One Step to Efficient Synthetic Data

Jordan Awan* and Zhanrui Cai†

Purdue University, The University of Hong Kong*†

Supplementary Material

Section S1 includes a brief introduction to differential privacy. Section S2 gives details for the derivation of the privatized beta estimates of Section 6.3. All other proofs and technical details are provided in Section S3.

S1 Background on Differential Privacy

In this section, we review the basics of differential privacy (DP), which was proposed by Dwork et al. (2006) as a framework to mathematically quantify the degree of privacy protection. To satisfy differential privacy, a method must introduce additional randomness into the analysis, and the constraint of DP requires that for all possible databases, the change in one person's data does not significantly change the distribution of outputs. Consequently, having observed the DP output, an adversary cannot accurately determine the input value of any single person in the database. Definition

1 gives a formal definition of DP. In Definition 1, $h : \mathcal{X}^n \times \mathcal{X}^n \to \mathbb{Z}^{\geq 0}$ represents the *Hamming metric*, defined by $h(\underline{x}, \underline{x}') = \#\{i \mid x_i \neq x_i'\}$.

Definition 1 (Differential privacy: Dwork et al. (2006))**.** *Let the* privacy parameter $\epsilon > 0$ *and the sample size* $n \in \{1, 2, \ldots\}$ *be given. Let* $\mathcal{X}$ *be any set, and* $(\mathcal{Y}, \mathcal{S})$ *a measurable space. Let* $\mathcal{M} = \{M_{\underline{x}} \mid \underline{x} \in \mathcal{X}^n\}$ *be a set of probability measures on* $(\mathcal{Y}, \mathcal{S})$*, which we call a* mechanism*. We say that* $\mathcal{M}$ *satisfies* ϵ-differential privacy *(*ϵ*-DP) if* $M_{\underline{x}}(S) \leq e^{\epsilon} M_{\underline{x}'}(S)$ *for all* $S \in \mathcal{S}$ *and all* $\underline{x}, \underline{x}' \in \mathcal{X}^n$ *such that* $h(\underline{x}, \underline{x}') = 1$.

An important property of differential privacy is that it is invariant to post-processing. Applying any data-independent procedure to the output of a DP mechanism preserves ϵ-DP (Dwork et al., 2014, Proposition 2.1). Furthermore, Smith (2011) demonstrated that under conditions similar to (R1)-(R3), there exist efficient DP estimators for parametric models. Using these techniques, the one-step procedure can produce DP synthetic data by using a DP efficient statistic.

Remark 6. *Besides Definition 1, there are many other variations of differential privacy, the majority of which are relaxations of Definition 1, which also allow for efficient estimators. For instance, approximate DP (Dwork et al., 2006), concentrated DP (Dwork and Rothblum, 2016; Bun and Steinke, 2016), truncated-concentrated DP (Bun et al., 2018), Renyi*

DP (Mironov, 2017), and Gaussian DP (Dong et al., 2022) all allow for efficient estimators. On the other hand, local differential privacy (Kasiviswanathan et al., 2011; Duchi et al., 2013) in general does not permit efficient estimators and would not fit in our framework. For an axiomatic treatment of formal privacy, see Kifer and Lin (2012).

One of the earliest and simplest privacy mechanisms is the *Laplace mechanism.* Given a statistic T, the Laplace mechanism adds independent Laplace noise to each entry of the statistic, with scale parameter proportional to the *sensitivity* of the statistic. Informally, the sensitivity of T is the largest amount that T changes, when one person's data is changed in the dataset.

Proposition 1 (Sensitivity and Laplace Mechanism: Dwork et al. (2006))**.** *Let the privacy parameter $\epsilon > 0$ be given, and let $T : \mathcal{X}^n \to \mathbb{R}^p$ be a statistic. The ℓ_1-*sensitivity *of T is $\Delta_n(T) = \sup \| T(\underline{x}) - T(\underline{x}') \|_1$, where the supremum is over all $\underline{x}, \underline{x}' \in \mathcal{X}^n$ such that $h(\underline{x}, \underline{x}') = 1$. Provided that $\Delta_n(T)$ is finite, releasing the vector $\{T_j(\underline{x}) + L_j\}_{j=1}^p$ satisfies ϵ-DP, where $L_1, \ldots, L_p \overset{i.i.d.}{\sim} \text{Laplace}\,\{\Delta_n(T)/\epsilon\}$.*

S2 Deriving an Efficient DP Estimator for the Beta Distribution

We assume that $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Beta}(\alpha, \beta)$, where $\alpha, \beta \geq 1$, and our goal is to produce differentially private (DP) synthetic data. Recall that X_i takes values in $[0, 1]$ and has pdf $f_X(x) = x^{\alpha-1}(1-x)^{\beta-1}/B(\alpha, \beta)$, where B is the Beta function.

Often, to ensure finite sensitivity, the data are clamped to artificial bounds $[a, b]$, introducing bias in the DP estimate. Naive bounds are fixed in n, resulting in asymptotically negligible noise, but $O_p(1)$ bias. However, we show that it is possible to increase the bounds in n to produce both noise and bias of order $o_p(n^{-1/2})$, resulting in an efficient DP estimator. We show through simulations that using this estimator along with Algorithm 1 results in a DP sample with optimal asymptotics. While we work with the beta distribution, this approach may be of value for other exponential family distributions as well. We note that the asymptotics of clamping bounds have appeared in other DP works, but which are not immediately applicable to our setting (e.g., Smith, 2011; Kamath et al., 2020).

Recall that $n^{-1} \sum_{i=1}^n \log(X_i)$ and $n^{-1} \sum_{i=1}^n \log(1 - X_i)$ are sufficient statistics for the beta distribution. We will add Laplace noise to each of these statistics to achieve differential privacy. However, the sensitivity of these quantities is unbounded. First we pre-process the data by setting $\widetilde{X}_i = \min\{\max(X_i, t), 1-t\}$, where t is a threshold that depends on n. Then the ℓ_1-sensitivity of the pair of sufficient statistics is $\Delta(t) = 2n^{-1}\left|\log(t) - \log(1-t)\right|$. We add independent noise to each of the statistics from the distribution Laplace$\{\Delta(t)/\epsilon\}$, which results in ϵ-DP versions of these statistics. Finally, we estimate $\theta = (\alpha, \beta)$ by plugging in the privatized sufficient statistics into the log-likelihood function and maximizing over θ. The resulting parameter estimate satisfies ϵ-DP by post-processing.

We must carefully choose the threshold t to ensure that the resulting estimate is efficient. The choice of t must satisfy $\Delta(t) = o(n^{-1/2})$ to ensure that the noise does not affect the asymptotics of the likelihood function. We also require that both $P(X_i < t) = o(n^{-1/2})$, and $P(X_i > 1-t) = o(n^{-1/2})$ to ensure that $\widetilde{X}_i = X_i + o_p(n^{-1/2})$, which limits the bias to $o_p(n^{-1/2})$. For the beta distribution, we can calculate that $P(X_i < t) = O(t^\alpha)$ and $P(X_i > 1-t) = O(t^\beta)$. Since we assume that $\alpha, \beta \geq 1$, so long as $t = o(n^{-1/2})$ the probability bounds will hold. Taking $t = \min[1/2, 10/\{\log(n)\sqrt{n}\}]$ satisfies $t = o(n^{-1/2})$, and we estimate the sensitivity as

$$\Delta(t) \leq 2n^{-1}\log(t^{-1}) \leq 2n^{-1}\log\{\log(n)\sqrt{n}\} = O\{\log(n)/n\} = o(n^{-1/2}),$$

which satisfies our requirement for Δ. While there are other choices of t

which would satisfy the requirements, our threshold was chosen to optimize the finite sample performance, so that the asymptotics could be illustrated with smaller sample sizes.

S3 Proofs and Technical Lemmas

For two distributions P and Q on $\mathbb{R}^k$, the *Kolmogorov-Smirnov distance* (KS-distance) is $\mathrm{KS}(P,Q) = \sup_{R \text{ rectangle}} |P(R) - Q(R)|$, where the supremum is over all axis-aligned rectangles. If F_P and F_Q are the multivariate cdfs of P and Q, then $\|F_P - F_Q\|_\infty \leq \mathrm{KS}(P,Q) \leq 2^k \|F_P - F_Q\|_\infty$, so convergence in distribution is equivalent to convergence in KS-distance (Smith, 2011). By definition, we have that $\mathrm{TV}(P,Q) \geq \mathrm{KS}(P,Q)$.

Proof of Theorem 1. First we will establish the asymptotic distribution of $\hat\theta_Z$. Recall that by efficiency, we know that $\sqrt{n}(\hat\theta_X - \theta) \overset{d}{\to} N\{0, I^{-1}(\theta)\}$ and $\sqrt{n}\{\hat\theta_Z - E(\hat\theta_Z \mid \hat\theta_X)\} \mid \hat\theta_X \overset{d}{\to} N\{0, I^{-1}(\hat\theta_X)\}$. Then by Slutsky's theorem, we have that $\sqrt{n}\{\hat\theta_Z - E(\hat\theta_Z \mid \hat\theta_X)\} \overset{d}{\to} N\{0, I^{-1}(\theta)\}$. We can easily compute that $\mathrm{Cov}\{\hat\theta_Z - E(\hat\theta_Z \mid \hat\theta_X), \hat\theta_X\} = 0$ using the law of total covariance. So, we have that $\sqrt{n}\{\hat\theta_Z - E(\hat\theta_Z \mid \hat\theta_X) + \hat\theta_X - \theta\} \overset{d}{\to} N\{0, 2I^{-1}(\theta)\}$. We also know that $\sqrt{n}\{E(\hat\theta_Z \mid \hat\theta_X) - \hat\theta_X\} = o_p(1)$, since $E(\hat\theta_Z \mid \hat\theta_X) = \hat\theta_X + o_p(n^{-1/2})$. Together, we have that $\sqrt{n}(\hat\theta_Z - \theta) \overset{d}{\to} N\{0, 2I^{-1}(\theta)\}$.

$$\begin{align}
\mathrm{TV}\left(\underline{X},\underline{Z}\right) &\geq \mathrm{TV}\left\{\sqrt{n}(\hat{\theta}_X-\theta),\sqrt{n}(\hat{\theta}_X-\theta)\right\} \tag{S3.1}\\
&\geq \mathrm{KS}\left\{\sqrt{n}(\hat{\theta}_X-\theta),\sqrt{n}(\hat{\theta}_X-\theta)\right\} \tag{S3.2}\\
&\geq \mathrm{KS}\left[N\{0,I^{-1}(\theta)\},N\{0,2I^{-1}(\theta)\}\right] \tag{S3.3}\\
&\quad -\mathrm{KS}\left[\sqrt{n}(\hat{\theta}_X-\theta),N\{0,I^{-1}(\theta)\}\right] \tag{S3.4}\\
&\quad -\mathrm{KS}\left[\sqrt{n}(\hat{\theta}_Z-\theta),N\{0,2I^{-1}(\theta)\}\right] \tag{S3.5}\\
&= \mathrm{KS}\left[N\{0,I^{-1}(\theta)\},N\{0,2I^{-1}(\theta)\}\right]+o(1) \tag{S3.6}\\
&\geq \Phi\left\{-\sqrt{\log(4)}/\sqrt{2}\right\}-\Phi\left\{-\sqrt{\log(4)}\right\}+o(1) \tag{S3.7}\\
&\geq .083+o(1) \tag{S3.8}
\end{align}$$

where (S3.1) is by the data processing inequality, (S3.2) uses the KS-distance as a lower bound on total variation, (S3.5) applies two triangle inequalities since KS-distance is a metric, and (S3.6) uses the asymptotic distributions of $\hat{\theta}_X$ and $\hat{\theta}_Z$.

To establish S3.7, consider the following. Denote $\sigma^2 = (I^{-1}(\theta))_{1,1}$. Then consider the sequence of rectangles $R_i = \{x \in \mathbb{R}^k \mid -(i+1)\sigma \leq x_1 \leq -\sqrt{\log(4)}\sigma, \text{ and } -i \leq x_j \leq i, \forall j \neq 1\}$. Note that $R_i \subset R_{i+1}$ and $\bigcup_{i=1}^{\infty} R_i = \{x \in \mathbb{R} \mid x_1 \leq -\sqrt{\log(4)}\sigma\}$. Denote by P the probability measure for $N\{0,2I^{-1}(\theta)\}$ and Q the probability measure for $N\{0,I^{-1}(\theta)\}$. Then

$$\begin{aligned}
\mathrm{KS}\left[N\{0,I^{-1}(\theta)\},N\{0,2I^{-1}(\theta)\}\right] &\geq \lim_{i\to\infty}|P(R_i)-Q(R_i)|\\
&= \left|\lim_{i\to\infty}P(R_i)-\lim_{i\to\infty}Q(R_i)\right|\\
&= \left|P\left(\bigcup_{i=1}^{\infty}R_i\right)-Q\left(\bigcup_{i=1}^{\infty}R_i\right)\right|\\
&= \Phi\left\{-\sqrt{\log(4)}/\sqrt{2}\right\}-\Phi\left\{-\sqrt{\log(4)}\right\}\\
&\geq .083,
\end{aligned}$$

where the value $\sqrt{\log(4)}$ was chosen as it is the maximizer of $\Phi(-t/\sqrt{2})-\Phi(t)$. □

For the following proofs, we will overload the $\frac{d}{d\theta}$ operator when working with multivariate derivatives. For a function $f:\mathbb{R}^p\to\mathbb{R}$, we write $\frac{d}{d\theta}f(\theta)$ to denote the $p\times 1$ vector of partial derivatives $(\frac{\partial}{\partial\theta_j}f(\theta))_{j=1}^{p}$. For a function $g:\mathbb{R}^p\to\mathbb{R}^q$, we write $\frac{d}{d\theta}g(\theta)$ to denote the $p\times q$ matrix $(\frac{\partial}{\partial\theta_j}g_k(\theta))_{j,k=1}^{p,q}$.

Lemmas 2 and 3 are used for the proof of Theorem 2. Parts 1 and 2 of Lemma 2 can be rephrased as the following: $\hat{\theta}$ is efficient if and only if it is consistent and $n^{-1}\sum_{i=1}^{n}S(\hat{\theta},X_i)=o_p(n^{-1/2})$. The third property of Lemma 2 is similar to many standard expansions used in asymptotics, for example in Van der Vaart (2000). However, we require the expansion for arbitrary efficient estimators, and include a proof for completeness.

Lemma 2. *Suppose $X_1, \ldots, X_n \overset{i.i.d.}{\sim} f_{\theta_0}$, and assume that (R1)-(R3) hold. Let $\hat{\theta}_X$ be an efficient estimator, which is a sequence of zeros of the score function. Suppose that $\widetilde{\theta}_X$ is a $\sqrt{n}$-consistent estimator of θ_0. Then*

1. *If $n^{-1}\sum_{i=1}^n S(\widetilde{\theta}_X, X_i) = o_p(n^{-1/2})$, then $\widetilde{\theta}_X - \hat{\theta}_X = o_p(n^{-1/2})$.*

2. *If $\widetilde{\theta}_X$ is efficient, then $n^{-1}\sum_{i=1}^n S(\widetilde{\theta}_X, X_i) = o_p(n^{-1/2})$.*

3. *If $\widetilde{\theta}_X$ is efficient, then $\widetilde{\theta}_X = \theta_0 + I^{-1}(\theta_0)n^{-1}\sum_{i=1}^n S(\theta_0, X_i) + o_p(n^{-1/2})$.*

Proof. As $\widetilde{\theta}_X$ and $\hat{\theta}_X$ are both $\sqrt{n}$-consistent, we know that $\widetilde{\theta}_X - \hat{\theta}_X = O_p(n^{-1/2})$. So, we may consider a Taylor expansion of the score function about $\widetilde{\theta}_X = \hat{\theta}_X$.

$$\begin{aligned}
&n^{-1}\sum_{i=1}^n S(\widetilde{\theta}_X, X_i)\\
&= n^{-1}\sum_{i=1}^n S(\hat{\theta}_X, X_i) + \left\{\frac{d}{d\theta} n^{-1}\sum_{i=1}^n S(\theta, X_i)\Big|_{\theta=\hat{\theta}_X}\right\}(\widetilde{\theta}_X - \hat{\theta}_X) + O_p(n^{-1})\\
&= 0 + \left\{\frac{d}{d\theta} n^{-1}\sum_{i=1}^n S(\theta, X_i)\big|_{\theta=\hat{\theta}_X} + O_p(n^{-1/2})\right\}(\widetilde{\theta}_X - \hat{\theta}_X)\\
&= \{-I(\theta_0) + o_p(1)\}(\widetilde{\theta}_X - \hat{\theta}_X),
\end{aligned} \tag{S3.9}$$

where we used assumptions (R1)-(R3) to justify that 1) the second derivative is bounded in a neighborhood about θ_0 (as both $\hat{\theta}_X$ and $\widetilde{\theta}_X$ converge to θ_0), 2) the derivative of the score converges to $-I(\theta_0)$ by Lehmann (2004, Theorem 7.2.1) along with the Law of Large Numbers, and 3) that $I(\theta_0)$ is finite, by (R3).

To establish property 1, note that the left hand side of Equation (S3.9) is $o_p(n^{-1/2})$ implying that $(\widetilde{\theta}_X - \hat{\theta}_X) = o_p(n^{-1/2})$. For property 2, recall that by Lehmann (2004, page 479), if $\widetilde{\theta}_X$ and $\hat{\theta}_X$ are both efficient, then $(\widetilde{\theta}_X - \hat{\theta}_X) = o_p(n^{-1/2})$. Plugging this into the right hand side of Equation (S3.9) gives $n^{-1}\sum_{i=1}^n S(\widetilde{\theta}_X, X_i) = o_p(n^{-1/2})$, establishing property 2.

For property 3, we consider a slightly different expansion:

$$\begin{aligned} o_p(n^{-1/2}) &= n^{-1}\sum_{i=1}^n S(\widetilde{\theta}, X_i) \\ &= n^{-1}\sum_{i=1}^n S(\theta_0, X_i) + \frac{d}{d\theta_0} n^{-1}\sum_{i=1}^n S(\theta_0, X_i)(\widetilde{\theta} - \theta_0) + O_p(n^{-1}), \\ &= n^{-1}\sum_{i=1}^n S(\theta_0, X_i) + \{-I(\theta_0) + o_p(1)\}(\widetilde{\theta} - \theta_0) + O_p(n^{-1}) \end{aligned}$$

where we used property 2 for the first equality, expanded the score about $\hat{\theta}_X = \theta_0$ for the second, and justify the $O_p(n^{-1})$ by (R2). By (R1)-(R2) and Law of Large Numbers along with Lehmann (2004, Theorem 7.2.1), we have the convergence of the derivative of score to $-I(\theta_0)$. By (R3), $I(\theta_0)$ is invertible. Solving the equation for $\widetilde{\theta}_X$ gives the desired result. □

Lemma 3. *Assume that (R0)-(R4) hold, and let* $\omega_1, \ldots, \omega_n \overset{i.i.d.}{\sim} P$. *Then*

$$n^{-1}\sum_{i=1}^n \frac{d}{d\theta} S\{\theta, X_\theta(\omega_i)\} = o_p(1).$$

Proof. First we can express the derivative as

$$
\begin{aligned}
&n^{-1}\sum_{i=1}^{n}\frac{d}{d\theta}S\{\theta, X_\theta(\omega_i)\}\\
&= n^{-1}\sum_{i=1}^{n}\left[\frac{d}{d\alpha}S\{\alpha, X_\theta(\omega_i)\} + \frac{d}{d\alpha}S\{\theta, X_\alpha(\omega_i)\}\right]\Big|_{\alpha=\theta}.
\end{aligned}
$$

The result follows from the Law of Large Numbers, provided that

$$E_{\omega\sim P}\left[\frac{d}{d\alpha}S\{\alpha, X_\theta(\omega)\} + \frac{d}{d\alpha}S\{\theta, X_\alpha(\omega)\}\right]\Big|_{\alpha=\theta} = 0.$$

The expectation of the first term is $-I(\theta)$, by Lehmann (2004, Theorem 7.2.1). For the second term, we compute

$$
\begin{aligned}
E_{\omega\sim P}\frac{d}{d\alpha}S\{\theta, X_\alpha(\omega)\}\Big|_{\alpha=\theta} &= \int_\Omega \frac{d}{d\alpha}S\{\theta, X_\alpha(\omega)\}\Big|_{\alpha=\theta}\pi(\omega)\ d\omega && \text{(S3.10)}\\
&= \frac{d}{d\alpha}\int_\Omega S\{\theta, X_\alpha(\omega)\}\ \pi(\omega)\ d\omega\Big|_{\alpha=\theta} && \text{(S3.11)}\\
&= \frac{d}{d\alpha}\int_{\mathbb{R}^d} S(\theta, x) f_\alpha(x)\ dx\Big|_{\alpha=\theta} && \text{(S3.12)}\\
&= \int_{\mathbb{R}^d} \frac{d}{d\alpha}S(\theta, x) f_\alpha(x)\Big|_{\alpha=\theta}\ dx && \text{(S3.13)}\\
&= \int_{\mathbb{R}^d} S(\theta, x)\left\{\frac{d}{d\alpha}f_\alpha(x)\Big|_{\alpha=\theta}\right\}^\top\ dx && \text{(S3.14)}\\
&= \int_{\mathbb{R}^d} S(\theta, x)\left\{\frac{\frac{d}{d\theta}f_\theta(x)}{f_\theta(x)}\right\}^\top f_\theta(x)\ dx && \text{(S3.15)}\\
&= \int_{\mathbb{R}^d} S(\theta, x) S^\top(\theta, x) f_\theta(x)\ dx && \text{(S3.16)}\\
&= E_{X\sim\theta}\left\{S(\theta, X) S^\top(\theta, X)\right\} && \text{(S3.17)}\\
&= I(\theta), && \text{(S3.18)}
\end{aligned}
$$

where for (S3.11) we use the boundedness of Ω from (R0) and (R4) to interchange the derivative and integral; for (S3.12), we apply a change of variables, using the fact that $f_\alpha(\omega)$ is the density for the random variable $X_\alpha(\omega)$; and for (S3.13), we use (R2) and the dominated convergence theorem to change the order of the derivative and integral again. □

Proof of Theorem 2. We expand $\hat{\theta}_Z$ about $\hat{\theta}_X$ using part 3 of Lemma 2:

$$\hat{\theta}_Z = \hat{\theta}_X + I^{-1}\{\hat{\theta}_X\} n^{-1} \sum_{i=1}^{n} S\{\hat{\theta}_X, X_{\hat{\theta}_X}(\omega_i)\} + o_p(n^{-1/2}) \tag{S3.19}$$

The score can be expanded about $\hat{\theta}_X = \theta_0$:

$$\begin{aligned} &n^{-1} \sum_{i=1}^{n} S\{\hat{\theta}_X, X_{\hat{\theta}_X}(\omega_i)\} \\ &= n^{-1} \sum_{i=1}^{n} S\{\theta_0, X_{\theta_0}(\omega_i)\} + \left[\frac{d}{d\widetilde{\theta}} n^{-1} \sum_{i=1}^{n} S\{\widetilde{\theta}, X_{\widetilde{\theta}}(\omega_i)\} \right] \{\hat{\theta}_X - \theta_0\} \\ &= n^{-1} \sum_{i=1}^{n} S\{\theta_0, X_{\theta_0}(\omega_i)\} + o_p(1) O_p(n^{-1/2}), \end{aligned}$$

where $\widetilde{\theta}$ is between $\hat{\theta}_X$ and θ_0; by Lemma 3, we justify that the derivative is $o_p(1)$.

Combining this derivation along with the fact that $I^{-1}(\hat{\theta}_X) = I^{-1}(\theta_0) + o_p(1)$ by the continuous mapping theorem, we have the following equation:

$$\hat{\theta}_Z = \hat{\theta}_X + I^{-1}(\theta_0) n^{-1} \sum_{i=1}^{n} S\{\theta_0, X_{\theta_0}(\omega_i)\} + o_p(n^{-1/2}). \tag{S3.20}$$

Using the same techniques, we do an expansion for $\hat{\theta}_Y$ about $\theta^* = 2\hat{\theta}_X - \hat{\theta}_Z$:

$$\begin{aligned}
\hat{\theta}_Y &= \theta^* + I^{-1}(\theta^*)n^{-1}\sum_{i=1}^{n} S\{\theta^*, X_{\theta^*}(\omega_i)\} + o_p(n^{-1/2}) && \text{(S3.21)}\\
&= \theta^* + I^{-1}(\theta_0)n^{-1}\sum_{i=1}^{n} S\{\theta_0, X_{\theta_0}(\omega_i)\} + o_p(n^{-1/2}) && \text{(S3.22)}\\
&= \theta^* + (\hat{\theta}_Z - \hat{\theta}_X) + o_p(n^{-1/2}) && \text{(S3.23)}\\
&= \hat{\theta}_X + o_p(n^{-1/2}), && \text{(S3.24)}
\end{aligned}$$

where line (S3.22) is a similar expansion as used for equation (S3.19), in line (S3.23) we substituted the expression from (S3.20), and line (S3.24) uses the fact that as $n \to \infty$, $\theta^* = 2\hat{\theta}_X - \hat{\theta}_Z$ with probability tending to one. Indeed, since $2\hat{\theta}_X - \hat{\theta}_Z$ is a consistent estimator of θ_0, we have that as $n \to \infty$, $P(2\hat{\theta}_X - \hat{\theta}_Z \in \Theta) \geq P\{2\hat{\theta}_X - \hat{\theta}_Z \in B(\theta_0)\} \to 1$. □

Proof of Lemma 1. For a fixed $\theta \in \Theta$, for $\omega \sim P$, the random variable $Y = X_\theta(\omega)$ is distributed with probability measure PX_θ^{-1}: for any measurable set E, $P(Y \in E) = PX_\theta^{-1}(E)$. We denote by P_Ω^n the joint probability measure on Ω^n, and $(PX_\theta^{-1})^n$ the joint probability measure on $\mathbb{R}^{d\times n}$.

Given $\theta^* \in \Theta$, our goal is to derive the probability distribution of the random variables $X_{\theta^*}(\omega_1), \ldots, X_{\theta^*}(\omega_n)$ conditioned on the event that

$\{\omega_1, \ldots, \omega_n \mid \hat{\theta}\{X_{\theta^*}(\omega_i)\} = \hat{\vartheta}\}$. However, this event may have zero probability. Instead, we will condition on $S^{\delta}_{\hat{\vartheta},\theta^*} = \{\omega_1, \ldots, \omega_n \mid \hat{\vartheta}\{X_{\theta^*}(\omega)\} \in B_{\delta}(\hat{\vartheta})\}$, where $B_{\delta}(\hat{\vartheta}) = \{\theta \mid \|\hat{\vartheta} - \theta\| \leq \delta\}$, which has positive probability. At the end, we will take the limit as $\delta \to 0$ to derive the desired distribution.

Let $E \subset \mathbb{R}^{d\times n}$be a measurable set. Then

$$\begin{aligned}
&P\{X_{\theta^*}(\omega_1), \ldots, X_{\theta^*}(\omega_n) \in E \mid \omega_1, \ldots, \omega_n \in S^{\delta}_{\theta^*,\hat{\vartheta}}\}\\
&= P(\omega_1, \ldots, \omega_n \in X^{-1}_{\theta^*}E \mid \omega_1, \ldots, \omega_n \in S^{\delta}_{\theta^*,\hat{\vartheta}})\\
&= \frac{P^n(X^{-1}_{\theta^*}E \cap S^{\delta}_{\theta^*,\hat{\vartheta}})}{P^n(S^{\delta}_{\theta^*,\hat{\vartheta}})}\\
&= \frac{(PX^{-1}_{\theta^*})^n(E \cap X_{\theta^*}S^{\delta}_{\theta^*,\hat{\vartheta}})}{(PX^{-1}_{\theta^*})^n(X_{\theta^*}S^{\delta}_{\theta^*,\hat{\vartheta}})},
\end{aligned}$$

where we used the definition of conditional probability and the fact that $X^{-1}_{\theta^*}X_{\theta^*}S^{\delta}_{\theta^*,\hat{\vartheta}} = S^{\delta}_{\theta^*,\hat{\vartheta}}$.

This last expression shows that $X_{\theta^*}(\omega_1), \ldots, X_{\theta^*}(\omega_n)$ conditioned on $\underline{\omega} \in S^{\delta}_{\theta^*,\hat{\vartheta}}$ is distributed as $f^n_{\theta^*}\{y_1, \ldots, y_n \mid \hat{\theta}(\underline{y}) \in B_{\delta}(\hat{\vartheta})\}$. This derivation is valid for all $\delta > 0$. Taking the limit as $\delta \to 0$ gives the desired formula:

$$Y_1^{\theta^*}, \ldots, Y_n^{\theta^*} \Big| \hat{\theta}(\underline{Y}^{\theta^*}) = \hat{\theta}(\underline{X}) \sim f^n_{\theta^*}\{y_1, \ldots, y_n \mid \hat{\theta}(\underline{y}) = \hat{\theta}(\underline{X})\}.$$

□

Proof of Theorem 3. While the distributions g depend on n, we will suppress this dependence for notational simplicity. We can then express the desired KL divergence as follows:

First, by the data processing inequality, we can add in the random variable $\hat{\theta}(\underline{X}) = \hat{\theta}(\underline{Y})$ to get an upper bound on the KL divergence. We then have closed formulas for the joint distributions $\{X_1, \ldots, X_n, \hat{\theta}(\underline{X})\}$ and $\{Y_1, \ldots, Y_n, \hat{\theta}(\underline{X})\}$.

$$
\begin{aligned}
&\mathrm{KL}\left(X_1, \ldots, X_n || Y_1, \ldots, Y_n\right) && \text{(S3.25)}\\
&\quad \leq \mathrm{KL}\left\{X_1, \ldots, X_n, \hat{\theta}(\underline{X}) \middle|\middle| Y_1, \ldots, Y_n, \hat{\theta}(\underline{X})\right\} && \text{(S3.26)}\\
&\quad = \mathrm{KL}\left[f_\theta^n\{\underline{x} \mid \hat{\theta}(\underline{x})\} g_\theta\{\hat{\theta}(x)\} \middle|\middle| f_{\theta_n}^n\{\underline{x} \mid \hat{\theta}(\underline{x})\} g_\theta\{\hat{\theta}(\underline{x})\}\right] && \text{(S3.27)}\\
&\quad = E_{\hat{\vartheta}\sim g(\cdot|\underline{X})} E_{\underline{X}\sim f_\theta} \log\left\{\frac{f_\theta(\underline{X} \mid \hat{\vartheta}) g_\theta(\hat{\vartheta})}{f_{\theta_n}(\underline{X} \mid \hat{\vartheta}) g_\theta(\hat{\vartheta})}\right\} && \text{(S3.28)}\\
&\quad = E_{\hat{\vartheta}\sim g(\cdot|\underline{X})} E_{\underline{X}\sim f_\theta} \log\left\{\frac{f_\theta(\underline{X}) g(\hat{\vartheta} \mid \underline{X})}{f_{\theta_n}(\underline{X} \mid \hat{\vartheta}) g_\theta(\hat{\vartheta})}\right\} && \text{(S3.29)}\\
&\quad = E_{\hat{\vartheta}\sim g(\cdot|\underline{X})} E_{\underline{X}\sim f_\theta} \log\left[\frac{f_\theta(\underline{X}) g(\hat{\vartheta} \mid \underline{X})}{\{f_{\theta_n}(\underline{X}) g(\hat{\vartheta} \mid \underline{X})/g_{\theta_n}(\hat{\vartheta})\} g_\theta(\hat{\vartheta})}\right] && \text{(S3.30)}\\
&\quad = E_{\hat{\vartheta}\sim g(\cdot|\underline{X})} E_{\underline{X}\sim f_\theta} \log\left\{\frac{f_\theta(\underline{X}) g(\hat{\vartheta} \mid \underline{X})}{f_{\theta_n}(\underline{X}) g(\hat{\vartheta} \mid \underline{X})}\right\} && \text{(S3.31)}\\
&\qquad + E_{\hat{\vartheta}\sim g(\cdot|\underline{X})} E_{\underline{X}\sim f_\theta} \log\left\{\frac{g_{\theta_n}(\hat{\vartheta})}{g_\theta(\hat{\vartheta})}\right\} && \text{(S3.32)}\\
&\quad = -E_{\underline{X}\sim f_\theta} \log\left\{\frac{f_{\theta_n}(\underline{X})}{f_\theta(\underline{X})}\right\} + E_{\hat{\vartheta}\sim g_\theta} \log\left\{\frac{g_{\theta_n}(\hat{\vartheta})}{g_\theta(\hat{\vartheta})}\right\}, && \text{(S3.33)}
\end{aligned}
$$

where line (S3.28) simply applies the definition of KL divergence, and line (S3.30) uses the definition of conditional distribution.

At this point, we need to compute the two expectations of line (S3.33), and show that everything cancels except for an $o_p(1)$ term.

We write $\ell(\theta \mid \underline{x}) = \sum_{i=1}^n \log f_\theta(x_i)$. Using our assumptions, we can expand $\ell(\theta_n \mid \underline{x})$:

$$\begin{aligned}\ell(\theta_n \mid \underline{x}) &= \ell(\theta \mid \underline{x}) + (\theta_n - \theta)^\top \nabla \ell(\theta \mid \underline{x}) + \frac{1}{2}(\theta_n - \theta)^\top \nabla^2 \ell(\theta \mid \underline{x})(\theta_n - \theta)\\ &\quad + \frac{1}{6}\xi^* \sum_{i,j,k} (\theta_n - \theta)_i (\theta_n - \theta)_j (\theta_n - \theta)_k \sum_{s=1}^n g_{ijk}(x_s),\end{aligned}$$

where $|\xi^*| \leq 1$ and $g_{ijk}(x)$ is an upper bound for $\left|\frac{\partial^3 \ell(\theta|\underline{x})}{\partial\theta_i\theta_j\theta_k}\right|$ for a ball about θ, which exists by (R3). These expansions are based on those from Serfling (1980). Applying $E_{\underline{X}\sim f_\theta}$ to this derivation gives

$$\begin{aligned}E_{\underline{X}\sim f_\theta} \log\left\{\frac{f_{\theta_n}(\underline{X})}{f_\theta(\underline{X})}\right\} &= 0 - \frac{n}{2}(\theta_n - \theta)^\top I(\theta)(\theta_n - \theta)\\ &\quad + O(1)\frac{n}{6}\sum_{i,j,k}\{E g_{i,j,k}(x)\}(\theta_n - \theta)_i(\theta_n - \theta)_j(\theta_n - \theta)_k,\\ &= \frac{-n}{2}(\theta_n - \theta)^\top I(\theta)(\theta_n - \theta) + O(n)\|\theta_n - \theta\|^3\end{aligned} \tag{S3.34}$$

where the first term is zero as the expected value of the score function is zero by (R3), the second term uses Lehmann (2004, Theorem 7.2.1) and (R3). The $O(1)$ factor in the third term is based on the fact that $|\xi^*| \leq 1$. Finally, note that $\sum_{i,j,k}\{E g_{i,j,k}(x)\}(\theta_n - \theta)_i(\theta_n - \theta)_j(\theta_n - \theta)_k \leq p^3 \sup_{i,j,k}\{E g_{i,j,k}(x)\}\|\theta_n - \theta\|_\infty^3 = O(1)\|\theta_n - \theta\|^3$. Note that all norms are equivalent in $\mathbb{R}^p$, so they can be interchanged up to a factor of $O(1)$.

Next, we will derive a similar formula for $\log g_{\theta^*}(\hat{\vartheta})$:

$$
\begin{aligned}
\log g_{\theta_n}(\hat{\vartheta}) = \log g_\theta(\hat{\vartheta}) &+ \nabla \log g_\theta(\hat{\vartheta})(\theta_n - \theta) \\
&+ \frac{1}{2}(\theta_n - \theta)^\top \nabla^2 \log g_\theta(\hat{\vartheta})(\theta_n - \theta) \\
&+ \frac{n}{6}\xi_2^* \sum_{i,j,k} (\theta_n - \theta)_i (\theta_n - \theta)_j (\theta_n - \theta)_k G_{i,j,k}(\hat{\vartheta}),
\end{aligned}
\tag{S3.35}
$$

where $|\xi_2^*| \leq 1$. In order to apply the expectation $E_{\hat{\vartheta}\sim\theta}$ to this equation, we will first show $E_{\hat{\vartheta}\sim\theta} \nabla \log g_\theta(\hat{\vartheta}) = 0$ and $E_{\hat{\vartheta}\sim\theta} \nabla^2 \log g_\theta(\hat{\vartheta}) = -nI(\theta) + o(n)$.

$$
\begin{aligned}
\left\{ E_{\hat{\theta}\sim\theta} \nabla \log g_\theta(\hat{\vartheta}) \right\}_j &= \int \left\{ \frac{\partial}{\partial \theta_j} \log g_\theta(\hat{\vartheta}) \right\} g_\theta(\hat{\vartheta}) \; d\hat{\vartheta} \\
&= \int \frac{\partial}{\partial \theta_j} g_\theta(\hat{\vartheta}) \; d\hat{\vartheta} \\
&= \int \frac{\partial}{\partial \theta_j} \int_x f_\theta(x) g(\hat{\vartheta} \mid x) \; dx \; d\hat{\vartheta} \\
&= \int_{\hat{\theta}} \int_x \frac{\partial}{\partial \theta_j} f_\theta(x) g(\hat{\vartheta} \mid x) \; dx \; d\hat{\vartheta} \\
&= \frac{\partial}{\partial \theta_j} \int \int f_\theta(x) g(\hat{\vartheta} \mid x) \; dx \; d\hat{\vartheta} \\
&= 0,
\end{aligned}
$$

where we use the assumption (R3) that $\left| \frac{\partial}{\partial \theta_j} f_\theta(x) \right|$ is bounded above by an integrable function, (R5) that $g(\hat{\vartheta} \mid x)$ is bounded, and the dominated convergence theorem to interchange the derivative and the integral.

Next we work on the second derivative:

$$
\begin{aligned}
&\left(E_{\hat{\vartheta}\sim\theta}\nabla^2 \log g_\theta(\hat{\vartheta})\right)_{j,k}\\
&= \int \left\{\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log g_\theta(\hat{\vartheta})\right\} g_\theta(\hat{\vartheta})\ d\hat{\vartheta}\\
&= \int \frac{g_\theta(\hat{\vartheta})\frac{\partial^2}{\partial\theta_j\partial\theta_k}g_\theta(\hat{\vartheta}) - \frac{\partial}{\partial\theta_j}g_\theta(\hat{\vartheta})\left\{\frac{\partial}{\partial\theta_k}g_\theta(\hat{\vartheta})\right\}^\top}{g_\theta^2(\hat{\vartheta})} g_\theta(\hat{\vartheta})\ d\hat{\vartheta}\\
&= 0 - E_{\hat{\vartheta}\sim\theta}\left(\nabla \log g_\theta(\hat{\vartheta})\nabla^\top \log g_\theta(\hat{\vartheta})\right)_{j,k},
\end{aligned}
$$

where we used (R3) along with the dominated convergence theorem to set the first term equal to zero. We see that $E\nabla^2 \log g_\theta(\hat{\vartheta}) = -I_{\hat{\theta}_X}(\theta)$, where $I_{\hat{\theta}_X}$ represents the Fisher information of the random variable $\hat{\theta}(\underline{X}) \sim g$. It is our current goal to show that $I_{\hat{\theta}_X}(\theta) = nI(\theta) + o(n)$, where $I(\theta)$ is the Fisher information for one sample $X \sim f_\theta$. First note that by the data processing inequality (Zamir, 1998), $I_{\hat{\theta}(\underline{X})}(\theta) \leq I_{X_1,\ldots,X_n}(\theta) = nI(\theta)$, where the inequality represents the positive-definite ordering of matrices. Next, we need to find a matching lower bound. By the Cramér Rao lower bound, we have that

$$\{I_{\hat{\theta}(\underline{X})}(\theta)\}^{-1} \leq \mathrm{Var}\{\hat{\theta}(\underline{X})\} + o(1/n),$$

where $\mathrm{Var}\{\hat{\theta}_X\}$ is the covariance matrix of the random variable $\hat{\theta}_X$, and we used the fact that $\hat{\theta}_X$ is asymptotically unbiased. By the efficiency of $\hat{\theta}(\underline{X})$, we have that

$$\mathrm{Var}\{\hat{\theta}(\underline{X})\} = n^{-1}I^{-1}(\theta) + o(1/n).$$

We then have

$$\begin{aligned} I_{\hat{\theta}(\underline{X})}(\theta) &\geq \left\{n^{-1}I^{-1}(\theta) + o(1/n)\right\}^{-1} \\ &= n\{I^{-1}(\theta) + o(1)\}^{-1} \\ &= n\{I(\theta) + o(1)\}, \end{aligned}$$

where for the last equality, we use the following matrix identity:

$$(A+B)^{-1} = A^{-1} - A^{-1}B(A+B)^{-1},$$

where we set $A = I^{-1}(\theta)$ and $B = o(1)$.

Combining our results, we have that

$$E_{\hat{\vartheta}\sim\theta}\nabla^2 \log g_\theta(\hat{\vartheta}) = -I_{\hat{\theta}(\underline{X})}(\theta) = n\{-I(\theta) + o(1)\}.$$

Finally, applying the expectation to equation (S3.35), we have

$$\begin{aligned} E_{\hat{\vartheta}\sim\theta} \log\left\{\frac{g_{\theta_n}(\hat{\vartheta})}{g_\theta(\hat{\vartheta})}\right\} &= 0 - \frac{n}{2}(\theta_n - \theta)\{I(\theta) + o(1)\}(\theta_n - \theta) \\ &\quad + O(1)\frac{n}{6}\sum_{i,j,k}\{EG_{i,j,k}(\hat{\vartheta})\}(\theta_n - \theta)_i(\theta_n - \theta)_j(\theta_n - \theta)_k \\ &= \frac{-n}{2}(\theta_n - \theta)^\top I(\theta)(\theta_n - \theta) + o(n)\|\theta_n - \theta\|^2 \\ &\quad + O(n)\|\theta_n - \theta\|^3. \end{aligned} \qquad \text{(S3.36)}$$

Combining equations (S3.34) and (S3.36), we have

$$
\begin{aligned}
&\mathrm{KL}\left(X_1,\ldots,X_n||Y_1,\ldots,Y_n\right)\\
&\quad\leq \mathrm{KL}\left[f_\theta^n\{\underline{x}\mid\hat{\theta}(\underline{x})\}g_\theta\{\hat{\theta}(x)\}\middle|\middle|f_{\theta_n}^n\{\underline{x}\mid\hat{\theta}(\underline{x})\}g_\theta\{\hat{\theta}(\underline{x})\}\right]\\
&\quad= -E_{\underline{X}\sim f_\theta}\log\left\{\frac{f_{\theta_n}(\underline{X})}{f_\theta(\underline{X})}\right\}+E_{\hat{\vartheta}\sim g_\theta(\hat{\vartheta})}\log\left\{\frac{g_{\theta_n}(\hat{\vartheta})}{g_\theta(\hat{\vartheta})}\right\}\\
&\quad= o(n)\|\theta_n-\theta\|^2+O(n)\|\theta_n-\theta\|^3.
\end{aligned}
$$

□

S4 Additional Simulation Results

S4.1 Differentially private beta synthetic data

In this section, we consider additional values of ϵ that are used in the Section 6.3 experiment on differentially private beta distributed synthetic data. All other simulation parameters are identical. We varied $\epsilon = .5, 2, 4, \infty$ (note that $\epsilon = 1$ appears in Figure 1(b) in the main paper). In Figure S.1, we see that at all values of ϵ, $\hat{\theta}_Y$ is very close to $\hat{\theta}_{DP}$. However, with smaller ϵ it requires a larger sample size before the performance of $\hat{\theta}_Y \approx \hat{\theta}_{DP}$ is similar to the MLE $\hat{\theta}_X$. Note that even with $\epsilon = \infty$, the performance of $\hat{\theta}_Z$ does not approach that of $\hat{\theta}_X$.

S4.2 DP two sample proportion test

In this section, we repeat the experiment of Section 6.4 with varying values of ϵ. All other simulation parameters are the same. We varied $\epsilon = .5, 2, 4, 10$ (note that $\epsilon = 1$ appears in Figure 1(b) in the main paper). In Figure S.2, we plot the p-values of both the one-step and parametric bootstrap tests. We see that the parametric bootstrap p-values are very conservative for all values of ϵ, whereas the one-step p-values are fairly well-calibrated, although sometimes slightly inflated. In Figure S.3 we plot the power of the two tests for the different ϵ values. We see that the relative performance of one-step compared to parametric bootstrap is unchanged as we vary ϵ.

Bibliography

Bun, M., C. Dwork, G. N. Rothblum, and T. Steinke (2018). Composable and versatile privacy via truncated cdp. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pp. 74–86.

Bun, M. and T. Steinke (2016). Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Theory of Cryptography Conference*, pp. 635–658. Springer.

Dong, J., A. Roth, and W. J. Su (2022). Gaussian differential privacy.

Journal of the Royal Statistical Society Series B: Statistical Methodology 84(1), 3–37.

Duchi, J. C., M. I. Jordan, and M. J. Wainwright (2013). Local privacy and statistical minimax rates. In *2013 IEEE 54th Annual Symposium on Foundations of Computer Science*, pp. 429–438. IEEE.

Dwork, C., K. Kenthapadi, F. McSherry, I. Mironov, and M. Naor (2006). Our data, ourselves: Privacy via distributed noise generation. In *Annual International Conference on the Theory and Applications of Cryptographic Techniques*, pp. 486–503. Springer.

Dwork, C., F. McSherry, K. Nissim, and A. Smith (2006). Calibrating noise to sensitivity in private data analysis. In *Theory of Cryptography Conference*, pp. 265–284. Springer.

Dwork, C., A. Roth, et al. (2014). The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science 9*(3–4), 211–407.

Dwork, C. and G. N. Rothblum (2016). Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*.

Kamath, G., V. Singhal, and J. Ullman (2020). Private mean estimation of heavy-tailed distributions. In *Conference on Learning Theory*, pp. 2204–2235. PMLR.

Kasiviswanathan, S. P., H. K. Lee, K. Nissim, S. Raskhodnikova, and A. Smith (2011). What can we learn privately? *SIAM Journal on Computing 40*(3), 793–826.

Kifer, D. and B.-R. Lin (2012). An axiomatic view of statistical privacy and utility. *Journal of Privacy and Confidentiality 4*(1).

Lehmann, E. L. (2004). *Elements of large-sample theory.* Springer Science & Business Media.

Mironov, I. (2017). Rényi differential privacy. In *2017 IEEE 30th Computer Security Foundations Symposium (CSF)*, pp. 263–275. IEEE.

Serfling, R. L. (1980). *Approximation Theorems in Mathematical Statistics.* New York: Wiley.

Smith, A. (2011). Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the Forty-Third Annual ACM Symposium on Theory of Computing*, pp. 813–822.

Van der Vaart, A. W. (2000). *Asymptotic statistics.* Cambridge University Press.

Zamir, R. (1998). A proof of the fisher information inequality via a data processing argument. *IEEE Transactions on Information Theory 44*(3), 1246–1250.

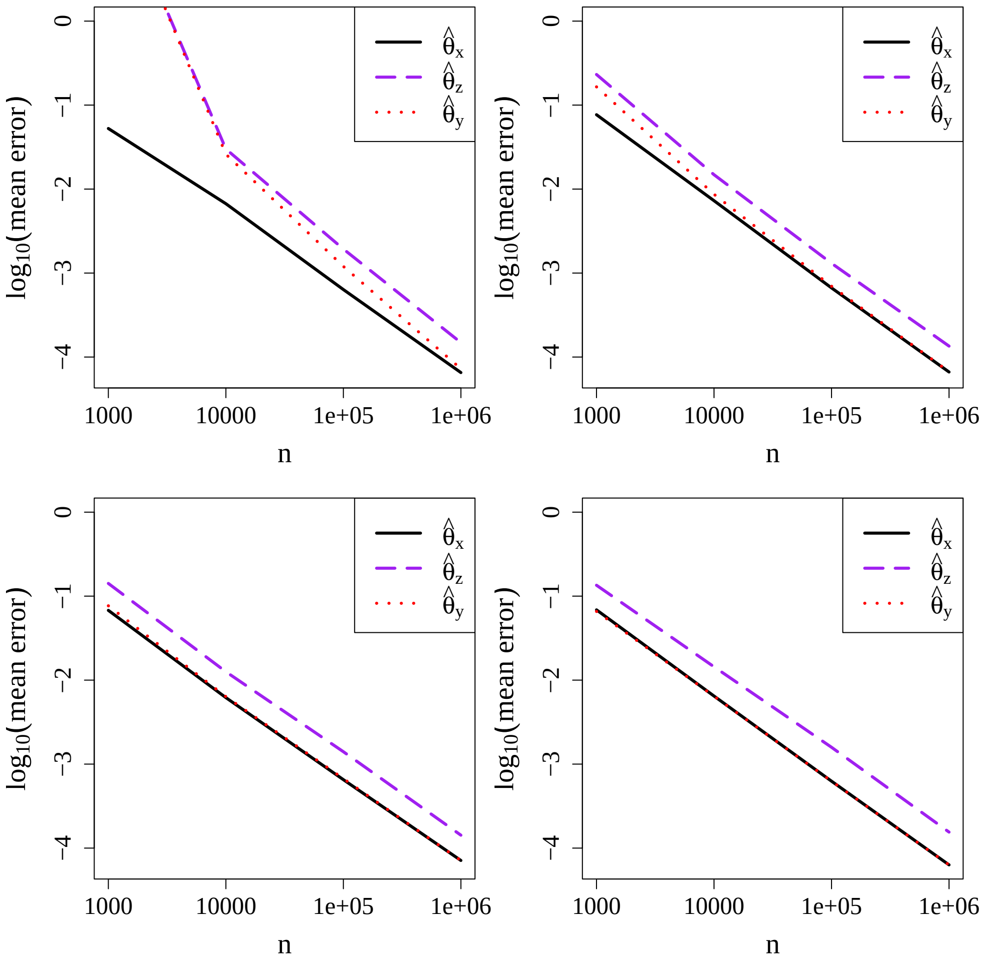

Figure S.1: Additional simulations for Section 6.3. In normal reading order, $\epsilon = .5, 2, 4, \infty$. Note that Figure 1(b) in the main paper is for $\epsilon = 1$

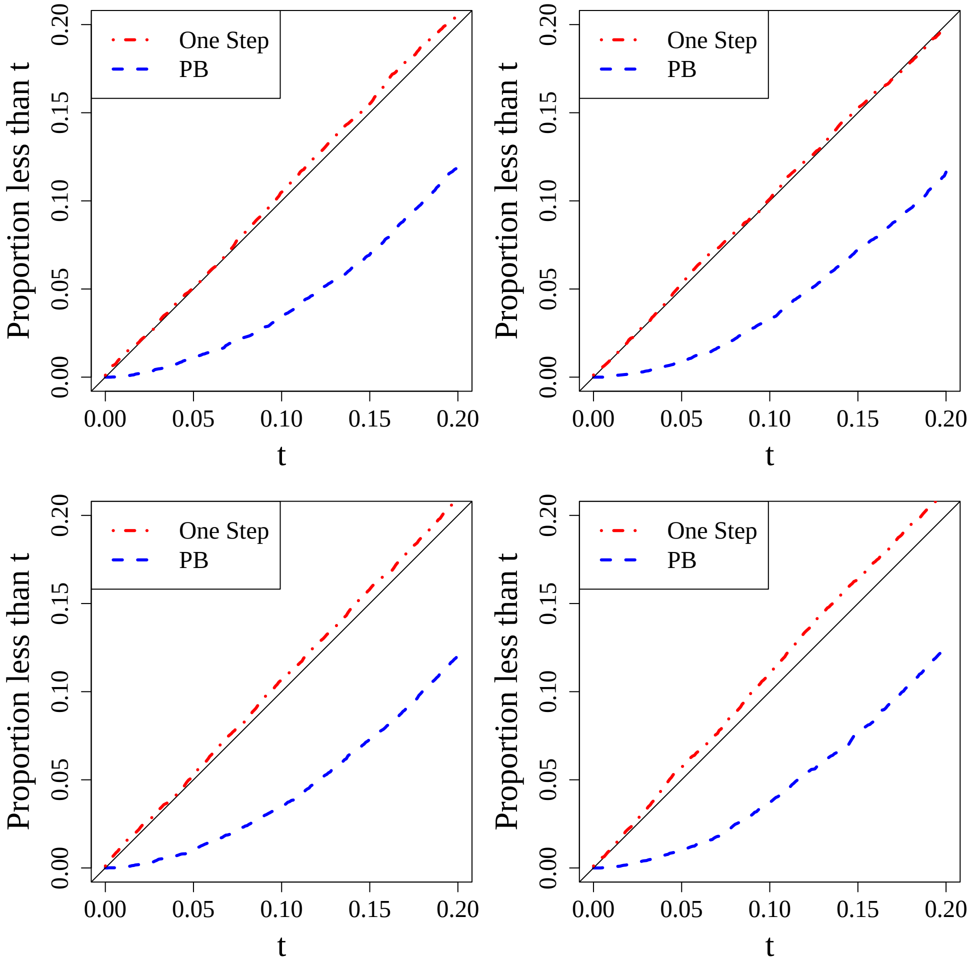

Figure S.2: Additional simulations for Section 6.4. In normal reading order, $\epsilon = .5, 2, 4, 10$. Note that Figure 2(a) in the main paper is for $\epsilon = 1$

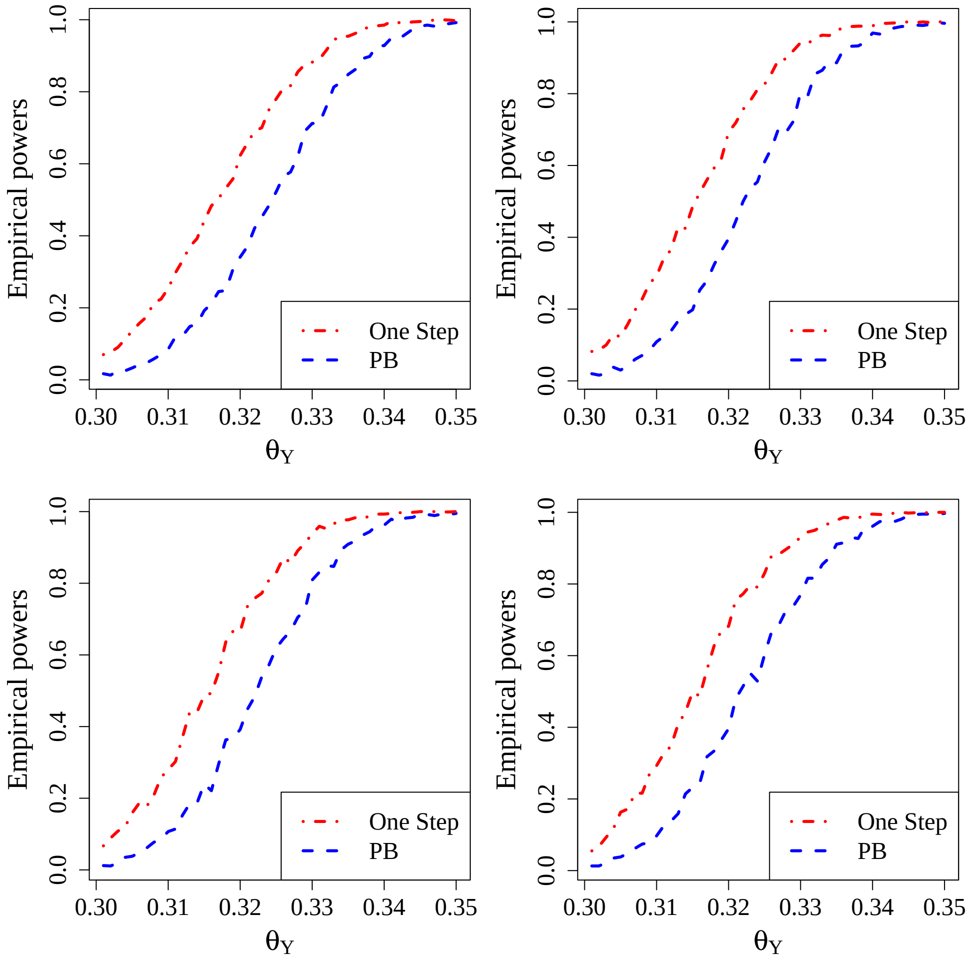

Figure S.3: Additional simulations for Section 6.4. In normal reading order, $\epsilon = .5, 2, 4, 10$. Note that Figure 2(b) in the main paper is for $\epsilon = 1$