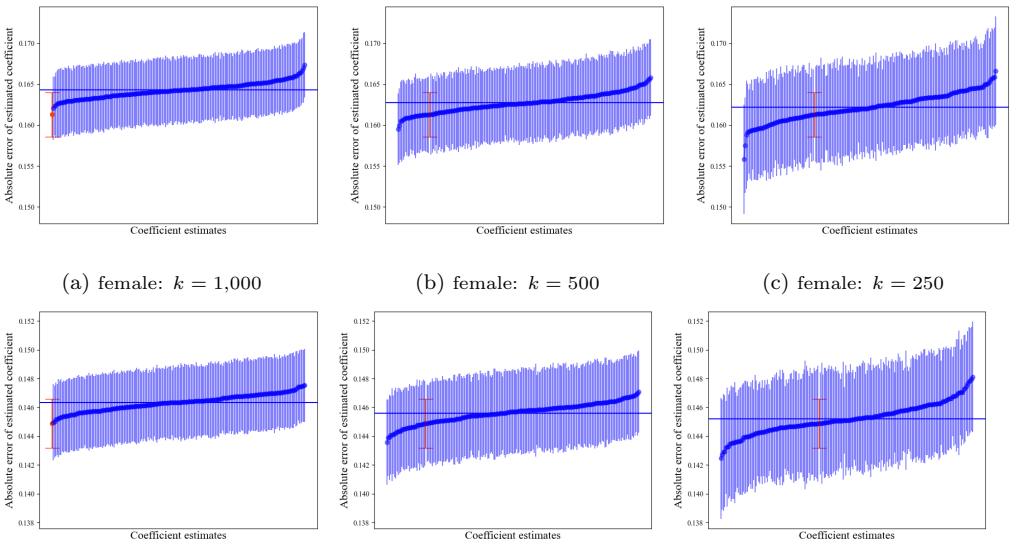
Inspired by the tests of Andrés F. Barrientos and Bowen (2024), we attempt to replicate the core analysis of Card (1999) under the constraints of DP. We use CPS ASEC data from 1994 to 1996 (Ruggles et al., 2021) and run OLS to estimate the following model:

$$\log(\text{inc\_wage}) = \beta_0 + \beta_1 \text{educ} + \beta_2 \text{PE} + \beta_3 \text{PE}^2 + \beta_4 \text{PE}^3 + \beta_5 \text{white} + \epsilon,$$

where *inc_wage* is an individual's total pre-tax wage and salary income, *educ* is years of education, *PE* is potential years of work experience, and *white* indicated whether or not the individual identifies as white. The effect of education on income is our question of interest, with the other variables serving as controls. This allows us to use the full OLS model within the bootstrap, but release and privately estimate only the estimated mean/variance of β_1.

Card (1999) runs this model separately for males and females; in Figure 4 we report the female results (n = 95,177) as well as for males and females combined (n = 197,756). To be consistent with Andrés F. Barrientos and Bowen (2024), we report results in *approximate DP* with $(\epsilon, \delta) = (5, 1/k)$, which translates to $\rho \approx \{0.879, 1.06, 1.23\}$ for $k = \{1000, 500, 250\}$. All results are run with $t = 5$ CoinPress iterations and an overestimation factor of 100, and we run the GVDP estimation algorithm 200 times to

show the long-run performance of the algorithm.



(a) female: $k = 1{,}000$ (b) female: $k = 500$ (c) female: $k = 250$

(d) combined: $k = 1{,}000$ (e) combined: $k = 500$ (f) combined: $k = 250$

Figure 4: Distribution of coefficient estimates and 95% confidence intervals for females only and both males and females (combined) with $\epsilon = 5$. The dot with a capped error bar represents the non-private estimate and confidence interval. The wider bars are the upper/lower bounds on the confidence intervals for the runs of GVDP. The horizontal line is the empirical mean of the GVDP estimates.

We note that our bootstrapped means do not always equal to the non-private mean in expectation, so although our algorithm's guarantees with respect to the bootstrapped distribution are met, they do not imply guarantees relative to the non-private answer as we hope they would. We see in these plots a clear trade-off. At $k = 1{,}000$, our confidence intervals are fairly tight, but are essentially centered around the upper end of the non-private confidence interval rather than the the true coefficient estimate. At

$k = 250$ we minimize bias by generating more a representative bootstrap distribution, but do so at the cost of wider confidence intervals.

S7 Notes on Assumptions and Analyst Choices

S7.1 Assumption 1

Assumption 1 essentially has three distinct pieces; we speak to the plausibility of each below.

First, we assume that the sampling distribution of the estimator is a member of a symmetric multivariate location-scale family, which we require because we want to be able to fully characterize the distribution by its mean and covariance. Given that X is of a reasonable sample size, we can appeal to the central limit theorem and argue that this assumption ought to hold. If the analyst cares only about confidence intervals for each element of the parameter, rather than a joint confidence region, it is sufficient for the marginal sampling distribution of each element in the parameter vector to belong to a symmetric univariate location-scale family.

Second, we assume that the BLB estimator in unbiased with respect to the estimand of interest in the non-private setting. The BLB shares many of the statistical properties of the traditional bootstrap, including

asymptotic consistency (as both $n \to \infty$ and $\frac{n}{k} \to \infty$), but also no finite-sample guarantee of unbiasedness. As such, this assumption may not hold in practice. However, if the BLB estimator exhibits low bias relative to the potential bias induced by poorly chosen clipping bounds, our method could still be an effective way to produce private estimates with lower bias than existing methods.

Third, we assume that the BLB estimates of the covariance are, with probability 1, a Löwner upper bound on the true covariance of the sampling distribution. This condition on $\hat{\Sigma}^{BLB}$ is onerous (especially in high dimensions) and seems unlikely to hold in general. In practice however, this condition can be dropped at the cost of a bit of extra fuzziness in the results. As stated above, we later make claims about our private estimator relative to $\tilde{\Sigma}^{BLB}$, which under Assumption 1 also hold relative to $\hat{\Sigma}$. This generalization to $\hat{\Sigma}$ is a higher bar than is typically set in applications of the bootstrap, where the bootstrap approximation is simply treated as a "good-enough" approximation of the sampling distribution. Moreover, if we care only about getting confidence intervals, rather than a confidence region, we can replace the Löwner condition with the condition that each element of the diagonal of $\hat{\Sigma}^{BLB}$ is at least as large as the corresponding element of $\hat{\Sigma}$.

S7.2 Assumption 2

Assumption 2 essentially follows from the first part of Assumption 1 where we assume the sampling distribution is from a symmetric location-scale family. If we can identify the location-scale family of the sampling distribution (again, we often appeal to the central limit theorem and say this is Multivariate Gaussian), then this same family trivially satisfies Assumption 2.

S7.3 Assumptions 3 and 4

Assumptions 3 and 4 state that the analyst can set bounds on the mean and covariance on the BLB estimates of both the mean and covariance of the sampling distribution. We believe that setting tight bounds would be very difficult in general, often more difficult than setting tight bounds on the data (the requirement we're trying to avoid). However, we suggest that the analyst aim to set very conservative bounds, unless they are very confident in their knowledge of the parameters. Because we use the CoinPress mean estimation algorithm to iteratively improve the bounds the analyst provides, the performance of the algorithm degrades slowly with more conservative bounds; e.g. see our results from Section 3 where the analyst's bounds are too conservative by a factor of 100 or Section S6.3 where we show performance when the analyst's bounds are too conservative by a factor of

10,000.

S7.4 Choosing k

Recall from our explanation of Algorithm 1 that k is the number of subsets into which we partition our original data, which in turn becomes the number of elements fed into our private mean estimation algorithm, Algorithm 2. This presents a trade-off for the user; when k is large, the sensitivity of our aggregator decreases (Line 7 of Algorithm 3) and thus so does the variance of the noise we need to add for privacy. On the other hand, we assume that the mean and covariance estimates we get from the BLB reasonably approximate the mean and covariance of the true sampling distribution of the parameters, which is provably true only as $n \to \infty$ and $\frac{n}{k} \to \infty$ for Hadamard differentiable estimators (Kleiner et al., 2014).

In the body of the paper, we argued that once $\frac{n}{k}$ is large enough that the BLB estimates have converged, there is no use in further increasing the ratio of n to k; we are better served by increasing k and reducing the noise needed for privacy. So, the best possible case for an analyst is that they choose the largest k such that the BLB estimates, operating over subsets of size $\frac{n}{k}$, approximates the true parameters of the sampling distribution.

As a final note, recall from Section 2.2 that we assume that n is public

knowledge or has been privately estimated. Thus, we can choose k in a way that depends on n with no extra privacy cost; if n is public knowledge then there's no dependence on the data at all and if it was privately estimated then this falls under the postprocessing property of zCDP.

Bibliography

Andrés F. Barrientos, Aaron R. Williams, J. S. and C. M. Bowen (2024). 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.

Bandeira, A. S. and R. Van Handel (2016). Sharp nonasymptotic bounds on the norm of random matrices with independent entries.

Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics 3*, 1801–1863.

Kleiner, A., A. Talwalkar, P. Sarkar, and M. I. Jordan (2014). A scalable bootstrap for massive data. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, 795–816.

Laurent, B. and P. Massart (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 1302–1338.

Rényi, A. (1961). On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*. The Regents of the University of California.

Ruggles, S., S. Flood, S. Foster, R. Goeken, J. Pacas, M. Schouweiler, and M. Sobek (2021). Ipums usa: Version 11.0 [dataset].

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. Proc. 3rd Berkeley Sympos. Math. Statist. Probability 1, 197-206 (1956).

Stein, C. and W. James (1961). Estimation with quadratic loss. In *Proc. 4th Berkeley Symp. Mathematical Statistics Probability*, Volume 1, pp. 361–379.

Vladimirova, M., S. Girard, H. Nguyen, and J. Arbel (2020, Jan). Sub-weibull distributions: Generalizing sub-gaussian and sub-exponential properties to heavier tailed distributions. *Stat 9*(1).

Wang, Y.-X. (2018). Revisiting differentially private linear regression: optimal and adaptive prediction & estimation in unbounded domain. In *Conference on Uncertainty in Artificial Intelligence*.