A Polynomial Time, Pure Differentially Private Estimator for Binary Product Distributions

Vikrant Singhal VIKRANT@SEAS.HARVARD.EDU
150 Western Ave
Boston, MA 02134

Editors: Claire Vernade and Daniel Hsu

Abstract

We present the first ε-differentially private, computationally efficient algorithm that estimates the means of product distributions over $\{0,1\}^d$ accurately in total-variation distance, whilst attaining the optimal sample complexity to within polylogarithmic factors. The prior work had either solved this problem efficiently and optimally under weaker notions of privacy, or had solved it optimally while having exponential running times.

Keywords: differential privacy, data privacy, statistics, machine learning, product distributions

1. Introduction

Machine learning and statistics aim to learn information about the population. The pertinent algorithms always involve using random samples from the relevant population to learn and release that information, but at the same time, often end up revealing sensitive information about the individuals in the datasets. *Differential privacy (DP)* Dwork et al. (2006) is now a *de facto* standard for preserving privacy in learning and testing algorithms. It informally guarantees that no adversary can infer anything more about an individual from the output of a differentially private algorithm, than they could have from its output if the individual were not present in the dataset.

In the last few years, a large body of work on differentially private statistics has emerged, which has shown that the privacy constraint almost always imposes an additional cost in the sample complexity for those tasks (see Appendix A). Depending on the strength of the privacy guarantee (e.g., pure, concentrated Bun and Steinke (2016); Dwork and Rothblum (2016), or approximate differential privacy), the running times of the private algorithms also tend to get affected greatly. It is quite often (but *not* always) the case that developing pure DP algorithms for a statistical task, which has the optimal sample complexity and a polyomial running time, is much more challenging than creating efficient and (sample) optimal, concentrated or approximate DP algorithms for the same task. For instance, mean estimation of heavy-tailed distributions in ℓ_2 distance under pure DP was solved optimally, but without computational efficiency, in Kamath et al. (2020), but obtaining a computationally efficient algorithm for the same remained an open problem until it was solved recently in Hopkins et al. (2022a). On the other hand, Kamath et al. (2020) did provide computationally efficient and optimal, concentrated and approximate DP algorithms for mean estimation of heavy-tailed distributions. Similarly, Kamath et al. (2019a) presented efficient and optimal, concentrated and approximate DP algorithms for mean estimation of multivariate Gaussians, but solving this problem optimally and with computational efficiency under pure DP remained open until Hopkins et al. (2022b) solved it recently.

Estimating binary product distributions (product distributions over $\{0,1\}^d$) in total-variation distance is another such example. In this case, too, optimal and computationally efficient, concentrated and approximate DP algorithms had been presented already in Kamath et al. (2019a), but no computationally efficient and optimal, pure DP algorithm was known after that, although multiple optimal algorithms for this problem under pure DP, but lacking the computational efficiency, have come up in the recent past (e.g., Bun et al. (2019)). In this work, we provide an optimal and a computationally efficient algorithm that satisfies pure DP, and close that gap for this problem.

1.1. Estimation in Total-Variation Distance

We would first like to remind the reader as to why estimating binary product distributions in total-variation distance is a much more difficult task than just simply estimating them in ℓ_2 distance. A small error in total-variation distance requires all the marginals to be estimated accurately with respect to the magnitudes of their respective means, that is, the error for each marginal needs to be scaled according to the magnitude of its mean. In certain situations, it could imply that each marginal needs to be estimated to within a small multiplicative error. Instead, if in the estimate, all the coordinates have similar additive errors, then unless that error is *very* small, the marginals with very small means would have significantly lower accuracy than the ones with much larger means. Therefore, the estimate would not be accurate in total-variation distance. On the other hand, if we have the additive errors to be very small for all the coordinates (say, if we are estimating the distribution to within ℓ_1 distance α or to within ℓ_∞ distance $\frac{\alpha}{d}$), then the cost of estimation would become very large in terms of the sample complexity. Therefore, we need to find a way to estimate product distributions accurately *direction-wise*.

Sans privacy, this is an easy task – we can just output the empirical mean of the samples. Under privacy constraints, however, this is hard to do without knowing even an approximate scale of the noise to add in each direction. The naïve way to estimate privately would just involve either adding a very small noise to each coordinate of the empirical mean, but that would increase the cost in the sample complexity dramatically (by $\mathrm{poly}(d)$). On the other hand, even if we use a sophisticated pure DP estimator, which simply estimates accurately in ℓ_2 distance α, this would not be accurate enough either. Thus, we need more non-trivial ways to privately estimate each coordinate to within an error that is scaled appropriately for that marginal.

1.2. Result

We informally state the main result of this work here. It essentially says that our pure DP algorithm is computationally efficient and estimates the mean of any product distribution over $\{0,1\}^d$ in total-variation distance using just $\widetilde{O}(d)$ samples.

Theorem 1 (Informal) *For every $\varepsilon, \alpha, \beta > 0$, there exists a polynomial-time, ε-DP algorithm that takes n i.i.d. samples from a product distribution P over $\{0,1\}^d$, and returns a product distribution Q, such that if*

$$n \geq \widetilde{O}\left(\frac{d}{\alpha^2} + \frac{d}{\varepsilon\alpha}\right),$$

where $\widetilde{O}(\cdot)$ hides all polylogarithmic factors, then with probability at least $1-\beta$, the total-variation distance between P and Q is at most α.

Note that this is the optimal sample complexity for this problem, and the upper bounds under approximate DP and the matching lower bounds (which also hold under pure DP) were proved in Kamath et al. (2019a). The first term in the sample complexity is the necessary term for attaining the desired accuracy without any privacy constraints, and the second term is the additive cost due to privacy. For $\varepsilon \geq \Omega(\alpha)$, privacy comes for "free", that is, there is only a small multiplicative cost in the non-private sample complexity. Additionally, as we will see in the formal version of the above theorem (see Theorem 4), there is also a multiplicative $\mathrm{polylog}(d, \frac{1}{\beta}, \frac{1}{\varepsilon})$ improvement in the sample complexity over that of Kamath et al. (2019a) due to our tighter analysis.

One more comparison that we would like to draw is with the work of Bun et al. (2019). It is true that the polylog factors in their results are better than ours, but their algorithm only provides guarantees for estimation within total-variation distance, whereas our work does that for parameter estimation within χ^2-distance (hence, KL-divergence, as well) for a very wide range of parameters. In that sense, our algorithm provides stronger results because the total-variation distance guarantee is implied from the above.

1.3. Overview of Techniques

Our techniques are similar to those in Kamath et al. (2019a) for estimating binary product distributions – private partitioning, followed by privately estimating.

The private partitioning is performed iteratively. In each round, we assume an upper bound on the marginals, and use the Laplace mechanism (scaled according to that upper bound and the number of marginals that remain to be partitioned) to get a rough private estimate of each marginal, and pick the ones that lie above a certain threshold. As we prove later, those coordinates are bound to have higher means than the ones with their noisy estimates below that threshold. We rescale those chosen heavier coordinates as per the assumed upper bound for that iteration, and mark them to be estimated later after the partitioning is complete. In the next round, we assume both a reduced upper bound on the marginals and a reduced threshold to filter out the next set of marginals, and repeat the process.

Once we have filtered and rescaled those heavier marginals, we estimate them in $\mathrm{poly}(n, d)$ time to within α in ℓ_2 distance under pure DP using the sub-Gaussian learner from Hopkins et al. (2022b). On inverse-rescaling the estimated marginals according to their respective original rescaling parameters, we get an accurate estimate for those heavier coordinates to within $O(\alpha)$ in total-variation distance. Note that we cannot simply invoke the estimator from Hopkins et al. (2022b) on the filtered out coordinates before rescaling them, since that would just give an ℓ_2 estimate of the original marginals, which would not be accurate direction-wise, hence, would not be accurate in total-variation distance, especially when there are marginals with very different magnitudes in that filtered set (see Section 1.1). Therefore, this combination of partitioning and rescaling those heavier marginals seems necessary for this kind of an approach.

Remark 2 *We would like to remark that in Hopkins et al. (2022b), the assumption is that the covariance of the distribution in question is $\Sigma = \mathbb{I}$, while what we require is that $\Sigma \preceq \mathbb{I}$. That said, their algorithm can still work with the latter assumption because their proof for mean estimation mainly relies on their Corollary 5.4 and Lemma B.1, which still hold under this relaxed assumption.*

There are two important aspects of our algorithm that make it different from the work by Kamath et al. (2019a).

- In the partitioning procedure, we filter out the "heavier" marginals iteratively, and group the ones with similar weights together, but unlike the algorithm in Kamath et al. (2019a), ours does not estimate them simultaneously while the partitioning is being performed. This is because we wanted to avoid the additional $\mathrm{poly}(d)$ cost in the sample complexity due to (basic) composition under pure DP. However, Kamath et al. (2019a) were able to both partition and estimate those heavy marginals at the same time because they were working under concentrated or approximate DP, and they could use the more sophisticated, advanced composition of privacy Dwork et al. (2010), which only provides concentrated or approximate DP guarantees.

- In order to estimate those heavier marginals after filtering and grouping them, we rescale them using their respective rough private estimates that we obtained while partitioning, and apply the pure DP sub-Gaussian mean estimator from Hopkins et al. (2022b), which is computationally efficient, as well. This gives us an estimate of the scaled marginals that is accurate to within α in ℓ_2 distance.

After the partitioning rounds, we have the final round, where we are just left with the lighter coordinates to estimate. For that, we simply use the Laplace mechanism again, but with a much lower sensitivity this time, and get an accurate estimate for those marginals directly in one shot.

With accurate estimates for both the heavy and the light marginals in hand, we finally combine the two via a simple concatenation, and this gives us an accurate estimate for the whole distribution, as we had originally desired.

Remark 3 *We also want to remark on the techniques of Bun et al. (2019) for comparison. The algorithm of Bun et al. (2019) involves an application of the* exponential mechanism *McSherry and Talwar (2007), but it is more than just that. It is also a very general-purpose algorithm. While simpler applications of the exponential mechanism could be made computationally efficient, doing that for the algorithm in Bun et al. (2019) for our case does not seem straightforward. Intuitively, in their work, the score function of a point in the output space is also based on its tournaments with all the other points in the output space (as opposed to just with respect to the dataset), which means that sampling a point from that space via the exponential mechanism becomes inefficient as there is no way to efficiently determine the score of that point itself.*

1.4. Motivation and Broader Impact

We would like to remark that in this recent line of work on computationally efficient, pure DP statistical estimation, our work deals with a very fundamental distribution under a tricky, direction-wise error metric (total-variation distance). Therefore, while it closes a long-standing open problem, it also contributes to the diverse body of algorithmic tools and ideas available in DP literature for statistical estimation tasks. Could the techniques in our work be directly applied to other distributions? The answer is not obvious because different families of distributions have different properties and the way distance metrics are characterised for them could be very different. However, the high-level idea of performing private preconditioning can be and has been certainly useful in DP statistical estimation tasks. Also, we believe that it might be possible to estimate certain families of distributions (say, a subset of those with finite domains) in metrics that are similar to χ^2-divergence using ideas from our algorithm, just like we do in this work.

Additionally, our work also shows that even though heavy and general-purpose machinery, such as Hopkins et al. (2022b), might be available at our disposal, many tasks (like the problem we address) may still not have simple solutions. As we point out in Section 1.3, a lot of non-trivial steps often need to be taken in order to effectively use these tools.

Computationally efficient statistical estimation under pure DP has been a topic of recent interest in the DP community, mostly because many pure DP algorithms, even though they may have optimal sample complexity, are not very practical as they tend to have exponential running times. Pure DP gives us much stronger privacy guarantees, and if we have as practical and sample-efficient algorithms as the ones under the weaker privacy notions (such as approximate DP) to solve problems under this regime, then we have the best of both worlds. Hence, our goal was also to fill in another gap in this literature and provide new algorithms in this broader line of work for the community interested in DP statistical estimation.

Estimating binary product distributions is a fundamental statistical problem, so we answered it in the contexts (1) of computationally efficient pure DP estimation and (2) of addressing a folklore statistical problem under DP constraints. Therefore, we believe that we addressed this question from both theoretical and practical perspectives.

1.5. Organization of Paper

We describe our main technical results in Section 2. Appendix A contains a detailed discussion of the relevant prior work. In Appendix B, we state the necessary preliminaries for our work. Then in Appendices C, D, and E, we state the missing results and proofs from Section 2. Finally, we provide a discussion of our work relative to some prior work in Appendix F.

2. A Pure DP Product Distribution Estimator

In this section we introduce and analyze our algorithm for learning a product distribution P over $\{0,1\}^d$ in total-variation distance. The pseudocode is stated in Algorithm 1, with some additional observations and information regarding the notations stated in Section 2.1. For simplicity of presentation, we assume that the product distribution has mean, whose marginals are bounded by $\frac{1}{2}$ (i.e., $\mathbb{E}[P] \preceq \frac{1}{2}$), however, we would like to clarify that this assumption is essentially without loss of any generality, and is easily removable at the cost of a meagre constant factor in the sample complexity. We make an additional assumption that $d \geq 2$ and that $\beta \leq \frac{1}{2}$.

The following is the main result of our work. We prove the privacy, the computational efficiency, and the accuracy guarantees of our algorithm in Appendices C and D, and in Section 2.2, respectively. We defer a few proofs from Section 2.2 to Appendix E.

Theorem 4 *For every $\varepsilon, \alpha, \beta > 0$, there exists an ε-DP algorithm (*PUREDPPDE*) that takes n i.i.d. samples from a product distribution P over $\{0,1\}^d$, and returns a product distribution Q over $\{0,1\}^d$ in $\mathrm{poly}(n, d, \frac{1}{\varepsilon}, \frac{1}{\alpha}, \log(\frac{1}{\beta}))$ time, such that if*

$$n \geq \widetilde{O}_\alpha\left(\frac{d\log^2(d/\beta)}{\alpha^2} + \frac{d\log^2(d/\varepsilon\beta)}{\varepsilon\alpha}\right),$$

where $\widetilde{O}_\alpha(\cdot)$ hides polylogarithmic factors in $\frac{1}{\alpha}$, then with probability at least $1-\beta$, $\mathrm{d_{TV}}(P,Q) \leq \alpha$.

2.1. The Algorithm

We first introduce a notation for the *truncated mean* of all the points in a dataset. Given a data point $x \in \{0,1\}^d$ and $B \geq 0$, we write

$$\mathsf{trunc}_B(x) = \begin{cases} x & \text{if } |x| \leq B \\ \frac{B}{|x|} \cdot x & \text{if } |x| > B \end{cases}$$

to denote the truncation (or clipping) of x to an ℓ_1-ball of radius B. Given a dataset $X = (X_1, \ldots, X_m) \in \{0,1\}^{m \times d}$ and $B > 0$, we use

$$\mathsf{tmean}_B(X) = \frac{1}{m} \sum_{i=1}^{m} \mathsf{trunc}_B(X_i)$$

to denote the mean of the truncated data points in X. If one of the data points in X does not satisfy the norm bound (i.e., its ℓ_1 norm is greater than B) when X is served as an input to tmean_B, then we will say, "truncation occurred," as a shorthand. We point out a couple of important observations.

- The ℓ_1-sensitivity of tmean_B is $\frac{B}{m}$, while the ℓ_1-sensitivity of the un-truncated mean is $\frac{d}{m}$.
- Unless $|X_i| > B$ for some $i \in [m]$, the truncated mean equals the un-truncated mean, i.e., $\mathsf{tmean}_B(X) = \frac{1}{m} \sum_{i=1}^{m} X_i$.

We also use the following notational conventions. Given a data point $X_i \in \{0,1\}^d$, we use $X_i[j]$ to refer to its j-th coordinate, and for a subset of coordinates $S \subseteq [d]$, the notation $X_i[S] = (X_i[j])_{j \in S}$ to refer to the vector X_i with coordinates restricted to S. Given a dataset $X = (X_1, \ldots, X_m)$, we use the notation $X[S] = (X_1[S], \ldots, X_m[S])$ to refer to the dataset consisting of each $X_i[S]$. Next, for a domain $\mathcal{X}$, a variable $x \in \mathcal{X}$, and a probability distribution D over $\mathcal{X}$, we write "$x \leftarrow_R D$" to indicate that a sample has been drawn from D, and assigned to x. Finally, in Algorithm 1, we use C_α to denote the $\mathrm{polylog}(1/\alpha)$ quantity in the sample complexity from Theorem 29.

We first briefly describe Algorithm 1 here. The algorithm runs in two phases, essentially – partitioning and final phases.

- **Partitioning:** This is an iterative phase. In each iteration, the algorithm truncates all the points in the dataset according to an upper bound, and computes the empirical mean of all the truncated points, and then adds independent Laplace noise to each of the remaining coordinates of the mean vector in that iteration, and works with only the noisy values of those coordinates. The coordinates with noisy values above a certain threshold are put aside, while those with noisy values below that threshold are the remaining coordinates to work with in the next iteration or in the final phase. The important observation is that the noisy values that appear large cannot actually be small without the noise with high probability, and vice versa, which we show later in the analysis. This helps us separate out the large coordinates with high confidence. In the next iteration, we reduce the upper bound and the threshold because the larger coordinates were separated out already, and we are now left with the ones with lower magnitudes. When the number of remaining coordinates is small enough, we exit the loop. At this point, we have all the batches of "similar" coordinates partitioned, so we scale

the coordinates of all the batches up according to their respective upper bounds that we had used in the previous iterations in which they were separated out, so that they all now have magnitudes that are within a constant factor of one another. Then we apply the learner from Theorem 29, which gives us an accurate estimate of that scaled vector in ℓ_2 distance. We rescale the coordinates of that estimate with the inverse of their respective, original scaling factors, which gives us an accurate total-variation estimate of those coordinates.

- **Final:** This is a "one-shot" phase. The coordinates, which were not estimated in the previous phase, are estimated in this phase. Here, the algorithm simply truncates all the points in the dataset as per a small upper bound, and computes their empirical mean, and then adds independent Laplace noise to all these remaining coordinates of the mean vector. The important observation here is that because the means of all these coordinates are small enough, we do not end up requiring a lot of noise to ensure privacy, so we are able to obtain an accurate ℓ_1 estimate, which is sufficient to get an accurate total-variation estimate, at a low cost in the sample complexity.

In the end, the estimates from both the phases are combined appropriately, and released.

2.2. Accuracy Analysis

In this section we prove the following proposition bounding the sample complexity required by PUREDPPDE to be accurate.

Proposition 5 *For every $d \in \mathbb{N}$, every product distribution P over $\{0,1\}^d$, and every $\varepsilon, \alpha, \beta > 0$, if $X = (X_1, \ldots, X_n)$ are independent samples from P, where*

$$n \geq \widetilde{O}_\alpha\left(\frac{d\log^2(d/\beta)}{\alpha^2} + \frac{d\log^2(d/\varepsilon\beta)}{\alpha\varepsilon}\right),$$

then with probability at least $1 - O(\beta)$, $\textsc{PureDPPDE}_{\varepsilon,\alpha,\beta}(X)$ outputs Q, such that $\mathrm{d_{TV}}(P,Q) \leq \alpha$. The notation $\widetilde{O}_\alpha(\cdot)$ hides polylogarithmic factors in $\frac{1}{\alpha}$.

2.2.1. ANALYSIS OF THE PARTITIONING ROUNDS

In this section we analyze the progress made during the partitioning rounds. We show two properties for any round r: (1) any coordinate j such that $q_r[j]$ was filtered out during the partitioning rounds has a large mean, and (2) any coordinate j, such that $q_r[j]$ was moved on to the next round, has a small mean. We capture the properties of the partitioning rounds that will be necessary for the proof of Theorem 5 in the following lemma.

Lemma 6 (Partitioning Rounds) *If $Y^1, \ldots, Y^R$ each contain at least*

$$m \geq 2048d\log(d/\beta) + \frac{2048d\log(d/\varepsilon\beta)}{\varepsilon}$$

i.i.d. samples from P, and Z contains

$$m_0 \geq \widetilde{O}_\alpha\left(\frac{d+\log(1/\beta)}{\alpha^2} + \frac{d+\log(1/\beta)}{\alpha\varepsilon} + \frac{d\log(d)}{\varepsilon}\right)$$

i.i.d. samples from P (where $\widetilde{O}_\alpha(\cdot)$ hides polylogarithmic factors in $\frac{1}{\alpha}$), then with probability at least $1 - O(\beta)$, in every partitioning round $r \in [R]$, we have the following.

Algorithm 1: Pure DP Product Distribution Estimator $\textsc{PureDPPDE}_{\varepsilon,\alpha,\beta}(X)$

Input: Samples $X_1, \dots, X_n \in \{0,1\}^d$ from an unknown product distribution P satisfying $\mathbb{E}[P] \preceq \frac{1}{2}$. Parameters $\varepsilon, \alpha, \beta > 0$.

Output: A product distribution Q over $\{0,1\}^d$ such that $\mathrm{d_{TV}}(P,Q) \le \alpha$.

Set parameters: $R \leftarrow \log_2(d/2) \quad m \leftarrow 2048d\log(d/\beta) + \frac{2048d\log(d/\varepsilon\beta)}{\varepsilon}$

$m_0 \leftarrow C_\alpha\left(\frac{d+\log(1/\beta)}{\alpha^2} + \frac{d+\log(1/\beta)}{\alpha\varepsilon} + \frac{d\log(d)}{\varepsilon}\right)$

$m_1 \leftarrow \frac{128d\log(d/\beta)}{\alpha^2} + \frac{256d\log(d/\varepsilon\alpha\beta)}{\varepsilon\alpha}$

Split X into three datasets Y, Y^F, and Z, of sizes mR, m_1, and m_0, respectively.

Split Y into R blocks $Y^1, \dots, Y^R$ of sizes m each, denoted by $Y^r = (Y^r_1, \dots, Y^r_m)$.

Let $q \in (0,1)^d$ with $q[j] \leftarrow 0$ for every $j \in [d]$, and let $S_1 = [d]$, $u_1 \leftarrow \frac{1}{2}$, $\tau_1 \leftarrow \frac{3}{16}$, and $r \leftarrow 1$.

// Partitioning rounds.

Let $T_P \leftarrow \emptyset$.

while $u_r |S_r| \ge 1$ *and* $r \le R$ **do**

  Let $S_{r+1}, T_r \leftarrow \emptyset$.

  Let $B_r \leftarrow 3u_r|S_r|\log(mR/\beta)$.

  Let $z_r \leftarrow_{\mathrm{R}} \mathsf{Lap}\left(\frac{B_r}{\varepsilon}\right)^{\otimes|S_r|}$ and $q_r[S_r] \leftarrow \mathsf{tmean}_{B_r}(Y^r[S_r]) + z_r$.

  for $j \in S_r$ **do**

    if $q_r[j] < \tau_r$ **then**

      Add j to S_{r+1}.

    else

      Add j to T_r.

    end

  end

  Set $Z[T_r] \leftarrow \frac{1}{\sqrt{u_r}} \cdot Z[T_r]$.

  Set $T_P \leftarrow T_P \cup T_r$.

  // Update the loop's parameters.

  Set $u_{r+1} \leftarrow \frac{1}{2}u_r$, $\tau_{r+1} \leftarrow \frac{1}{2}\tau_r$, and $r \leftarrow r+1$.

end

// Run the sub-Gaussian learner from Hopkins et al. (2022b) restricted to T_P.

Let $\widehat{q}[T_P] \leftarrow_{\mathrm{R}} \textsc{DPSGLearner}_{\varepsilon,\frac{\alpha}{5},\beta,\sqrt{d}}(Z[T_P])$.

for $i \in [r-1]$ **do**

  Set $q[T_i] \leftarrow \sqrt{u_i} \cdot \widehat{q}[T_i]$.

end

// Final round.

Let $S_F \leftarrow [d] \setminus T_P$.

if $|S_F| \ge 1$ **then**

  Let $B_F \leftarrow 4\log(m_1/\beta)$.

  Let $z \leftarrow_{\mathrm{R}} \mathsf{Lap}\left(\frac{B_F}{\varepsilon}\right)^{\otimes|S_F|}$ and $q[S_F] \leftarrow \mathsf{tmean}_{B_F}(Y^F[S_F]) + z$.

end

// Return the final estimate.

return $Q = \mathrm{Ber}(q[1]) \otimes \cdots \otimes \mathrm{Ber}(q[d])$.

1. *If a coordinate j is filtered out in round r (i.e., $q_r[j] \geq \tau_r$), then $p[j]$ is large:*
$$p[j] \geq \tfrac{15\tau_r}{17}.$$
2. *If a coordinate j is not filtered out in round r (i.e., $q_r[j] < \tau_r$), then $p[j]$ is small:*
$$p[j] \leq u_{r+1} = \frac{u_r}{2}.$$

Therefore, if $S_P \subseteq [d]$ is the set of all the coordinates estimated in the partitioning rounds, then with probability at least $1 - O(\beta)$, $\mathrm{d_{TV}}(P[S_P], Q[S_P]) \leq \frac{\alpha}{2}$.

Proof We will prove the lemma by induction on r. Therefore, we will assume that in every round r, $p[j] \leq u_r$ for every $j \in S_r$ and prove that if this bound holds, then the two claims in the lemma hold. For the base of the induction, observe that, by assumption, $p[j] \leq u_1 = \frac{1}{2}$ for every $j \in S_1 = [d]$. In what follows we fix an arbitrary round $r \in [R]$. Throughout the proof, we will use the notation $\tilde{p}_r = \frac{1}{m}\sum_{i=1}^{m} Y_i^r$ to denote the empirical mean of the r-th block of samples.

Claim 7 (Sampling Error in Partioning Rounds) *If $\tilde{p}_r[j] = \frac{1}{m}\sum_{i=1}^{m} Y_i^r[j]$ and $m \geq 1024d\log(dR/\beta)$, then with probability at least $1 - \frac{2\beta}{R}$,*
$$\forall j \in S_r \quad |p[j] - \tilde{p}_r[j]| \leq \sqrt{\frac{4p[j]\log\left(\frac{dR}{\beta}\right)}{m}}.$$
Furthermore, if $p[j] \geq \frac{1}{d}$, then
$$|p[j] - \tilde{p}_r[j]| \leq \frac{p[j]}{16}.$$

Claim 8 (No Truncation in Partioning Rounds) *In round r, with probability at least $1 - \frac{\beta}{R}$, for every $Y_i^r \in Y^r$, we have that $|Y_i^r| \leq B_r$. So, no rows of Y^r are truncated while computing $\mathsf{tmean}_{B_r}(Y^r)$.*

Claim 9 (Error due to Privacy in Partioning Rounds) *With probability at least $1 - \frac{2\beta}{R}$,*
$$\forall j \in S_r \quad |\tilde{p}_r[j] - q_r[j]| \leq \frac{3u_r|S_r|\log\left(\frac{mR}{\beta}\right)\log\left(\frac{dR}{\beta}\right)}{\varepsilon m}.$$

Applying the triangle inequality and simplifying, using our choice of m, and noting that in our algorithm, $\tau_r = \frac{3u_r}{8}$, we get that in each round r, with probability at least $1 - \frac{4\beta}{R}$,
$$\begin{aligned}
|p[j] - q_r[j]| &\leq \sqrt{\frac{4p[j]\log\left(\frac{dR}{\beta}\right)}{m}} + \frac{3u_r|S_r|\log\left(\frac{mR}{\beta}\right)\log\left(\frac{dR}{\beta}\right)}{\varepsilon m} \\
&\leq \frac{p[j]}{16} + \frac{3u_r}{128} \qquad (1) \\
&= \frac{p[j]}{16} + \frac{\tau_r}{16}. \qquad (2)
\end{aligned}$$
To simplify our calculations, we will define $e_{r,j}$ to be the quantity on the right-hand-side of Inequality 2.

Claim 10 (Noisy Estimates above Threshold) *With probability at least* $1 - \frac{4\beta}{R}$*, for every* $j \in S_r$*,*

$$q_r[j] \geq \tau_r \implies p[j] \geq \frac{15\tau_r}{17}.$$

Proof We know that $|p_j - q_r[j]| \leq e_{r,j}$ with high probability, which implies that $p[j] \geq q_r[j] - e_{r,j}$. Now, given that $q_r[j] \geq \tau_r$ and the bound on $e_{r,j}$ from Inequality 2, we have the following.

$$p[j] \geq q_r[j] - e_{r,j} \geq \tau_r - \frac{p[j]}{16} - \frac{\tau_r}{16} \implies p[j] \geq \frac{15\tau_r}{17}$$

This completes the proof. ■

Claim 11 (Noisy Estimates below Threshold) *With probability at least* $1 - \frac{4\beta}{R}$*, for every* $j \in S_r$*,*

$$q_r[j] < \tau_r \implies p[j] \leq u_{r+1} = \frac{u_r}{2}.$$

Proof We know that with high probability, $p_j \leq q_r[j] + e_{r,j}$. But since $q_r[j] < \tau_j$, we know that $p[j] < \tau_r + e_{r,j}$. Also, we set $\tau_r = \frac{3}{4}u_{r+1}$ in our algorithm. Using the bound on $e_{r,j}$ from Inequality 1, this gives us the following.

$$p[j] < \tau_r + e_{r,j} \leq \frac{3u_{r+1}}{4} + \frac{p[j]}{16} + \frac{3u_r}{128} = \frac{3u_{r+1}}{4} + \frac{p[j]}{16} + \frac{3u_{r+1}}{256} \implies p[j] \leq \frac{13u_{r+1}}{16} < u_{r+1} = \frac{u_r}{2}$$

Our proof is complete. ■

Claim 11 completes the inductive step of the proof. It establishes that at the beginning of round $r+1$, $p_j \leq u_{r+1}$ for all $j \in S_{r+1}$.

Finally, we analyse the accuracy of the coordinates collected in T_P over all the partitioning rounds by the call to DPSGLEARNER. We can bound the total-variation distance between $P[S_P]$ and $Q[S_P]$ by computing the χ^2 divergence between the two. We make two key observations here.

- In round r, we scale $Z[T_r]$ by $\frac{1}{\sqrt{u_r}}$, which means that the (scaled product) distribution of that subset of the coordinates (denoted by $\widehat{P}[T_r]$) is over $\left\{0, \frac{1}{\sqrt{u_r}}\right\}$. Let Σ_{T_r} be its covariance matrix. Then we can observe that the eigenvalues of Σ_{T_r} lie between $\frac{\tau_r(1-\tau_r)}{u_r} = \frac{3(1-\frac{3u_r}{8})}{8} \geq \frac{39}{64}$ and $\frac{u_r(1-u_r)}{u_r} = 1 - u_r \leq 1$. Since this is true for any r, this must be true for the entire $\widehat{P}[T_P]$, as well.

- Let the output product distribution from invoking DPSGLEARNER on $Z[T_P]$ be $\widehat{Q}[T_P]$ (which has mean $\widehat{q}[T_P]$). Suppose the covariance matrix of $\widehat{P}[T_P]$ is Σ_{P_P}, and the value of r before the update in the end of the final iteration of the **While**-loop was r^*. Then from Theorem 29 (accuracy guarantees of DPSGLEARNER) and our setting of the accuracy parameters in the call to DPSGLEARNER in Algorithm 1, with probability at least $1-\beta$, the squared ℓ_2 distance

between $\widehat{P}[T_P]$ and $\widehat{Q}[T_P]$ is given by,

$$\begin{aligned}
\frac{\alpha^2}{25} &\geq \|\widehat{q}[T_P] - \widehat{p}[T_P]\|_2^2 \\
&= \sum_{i\in[r^*]} \|\widehat{q}[T_i] - \widehat{p}[T_i]\|_2^2 \\
&= \sum_{i\in[r^*]} \frac{\|q[T_i] - p[T_i]\|_2^2}{u_i} \\
&= \sum_{i\in[r^*]} \frac{3\|q[T_i] - p[T_i]\|_2^2}{8\tau_i} && (\tau_i = \tfrac{3u_i}{8}) \\
&\geq \sum_{i\in[r^*]} \sum_{j\in T_i} \frac{45\cdot(q[j]-p[j])^2}{136\cdot p[j]} && \text{(Claim 10)} \\
&= \frac{45}{544}\cdot \sum_{j\in T_P} \frac{4(q[j]-p[j])^2}{p[j]}.
\end{aligned}$$

On rearranging the above, this implies that,

$$\sum_{j\in T_P} \frac{4(q[j]-p[j])^2}{p[j]} \leq \frac{\alpha^2}{2}.$$

The above, on combining with Lemmata 18, 19, and 20, implies that $\mathrm{d_{TV}}(P[T_P], Q[T_P]) \leq \frac{\alpha}{2}$.

Now, we can take the union bound over all the failure events in all the rounds and over the failure of DPSGLEARNER, so that the conclusions of Lemma 6 hold with probability $1-5\beta$. This completes the proof. ■

2.2.2. ANALYSIS OF THE FINAL ROUND

In this section we show that the TV error of the coordinates j, such that $q[j]$ was set in the final round, is small.

Lemma 12 (Final Round) *In the final round, let $k \in [R+1]$ for which $u_k|S_k| < 1$. If $p[j] \leq u_k$ for every $j \in S_F$, and Y^F contains at least*

$$m_1 = \frac{128d\log(d/\beta)}{\alpha^2} + \frac{256d\log(d/\varepsilon\alpha\beta)}{\alpha\varepsilon}$$

i.i.d. samples from P, then with probability at least $1-O(\beta)$, then $\mathrm{d_{TV}}(P[S_F], Q[S_F]) \leq \frac{\alpha}{2}$.

Proof Again, we use the notation, $\tilde{p} = \frac{1}{m_1}\sum_{i=1}^{m_1} Y_i^F$, for the rest of this proof. First, we have two claims that bound the difference between $p[j]$ and $\tilde{p}[j]$.

Claim 13 (Sampling Error for Large Coordinates in Final Round) *For each $j \in S_F$, such that $p_j > \frac{1}{d}$, with probability at least $1-2\beta/d$, we have,*

$$|p[j] - \tilde{p}[j]| \leq \sqrt{\frac{4p[j]\log\left(\frac{d}{\beta}\right)}{m_1}}.$$

Claim 14 (Sampling Error for Small Coordinates in Final Round) *For each $j \in S_F$, such that $p_j \leq \frac{1}{d}$, with probability at least $1 - 2\beta/d$, we have,*

$$|p[j] - \tilde{p}[j]| \leq \frac{\alpha}{8d}.$$

Claim 15 (No Truncation in Final Round) *With probability at least $1 - \beta$, for every $Y_i^F \in Y^F$, $\left|Y_i^F\right| \leq B_F$, so no rows of Y^F are truncated in the computation of $\mathsf{tmean}_{B_F}(Y^F)$.*

Claim 16 (Error due to Privacy in Final Round) *With probability at least $1 - 2\beta$,*

$$\forall j \in S_r \quad |\tilde{p}[j] - q_r[j]| \leq \frac{4 \log\left(\frac{m_1}{\beta}\right) \log\left(\frac{d}{\beta}\right)}{\varepsilon m_1}.$$

Note that because no truncation happens in this round, the sampling error without the Laplace noise is bounded by the quantity as specified in Claims 13 and 14. Because of the possible difference in magnitudes of the means of the marginals in S_F, we establish their error guarantees separately. Let $H \subset S_F$ be the set of all coordinates, whose means are greater than $\frac{1}{d}$, i.e., $H \coloneqq \{j \in S_F : p[j] > 1/d\}$. Likewise, let $L \subseteq S_F$ be the set of lighter coordinates, i.e., $L \coloneqq S_F \setminus H$.

We analyse the heavier coordinates in H first. Let $\widetilde{P}[H]$ be the product distribution over the coordinates in H with mean $\tilde{p}[H]$. For all $j \in H$, using Claim 13 and our choice of m_1, we know that

$$|\tilde{p}[j] - p[j]| \leq \sqrt{\frac{4p[j] \log\left(\frac{d}{\beta}\right)}{m_1}} \implies \frac{4(p[j] - \tilde{p}[j])^2}{p[j]} \leq \frac{16 \log\left(\frac{d}{\beta}\right)}{m_1} \leq \frac{\alpha^2}{32d},$$

which (by Lemma 18) means that $\mathrm{d}_{\chi^2}(\widetilde{P}[j], P[j]) \leq \frac{\alpha^2}{32}$. Combined with Lemma 20, this implies that $\mathrm{d}_{\mathrm{TV}}(P[j], \widetilde{P}[j]) \leq \frac{\alpha}{8d}$. Now, from Claim 16 and our choice of m_1, we know that for all $j \in H$,

$$|\tilde{p}[j] - q[j]| \leq \frac{4 \log\left(\frac{m_1}{\beta}\right) \log\left(\frac{d}{\beta}\right)}{\varepsilon m_1} \leq \frac{\alpha}{8d},$$

which implies that $\mathrm{d}_{\mathrm{TV}}(\widetilde{P}[j], Q[j]) \leq \frac{\alpha}{8d}$. Therefore, by triangle inequality, $\mathrm{d}_{\mathrm{TV}}(P[j], Q[j]) \leq \frac{\alpha}{4d}$. Finally, from Lemma 19, we have that $\mathrm{d}_{\mathrm{TV}}(P[H], Q[H]) \leq \frac{\alpha}{4}$.

Next, we bound the error on the lighter coordinates in L. Claims 14 and 16, the triangle inequality, and our choice of m_1 show that for all $j \in L$,

$$\begin{aligned} |p[j] - q[j]| &\leq \frac{\alpha}{8d} + \frac{4 \log\left(\frac{m_1}{\beta}\right) \log\left(\frac{d}{\beta}\right)}{\varepsilon m_1} \\ &\leq \frac{\alpha}{8d} + \frac{\alpha}{8d} \\ &= \frac{\alpha}{4d}. \end{aligned}$$

This implies that for all $j \in L$, $\mathrm{d}_{\mathrm{TV}}(P[j], Q[j]) \leq \frac{\alpha}{4d}$. Therefore, from Lemma 19, $\mathrm{d}_{\mathrm{TV}}(P[L], Q[L]) \leq \frac{\alpha}{4}$.

Finally, through an application of Lemma 19 again, we obtain that

$$\mathrm{d}_{\mathrm{TV}}(P[S_F], Q[S_F]) \leq \mathrm{d}_{\mathrm{TV}}(P[H], Q[H]) + \mathrm{d}_{\mathrm{TV}}(P[L], Q[L]) \leq \frac{\alpha}{2}.$$

This completes our proof. ■

2.2.3. PUTTING IT ALL TOGETHER

In this section, we combine Lemmata 6 and 12 to prove Proposition 5. First, by Lemma 6, with probability at least $1 - O(\beta)$, if S_P is the set of coordinates j, such that $q[j]$ was set in any of the partitioning rounds, then,

1. $\mathrm{d_{TV}}(P[S_P], Q[S_P]) \leq \frac{\alpha}{2}$ and
2. if $j \notin S_F$ and k is the index of the final round, then $p[j] \leq u_k$ and $u_k \left|S_F\right| < 1$.

Next, due to the second consequence listed above, we can apply Lemma 12 to obtain that if S_F consists of all coordinates set in the final round, then with probability at least $1 - O(\beta)$, $\mathrm{d_{TV}}(P[S_F], Q[S_F]) \leq \frac{\alpha}{2}$. Finally, we use the union bound and Lemma 19 to conclude that, with probability at least $1 - O(\beta)$,

$$\mathrm{d_{TV}}(P, Q) \leq \mathrm{d_{TV}}(P[S_P], Q[S_P]) + \mathrm{d_{TV}}(P[S_F], Q[S_F]) \leq \alpha.$$

This completes the proof of Proposition 5.

Acknowledgments

We would like to thank Gautam Kamath for their helpful discussion on this problem. The author was supported by an NSERC Discovery Grant throughout the duration of this project.

References

Jayadev Acharya, Ziteng Sun, and Huanyu Zhang. Differentially private assouad, fano, and le cam. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, ALT '21, pages 48–78. JMLR, Inc., 2021.

Ishaq Aden-Ali, Hassan Ashtiani, and Gautam Kamath. On the sample complexity of privately learning unbounded high-dimensional gaussians. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory*, ALT '21, pages 185–216. JMLR, Inc., 2021a.

Ishaq Aden-Ali, Hassan Ashtiani, and Christopher Liaw. Privately learning mixtures of axis-aligned gaussians. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021b.

Daniel Alabi, Pravesh K Kothari, Pranay Tankala, Prayaag Venkat, and Fred Zhang. Privately estimating a Gaussian: Efficient, robust and optimal. *arXiv preprint arXiv:2212.08018*, 2022.

Hassan Ashtiani and Christopher Liaw. Private and polynomial time algorithms for learning Gaussians and beyond. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 1075–1076, 2022.

Marco Avella-Medina and Victor-Emmanuel Brunel. Differentially private sub-Gaussian location estimators. *arXiv preprint arXiv:1906.11923*, 2019.

Brendan Avent, Yatharth Dubey, and Aleksandra Korolova. The power of the hybrid model for mean estimation. *Proceedings on Privacy Enhancing Technologies*, 2020(4):48–68, 2019.

Rina Foygel Barber and John C Duchi. Privacy and statistical risk: Formalisms and minimax bounds. *arXiv preprint arXiv:1412.4451*, 2014.

Raef Bassily, Kobbi Nissim, Adam Smith, Thomas Steinke, Uri Stemmer, and Jonathan Ullman. Algorithmic stability for adaptive data analysis. In *Proceedings of the 48th Annual ACM Symposium on the Theory of Computing*, STOC '16, pages 1046–1059, New York, NY, USA, 2016. ACM.

Omri Ben-Eliezer, Dan Mikulincer, and Ilias Zadik. Archimedes meets privacy: On privately estimating quantiles in high dimensions under minimal assumptions. In *Advances in Neural Information Processing Systems 35*, NeurIPS '22. Curran Associates, Inc., 2022.

Alex Bie, Gautam Kamath, and Vikrant Singhal. Private estimation with public data. In *Advances in Neural Information Processing Systems 35*, NeurIPS '22. Curran Associates, Inc., 2022.

Sourav Biswas, Yihe Dong, Gautam Kamath, and Jonathan Ullman. Coinpress: Practical private mean and covariance estimation. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20, pages 14475–14485. Curran Associates, Inc., 2020.

Gavin Brown, Marco Gaboardi, Adam Smith, Jonathan Ullman, and Lydia Zakynthinou. Covariance-aware private mean estimation without private covariance estimation. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

Mark Bun and Thomas Steinke. Concentrated differential privacy: Simplifications, extensions, and lower bounds. In *Proceedings of the 14th Conference on Theory of Cryptography*, TCC '16-B, pages 635–658, Berlin, Heidelberg, 2016. Springer.

Mark Bun and Thomas Steinke. Average-case averages: Private algorithms for smooth sensitivity and mean estimation. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 181–191. Curran Associates, Inc., 2019.

Mark Bun, Jonathan Ullman, and Salil Vadhan. Fingerprinting codes and the price of approximate differential privacy. In *Proceedings of the 46th Annual ACM Symposium on the Theory of Computing*, STOC '14, pages 1–10, New York, NY, USA, 2014. ACM.

Mark Bun, Kobbi Nissim, Uri Stemmer, and Salil Vadhan. Differentially private release and learning of threshold functions. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 634–649, Washington, DC, USA, 2015. IEEE Computer Society.

Mark Bun, Thomas Steinke, and Jonathan Ullman. Make up your mind: The price of online queries in differential privacy. In *Proceedings of the 28th Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '17, pages 1306–1325, Philadelphia, PA, USA, 2017. SIAM.

Mark Bun, Gautam Kamath, Thomas Steinke, and Zhiwei Steven Wu. Private hypothesis selection. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 156–167. Curran Associates, Inc., 2019.

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

Hongjie Chen, Vincent Cohen-Addad, Tommaso d'Orsi, Alessandro Epasto, Jacob Imola, David Steurer, and Stefan Tiegel. Private estimation algorithms for stochastic block models and mixture models. *arXiv preprint arXiv:2301.04822*, 2023.

Christian Covington, Xi He, James Honaker, and Gautam Kamath. Unbiased statistical estimation and valid confidence intervals under differential privacy. *arXiv preprint arXiv:2110.14465*, 2021.

Rachel Cummings and David Durfee. Individual sensitivity preprocessing for data privacy. In *Proceedings of the 31st Annual ACM-SIAM Symposium on Discrete Algorithms*, SODA '20, Philadelphia, PA, USA, 2020. SIAM.

Ilias Diakonikolas, Moritz Hardt, and Ludwig Schmidt. Differentially private learning of structured discrete distributions. In *Advances in Neural Information Processing Systems 28*, NIPS '15, pages 2566–2574. Curran Associates, Inc., 2015.

Wenxin Du, Canyon Foot, Monica Moniot, Andrew Bray, and Adam Groce. Differentially private confidence intervals. *arXiv preprint arXiv:2001.02285*, 2020.

John Duchi, Saminul Haque, and Rohith Kuditipudi. A fast algorithm for adaptive private mean estimation. *arXiv preprint arXiv:2301.07078*, 2023.

John C. Duchi, Michael I. Jordan, and Martin J. Wainwright. Local privacy and statistical minimax rates. In *Proceedings of the 54th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '13, pages 429–438, Washington, DC, USA, 2013. IEEE Computer Society.

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

Cynthia Dwork and Guy N. Rothblum. Concentrated differential privacy. *arXiv preprint arXiv:1603.01887*, 2016.

Cynthia Dwork, Frank McSherry, Kobbi Nissim, and Adam Smith. Calibrating noise to sensitivity in private data analysis. In *Proceedings of the 3rd Conference on Theory of Cryptography*, TCC '06, pages 265–284, Berlin, Heidelberg, 2006. Springer.

Cynthia Dwork, Guy N. Rothblum, and Salil Vadhan. Boosting and differential privacy. In *Proceedings of the 51st Annual IEEE Symposium on Foundations of Computer Science*, FOCS '10, pages 51–60, Washington, DC, USA, 2010. IEEE Computer Society.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015a.

Cynthia Dwork, Adam Smith, Thomas Steinke, Jonathan Ullman, and Salil Vadhan. Robust traceability from trace amounts. In *Proceedings of the 56th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '15, pages 650–669, Washington, DC, USA, 2015b. IEEE Computer Society.

Vitaly Feldman and Thomas Steinke. Generalization for adaptively-chosen estimators via stable median. In *Conference on Learning Theory*, pages 728–757. PMLR, 2017.

Anand Jerry George, Lekshmi Ramesh, Aditya Vikram Singh, and Himanshu Tyagi. Continual mean estimation under user-level privacy. *arXiv preprint arXiv:2212.09980*, 2022.

Kristian Georgiev and Samuel B Hopkins. Privacy induces robustness: Information-computation gaps and sparse mean estimation. In *Advances in Neural Information Processing Systems 35*, NeurIPS '22. Curran Associates, Inc., 2022.

Moritz Hardt and Jonathan Ullman. Preventing false discovery in interactive data analysis is hard. In *Proceedings of the 55th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '14, pages 454–463, Washington, DC, USA, 2014. IEEE Computer Society.

Samuel B Hopkins, Gautam Kamath, and Mahbod Majid. Efficient mean estimation with pure differential privacy via a sum-of-squares exponential mechanism. 2022a.

Samuel B. Hopkins, Gautam Kamath, Mahbod Majid, and Shyam Narayanan. Robustness implies privacy in statistical estimation, 2022b.

Ziyue Huang, Yuting Liang, and Ke Yi. Instance-optimal mean estimation under differential privacy. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

Gautam Kamath and Jonathan Ullman. A primer on private statistics. *arXiv preprint arXiv:2005.00010*, 2020.

Gautam Kamath, Jerry Li, Vikrant Singhal, and Jonathan Ullman. Privately learning high-dimensional distributions. In *Proceedings of the 32nd Annual Conference on Learning Theory*, COLT '19, pages 1853–1902, 2019a.

Gautam Kamath, Or Sheffet, Vikrant Singhal, and Jonathan Ullman. Differentially private algorithms for learning mixtures of separated Gaussians. In *Advances in Neural Information Processing Systems 32*, NeurIPS '19, pages 168–180. Curran Associates, Inc., 2019b.

Gautam Kamath, Vikrant Singhal, and Jonathan Ullman. Private mean estimation of heavy-tailed distributions. In *Proceedings of the 33rd Annual Conference on Learning Theory*, COLT '20, pages 2204–2235, 2020.

Gautam Kamath, Xingtu Liu, and Huanyu Zhang. Improved rates for differentially private stochastic convex optimization with heavy-tailed data. In *Proceedings of the 39th International Conference on Machine Learning*, ICML '22, pages 10633–10660. JMLR, Inc., 2022a.

Gautam Kamath, Argyris Mouzakis, and Vikrant Singhal. New lower bounds for private estimation and a generalized fingerprinting lemma. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24405–24418. Curran Associates, Inc., 2022b. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/9a6b27821896649919491f55ccf8b75-Paper-Conference.pdf.

Gautam Kamath, Argyris Mouzakis, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. A private and computationally-efficient estimator for unbounded gaussians. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 544–572, 2022c.

Gautam Kamath, Argyris Mouzakis, Matthew Regehr, Vikrant Singhal, Thomas Steinke, and Jonathan Ullman. A bias-variance-privacy trilemma for statistical estimation, 2023.

Vishesh Karwa and Salil Vadhan. Finite sample differentially private confidence intervals. In *Proceedings of the 9th Conference on Innovations in Theoretical Computer Science*, ITCS '18, pages 44:1–44:9, Dagstuhl, Germany, 2018. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik.

Pravesh K Kothari, Pasin Manurangsi, and Ameya Velingker. Private robust estimation by stabilizing convex relaxations. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 723–777, 2022.

Daniel Levy, Ziteng Sun, Kareem Amin, Satyen Kale, Alex Kulesza, Mehryar Mohri, and Ananda Theertha Suresh. Learning with user-level privacy. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

Xiyang Liu, Weihao Kong, Sham Kakade, and Sewoong Oh. Robust and differentially private mean estimation. In *Advances in Neural Information Processing Systems 34*, NeurIPS '21. Curran Associates, Inc., 2021.

Xiyang Liu, Weihao Kong, and Sewoong Oh. Differential privacy and robust statistics in high dimensions. In *Proceedings of the 35th Annual Conference on Learning Theory*, COLT '22, pages 1167–1246, 2022.

Yuhan Liu, Ananda Theertha Suresh, Felix Yu, Sanjiv Kumar, and Michael Riley. Learning discrete distributions: User vs item-level privacy. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20. Curran Associates, Inc., 2020.

Frank McSherry and Kunal Talwar. Mechanism design via differential privacy. In *Proceedings of the 48th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '07, pages 94–103, Washington, DC, USA, 2007. IEEE Computer Society.

Aleksandar Nikolov and Haohua Tang. Gaussian noise is nearly instance optimal for private unbiased mean estimation. *arXiv preprint arXiv:2301.13850*, 2023.

Kelly Ramsay and Shoja'eddin Chenouri. Differentially private depth functions and their associated medians. *arXiv preprint arXiv:2101.02800*, 2021.

Kelly Ramsay, Aukosh Jagannath, and Shoja'eddin Chenouri. Concentration of the exponential mechanism and differentially private multivariate medians. *arXiv preprint arXiv:2210.06459*, 2022.

Ryan Rogers, Aaron Roth, Adam Smith, and Om Thakkar. Max-information, differential privacy, and post-selection hypothesis testing. In *Proceedings of the 57th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '16, pages 487–494, Washington, DC, USA, 2016. IEEE Computer Society.

Adam Smith. Privacy-preserving statistical estimation with optimal convergence rates. In *Proceedings of the 43rd Annual ACM Symposium on the Theory of Computing*, STOC '11, pages 813–822, New York, NY, USA, 2011. ACM.

Thomas Steinke and Jonathan Ullman. Interactive fingerprinting codes and the hardness of preventing false discovery. In *Proceedings of the 28th Annual Conference on Learning Theory*, COLT '15, pages 1588–1628, 2015.

Thomas Steinke and Jonathan Ullman. Between pure and approximate differential privacy. *The Journal of Privacy and Confidentiality*, 7(2):3–22, 2017a.

Thomas Steinke and Jonathan Ullman. Tight lower bounds for differentially private selection. In *Proceedings of the 58th Annual IEEE Symposium on Foundations of Computer Science*, FOCS '17, pages 552–563, Washington, DC, USA, 2017b. IEEE Computer Society.

Eliad Tsfadia, Edith Cohen, Haim Kaplan, Yishay Mansour, and Uri Stemmer. Friendlycore: Practical differentially private aggregation. In *Proceedings of the 39th International Conference on Machine Learning*, ICML '22, pages 21828–21863. JMLR, Inc., 2022.

Christos Tzamos, Emmanouil-Vasileios Vlatakis-Gkaragkounis, and Ilias Zadik. Optimal private median estimation under minimal distributional assumptions. In *Advances in Neural Information Processing Systems 33*, NeurIPS '20, pages 3301–3311. Curran Associates, Inc., 2020.

Duy Vu and Aleksandra Slavković. Differential privacy for clinical trial data: Preliminary evaluations. In *2009 IEEE International Conference on Data Mining Workshops*, ICDMW '09, pages 138–143. IEEE, 2009.

Di Wang, Hanshen Xiao, Srinivas Devadas, and Jinhui Xu. On differentially private stochastic convex optimization with heavy-tailed data. In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, pages 10081–10091. JMLR, Inc., 2020.

Larry Wasserman and Shuheng Zhou. A statistical framework for differential privacy. *Journal of the American Statistical Association*, 105(489):375–389, 2010.

Huanyu Zhang, Gautam Kamath, Janardhan Kulkarni, and Zhiwei Steven Wu. Privately learning Markov random fields. In *Proceedings of the 37th International Conference on Machine Learning*, ICML '20, pages 11129–11140. JMLR, Inc., 2020.

Appendix A. Related Work

Besides a large selection of folklore work in non-private estimation of distributions, there has been a lot of work in recent years on differentially private statistical estimation. Mean estimation is possibly the most fundamental question in this space, enjoying significant attention (e.g., Barber and Duchi (2014); Duchi et al. (2013); Karwa and Vadhan (2018); Bun and Steinke (2019); Kamath et al. (2019a, 2020); Wang et al. (2020); Du et al. (2020); Biswas et al. (2020); Cai et al. (2021); Brown et al. (2021); Huang et al. (2021); Liu et al. (2021, 2022); Kamath et al. (2022a); Hopkins et al. (2022a); Kothari et al. (2022); Tsfadia et al. (2022); Duchi et al. (2023); Covington et al. (2021);

Nikolov and Tang (2023); Kamath et al. (2023)). Other related problems include private covariance or density estimation Bun et al. (2015, 2019); Aden-Ali et al. (2021a); Kamath et al. (2022c); Ashtiani and Liaw (2022); Alabi et al. (2022); Hopkins et al. (2022b); Kothari et al. (2022); Tsfadia et al. (2022); Liu et al. (2022); Biswas et al. (2020). Beyond these settings, other works have examined statistical estimation under differential privacy constraints for mixtures of Gaussians Kamath et al. (2019b); Aden-Ali et al. (2021b); Chen et al. (2023), graphical models Zhang et al. (2020), discrete distributions Diakonikolas et al. (2015), median estimation Avella-Medina and Brunel (2019); Tzamos et al. (2020); Ramsay and Chenouri (2021); Ramsay et al. (2022); Ben-Eliezer et al. (2022); Cummings and Durfee (2020), and more. Several recent works have explored the connections between privacy and robustness Liu et al. (2021); Hopkins et al. (2022a); Georgiev and Hopkins (2022); Liu et al. (2022); Kothari et al. (2022); Alabi et al. (2022); Hopkins et al. (2022b); Chen et al. (2023), and between privacy and generalization Hardt and Ullman (2014); Dwork et al. (2015a); Steinke and Ullman (2015); Bassily et al. (2016); Rogers et al. (2016); Feldman and Steinke (2017). Upcoming directions of interest include ensuring privacy when one individual may contribute multiple data points Liu et al. (2020); Levy et al. (2021); George et al. (2022) (or what is known as, *user-level differential privacy*), a combination of local and central DP for different users Avent et al. (2019), and estimation with access to trace amounts of public data Bie et al. (2022). We refer the reader to Kamath and Ullman (2020) for more coverage of the recent work on differentially private statistical estimation. Differentially private statistical inference also has been an active area of research for over a decade (e.g., Dwork and Lei (2009); Vu and Slavković (2009); Wasserman and Zhou (2010); Smith (2011)), but the literature is too broad to fully summarize here.

Another broader line of work includes those on the minimax sample complexities for various differentially private statistical estimation tasks. The first minimax sample complexity bounds to show an asymptotic separation between private and non-private estimation for private mean estimation were proved in Bun et al. (2014), and subsequently sharpened and generalized in several ways Dwork et al. (2015b); Bun et al. (2017); Steinke and Ullman (2017a,b); Kamath et al. (2019a). Recently, Cai et al. (2021) extended these bounds to sparse estimation and regression problems. Acharya et al. (2021) provides an alternative, user-friendly approach to proving sample complexity bounds, which is directly analogous to the classical approaches in statistics and learning theory for proving minimax lower bounds. Kamath et al. (2022b) also provides a generalized version of the well-known "fingerprinting technique" that is used to prove hardness results for these kinds of problems under approximate DP.

The two prior works most relevant to ours among the above are those by Kamath et al. (2019a); Hopkins et al. (2022b) on learning binary product distributions under concentrated and approximate DP, and estimating means of sub-Gaussian distributions under pure DP, respectively. The question of learning binary product distributions optimally in polynomial time under pure DP has stayed open for a while now. In our work, we modify and use a combination of their techniques to solve this problem.

Appendix B. Preliminaries

For the utility analysis, we mostly rely on the concentration properties of Bernoulli distributions and the Laplace distribution, and on the relationships among different distance metrics for probability

distributions. The privacy analysis is much simpler, and we use the privacy guarantees of the existing differentially private mechanisms. We describe all these notions and results in this section.

B.1. Statistics Preliminaries

Here, we state the essential definitions and results from statistics that would be used throughout the draft. They include descriptions of various metrics for probability distributions, and a few useful concentration inequalities.

Notations. Let $P_1, \ldots, P_k$ be distributions over domains $\mathcal{X}_1, \ldots, \mathcal{X}_k$, respectively. Then we say that $P = P_1 \otimes \cdots \otimes P_k$ is a product distribution over $\mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_k$. Next, for a distribution P over domain $\mathcal{X}$, we use $P^{\otimes k} = P \otimes \cdots \otimes P$ (k times) to denote the product distribution over $\mathcal{X} \otimes \cdots \otimes \mathcal{X}$ (k times), where each marginal is P. Also, for $0 \leq p \leq 1$, we use $\mathrm{Ber}(p)$ to denote a Bernoulli random variable with mean p. Finally, for any $v = (v_1, \ldots, v_d) \in \mathbb{R}^d$, we use $|v|$ to denote its ℓ_1 norm, i.e., $|v| = \sum_{i=1}^{d} |v_i|$.

B.1.1. Distances Between Distributions

We use several notions of distance metrics between distributions.

Definition 17 *If P, Q are distributions, then,*

- *the* statistical distance *or the* total-variation distance *is* $\mathrm{d}_{\mathrm{TV}}(P, Q) = \frac{1}{2} \sum_x |P(x) - Q(x)|$,
- *the* χ^2-divergence *is* $\mathrm{d}_{\chi^2}(P \| Q) = \sum_x \frac{(P(x) - Q(x))^2}{Q(x)}$, *and*
- *the* KL-divergence *is* $\mathrm{d}_{\mathrm{KL}}(P \| Q) = \sum_x P(x) \log \frac{P(x)}{Q(x)}$.

Next, we have the following bound on the χ^2-divergence between two Bernoulli distributions.

Lemma 18 (χ^2-Divergence between Bernoulli Distributions) *For Bernoulli distributions P and Q over $\{0, 1\}$ (with means p and q, respectively), such that $|p - q| \leq \frac{1}{4}$ and $p \leq \frac{1}{2}$,*

$$\mathrm{d}_{\chi^2}(P, Q) \leq \frac{4(p - q)^2}{q}.$$

Proof The proof is identical to that of Claim 5.12 from Kamath et al. (2019a). ■

For product distributions $P = P_1 \otimes \cdots \otimes P_k$ and $Q = Q_1 \otimes \cdots \otimes Q_k$, the KL-divergence is additive, and the total-variation distance and the χ^2-divergence are sub-additive. In particular, we have the following.

Lemma 19 (Sub-Additivity under Product Distributions) *Let $P = P_1 \otimes \cdots \otimes P_k$ and $Q = Q_1 \otimes \cdots \otimes Q_k$ be two product distributions. Then,*

- $\mathrm{d}_{\mathrm{TV}}(P, Q) \leq \sum_{j=1}^{d} \mathrm{d}_{\mathrm{TV}}(P_j, Q_j)$,
- $\mathrm{d}_{\chi^2}(P \| Q) \leq \sum_{j=1}^{d} \mathrm{d}_{\chi^2}(P_j \| Q_j)$, *and*

- $\mathrm{d}_{\mathrm{KL}}(P\|Q) = \sum_{j=1}^{d} \mathrm{d}_{\mathrm{KL}}(P_j\|Q_j)$.

The three metrics are related to each other in a clean way as follows.

Lemma 20 (Pinsker's Inequality) *For any two distributions P and Q, we have,*

$$2 \cdot \mathrm{d}_{\mathrm{TV}}(P,Q)^2 \leq \mathrm{d}_{\mathrm{KL}}(P\|Q) \leq \mathrm{d}_{\chi^2}(P\|Q).$$

B.1.2. TAIL BOUNDS

We use a few tail bounds for sums of independent Bernoulli random variables. The first lemma is an additive form of the Chernoff bound.

Lemma 21 (Bernstein's Inequality) *For every $p > 0$, if $X_1, \ldots, X_m$ are i.i.d. samples from $\mathrm{Ber}(p)$, then for every $\varepsilon > 0$,*

$$\mathbb{P}\left[\frac{1}{m}\sum_{i=1}^{m} X_i \geq p + \varepsilon\right] \leq e^{-\mathrm{d}_{\mathrm{KL}}(p+\varepsilon\|p)\cdot m} \quad \textit{and} \quad \mathbb{P}\left[\frac{1}{m}\sum_{i=1}^{m} X_i \leq p - \varepsilon\right] \leq e^{-\mathrm{d}_{\mathrm{KL}}(p-\varepsilon\|p)\cdot m}$$

The second lemma is an multiplicative form of the Chernoff bound.

Lemma 22 (Multiplicative Chernoff Bound) *For every $p > 0$, if $X_1, \ldots, X_m$ are i.i.d. samples from $\mathrm{Ber}(p)$, then for every $\delta \geq 0$,*

$$\mathbb{P}\left[\sum_{i=1}^{m} X_i \geq (1+\delta)pm\right] \leq e^{-\frac{\delta^2 pm}{2+\delta}}$$

The next lemma follows from Lemma 22, and bounds the norms of points sampled from a binary product distribution with bounded marginals.

Lemma 23 (Bounded Norms of Rows) *Suppose $X_1, \ldots, X_m$ are sampled i.i.d. from a product distribution P over $\{0,1\}^t$, where the mean of each coordinate is upper bounded by p (i.e., $\mathbb{E}[P] \preceq p$). Then,*

1. *if $pt \geq 1$, then for each i, $\mathbb{P}\left[|X_i| \geq pt\left(1 + 2\log(\frac{m}{\beta})\right)\right] \leq \frac{\beta}{m}$, and*
2. *if $pt < 1$, then for each i, $\mathbb{P}\left[|X_i| \geq 4\log(\frac{m}{\beta})\right] \leq \frac{\beta}{m}$.*

Proof In the first case, we apply Lemma 22 after setting $\delta = 2\log(m/\beta)$ and $\mu = pm$, and by noting that $\log(m/\beta) \geq 1$. In the second case, we do the same as in the first case, but set $\delta = \frac{2\log(m/\beta)}{pm}$. ■

Finally, we describe the concentration of Laplace random variables with mean 0.

Lemma 24 (Laplace Concentration) *Let $Z \sim \mathsf{Lap}(t)$. Then $\mathbb{P}[|Z| > t \cdot \ln(1/\beta)] \leq \beta$.*

B.2. Privacy Preliminaries

We start with the definition of differential privacy.

Definition 25 (Differential Privacy (DP) Dwork et al. (2006)) *A randomized algorithm $M : \mathcal{X}^n \to \mathcal{Y}$ satisfies (ε, δ)-differential privacy ((ε, δ)-DP) if for every pair of neighboring datasets $X, X' \in \mathcal{X}^n$ (i.e., datasets that differ in at most one entry, denoted by $X \sim X'$),*

$$\forall\, Y \subseteq \mathcal{Y}, \quad \mathbb{P}\left[M(X) \in Y\right] \leq e^{\varepsilon} \cdot \mathbb{P}\left[M(X') \in Y\right] + \delta.$$

When $\delta = 0$, we say that M satisfies ε-differential privacy or pure differential privacy.

These definitions of DP are closed under post-processing.

Lemma 26 (Post-Processing Dwork et al. (2006)) *If $M : \mathcal{X}^n \to \mathcal{Y}$ is (ε, δ)-DP, and $P : \mathcal{Y} \to \mathcal{Z}$ is any randomized function, then the algorithm $P \circ M$ is (ε, δ)-DP.*

B.2.1. KNOWN DIFFERENTIALLY PRIVATE MECHANISMS

We state a standard result on achieving differential privacy via noise addition proportional to the *sensitivity* of the function being computed Dwork et al. (2006).

Definition 27 (Sensitivity) *Let $f : \mathcal{X}^n \to \mathbb{R}^d$ be a function, its* ℓ_1-sensitivity *is*

$$\Delta_{f,1} \coloneqq \max_{X \sim X' \in \mathcal{X}^n} \left| f(X) - f(X') \right|.$$

For a function with bounded ℓ_1-sensitivity, we can achieve ε-DP by adding noise from a Laplace distribution scaled to its ℓ_1-sensitivity.

Lemma 28 (Laplace Mechanism) *Let $f : \mathcal{X}^n \to \mathbb{R}^d$ be a function with ℓ_1-sensitivity $\Delta_{f,1}$. Then the Laplace mechanism*

$$M(X) \coloneqq_R f(X) + \mathsf{Lap}\left(\frac{\Delta_{f,1}}{\varepsilon}\right)^{\otimes d}$$

satisfies ε-DP, where "$\coloneqq_R$" is a notation we use to define a randomized mechanism.

We finally state the result from Hopkins et al. (2022b) about estimating the means of sub-Gaussian distributions under pure DP in polynomial time.

Theorem 29 (Sub-Gaussian Learner from (Hopkins et al., 2022b, Theorem 5.1)) *Assume that $0 < \alpha, \beta, \varepsilon < 1$ and $R > 0$. Let $\mu \in \mathbb{R}^d$, where $\|\mu\|_2 \leq R$, be unknown. There exists an ε-DP algorithm (*DPSGLEARNER*) that takes n i.i.d. samples from a sub-Gaussian distribution with mean μ and covariance $0 \preceq \Sigma \preceq \mathbb{I}$, such that,*

$$n \geq \widetilde{O}_{\alpha}\left(\frac{d + \log(1/\beta)}{\alpha^2} + \frac{d + \log(1/\beta)}{\alpha\varepsilon} + \frac{d \log(R)}{\varepsilon}\right),$$

where $\widetilde{O}_{\alpha}(\cdot)$ hides polylogarithmic factors in $\frac{1}{\alpha}$, runs in time $\mathrm{poly}(n, d)$, and with probability at least $1 - \beta$, outputs $\widehat{\mu}$ such that $\|\mu - \widehat{\mu}\|_2 \leq \alpha$.

Appendix C. Privacy Analysis

The privacy analysis of Algorithm 1 is based on the privacy guarantees of the Laplace mechanism and bounded sensitivity of the truncated mean, along with the privacy guarantees of DPSGLEARNER (Theorem 29).

Proposition 30 *For every $\varepsilon, \alpha, \beta > 0$,* $\text{PUREDPPDE}_{\varepsilon,\alpha,\beta}(X)$ *satisfies ε-DP.*

Proof Each individual's data is used only once in exactly one of these three situations – while computing $\mathsf{tmean}_{B_r}(Y^r)$ in some round r of the partitioning rounds, in the call to DPSGLEARNER at the end of the **While**-loop of the partitioning rounds, or in the final round while computing $\mathsf{tmean}_{B_F}(Y^F)$. In other words, since all the datasets – Y^r (for all $i \in [R]$), Y^F, and Z – are disjoint and only used once in the entire algorithm, we do not need to apply composition, but instead, we just have to show that each individual computation is ε-DP.

In each partitioning round r, we perform an ℓ_1 truncation on all rows to within B_r, and add Laplace noise scaled to $\frac{B_r}{\varepsilon}$ to every coordinate of $\tilde{p}_r[S_r]$. By Lemma 28, this satisfies ε-DP. By similar reasoning, the final round also satisfies ε-DP. Finally, the mean of P has an ℓ_2 norm of at most $\sqrt{d}$. Therefore, from the privacy guarantees of DPSGLEARNER (Theorem 29), we have ε-DP for this step, as well. ■

Appendix D. Efficiency Analysis

Here, we prove the computational efficiency of Algorithm 1, which is a key feature of our work.

Proposition 31 *There exists a polynomial $f : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$, such that for every $\varepsilon, \alpha, \beta, n, d > 0$ and $X \in \{0,1\}^{n \times d}$,* $\text{PUREDPPDE}_{\varepsilon,\alpha,\beta}(X)$ *has a running time of $f(n,d)$.*

Proof There are at most $\log_2(d)$ iterations in the partitioning rounds, and in each iteration, we perform $O(d)$ operations. The call to DPSGLEARNER after that costs another fixed polynomial $g(n,d)$ time, where $g : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$ (Theorem 29). The final round has an $O(d)$ running time, as well. Therefore, we have an $f(n,d) = O(d \log(d)) + g(n,d) + O(d)$ running time, which is polynomial in n, d. Note that we are assuming that the cost in the running time due to sampling from Laplace distribution is very low. ■

Appendix E. Missing Proofs from Section 2.2

Proof [Proof of Claim 7] We use Lemma 21 and the facts that for all $\gamma > 0$ and $0 < p \leq 1$,

$$\mathrm{d}_{\mathrm{KL}}(p+\gamma||p) \geq \frac{\gamma^2}{2(p+\gamma)} \quad \text{and} \quad \mathrm{d}_{\mathrm{KL}}(p-\gamma||p) \geq \frac{\gamma^2}{2p},$$

and set

$$\gamma = \sqrt{\frac{4p[j]\log\left(\frac{dR}{\beta}\right)}{m}}.$$

Note that when $p[j] \geq \frac{1}{d}$, due to our choice of parameters, $\gamma \leq \frac{p[j]}{16}$. Therefore, $2(p[j]+\gamma) \leq 4p[j]$. Finally, taking a union bound over the cases when $\tilde{p}_r[j] \leq p[j] - \gamma$ and when $\tilde{p}_r[j] \geq p[j] + \gamma$, we prove the claim. ■

Proof [Proof of Claim 8] By assumption, all marginals specified by S_r are upper bounded by u_r. Now, the expected value of $|X_i^r|$ is at most $u_r|S_r|$. Since $B_r = 3u_r|S_r|\log(mR/\beta)$, we know that $B_r \geq u_r|S_r|\left(1 + 2\log(mR/\beta)\right)$. The claim now follows from Lemma 23 and a union bound over the rows of Y^r. ■

Proof [Proof Claim 9] We assume that all marginals specified by S_r are upper bounded by u_r. From Claim 8, we know that, with probability at least $1-\beta/R$, there is no truncation, so $\mathsf{tmean}_{B_r}(Y^r[S_r]) = \frac{1}{m}\sum_{Y_i^r \in Y^r} Y^r[S_r] = \tilde{p}_r[S_r]$. So, the Laplace noise is added to $\tilde{p}_r[j]$ for each $j \in S_r$. Therefore, the only source of error here is the Laplace noise. Using the standard tail bound for Laplace distributions (Lemma 24) after setting,

$$t = \frac{3u_r|S_r|\log\left(\frac{mR}{\beta}\right)}{\varepsilon m},$$

and taking a union bound over all coordinates in S_r, and the event of truncation, we obtain the claim. ■

Proof [Proof of Claim 13] The proof is identical to that of Claim 7. ■

Proof [Proof of Claim 14] We use Lemma 21 and facts that

$$\forall \gamma > 0 \quad \mathrm{d_{KL}}(p+\gamma||p) \geq \frac{\gamma^2}{2(p+\gamma)} \quad \text{and} \quad \mathrm{d_{KL}}(p-\gamma||p) \geq \frac{\gamma^2}{2p}.$$

Set $\gamma = \frac{\alpha}{8d}$. Then from Lemma 21 and our choice of m_1, we have the following.

- For all $j \in S_F$, with probability at least $1 - \frac{\beta}{d}$, $\tilde{p}[j] \leq p[j] + \gamma$.
- For all $j \in S_F$, with probability at least $1 - \frac{\beta}{d}$, $\tilde{p}[j] \geq \max\{0, p[j] - \gamma\}$.

Applying the union bound, we get the required result. ■

Proof [Proof of Claim 15] Note that all the marginals specified by S_F are upper bounded by u_k (where k is the index as specified in Lemma 12) and that $u_k|S_F| < 1$. With this, we use Lemma 23 and get the required result because we set the truncation radius $B_F = 4\log(m_1/\beta)$. ■

Proof [Proof of Claim 16] Using the standard tail bound for Laplace random variables (Lemma 24) with the following parameters,

$$t = \frac{4\log\left(\frac{m_1}{\beta}\right)}{\varepsilon m_1},$$

and taking the union bound over all the columns of the dataset in that round and the event of truncation, we obtain the claim. ■

Appendix F. Discussion

In this work, we solved a fundamental statistical problem of estimating the means of binary product distributions in total-variation distance under pure DP with optimal sample complexity and under polynomial running time.

However, we would like to mention again that our techniques hold similarities with those in Kamath et al. (2019a). That said, we note that private preconditioning steps are commonly seen in DP statistics now, especially when trickier, direction-wise metrics, such as total-variation distance, are involved. The goal in these cases is to have direction-wise accuracy, so the choice of the error metric is crucial. That said, the way this preconditioning is done could depend on the family of distributions in question according to the way total-variation distance is characterised for those distributions and on other factors, such as concentration properties and domain. For example, in Kamath et al. (2019a) and Kamath et al. (2022c), such steps were performed for estimation of covariances of Gaussians – the idea was to make all the directions of the Gaussian similar to one another privately, and then estimate them accurately, before reverting the transformation. There, the preconditioning looked different from what we did here, but the high-level goal was still the same – adding appropriate amounts of noise in all directions. The problem being solved in our work is also under total-variation distance, which is why we went back to a preconditioning-style algorithm. Is there a different (but a more direct) approach to solving this problem efficiently under pure DP that did not involve any private preconditioning? We do not know the answer yet, but it is an interesting question to think about.

Now, the steps in our algorithm were motivated by the current tools available to solve this problem. We realised that the sub-Gaussian mean estimator from Hopkins et al. (2022b) could not be directly applied to our problem, otherwise the problem would become quite trivial to solve. Since estimating in total-variation distance requires direction-wise accuracy, we had to adapt the preconditioning approach in Kamath et al. (2019a) for the estimator from Hopkins et al. (2022b) to give us anything useful. This also led to requiring different technical lemmata at different stages to prove the accuracy guarantees of our algorithm, thereby creating important and non-trivial differences in our analyses from those in Kamath et al. (2019a).

We also admit that the recent development of Hopkins et al. (2022b) was an important factor in our work. Before that, there was no efficient, pure DP method to estimate the means of sub-Gaussians. However, as we mentioned above, this was not enough by itself because it can only give an estimate that is accurate to within ℓ_2 distance, so our preconditioning approach seemed necessary to us if we were to use the algorithm from Hopkins et al. (2022b) as a black box.