Differentially Private Synthetic Data Using KD-Trees
(Supplementary Material)

Eleonora Kreačić[1] **Navid Nouri**[1] **Vamsi K. Potluru**[1] **Tucker Balch**[1] **Manuela Veloso**[1]

[1] JP Morgan AI Research

A SOME AUXILIARY RESULTS

Remark 1. *For* $X \sim \textsc{Lap}(2/\epsilon)$*, we have*

$$\Pr[X \geq \alpha] = \frac{1}{2} e^{-\frac{\alpha}{2/\epsilon}}.$$

Remark 2. *As a consequence of Remark 1, for* $X \sim \textsc{Lap}(2/\epsilon)$*, we have*

$$\Pr\left[|X| \geq \frac{4C \log n}{\epsilon}\right] = n^{-2C}.$$

Remark 3. *Let* $X \sim \textsc{Lap}(2/\epsilon)$*, then we have*

$$\mathbf{E}[|X|] = \frac{1}{\epsilon}.$$

Remark 4 (Laurent and Massart [2000])**.** *Let* $Y := \sum_{k=1}^{d} Z_k^2$*, where* $Z_k \sim \mathcal{N}(0,1)$ *are i.i.d. random variables. Then*

$$\Pr\left[Y \geq d + 2\sqrt{dx} + 2x\right] \leq e^{-x}, \forall x > 0$$

Corollary 1. *Let* $Y := \sum_{k=1}^{d} Z_k^2$*, where* $Z_k \sim \mathcal{N}(0,\sigma^2)$ *are i.i.d. random variables. Then*

$$\Pr\left[Y/\sigma^2 \geq 2d + 3x\right] \leq \Pr\left[Y/\sigma^2 \geq d + 2\sqrt{dx} + 2x\right] \leq e^{-x}, \forall x > 0$$

Corollary 2. *For any* $X \sim \mathcal{N}(\mathbf{c}, \sigma^2 I)$*, we have*

$$\Pr\left[||X - \mathbf{c}||_2 \geq \sigma\sqrt{2d + 3x}\right] \leq e^{-x}$$

Lemma 3 (Chernoff bounds)**.** *Let* $X_1, X_2, \ldots, X_n$ *be independent binary random variables. Define* $Y := \sum_{i=1}^{n} X_i$ *and* $\mu := \mathbb{E}[Y]$*. Then, for any* $\delta > 0$ *and*

$$\Pr[|Y - \mu| > \delta\mu] \leq 2\exp(-\delta^2\mu/4).$$

B DATA INDEPENDENT APPROACH: $t = 0$ CASE

Lemma 4 (Perturbation bound for Gaussian kernel)**.** *For Gaussian kernel given by* $K(\mathbf{x}, \mathbf{y}) = e^{-\frac{||\mathbf{x}-\mathbf{y}||_2^2}{2}}$*,* $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$*, if* $\mathbf{x} \in \mathbb{R}^d$ *and* $\mathbf{x}' \in \mathbb{R}^d$ *are such that* $||\mathbf{x} - \mathbf{x}'||_2 \leq \alpha$*, then:*

$$\max_{\mathbf{y} \in \mathbb{R}^d} |K(\mathbf{x}, \mathbf{y}) - K(\mathbf{x}', \mathbf{y})| \leq \min(1, \frac{\alpha}{\sqrt{e}})$$

Accepted for the 39th Conference on Uncertainty in Artificial Intelligence (UAI 2023).

Proof. First, by triangle inequality and the assumption that $||x - x'||_2 \leq \alpha$, we have

$$||\mathbf{x} - \mathbf{y}||_2 - ||\mathbf{x}' - \mathbf{y}||_2 \leq ||\mathbf{x} - \mathbf{x}'||_2 \leq \alpha. \tag{1}$$

For $f(x) = e^{-\frac{x^2}{2}}$, we have

$$\max_{x \in \mathbb{R}} |f'(x)| = \frac{1}{\sqrt{e}}, \tag{2}$$

and thus

$$\max_{\mathbf{y} \in \mathbb{R}^d} |K(\mathbf{x}, \mathbf{y}) - K(\mathbf{x}', \mathbf{y})| \leq \frac{\alpha}{\sqrt{e}}.$$

It remains to note that $0 \leq K(\mathbf{x}, \mathbf{y}) \leq 1$ and thus also $\max_{\mathbf{y} \in \mathbb{R}^d} |K(\mathbf{x}, \mathbf{y}) - K(\mathbf{x}', \mathbf{y})| \leq 1$. ☐

Lemma 5 (Error analysis of rounding to centers)**.** *Let $P \subset \mathbb{R}^d$ be the input dataset and let $P' := \{(c_1, v_1), \dots, (c_J, v_J)\}$, where vector of J point counts $\mathbf{v} \in \mathbb{R}^J$, and centers of J bins $c_1, \dots, c_J$ are defined in lines 7 and 8 of Algorithm 1, respectively. Then for the KDE metric between P' and P we have*

$$\sup_{x \in \mathbb{R}^d} |\mathrm{KD}_{P'}(x) - \mathrm{KD}_P(x)| \leq \max\left(\frac{w\sqrt{d}}{2\sqrt{e}}, 1\right). \tag{3}$$

Proof. For any point $x \in \mathbb{R}^d$ that belongs to a bin with a center $c \in \mathbb{R}^d$ we have

$$||x - c||_2 \leq \frac{w\sqrt{d}}{2}.$$

Plugging the above bound in Lemma 4 completes the proof. ☐

Lemma 6. *If total number of bins J in Algorithm 1 is such that $J \leq n^C$, then for each bin $i \in [J]$, the noise term $\mathbf{w}_i = \widetilde{\mathbf{v}}_i - \mathbf{v}_i$ added in step 9 is such that*

$$\mathbf{w}_i \leq \frac{4C \log n}{\epsilon},$$

with probability at least $1 - \frac{1}{n^C}$.

Proof. This is a consequence of Remark 2 and union bound argument. ☐

Lemma 7 (Bound on the total noisy count)**.** *For J-dimensional vector of noisy point counts $\widetilde{\mathbf{v}}$ defined in line 9 of Algorithm 1, with probability at least $1 - \frac{1}{n^C}$ we have*

$$n - \frac{4CJ \log n}{\epsilon} \leq |\widetilde{\mathbf{v}}| \leq n + \frac{4CJ \log n}{\epsilon}.$$

Proof. This is a consequence of Lemma 6 and the fact that initial size of dataset is n. ☐

Lemma 8 (Error analysis of noise addition)**.** *Let $P' := \{(c_1, v_1), \dots, (c_J, v_J)\}$, where vector of J point counts $\mathbf{v} \in \mathbb{R}^J$, and centers of J bins $c_1, \dots, c_J$ are defined in lines 7 and 8 of Algorithm 1, respectively. Let $Q := \{(c_1, \widetilde{\mathbf{v}}_1), (c_2, \widetilde{\mathbf{v}}_2), \dots, (c_J, \widetilde{\mathbf{v}}_J)\}$ be the noisy output of the algorithm. Then with probability at least $1 - \frac{1}{n^C}$ we have*

$$\sup_{x \in \mathbb{R}^d} |\mathrm{KD}_Q(x) - \mathrm{KD}_{P'}(x)| \leq \frac{8CJ \log n}{\epsilon n - 4CJ \log n}. \tag{4}$$

Proof. For any $x \in \mathbb{R}^d$ we have

$$\begin{aligned}
|\mathrm{KD}_Q(x) - \mathrm{KD}_{P'}(x)| &= \left| \frac{1}{|\widetilde{\mathbf{v}}|} \sum_{i=1}^{J} \widetilde{\mathbf{v}}_i K(c_i, x) - \frac{1}{|\mathbf{v}|} \sum_{i=1}^{J} \mathbf{v}_i K(c_i, x) \right| \\
&= \frac{1}{n} \left| \sum_{i=1}^{J} K(c_i, x) \left(\frac{n}{|\widetilde{\mathbf{v}}|} \widetilde{\mathbf{v}}_i - \mathbf{v}_i \right) \right| \\
&\le \frac{1}{n} \sum_{i=1}^{J} \left| \frac{n}{|\widetilde{\mathbf{v}}|} \widetilde{\mathbf{v}}_i - \mathbf{v}_i \right|
\end{aligned}$$

where the second equality is the consequence of the fact that the point count in the original dataset is n, i.e. $|\mathbf{v}| = n$, and the inequality follows from $K(c_i, x) \le 1$. Let $\mathbf{w}_i$ denote the noise added to the ith bin's point count in step 9 of Algorithm 1, so that $\mathbf{w}_i = \widetilde{\mathbf{v}}_i - \mathbf{v}_i$. Then we have

$$|\mathrm{KD}_Q(x) - \mathrm{KD}_{P'}(x)| \le \frac{1}{n} \sum_{i=1}^{J} \left| \mathbf{v}_i \left(\frac{n}{|\widetilde{\mathbf{v}}|} - 1 \right) + \frac{n}{|\widetilde{\mathbf{v}}|} \mathbf{w}_i \right|$$

$$\le \frac{1}{n} \sum_{i=1}^{J} \left| \mathbf{v}_i \left(\frac{n}{|\widetilde{\mathbf{v}}|} - 1 \right) \right| + \frac{1}{n} \sum_{i=1}^{J} \left| \frac{n}{|\widetilde{\mathbf{v}}|} \mathbf{w}_i \right| \tag{5}$$

$$\le \left| \frac{n}{|\widetilde{\mathbf{v}}|} - 1 \right| + \sum_{i=1}^{J} \left| \frac{1}{|\widetilde{\mathbf{v}}|} \mathbf{w}_i \right| \tag{6}$$

where the second inequality is the triangle inequality and the last one follows as $|\mathbf{v}| = n$. Since $\mathbf{w}_i \sim \textsc{Lap}(2/\epsilon)$, by Lemma 7 we have

$$|\mathrm{KD}_Q(x) - \mathrm{KD}_{P'}(x)| \le \left| \frac{n}{n - \frac{4CJ\log n}{\epsilon}} - 1 \right| + \frac{1}{n - \frac{4CJ\log n}{\epsilon}} \sum_{i=1}^{J} |\mathbf{w}_i|$$

$$\le \frac{n}{n - \frac{4CJ\log n}{\epsilon}} - 1 + \frac{1}{n - \frac{4CJ\log n}{\epsilon}} \frac{4CJ\log n}{\epsilon} \tag{7}$$

$$= \frac{8CJ\log n}{\epsilon n - 4CJ\log n}. \tag{8}$$

with probability $1 - \frac{1}{n^C}$. The second inequality is the consequence of Lemma 6. ☐

Proof of Theorem 3: This is a consequence of Lemma 5 and Lemma 8 for C such that $\delta = \frac{1}{n^C}$. Triangle inequality completes the proof. ☐

C DATA INDEPENDANT APPROACH: $t > 0$ CASE

Lemma 9 (Algorithm 1 filters out all $t/2$-light bins)**.** *If for $t = \frac{8C\log n}{\epsilon}$ and the total number of bins J we have $J \le n^C$ for some constant C, then with probability at least $1 - \frac{1}{2}n^{-C}$ all $t/2$-light bins will be filtered out by step 10 of Algorithm 1.*

Proof. For $t = \frac{8C\log n}{\epsilon}$, since we are adding Lap$(2/\epsilon)$ noise the probability of a bin with point count less than $t/2$ having noisy point count more than t is upper bounded by $\frac{1}{2}n^{-2C}$ (see Remark 2).

Union bound over $J \le n^C$ bins completes the proof. ☐

Lemma 10 (Algorithm 1 does not filter any $3t/2$-heavy bins)**.** *If for $t = \frac{8C\log n}{\epsilon}$ and the total number of bins J we have $J \le n^C$ for some constant C, then with probability at least $1 - \frac{1}{2}n^{-C}$ no $3t/2$-heavy bin gets filtered out by step 10 of Algorithm 1. Algorithm 1 does not filter any $3t/2$-heavy bin with probability at least $1 - \frac{1}{2}n^{-C}$.*

Proof. For $t = \frac{8C\log n}{\epsilon}$, since we are adding Lap$(2/\epsilon)$ noise the probability of a bin with point count at least $3t/2$ having noisy point count less than t is upper bounded by $\frac{1}{2}n^{-2C}$ (see Remark 2). Union bound argument over $J < n^C$ bins completes the proof. ☐

Lemma 11 (Noisy point counts). *If for $t = \frac{8C\log n}{\epsilon}$ and the total number of bins J we have $J \leq n^C$ for some constant C, then for J-dimensional vector of noisy point counts $\widetilde{\mathbf{v}}$ defined in line 9 of Algorithm 1, with probability at least $1 - \frac{1}{n^C}$ we have*

$$n - m - \frac{4CM\log n}{\epsilon} \leq |\widetilde{\mathbf{v}}| \leq n + \frac{4CM\log n}{\epsilon},$$

where M and m denote the total number of $t/2$-heavy bins and the total number of points in $3t/2$-light bins, respectively.

Proof. Let $F := \{i : \mathbf{v}_i > 0, \widetilde{\mathbf{v}}_i = 0\}$, $Z := \{i : \mathbf{v}_i = 0, \widetilde{\mathbf{v}}_i = 0\}$ and $H := \{i : \widetilde{\mathbf{v}}_i > 0\}$ denote the set of non empty bins that are filtered out, the set of empty bins that are filtered out and the set of bins that survive filtering, respectively. Note that every bin belongs to one of the three sets i.e. $[J] = F \cup Z \cup H$. We have

$$\begin{aligned}
|\widetilde{\mathbf{v}}| &= \sum_{i=1}^{J} \widetilde{\mathbf{v}}_i \\
&= \sum_{i\in F} \widetilde{\mathbf{v}}_i + \sum_{i\in Z} \widetilde{\mathbf{v}}_i + \sum_{i\in H} \widetilde{\mathbf{v}}_i \\
&= \sum_{i\in H} \widetilde{\mathbf{v}}_i \\
&\leq |\mathbf{v}| + |H| \cdot \frac{4C\log n}{\epsilon} \\
&\leq n + \frac{4CM\log n}{\epsilon},
\end{aligned}$$

where the third equality follows by definition of F and Z, and the first inequality is the consequence of Lemma 6. The last inequality follows from $|\mathbf{v}| = n$ and the consequence of Lemma 9 which gives that with probability at least $1 - \frac{1}{2}n^{-C}$ any bin that survives filtering is $t/2$-heavy i.e. $|H| \leq M$. On the other hand, we also have

$$\begin{aligned}
|\widetilde{\mathbf{v}}| &= \sum_{i=1}^{J} \widetilde{\mathbf{v}}_i \\
&= \sum_{i\in H} \widetilde{\mathbf{v}}_i \\
&\geq \sum_{i\in H} \mathbf{v}_i - |H| \cdot \frac{4C\log n}{\epsilon} \\
&\geq \sum_{i\in H} \mathbf{v}_i - \frac{4CM\log n}{\epsilon} \\
&\geq n - m - \frac{4CM\log n}{\epsilon}
\end{aligned}$$

where again second equality comes from the definition of F and Z, and the first inequality is the consequence of Lemma 6 and the second inequality follows from $|H| \leq M$ as above. Finally, the last inequality is the consequence of Lemma 10 which gives us that with probability at least $1 - \frac{1}{2}n^{-C}$ any bin that gets filtered out is $3t/2$-light and so the total number of filtered out points is upper bounded by m. This means that the total number of points in bins that survive filtering is at least $n - m$ i.e. $\sum_{i\in H} \mathbf{v}_i \geq |\mathbf{v}| - m \geq n - m$. Union bound argument completes the proof. ☐

Now, we analyze the error between kernel density induced by Q and P'. For any $q \in \mathbb{R}^d$ we have

Lemma 12. *Let $P' := \{(c_1, v_1), \ldots, (c_J, v_J)\}$, where v, c are defined as in lines 7 and 8 of Algorithm 1. For $t = \frac{8C\log n}{\epsilon}$ and the total number of bins $J \leq n^C$, with probability $1 - n^{-C}$ we have*

$$\sup_{x\in\mathbb{R}^d} |\mathrm{KD}_Q(x) - \mathrm{KD}_{P'}(x)| \leq \frac{\epsilon m + 8CM\log n}{\epsilon n - \epsilon m - 4CJ\log n} + \frac{m}{n}. \tag{9}$$

Proof. Let $H := \{i : \widetilde{\mathbf{v}}_i > 0\}$ denote the bins that survive filtering step. We have

$$\begin{aligned}
|\mathrm{KD}_Q(x) - \mathrm{KD}_{P'}(x)| &= \left| \frac{1}{|\widetilde{\mathbf{v}}|} \sum_{i=1}^{J} \widetilde{\mathbf{v}}_i K(c_i, x) - \frac{1}{|\mathbf{v}|} \sum_{i=1}^{J} \mathbf{v}_i K(c_i, x) \right| \\
&= \frac{1}{n} \left| \sum_{i=1}^{J} K(c_i, x) \left(\frac{n}{|\widetilde{\mathbf{v}}|} \widetilde{\mathbf{v}}_i - \mathbf{v}_i \right) \right| \\
&\leq \frac{1}{n} \sum_{i=1}^{J} \left| \frac{n}{|\widetilde{\mathbf{v}}|} \widetilde{\mathbf{v}}_i - \mathbf{v}_i \right| \\
&\leq \frac{1}{n} \left(\sum_{i \in H} \left| \frac{n}{|\widetilde{\mathbf{v}}|} \widetilde{\mathbf{v}}_i - \mathbf{v}_i \right| + \sum_{i \notin H} |\mathbf{v}_i| \right) \\
&\leq \frac{1}{n} \left(\sum_{i \in H} \left| \frac{n}{|\widetilde{\mathbf{v}}|} \widetilde{\mathbf{v}}_i - \mathbf{v}_i \right| + m \right)
\end{aligned}$$

where the second equality holds since $|\mathbf{v}| = n$, and the first inequality is the consequence of $K(c_i, x) \leq 1$. The second inequality follows by the definition of H, and the final one is the consequence of Lemma 10, as with probability at least $1 - \frac{1}{2} n^{-C}$ any bin that is filtered out must be $3t/2$-light. Let $\mathbf{w}_i = \widetilde{\mathbf{v}}_i - \mathbf{v}_i$, then $\mathbf{w}_i \sim \textsc{Lap}(2/\epsilon)$. Then we further have

$$\begin{aligned}
|\mathrm{KD}_Q(x) - \mathrm{KD}_{P'}(x)| &\leq \frac{1}{n} \left(\sum_{i \in H} \left| \frac{n}{|\widetilde{\mathbf{v}}|} \widetilde{\mathbf{v}}_i - \mathbf{v}_i \right| + m \right) \\
&\leq \frac{1}{n} \left(\sum_{i \in H} \left| \mathbf{v}_i \left(\frac{n}{|\widetilde{\mathbf{v}}|} - 1 \right) + \frac{n}{|\widetilde{\mathbf{v}}|} \mathbf{w}_i \right| + m \right) \\
&\leq \frac{1}{n} \left(\sum_{i \in H} \left| \mathbf{v}_i \left(\frac{n}{|\widetilde{\mathbf{v}}|} - 1 \right) \right| + \frac{1}{n} \sum_{i \in H} \left| \frac{n}{|\widetilde{\mathbf{v}}|} \mathbf{w}_i \right| + m \right) \\
&\leq \left| \frac{n}{|\widetilde{\mathbf{v}}|} - 1 \right| + \sum_{i \in H} \left| \frac{1}{|\widetilde{\mathbf{v}}|} \mathbf{w}_i \right| + \frac{m}{n}
\end{aligned} \tag{10}$$

where the third is the triangle inequality and the last one follows since $|\mathbf{v}| = n$. Let $n' := n - m - \frac{4CM \log n}{\epsilon}$. By Lemma 11 with probability $1 - n^{-C}$ we have

$$\begin{aligned}
|\mathrm{KD}_Q(x) - \mathrm{KD}_{P'}(x)| &\leq \left| \frac{n}{n'} - 1 \right| + \frac{1}{n'} \sum_{i=1}^{J} |\mathbf{w}_i| + \frac{m}{n} \\
&\leq \frac{n}{n'} - 1 + \frac{4CM \log n}{n' \epsilon} + \frac{m}{n} \\
&= \frac{\epsilon m + 8CM \log n}{\epsilon n - \epsilon m - 4CJ \log n} + \frac{m}{n}.
\end{aligned} \tag{11}$$

where the second inequality holds by Lemma 6 □

Proof of Theorem 5:

As a consequence of triangle inequality and Lemmas 5 and 12, for C such that $\delta = n^{-C}$ we have

$$\begin{aligned}
\sup_{x \in \mathbb{R}^d} |\mathrm{KD}_P(q) - \mathrm{KD}_Q(q)| &\leq \frac{\epsilon m + 8CM \log n}{\epsilon n - \epsilon m - 4CM \log n} + \frac{m}{n} + \frac{w\sqrt{d}}{2\sqrt{e}} \\
&= \frac{\epsilon m + 8M \log \frac{1}{\delta}}{\epsilon n - \epsilon m - 4M \log \frac{1}{\delta}} + \frac{m}{n} + \frac{w\sqrt{d}}{2\sqrt{e}}.
\end{aligned}$$

This completes the proof. □

D SPECIAL CASE: ORIGINAL DATASET FROM MIXTURE OF GAUSSIANS

Lemma 13. *If* $r > 0$, $C > 0$ *are such that* $\frac{nw^d}{(2\pi\sigma^2)^{d/2}} e^{-\frac{(r/2)^2}{2\sigma^2}} > \frac{2C\log n}{\epsilon}$, *then* $r \leq 3\sigma\sqrt{\log n + d\log\left(\frac{w}{\sigma\sqrt{2\pi}}\right)}$.

Proof. The condition of the lemma translates to

$$
\begin{aligned}
r &\leq 2\sqrt{2}\sigma\sqrt{\log\left(\frac{n}{2C\log n}\cdot\left(\frac{w}{\sigma\sqrt{2\pi}}\right)^d\right)} \\
&= 2\sqrt{2}\sigma\sqrt{\log n - \log\log n - \log 2C + d\log\left(\frac{w}{\sigma\sqrt{2\pi}}\right)} \\
&\leq 3\sigma\sqrt{\log n + d\log\left(\frac{w}{\sigma\sqrt{2\pi}}\right)}
\end{aligned}
$$

□

Lemma 14. *If* $r > 0$, $C > 0$ *are such that* $\frac{nw^d}{(2\pi\sigma^2)^{d/2}} e^{-\frac{r^2}{\sigma^2}} < \frac{16C\log n}{\epsilon}$, *then we have* $r \geq \sigma\sqrt{\log(\epsilon n) - \log(16C\log n) + d\log\left(\frac{w}{\sigma\sqrt{2\pi}}\right)}$

Proof. Condition on r translates to

$$
\begin{aligned}
r &\geq \sqrt{-\sigma^2\log\left(\frac{16C\log n}{\epsilon n}\cdot\left(\frac{\sigma\sqrt{2\pi}}{w}\right)^d\right)} \\
&\geq \sigma\sqrt{\log\left(\frac{\epsilon n}{16C\log n}\cdot\left(\frac{w}{\sigma\sqrt{2\pi}}\right)^d\right)} \\
&= \sigma\sqrt{\log(\epsilon n) - \log(16C\log n) + d\log\left(\frac{w}{\sigma\sqrt{2\pi}}\right)}.
\end{aligned}
$$

□

Definition 15. *For a bin (hypercube)* B *and a point* $y \in \mathbb{R}^d$, *we define their distance as the* ℓ_2 *distance of the center of* B *to* y.

Lemma 16 (Upper bound on M)**.** *For a dataset coming from a multivariate Gaussian with variance* $\sigma^2 I$ *in* $\mathbb{R}^d$ *there are at most*

$$
\left(\frac{6\sigma}{w}\sqrt{\log n + d\log\left(\frac{w}{\sigma\sqrt{2\pi}}\right)} + 2\right)^d
$$

$t/2$*-heavy bins with arbitrary high probability.*

Proof. Without loss of generality we assume that the mean of the distribution is the origin. Let B be a bin at distance r from the origin, and $x \in \mathbb{R}^d$ be a point inside B. For X from multivariate Gaussian we have

$$
f(X = x) \leq \frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{(r - \frac{w\sqrt{d}}{2})^2}{2\sigma^2}}.
$$

Hence, the expected number of points within B is upper bounded by $\frac{nw^d}{(2\pi\sigma^2)^{d/2}} e^{-\frac{(r-\frac{w\sqrt{d}}{2})^2}{2\sigma^2}}$. For r such that $w\sqrt{d} \leq r$, this is further upper bounded by $\frac{nw^d}{(2\pi\sigma^2)^{d/2}} e^{-\frac{(r/2)^2}{2\sigma^2}}$.

For simplicity of notation, let us introduce $\mu(r) = \frac{nw^d}{(2\pi\sigma^2)^{d/2}} e^{-\frac{(r/2)^2}{2\sigma^2}}$. If we assume that r is large enough so that $\mu(r) \leq 2C \log n$, Chernoff bounds (see Lemma 3) give us

$$\begin{aligned} \Pr\left[|B| \geq 4C \log n\right] &\leq 2e^{-\mu} \\ &\leq \frac{1}{n^{2C}}, \end{aligned}$$

where $|B|$ denotes the number of the points from the dataset within B. Thus if r is such that $\mu(r) > 2C \log n$, then bins at distance at least r from the origin are $t/2$-light with probability at least $1 - \frac{1}{n^{2C}}$. Equivalently this means that all $t/2$-heavy bins are at distance at most r from the origin.

By Lemma 13 the furthest bin that can be $t/2$-heavy is at distance at most

$$3\sigma \sqrt{\log n + d \log\left(\frac{w}{\sigma\sqrt{2\pi}}\right)}$$

from the origin. It remains to note that the area of diameter α is covered by at most $\alpha/w + 2$ bins of width on a single axis. Thus, the number of $t/2$-heavy bins is bounded by

$$\left(\frac{6\sigma}{w} \sqrt{\log n + d \log\left(\frac{w}{\sigma\sqrt{2\pi}}\right)} + 2\right)^d$$

with probability at least $1 - \frac{1}{n^{2C}}$. ☐

Lemma 17 (Upper bound on m)**.** *For a dataset coming from a multivariate Gaussian with variance $\sigma^2 I$ in $\mathbb{R}^d$, when binning is done with widths such that $8 \leq \log(\frac{w}{\sigma\sqrt{2\pi}})$, there are at most $n^{2/3+o(1)}$ points in $3t/2$-light bins with arbitrary high probability.*

Proof. Without loss of generality we assume that the mean of the distribution is the origin. Let B be a bin at distance r from the origin, and $x \in \mathbb{R}^d$ be a point inside B. For X from multivariate Gaussian we have

$$\frac{1}{(2\pi\sigma^2)^{d/2}} e^{-\frac{(r+\frac{w\sqrt{d}}{2})^2}{2\sigma^2}} \leq f(X = x).$$

Hence, the expected number of points within B is lower bounded by $\frac{nw^d}{(2\pi\sigma^2)^{d/2}} e^{-\frac{(r+\frac{w\sqrt{d}}{2})^2}{2\sigma^2}}$. For r such that $\frac{w\sqrt{d}}{2} \leq (\sqrt{2}-1)r$ we further have the lower bound of $\mathbb{E}[|B|] \geq \frac{nw^d}{(2\pi\sigma^2)^{d/2}} e^{-\frac{2r^2}{2\sigma^2}}$, where B denotes the number of points from the dataset within B.

For simplicity of notation let $\mu(r) = \frac{nw^d}{(2\pi\sigma^2)^{d/2}} e^{-\frac{r^2}{\sigma^2}}$. For r such that $\mu(r) \geq \frac{16C \log n}{\epsilon}$, Chernoff bounds (see Lemma 3) we have

$$\Pr\left[|B| \leq \frac{12C \log n}{\epsilon}\right] \leq \frac{1}{n^{2C}}.$$

Thus if r is such that $\mu(r) \geq \frac{16C \log n}{\epsilon}$ for some r, then bins at distance at most r are $3t/2$-heavy with probability at least $1 - \frac{1}{n^{2C}}$. In other words, all $3t/2$-light bins are at distance at least r. By Lemma 14, the smallest r such that all $3t/2$-light bins are at distance at least r is $\sigma\sqrt{\log(\epsilon n) - \log(16C \log n) + d \log\left(\frac{w}{\sigma\sqrt{2\pi}}\right)}$.

From Corollary 2 it follows that the probability of a point taking distance at least $\sigma\sqrt{\log(\epsilon n) - \log(16C\log n) + d\log\left(\frac{w}{\sigma\sqrt{2\pi}}\right)}$ from the cluster center is

$$\Pr\left[X \geq \sigma\sqrt{2d + 3\left(\frac{\log(\epsilon n) - \log(16C\log n) + d\log(\frac{w}{\sigma\sqrt{2\pi}}) - 2d}{3}\right)}\right] \leq \exp\left(-\frac{(\log(\epsilon n) - \log(16C\log n) + d(\log(\frac{w}{\sigma\sqrt{2\pi}}) - 2)}{3}\right.$$

$$\leq \frac{(16C\log n)^{1/3}\cdot e^{-\frac{d}{3}(\log\frac{w}{\sigma\sqrt{2\pi}}-2)}}{(\epsilon n)^{1/3}}$$

Let $\mathcal{C}'$ be the set of points that are at least $\sigma\sqrt{2d + 3\left(\frac{\log(\epsilon n)-\log(16C\log n)+d\log(\frac{w}{\sigma\sqrt{2\pi}})-2d}{3}\right)}$ far from the mean of the ditribution i.e. origin. Any point in $3t/2$-light bins belongs to $\mathcal{C}'$ with probability at least $1 - \frac{1}{n^{2C}}$. Chernoff bound (see Lemma 3) gives us

$$|\mathcal{C}'| \leq \epsilon^{-1/3}n^{2/3}(16C\log n)^{1/3}\cdot e^{-\frac{d}{3}(\log\frac{w}{\sigma\sqrt{2\pi}}-2)}$$

with high probability. This completes the proof. ☐

Proof of Theorem 6: From above analysis we have that the number of $t/2$-heavy bins is bounded by

$$M = \left(\frac{6\sigma}{w}\sqrt{\log n + d\log\left(\frac{w}{\sigma\sqrt{2\pi}}\right)} + 2\right)^d$$

with high probability. We also have that the total number of points in $3t/2$-light bins is upper bounded by $m = \epsilon^{-1/3}n^{2/3}(16C\log n)^{1/3}\cdot e^{-\frac{d}{3}(\log\frac{w}{\sigma\sqrt{2\pi}}-2)}$. Thus we have

$$\begin{aligned} M &= \left(\frac{6\sigma}{w}\sqrt{\log n + d\log\left(\frac{w}{\sigma\sqrt{2\pi}}\right)} + 2\right)^d && (12)\\ &\leq \left(\frac{6\sigma}{w}\sqrt{2}\sqrt{\log n} + 2\right)^d && (13)\\ &\leq \left(\frac{12\sigma}{w}\right)^d(\log n)^{d/2} && (14) \end{aligned}$$

where the first inequality follows by $n \geq \left(\frac{w}{\sigma\sqrt{2\pi}}\right)^d$ and the second also follows for large n. We have

$$\begin{aligned} \epsilon m + 8M\log\frac{1}{\delta} &\leq \epsilon m + 8\log\left(\frac{1}{\delta}\right)\cdot\left(\frac{12\sigma}{w}\right)^d(\log n)^{d/2} && (15)\\ &\leq \frac{1}{2}\epsilon n && (16) \end{aligned}$$

where the second inequality follows from condition $\frac{n}{(\log n)^{d/2}} > 16\log\left(\frac{1}{\delta}\right)\cdot\left(\frac{12\sigma}{w}\right)^d$. Thus remains to apply Theorem 5 and we get

$$\frac{\epsilon m + 8M\log\frac{1}{\delta}}{\epsilon n - \epsilon m - 4M\log\frac{1}{\delta}} + \frac{m}{n} + \frac{w\sqrt{d}}{2\sqrt{e}} \leq \frac{\epsilon(\epsilon^{-1/3}n^{2/3}(16C\log n)^{1/3}\cdot e^{-\frac{d}{3}(\log\frac{w}{\sigma\sqrt{2\pi}}-2)}) + 8\log\left(\frac{1}{\delta}\right)\cdot\left(\frac{12\sigma}{w}\right)^{d}(\log n)^{d/2}}{\frac{1}{2}\epsilon n} \tag{17}$$

$$+\frac{n^{2/3}(16C\log n)^{1/3}\cdot e^{-\frac{d}{3}(\log\frac{w}{\sigma\sqrt{2\pi}}-2)}}{n} + \frac{w\sqrt{d}}{2\sqrt{e}} \tag{18}$$

$$\leq \frac{3(16C\log n)^{1/3}\cdot e^{-\frac{d}{3}(\log\frac{w}{\sigma\sqrt{2\pi}}-2)}}{(\epsilon n)^{1/3}} + \frac{16\log\left(\frac{1}{\delta}\right)\cdot\left(\frac{12\sigma}{w}\right)^{d}(\log n)^{d/2}}{\epsilon n} + \frac{w\sqrt{d}}{2\sqrt{e}} \tag{19}$$

□

E DATA DEPENDENT ALGORITHM

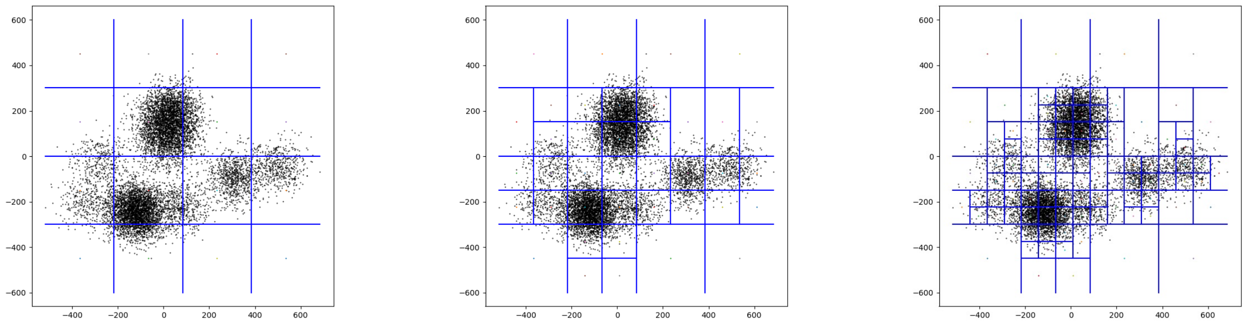

Figure 1: Example of stages of data dependent partitioning of the dataset in $\mathbb{R}^2$.

E.1 IMPLICIT SAMPLING FOR DATA INDEPENDENT ALGORITHM 1

For the data independent algorithm, implicit implementation of empty bins sampling is straightforward. Instead of storing $(\frac{R}{w})^d$ bins centers, it is enough to store only those corresponding to non empty bins, and sample empty centers via independent uniform random sampling of each coordinate from the set of possible values (and rejection if gluing them together gives center corresponding to a non empty bin). Note that this requires storing $O(\frac{R}{w}\cdot d)$ values for non empty bins instead of $O(\frac{R}{w})^d$.

F IMPLICIT SAMPLING OF EMPTY BINS

Proof of Lemma 9: Equivalence of Algorithm 4 and Algorithm 5 is the consequence of independence of Bernoulli indicators (as Laplace noise are independent for different bins) and the fact that Binomial can be represented as the sum of independent Bernoullis. □

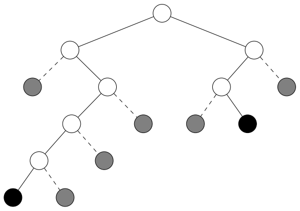

Figure 2: Tree with $h = 2$ and $h' = 5$. Black nodes are non-empty bins, gray nodes are empty bins we need to sample.

Lemma 18. *For a dataset of size n, if the decision tree is data independent up to depth level h and data dependent in the remaining part, then the number of empty bins is upper bounded by*

$$2^h + n(h' - h)$$

where h' denotes the total number of levels.

Proof. We need to upper bound the number of leaves for such binary tree. Since the binary tree is complete up to depth h, we have at most 2^h leaves at level h. Furthermore, for any non empty bin we can have at most $h' - h$ empty bins between depth $h + 1$ and h'. Thus, we have at most $n(h' - h)$ empty bins between depth $h + 1$ and h'. ☐

Proof of Lemma 10: For a single empty bin, by Remark 1 we have

$$\Pr[\text{Lap}(\frac{2(h'-h)}{\epsilon}) \geq \tau] = \frac{1}{2}e^{-\frac{\tau}{2(h'-h)/\epsilon}}.$$

Thus for

$$\tau = \frac{2(h'-h)}{\epsilon}\log\left(\frac{1}{\delta}\cdot\left(2^h + n(h'-h)\right)\right)$$

the right hand side is less than $\frac{\delta}{2^h+n(h'-h)}$. As a consequence of Lemma 18 there are at most $2^h + n(h' - h$ empty bins and thus union bound argument completes the proof. ☐

More implementation details: Our algorithm for implicitly sampling the empty bins proceeds as follows: in the recursion tree, from left to right, our algorithm first finds common ancestor for any two consecutive non-empty bins, say i'th and $i + 1$'th and calls it *i'th common ancestor*. Then, it calculates the number of empty bins between i'th non-empty bin and i'th common ancestor. And similarly, it calculates the number of empty bins between $i + 1$'th non-empty bin and i'th common ancestor. Note that by the above-mentioned binomial distribution implicit sampling argument, one can apply a binomial distribution sampling technique to each of these numbers, and it is not hard to locate the sampled empty bins. Finally, we need to apply an implicit sampling technique to empty bins at level h. This is done similarly to the rejection sampling technique we mentioned in implicit implementation of data independent algorithm.

G EXPERIMENTS

G.1 EXPERIMENTAL SETTING FOR COMPARISON WITH Balog et al. [2018]

For both dimension 2 and 5, the dataset consists of $n = 100,000$ samples from a multivariate mixture of Gaussians. The mixture has 10 components, with mixing weights proportional to $(1, 1/2, \ldots, 1/10)$, and their means are chosen from spherical Gaussian with mean $[100, \ldots, 100]$ and covariance $200I$. Each point is simulated by first sampling the mixture component, and then sampling from a spherical Gaussian centered at the mean of the chosen mixture component and with covariance $30I$. For accuracy of the comparison, we did not re-sample the dataset and used the exact version in the code of Balog et al. [2018].

G.2 DEPENDENCY ON DATASET SIZE

Intuitively, it is easier to hide the contribution of an individual in a large set, compared to a small set, in kernel density. We show this empirically using three datasets with different dataset sizes, but the same underlying mixture of Gaussians parameters as the 5 dimensional datasets in the experiments section, see Figure 3. This experiment also shows that our algorithm is able to produce synthetic datasets with better minimum error when the size of the dataset, N, is larger. This dependency in N was also evident in our theoretical results in Theorem 5 and Theorem 6.

G.3 DEPENDENCY ON VARIANCE

Next, we show the effect of σ (see Theorem 6 for the definition of σ) in the mixture of Gaussians datasets. We consider two datasets with underlying mixture of Gaussians distributions in dimension 10, with $\sigma = 30$ and $\sigma = 3$ for each cluster. As expected, the simulation results presented in Figure 4 confirm that our algorithm performs better in the setting where clusters are more concentrated around a center, i.e., small σ case.

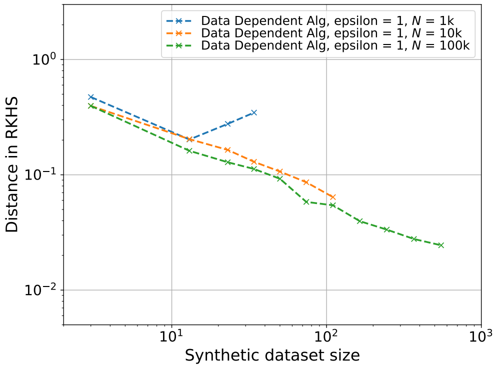

Figure 3: Using three 5 dimensional datasets with 1k, 10k and 100k data points, yet with the same underlying distribution parameters, we show that larger dataset size, N, naturally results in a better performance. Moreover, for larger N, our algorithm is capable of achieving better minimum error.

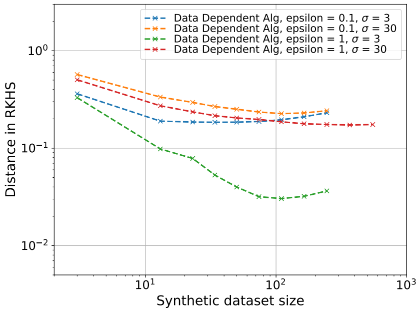

Figure 4: Performance comparison of our data dependent algorithm on 10-dimensional datasets with underlying mixture of Gaussians distribution with $\sigma = 3$ and $\sigma = 30$. Note that our algorithm performs better with smaller σ as predicted by our theory.

G.4 BINARY CLASSIFICATION

We use a dataset from a Kaggle competition `https://www.kaggle.com/datasets/mlg-ulb/creditcardfraud` with information of credit card transactions which were either fraudelent or not, which was also used in Harder et al. [2021]. This dataset has 31 categories, 30 numerical features and a binary label. Similarly to Harder et al. [2021], we use all but the first feature (Time).

For both our data dependent algorithm and DP-MERF [Harder et al., 2021], we use 80% of input data for synthetic data generation, for various privacy budgets ϵ. Synthetic data is then used to train 12 classifiers (see Table 1), which are tested on remaining 80% of input data.

For training DP-MERF synthesizers, we set parameters as in Harder et al. [2021], i.e. number of epochs 4000, number of Furier features 5000, mini-batch side 0.5, undersampling rate 0.005. For our data dependent algorithm, we use undersampling rate of 0.005, and set the number of data independent levels to be equal to 30 and maximal number of levels to 60.

As comparison metrics, we use ROC (area under the receiver operating curve). Table 1 shows average ROC for our data dependent algorithm over 20 repetitions for each classifier, as well as average ROC over the classifiers. Table 2 shows average ROC over classifiers for DP-MERF with 5 repetitions for each classifier.

Although our data dependent algorithm does not outperform DP-MERF, its performance degrades slower for increasing privacy.

Table 1: Data dependent algorithm: ROC for various levels of privacy. Average over 20 repetitions.

	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 0.1$	$\epsilon = 0.01$
Logistic Regression	0.705	0.545	0.481	0.527
Gaussian Naive Bayes	0.562	0.563	0.479	0.547
Bernoulli Naive Bayes	0.495	0.564	0.497	0.521
Linear SVM	0.758	0.524	0.508	0.546
Decision Tree	0.676	0.611	0.519	0.532
LDA	0.518	0.580	0.480	0.542
Ada Boost	0.632	0.572	0.485	0.521
Bagging	0.673	0.579	0.518	0.508
Random Forest	0.663	0.594	0.530	0.543
GBM	0.631	0.582	0.521	0.523
Multi-layer percepton	0.625	0.553	0.486	0.525
XGBoost	0.588	0.598	0.527	0.478
Average	0.627	0.572	0.503	0.526

Table 2: DP-MERF: ROC for various levels of privacy. Average over 5 repetitions.

	$\epsilon = 10$	$\epsilon = 1$	$\epsilon = 0.1$
Average	0.880	0.792	0.564

H APPROXIMATING GAUSSIAN DISTRIBUTION BY MIXTURE OF UNIFORMS

H.1 MMD

Let us assume that the data is arising from a Gaussian distribution and we are estimating with the samples $\{z_i\}_1^n$. The maximum-mean discrepance (MMD) between the population P and the samples is given by:

$$\mathrm{MMD}_u^2(\mathcal{N}_d, Q_n) = E_{x,x'\sim\mathcal{N}_d}[k(x,x')] - \frac{2}{n}\sum_{i=1}^{n} E_{x\sim\mathcal{N}_d}[k(x,z_i)] + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} k(z_i,z_j). \tag{20}$$

The expectations in the expression above can be computed analytically to yield the formula Rustamov [2021]:

$$\mathrm{MMD}_u^2(\mathcal{N}_d, Q_n) = \left(\frac{\gamma^2}{2+\gamma^2}\right)^{d/2} - \frac{2}{n}\left(\frac{\gamma^2}{1+\gamma^2}\right)^{d/2}\sum_{i=1}^{n} e^{-\frac{\|z_i\|^2}{2(1+\gamma^2)}} + \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i}^{n} e^{-\frac{\|z_i-z_j\|^2}{2\gamma^2}}.$$

H.2 WIDTHS OF BINS

Let us assume that we have $2k+1$ boxes to approximate the Gaussian distribution where we assume an odd number of boxes to apply symmetry arguments. The distributions are given by:

$$P(x) = \frac{1}{\sqrt{2\pi}}e^{-x^2/2} Q(x) = w_0 I_0 + \sum_{i=1}^{k} w_i I_i + \sum_{i=1}^{k} w_{-i} I_{-i} \tag{21}$$

where $I_i = I[(2i-1)C <= x < (2i+1)C]$ and $\sum_{i=-k}^{k} w_i = 1$. The KL divergence between distributions is given by:

$$D_{KL}(P||Q) := -\int_{\infty}^{\infty} p(x)\log\frac{p(x)}{q(x)}dx \tag{22}$$

and in particular for our setting is given by:

$$D_{KL}(Q||P) = \sum_{i=-k}^{k} w_i \log\frac{w_i}{2C} + \frac{1}{2}\log(2\pi) + \sum_{i=-k}^{k}\frac{w_i C^2}{6}(12i^2+1) \tag{23}$$

Using the KL bounds for measuring divergence, we are able to obtain the following weights and size of the boxes. Each box is of size given by $2c$ and placed at location $2ci$ for $i \in [-k,k]$. Applying the method of Lagrange multipliers, we can obtain the optimal box weights for a mixture of uniform distributions with respect to the Gaussian distribution:

$$w_0 = \frac{1}{1+2\sum_{i=1}^{k} e^{-2i^2c^2}} \tag{24}$$

$$w_i = w_0 e^{-2i^2c^2} \qquad \forall i \in [-k,k] \text{ and } i \neq 0 \tag{25}$$

References

Matej Balog, Ilya O. Tolstikhin, and Bernhard Schölkopf. Differentially private database release via kernel mean embeddings. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 423–431. PMLR, 2018. URL `http://proceedings.mlr.press/v80/balog18a.html`.

Frederik Harder, Kamil Adamczewski, and Mijung Park. DP-MERF: differentially private mean embeddings with randomfeatures for practical privacy-preserving data generation. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 1819–1827. PMLR, 2021. URL `http://proceedings.mlr.press/v130/harder21a.html`.

Béatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, 28:1302–1338, 2000.

Raif M. Rustamov. Closed-form expressions for maximum mean discrepancy with applications to wasserstein auto-encoders. *Stat*, 10(1):e329, 2021. doi: https://doi.org/10.1002/sta4.329. URL `https://onlinelibrary.wiley.com/doi/abs/10.1002/sta4.329`. e329 sta4.329.