A Missing Definitions

We provide the definition for the distribution version of maximum mean discrepancy.

Definition 1. *Maximum mean discrepancy (MMD) between two distributions P and Q is defined as:*

$$\begin{aligned}\mathrm{MMD}^2(P,Q) = \mathbb{E}_{X,X'\sim P}[k(X,X')] + \mathbb{E}_{Y,Y'\sim Q}[k(Y,Y')] \\ - 2\mathbf{E}_{X\sim P,Y\sim Q}[k(X,Y)]\end{aligned} \tag{4}$$

where $k(\cdot,\cdot)$ is a kernel function underlying an RKHS (Reproducing Kernel Hilbert Space) function space such that $k(x,y) = k(y,x)$ and $k(\cdot,\cdot)$ is positive definite.

B Quantization

Quantization Function Define a grid S of points $S = \{-1, -1+\eta, -1+2\eta \ldots \ldots 1-\eta, 1\}$, where we assume $2/\eta$ is an integer for convenience. Define a random quantization function $Q : [-1,1] \to S$ as follows:

$$Q(x) = \left\{ \begin{array}{cc} -1+k\eta, & w.p. \frac{(k+1)\eta-1-x}{\eta} \\ -1+(k+1)\eta & w.p. \frac{x+1-k\epsilon}{\eta} \end{array} \right\} \tag{5}$$

where $k = \lfloor (x+1)/\eta \rfloor$. Here, the value x is quantized to one of the two nearest points from S with probabilities chosen carefully to make sure that the expected quantization error is 0. Now, we consider the quantized data set $D_Q = \left\{ Q\left(\sqrt{\frac{d}{2}}\mathbf{v}_i\right) \right\}_{i=1}^{q}$. Observe that $D_Q \in S^{q\times d}$. Let $\tilde{v}_1, \tilde{v}_2 \ldots \tilde{v}_q \in S^{1\times d}$ be the quantized vectors in D_Q. Let $w(D_Q, i) = \sum_{j=1}^{q} \tilde{v}_{ji}$.

C Approximation, Efficiency and Privacy Guarantees for the Protocol

Guarantees for $\mathfrak{h}_2(\cdot)$: Now, we prove approximation and privacy guarantees for the hash function $\mathfrak{h}_2(\cdot)$ with respect to the input dataset it operates on. We observe that computing $\mathfrak{h}_2(\cdot)$ involves maintaining a distribution over $|\frac{2}{\eta}|^d$ variables which is exponentially large. We first prove that we need only linear $O\left(\frac{d}{\eta}\right)$ memory and update time to maintain the different distributions.

Lemma 1. *In Algorithm 2 for all $0 \le t \le T$, $P_t(\mathbf{s})$ needs $O\left(\frac{d}{\eta}\right)$ memory and update time.*

Proof. It is enough to prove that distribution $P_t(\mathbf{s})$ satisfies the following two properties:

a) (Product Distribution): $P_t(\mathbf{s}) = \prod_{i=1}^{d} P_t(s_i)$, $\forall t$ here $P_t(s_i)$ is the marginal distribution on the coordinate i.

b) (Marginal Update): $P_t(s_j) = P_{t-1}(s_j)$, $j \neq i(t)$. $P_t(s_j) = P_{t-1}(s_j)\exp\left[s_j\left(\mu_j - w(P_{t-1}, j)/2q\right)\right], j = i(t)$.

We first prove (a) by induction. The base case is true since the initial distribution is uniform. Now suppose it is true for some $t-1$, with $t > 1$.

$$\begin{aligned} P_t(\mathbf{s}) &= \frac{P_{t-1}(\mathbf{s})\exp(s_{i(t)}\frac{\mu_{i(t)} - w(P_{t-1}, i(t))}{2q})}{\sum_{\mathbf{s}} P_{t-1}(\mathbf{s})\exp(s_{i(t)}\frac{\mu_{i(t)} - w(P_{t-1}, i(t))}{2q})} \\ &= \left[\Pi_{i\neq i(t)} P_{t-1}(s_i)\right] * \\ &\quad \left[\frac{P_{t-1}(s_{i(t)})\exp(s_{i(t)}\frac{\mu_{i(t)} - w(P_{t-1}, i(t))}{2q})}{\sum_{\mathbf{s}} P_{t-1}(\mathbf{s})\exp(s_{i(t)}\frac{\mu_{i(t)} - w(P_{t-1}, i(t))}{2q})}\right] \end{aligned} \tag{6}$$

Now,

$$\begin{aligned}
&\sum_{\mathbf{s}} P_{t-1}(\mathbf{s}) \exp(s_{i(t)}(\mu_{i(t)} - w(P_{t-1}, i(t)))/2q) \\
&= \sum_{s} \sum_{\mathbf{s}(i(t))=s} P_{t-1}(\mathbf{s}|\mathbf{s}(i(t)) = s) \exp(s\frac{\mu_{i(t)} - w(P_{t-1}, i(t))}{2q}) \\
&= \sum_{s} \exp(s\frac{\mu_{i(t)} - w(P_{t-1}, i(t))}{2q}) \sum_{\mathbf{s}_{i(t)}=s} P_{t-1}(\mathbf{s}|s_{i(t)} = s) \\
&= \sum_{s} \exp(s(\mu_{i(t)} - w(P_{t-1}, i(t)))/2q) P_{t-1}(s_{i(t)} = s) \qquad (7)
\end{aligned}$$

It follows that the summation expression only depends on the coordinate $i(t)$ and hence we have decomposed $P_t(\mathbf{s})$ into distributions that are dependent only on the coordinates. Now (b) follows by computing the marginal distributions on each coordinate. □

Now, we prove an additive approximation guarantee for every coordinate of $\mathfrak{h}_2(D, \varepsilon)$.

Theorem 3. *(Expected Approximation Guarantee) Algorithm 2 has the following approximation guarantee:*

$$\mathbb{E}\left[\max_{i\in[d]} \left|\frac{1}{q} w(D, i) - \sqrt{\frac{2}{d}}\frac{1}{q} w(P_{\text{avg}}, i)\right|\right] \leq 2\sqrt{\frac{2\log(2/\eta)}{d^2}} + 11\sqrt{2}\frac{\log d}{q\varepsilon\sqrt{d}} + \frac{4}{d} + 2d\exp(-q/4) + \eta$$

After the quantization step, the algorithm for $\mathfrak{h}_2(\cdot)$ (Algorithm 2) follows steps similar to the MWEM algorithm of Hardt et al. (2012) but applied to the vectors in dataset D_Q. The different scalar queries on this data set are essentially the sums of the vectors in D_Q along each of the d coordinates. Therefore, we have the following theorem from Hardt et al. (2012), adapted to our case where the data set is D_Q and the set of queries are the marginal sums $w(D_Q, i)$. This gives the following guarantee:

Theorem 4. *Hardt et al. (2012) For any constant $c \geq 1$, with probability at least $1 - \frac{2T}{d^c}$, Algorithm 2 produces P_{avg} such that: $\max_{i\in[d]} |w(D_Q, i) - w(P_{\text{avg}}, i)| \leq 2q\sqrt{\frac{d\log|S|}{T}} + (3c + 2)\frac{\log d}{\varepsilon}$.*

Proof. This follows directly from Hardt et al. (2012), where we set the distribution support to be $|S|^d$ and support of every entry in D_Q to be from $[-q, q]$. □

Now, we provide an approximation guarantee for the quantization step using the Q function.

Lemma 2. $\mathbb{E}[w(D_Q, i)] = \sqrt{\frac{d}{2}} w(D, i)$. *With probability at least $1 - 2d\exp(-\frac{q}{4})$, we have the following approximation: $\left|\frac{1}{q}\sqrt{\frac{2}{d}} w(D_Q, i) - \frac{1}{q} w(D, i)\right| \leq \eta$*

Proof. Every variable $\tilde{v}_{ji} - \sqrt{2}dv_{ji}$ is an independent mean zero random variable bounded in the interval $[-\eta, \eta]$. Therefore, applying Chernoff Jukna (2011) bounds for bounded random variables with deviation $q\eta$ to the sum random variable $w(D_Q, i)$ and combining it with a union bound on the d coordinates yields the result. □

Proof of Theorem 3. The theorem statement follows from the following: a) $|\frac{1}{q} w(D, i) - \sqrt{\frac{2}{d}}\frac{1}{q} w(P_{\text{avg}}, i)| \leq 2$ in the worst case and b) Lemma 2 and choosing the parameters $T = d^2, c = 3$ in Theorem 4. □

Final Differential Privacy and Approximation Guarantees: We now describe the choices of different parameters in our protocol, including, $\epsilon_{\ell,T}$ over various epochs. In each of the p epochs (note that p is the final summary size), we apply Algorithm 2. In Theorem 2, we prove that releases of aggregator to any data owner i in our protocol are ϵ-differentially private (using the composition theorem from Kairouz et al. (2017)) with respect to data sets of all other data owners except i. Further,

we also bound the final expected additive error of our protocol over multiple rounds. Hence, using the following corollary (of a theorem due to Nemhauser, Wolsey and Fisher) we obtain approximation guarantees closely matching the greedy algorithm.

Theorem 5. *(Corollary of Nemhauser et al. (1978)) Given a non-negative, monotone, submodular function $f : 2^U \to \mathbb{R}^+ \cup \{0\}$. Let OPT be the optimal subset maximizing f such that $|OPT| \leq p$. Similarly, let A be the subset produced by greedy algorithm such that the additive error in the marginal gain in iteration i is Δ_i. Then, $f(A) \geq (1 - e^{-1})f(OPT) - \sum_{i\in[p]} \Delta_i$*

D Proof of Theorem 2

We first prove the following differential privacy guarantees on various participant releases:

Theorem 6. *For any fixed $\frac{1}{e} > \tilde{\delta} > 0$, the releases of the aggregator during Algorithm 3 to the any data owner i is $(\varepsilon, \tilde{\delta})$-differentially private over all the iterations/epochs with respect to $\cup_{j\neq i} D_i$ when we set $\epsilon_{\ell,T} = \frac{\epsilon}{\sqrt{16T\ell \log\left(\frac{1}{\tilde{\delta}}\right) \log p}}$. Similarly, we have $(\epsilon, \tilde{\delta})$-differentially privacy over all the iterations with respect to validation set D_v, when we set $\varepsilon_v = \frac{\epsilon}{\sqrt{16T}}$.*

We quote a recent result on composition theorems for differential privacy first.

Theorem 7. *Kairouz et al. (2017) For any $\epsilon_\ell > 0$, $\delta_\ell \in [0, 1]$ for any $\ell \in \{1, 2, \ldots, k\}$ and $\tilde{\delta} \in [0, 1/e]$, the class $(\epsilon_\ell, \delta_\ell)$-differentially private mechanisms satisfy $(\tilde{\epsilon}_{\tilde{\delta}}, 1 - (1 - \tilde{\delta})\Pi_{\ell=1}^k (1 - \delta_\ell))$-differential privacy under k-fold adaptive composition, for $\tilde{\epsilon}_{\tilde{\delta}} =$*

$$\min \left\{ \sum_{\ell=1}^k \epsilon_\ell, \sum_{\ell=1}^k \frac{(e^{\epsilon_\ell} - 1)\epsilon_\ell}{e^{\epsilon_\ell} + 1)} + \sqrt{\sum_{\ell=1}^k 2\epsilon_\ell^2 \log\left(\frac{1}{\tilde{\delta}}\right)}, \right.$$
$$\left. \sum_{\ell=1}^k \frac{(e^{\epsilon_\ell} - 1)\epsilon_\ell}{(e^{\epsilon_\ell} + 1)} + \sqrt{\sum_{\ell=1}^k 2\epsilon_\ell^2 \log\left(e + \frac{\sqrt{\sum_{\ell=1}^k 2\epsilon_\ell^2}}{\tilde{\delta}}\right)} \right\}$$

Proof of Theorem 6. There are two types of releases by the aggregator to the data providers, over various iterations and we bound the differential privacy for these releases individually.

1. Releases of hashes $\mathfrak{h}_1(\cdot)$ and $\mathfrak{h}_2(\cdot)$ over multiple iterations.
2. Release of information in the process of collecting data points from "winner" data sources.

Let us now analyze differential privacy of releases of type 2. We set $\epsilon_{auc} = \frac{\epsilon}{3\sqrt{2}\log\frac{1}{\delta}} K^{\frac{-1}{3}}$ and $\tau = K^{\frac{2}{3}}$. For the analysis of differential privacy from the perspective of data source j, consider two neighboring datasets $D = \cup_{j\neq i} D_i$ and $D' = D \cup \{x\}$. Let us assume that x belongs to data source $i' \neq j$. When D' is involved, define an iteration as *bad* if (a) x is chosen by a data owner as marginally the best point in $D_{i'}$ (b) x is not chosen by the aggregator. By the virtue of our auction mechanism, there are at most τ such bad iterations, beyond which the point x is chosen by the aggregator.

The key point to note is that if an iteration is not bad, then the output distribution, i.e., the probabilities of chosen points by the aggregator, remains unchanged compared to the case when D is involved.

Further, in a bad iteration, the bid-value position of the data source j can change by at most 1, say from j to $j + 1$ and thus the probability of choosing the data source D_j's point can change by a factor of at most $\frac{e^{-(j-1)\epsilon_{auc}}}{e^{-j\epsilon_{auc}}} = e^{\epsilon_{auc}}$. Thus, the aggregator's queries for the private auction to data source j for these iterations are ϵ_{auc}-differentially private.

Now, applying Theorem 7, we have:

$$\sum_{b=1}^{\tau} \frac{(e^{\epsilon_{auc}}-1)\epsilon_{auc}}{(e^{\epsilon_{auc}}+1)} \leq \sum_{b=1}^{\tau} \epsilon_{auc}^2 = \frac{\epsilon^2}{18\log\frac{1}{\tilde{\delta}}} \leq \epsilon^2/18 \tag{8}$$

and

$$\sqrt{\sum_{b=1}^{\tau} 2\epsilon_{auc}^2 \log\left(\frac{1}{\tilde{\delta}}\right)} = \sqrt{\frac{2\epsilon^2 \log\frac{1}{\tilde{\delta}}}{18\log^2\left(\frac{1}{\tilde{\delta}}\right)}} \leq \frac{\epsilon}{3} \tag{9}$$

In Algorithm 2, steps 6 and 7 together release $i(t)$ and $\mu_i(t)$ (that are function of the final summary D_s) which is used in the computation of $P_t(\mathbf{s})$ which is used in the release from the aggregator to the data owners. Each of them is ε differentially private. However, the ℓ-th call to Algorithm 2 by the protocol 3 uses $\varepsilon = \epsilon_{\ell,T}$. There are T steps inside each call.

We now set $\epsilon_{\ell,T} = \frac{\epsilon}{\sqrt{36T\ell\log\left(\frac{1}{\tilde{\delta}}\right)\log p}}$, for iteration ℓ and apply Theorem 7 over all iterations. Firstly, note that by basic calculus, for $x \geq 0$, $\frac{e^x-1}{e^x+1} \leq x$. This is because, setting $f(x) = (e^x-1) - x(e^x+1)$ has $f(0) = 0$ and $f'(x) = e^x - (e^x+1) - x(e^x+1) <= 0$.

Thus, we have,

$$\begin{aligned}\sum_{\ell=1}^{p}\sum_{t=1}^{2T} \frac{(e^{\epsilon_{\ell,T}}-1)\epsilon_{\ell,T}}{(e^{\epsilon_{\ell,T}}+1)} &\leq \sum_{\ell=1}^{p}\sum_{t=1}^{2T} \epsilon_{\ell,T}^2 \\ &= \sum_{\ell=1}^{p} \frac{\epsilon^2}{18\log\left(\frac{1}{\tilde{\delta}}\right)\ell\log p} \\ &\leq \frac{\epsilon^2}{18\log\left(\frac{1}{\tilde{\delta}}\right)\log p}\left(\sum_{\ell=1}^{p}\frac{1}{\ell}\right) \leq \frac{\epsilon^2}{18}\end{aligned} \tag{10}$$

and

$$\sqrt{\sum_{\ell=1}^{p}\sum_{t=1}^{2T} 2\epsilon_{\ell,T}^2 \log\left(\frac{1}{\tilde{\delta}}\right)} = \sqrt{\sum_{\ell=1}^{p} \frac{2\log\frac{1}{\tilde{\delta}}\epsilon^2}{18\log\left(\frac{1}{\tilde{\delta}}\right)\ell\log p}} \leq \frac{\epsilon}{3} \tag{11}$$

By Theorem 7, the protocol releases to any data owner is $(\tilde{\epsilon}, \tilde{\delta})$-differentially private with respect to $D_s - D_i$ where $\tilde{\epsilon} \leq \frac{2\epsilon}{3} + \frac{\epsilon^2}{9}$. A similar computation shows that the releases of the aggregator during the protocol is $(\epsilon, \tilde{\delta})$-differential private with respect to the validation set D_v. □

Now, we bound the overall expected additive error of our protocol. Define $\mathsf{err}(E)$ as the expected additive error in computing an expression E.

Lemma 3. *Suppose in the greedy algorithm, S_q is the set of points chosen until iteration q and x_{q+1} be the new point chosen in iteration $q+1$. Let D_v be the validation set. Let ξ denote the maximum expected error in computing the terms, $\mathsf{err}(\frac{\sum_{i\in D_v} k(x_{q+1},y_i)}{m}) \leq \xi$ and $\mathsf{err}(\frac{\sum_{j\in S_q} k(x_{q+1},y_j)}{q}) \leq \xi$. Then the overall expected additive error of the algorithm is bounded by $\Delta \leq 7\xi \ln p$.*

Proof. Consider the marginal increment in $J(.)$ in iteration $q+1$:

$$
\begin{aligned}
& J(S_q \cup x_{q+1}) - J(S_q) \\
& = \frac{2}{m}\left\{\frac{1}{q+1}\sum_{i\in D_v, j\in S_{q+1}} k(y_i,x_j) - \frac{1}{q}\sum_{i\in D_v, j\in S_q} k(y_i,x_j)\right\} \\
& - \left\{\frac{1}{(q+1)^2}\sum_{i,j\in S_{q+1}} k(x_i,x_j) - \frac{1}{q^2}\sum_{i,j\in S_q} k(x_i,x_j)\right\} \\
& = \frac{1}{q+1}\left\{\frac{2\sum_{i\in D_v} k(x_{q+1},y_i)}{m}\right. \\
& \left. - \frac{q}{q+1}\frac{(1+2\sum_{j\in S_q} k(x_{q+1},x_i))}{q}\right\} \\
& + \left(\frac{1}{q+1} - \frac{1}{q}\right)\frac{\sum_{q\in S_q}\sum_{i\in D_v} k(x_q,y_i)}{m} + \\
& \left(\frac{1}{(q+1)^2} - \frac{1}{q^2}\right)\sum_{i,j\in S_q} k(x_i,x_j)
\end{aligned} \tag{12}
$$

Now, we bound the additive error in computing this marginal increment as follows.

$$
\begin{aligned}
& \mathsf{err}(J(S_q \cup x_{q+1}) - J(S_q)) \\
& \leq \frac{1}{q+1}\mathsf{err}\left(\frac{2\sum_{i\in D_v} k(x_{q+1},y_i)}{m}\right) + \\
& \frac{q}{(q+1)^2}\mathsf{err}\left(\frac{2\sum_{j\in S_q} k(x_{q+1},x_i)}{q}\right) \\
& + \left(\frac{1}{q} - \frac{1}{q+1}\right)\sum_{q\in S_q}\mathsf{err}\left(\frac{\sum_{i\in D_v} k(x_q,y_i)}{m}\right) + \\
& \left(\frac{q}{q^2} - \frac{q}{(q+1)^2}\right)\sum_{q\in S_q}\mathsf{err}\left(\frac{\sum_{j\in S_q} k(x_q,x_j)}{q}\right) \\
& \leq \frac{2\xi}{q+1} + \frac{2q\xi}{(q+1)^2} + \frac{1}{q(q+1)}\sum_{q\in S_q}\xi + \frac{2q+1}{q(q+1)^2}\sum_{q\in S_q}\xi \\
& = \frac{2\xi}{q+1} + \frac{2q\xi}{(q+1)^2} + \frac{1}{q(q+1)}q\xi + \frac{2q+1}{q(q+1)^2}q\xi \leq \frac{7\xi}{q+1}
\end{aligned} \tag{13}
$$

By Theorem 5, the overall expected additive error in the greedy algorithm is bounded by $\Delta \leq \sum_{q\in[p]}\Delta_q \leq \sum_{q\in[p]}\frac{7\xi}{q+1} \leq 7\xi\ln p$ □

Lemma 4. *Let $0 < a < 1$ be a small fixed constant. Let $|D_v| \geq \frac{11*4\sqrt{2}\sqrt{d}\log d\log^2 p}{\varepsilon_v}$, $|D_{\text{init}}| \geq 121*8d^2\log^2 d\log(\frac{1}{\delta})\log^5 p$, $d \geq \frac{16(\log 2N)(\log p)^2}{a^2}$, $\eta \leq \frac{1}{d}$, we have $\Delta \leq 7\xi\ln p < O(\frac{\log p\sqrt{\ln d}}{\sqrt{d}}) + a + \frac{1}{\epsilon\log p} < 1$.*

Proof. Let N be total number of points in the system. First, we use a theorem from Rahimi & Recht (2008), to show that $\mathbb{P}(\sup_{x_i,x_j}|\mathfrak{h}_1(x_i)\cdot\mathfrak{h}_1(x_j) - k(x_i,x_j)| \geq \varepsilon_{rr}) \leq \frac{1}{N^2}$. Indeed, for a *fixed* pair of points, x_i, x_j, it holds that: $\mathbb{P}(|\mathfrak{h}_1(x_i)\cdot\mathfrak{h}_1(x_j) - k(x_i,x_j)| \geq \varepsilon_{rr}) \leq \exp(-\frac{d\varepsilon_{rr}^2}{4})$. Thus, by union bound, and setting $d \geq \frac{16\log 2N}{\varepsilon_{rr}^2}$, we have the above claim.

Now, from Theorem 3, for iteration ℓ in the protocol, we have the following guarantee:

$$\mathsf{err}\left(\frac{w(\mathfrak{h}_1(D_s),i)}{q}\right) \leq 2\sqrt{\frac{2\log(2/\eta)}{d^2}} + 11\sqrt{2}\frac{\log d}{q\epsilon_{\ell,T}\sqrt{d}} + \frac{4}{d} + 2d\exp(-q/4) + \eta \tag{14}$$

Observe that at iteration ℓ, $q = \ell + |D_{\text{init}}|$ since this is the effective size of the summary. By the inequality between the arithmetic and geometric mean, we have: $q \geq \sqrt{4\ell|D_{\text{init}}|}$. Now, we let $|D_{\text{init}}| \geq 121 * 8d^2\log^2 d\log(\frac{1}{\delta})\log^5 p$. Now, we set $\eta \leq \frac{1}{d}$. Then,

$$\begin{aligned}\mathsf{err}\left(\frac{w(\mathfrak{h}_1(D_s),i)}{q}\right) &\leq 2\sqrt{\frac{2\log(2/\eta)}{d^2}} + \frac{6}{d} + \frac{1}{\sqrt{d}\epsilon\log^2 p} \\ &= \Delta_{\max}\end{aligned}$$

Let $\mathfrak{h}_1(\mathbf{x}_{q+1})[i]$ be the i-th coordinate of $\mathfrak{h}_1(\mathbf{x}_{q+1})$. Observe that $|\mathfrak{h}_1(\mathbf{x}_{q+1})[i]| \leq \sqrt{\frac{2}{d}}$.

We have the following expected additive error:

$$\begin{aligned}&\mathsf{err}\left(\frac{\sum_{j\in D_s} k(x_{q+1},x_j)}{q}\right) \leq \varepsilon_{rr} + \\ &\mathsf{err}\left(\frac{\sum_{j\in D_s} \mathfrak{h}_1(x_{q+1})\cdot\mathfrak{h}_1(x_j)}{q}\right) \leq \varepsilon_{rr} + \Delta_{\max}\sqrt{2d}.\end{aligned} \tag{15}$$

We set $\epsilon_{err} = a/\log p$ for some small constant $a > 0$. Therefore, $d \geq 16\log 2N(\log p)^2/a^2$. Now, we have:

$$\begin{aligned}\mathsf{err}\left(\frac{\sum_{j\in D_s} k(x_{q+1},x_j)}{q}\right) &\leq \varepsilon_{rr} + \Delta_{\max}\sqrt{2d} \\ &= \frac{a}{\log p} + 4\sqrt{\frac{\log(2/\eta)}{d}} + \frac{6\sqrt{2}}{\sqrt{d}} + \frac{1}{\epsilon\log^2 p} \\ &= O(\frac{\sqrt{\ln d}}{\sqrt{d}}) + \frac{a}{\log p} + \frac{1}{\epsilon\log^2 p}\end{aligned} \tag{16}$$

Similarly, we can show that for validation set D_v, we need $|D_v| \geq 11 * 4\sqrt{2}\sqrt{d}\log d\log^2 p$. Now, since $\varepsilon_v = \frac{\varepsilon}{\sqrt{16d^2}}$, the $\Delta_{\max}$ bound holds for the validation term too. □

Lemma 5. *In Algorithm 3 the expected number of points accessed by the aggregator is* $p(\frac{K}{\tau} + \frac{1}{\epsilon_{auc}}) = (1 + \frac{3\sqrt{2}\log\frac{1}{\delta}}{\epsilon})pK^{\frac{1}{3}} = O(\frac{p\log\frac{1}{\delta}}{\epsilon}K^{\frac{1}{3}})$

Proof. In the Step 13 of the mechanism, the expected number of points chosen in each iteration is $\sum_{i\in[K]}\mathbb{P}(x_i) = \sum_{i\in[K]} e^{(i-1)\epsilon_{auc}} = (1-e^{-K\epsilon_{auc}})/(1-e^{-\epsilon_{auc}}) \leq \frac{1}{1-e^{-\epsilon_{auc}}} \leq \frac{1}{\epsilon_{auc}}$. Thus in p iterations, the expected number of points chosen $= \frac{p}{\epsilon_{auc}}$. In the second step, the maximum number of points that are the best for a data source more than τ times is $\frac{pK}{\tau}$. Thus the lemma follows. □

Proof of Theorem 2. The proof follows from the results in this section.

□

E Proof of Theorem 1

We begin by quoting a known result from Kim et al. (2016).

Theorem 8. *Kim et al. (2016) Let $\mathbf{H} \in \mathbb{R}^{g \times g}$ be element-wise non-negative and bounded with $h^* = \max i, j \in [g] h_{i,j} > 0$. Define a binary matrix $\mathbf{E}$ with entries $e_{i,j} = 1$ if $h_{i,j} = h^*$ and 0 otherwise. Similarly define $\mathbf{E}' = \mathbf{1} - \mathbf{E}$. Given the ground set $\mathcal{S} \subseteq 2^{[g]}$ consider the linear form: $F(\mathbf{H}, S) = \langle A(S), \mathbf{H} \rangle \; \forall S \in \mathcal{S}$. Given $s = |S|$, define the functions:*

$\alpha(g,s) = \frac{a(S \cup \{u\}) - a(S)}{b(S)}$, $\beta(g,s) = \frac{a(S \cup \{u\}) + a(S \cup \{v\}) - a(S \cup \{u,v\}) - a(S)}{b(S) + b(S \cup \{u,v\})}$, *where* $a(S) = F(\mathbf{E}, S)$ *and* $b(S) = F(\mathbf{E}', S)$ *for all* $u, v \in S$. *Let* $s^* = \max_{S \in \mathcal{S}} |S|$, *we have*

1. $F(\mathbf{H}, S)$ *is monotone, if* $h_{i,j} \leq h^* \alpha(g,s), \forall 0 \leq s \leq s^*$
2. $F(\mathbf{H}, S)$ *is submodular, if* $h_{i,j} \leq h^* \beta(g,s), \forall 0 \leq s \leq s^*$

Proof of Theorem 1. Firstly, we show that the function $J(S)$ can be written in a linear form. Note that the same linear form used by Kim et al. (2016) would not work for our case.

We define $\mathbf{U}$ as the kernel matrix of all the points in $D_1 \cup D_2 \ldots D_k \cup D_v$.

Now, we observe that our $J(S) = \langle \mathbf{A}(S), \mathbf{U} \rangle$, where $\mathbf{A}(S) = \frac{2}{m|S|} \mathbf{1}_{[i \in S]} \mathbf{1}_{[j \in V]} - \frac{1}{|S|^2} \mathbf{1}_{[i \in S]} \mathbf{1}_{[j \in S]}$. Let $\mathbf{E}$ as the binary matrix defined in Theorem 8 with $\mathbf{H} = \mathbf{U}$.

We now compute $a(S) = \langle \mathbf{A}(S), \mathbf{E} \rangle$ and $b(S) = \langle \mathbf{A}(S), \mathbf{1} - \mathbf{E} \rangle$ values.

Computing $a(S)$:

$$a(S) = \langle \mathbf{A}(S), \mathbf{I} \rangle = \frac{2}{m|S|} 0 - \frac{1}{|S|^2} |S| = -\frac{1}{|S|} \tag{17}$$

Computing $b(S)$:

$$\begin{aligned} b(S) &= \langle \mathbf{A}(S), \mathbf{1} - \mathbf{I} \rangle = \langle \mathbf{A}(S), \mathbf{1} \rangle - \langle \mathbf{A}(S), \mathbf{I} \rangle \\ &= \frac{2}{m|S|} |S| m - \frac{1}{|S|^2} |S|^2 + \frac{1}{|S|} = \frac{1}{|S|} + 1 \end{aligned} \tag{18}$$

Now, we show that the bounds on $\alpha(g,s)$ and $\beta(g,s)$ hold:

$$\alpha(g,s) = \frac{a(S \cup \{u\}) - a(S)}{b(S)} = \frac{\frac{1}{|S|} - \frac{1}{|S|+1}}{\frac{1}{|S|} + 1} = \frac{1}{(1+|S|)^2} \tag{19}$$

Further,

$$\begin{aligned} \beta(g,s) &= \frac{a(S \cup \{u\}) + a(S \cup \{v\}) - a(S \cup \{u,v\}) - a(S)}{b(S) + b(S \cup \{u,v\})} \\ &= \frac{-\frac{2}{|S|+1} + \frac{1}{|S|+2} + \frac{1}{|S|}}{\frac{1}{|S|} + 1 + \frac{1}{|S|+2} + 1} = \frac{1}{n^3 + 3n^2 + n} \end{aligned} \tag{20}$$

Thus, we have $k_{i,j} \leq \beta(g,s) k^*$ and hence the conditions of the Theorem 8 are satisfied. Therefore, $J(S)$ is a monotone and submodular function. □

F Additional Privacy Properties of $\mathfrak{h}_1(\cdot)$:

Consider any data set D_r. Over the course of p epochs, suppose the data source r contributes p_r points by winning bids at Line 7 of Algorithm 3. Let the points be $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{p_r}$. We show that the joint probability density function of the random variables $\mathfrak{h}_1(\mathbf{x}_1), \ldots, \mathfrak{h}_1(\mathbf{x}_{p_r})$ depends only on the pairwise distances between the points, i.e. $\|\mathbf{x}_u - \mathbf{x}_v\| \; \forall u, v \in [1 : p_r]$. In a strong information theoretic sense, this implies that the only information that can be gained about these points by the aggregator are the pairwise distances between the data points. The

intention of usage of the hashes is to compute $k(\mathbf{x}_u, \mathbf{x}_v) = k(\|\mathbf{x}_u - \mathbf{x}_v\|_2)$ approximately at the aggregator. Hence, the aggregator gains strictly no more information than it needs. Consider the matrix $\left[\mathfrak{h}_1(\mathbf{x}_1)^T, \ldots, \mathfrak{h}_1(\mathbf{x}_{p_r})^T\right]^T$. Each column of this matrix is an i.i.d sample drawn from the distribution on the variables: $\left[\sqrt{2/d}\cos(\mathbf{w}^T\mathbf{x}_1 + b) \ldots \sqrt{2/d}\cos(\mathbf{w}^T\mathbf{x}_{p_r} + b)\right]$ where $\mathbf{w} \sim N(0, 2\gamma\mathbf{I}_n),\ b \sim \text{Uniform}[0, 2\pi]$. In fact, we will analyze the joint characteristic function of the angles in a single column given by: $\left[(\mathbf{w}^T\mathbf{x}_1 + b) \mod 2\pi \ldots (\mathbf{w}^T\mathbf{x}_{p_r} + b) \mod 2\pi\right]^T$. In an intuitive sense, these variables represent a randomly shifted jointly Gaussian variables 'wrapped' around a unit circle (usually called the wrapped distribution Mardia & Jupp (2009)). The next theorem shows that the characteristic function depends only on the pairwise distance of the data points.

Theorem 9. *Let $\phi_w(\cdot)$ be the characteristic function of the wrapped distribution of the variables $\left[(\mathbf{w}^T\mathbf{x}_1 + b) \mod 2\pi, \ldots, (\mathbf{w}^T\mathbf{x}_{p_r} + b) \mod 2\pi\right]$. Then, we have: a) $\forall \mathbf{s} \in \mathbb{R}^{p_r} - \mathbb{Z}^{p_r}$, $\phi_w(\mathbf{s}) = 0$. b) $\forall \mathbf{k} \in \mathbb{Z}^{p_r},\ \mathbf{1}^\mathbf{k} \neq 0,\ \phi_w(\mathbf{k}) = 0$. c) $\forall \mathbf{k} \in \mathbb{Z}^{p_r},\ \mathbf{1}^\mathbf{k} \neq 0$, $\phi_w(\mathbf{k}) = \phi(\mathbf{k}) = \prod_{i,j \in [1:p_r]} \left(\phi^{(i,j)}\right)^{m_{i,j}}$ where $m_{i,j}$ are some integers that depend on the vector $\mathbf{k}$ alone. Here, $\phi^{(i,j)} = \exp(-2\gamma\|\mathbf{x}_i - \mathbf{x}_j\|_2^2) = k(\|\mathbf{x}_i - \mathbf{x}_j\|_2)$.*

Remark: We are not aware of any analysis of the joint distribution of multiple data point releases using Rahimi-Recht random features method for the RBF kernel. We use Fourier analysis, properties of multi-dimensional Dirac-combs Giraud & Peschanski (2015) to prove the above theorem.

F.1 Proof of Theorem 9

We first review results relating characteristic function of unwrapped distributions and the wrapped distributions. This relationships is due to some facts known about multi-dimensional Dirac Comb in standard Fourier Analysis. Let $p(\mathbf{v})$ be a density function defined on $\mathbb{R}^s$. Here $p(\cdot)$ is the unwrapped joint density function of the variables, $\left[\mathbf{w}^T\mathbf{x}_1 + b \ldots \mathbf{w}^T\mathbf{x}_{p_r} + b\right]$. Here, $\mathbf{v} \in \mathbb{R}^s$. The wrapped distribution of this density function is given by: $p_w(\mathbf{v}) = \sum_{\mathbf{k} \in \mathbb{Z}^n} p(\mathbf{v} + 2\pi\mathbf{k})$. Define the Dirac comb as: $\Delta_{2\pi}(\mathbf{v}) = \sum_{\mathbf{k} \in \mathbb{Z}^s} \delta(\mathbf{v} - 2\pi\mathbf{k})$ where $\delta(\mathbf{v}) = \prod_i \delta(v_i)$ and $\delta(\cdot)$ is a single dimensional Dirac-delta function. Although Dirac-delta functions are not rigorous as a real function, as a measure on the space $\mathbb{R}^s$, they are very well defined and rigorous.

It is known that the Fourier Series of the Dirac comb is given by:

$$\Delta_{2\pi}(\mathbf{v}) = \frac{1}{(2\pi)^s} \sum_{\mathbf{k} \in \mathbb{Z}^s} \exp(-i\mathbf{k}^T\mathbf{v}) \tag{21}$$

Therefore, any wrapped distribution can be written in the following way:

$$\begin{aligned} p_w(\mathbf{v}) &= \int p(\mathbf{v}')\Delta_{2\pi}(\mathbf{v} - \mathbf{v}')d\mathbf{v}' \\ &= \frac{1}{(2\pi)^s} \int p(\mathbf{v}') \sum_{\mathbf{k} \in \mathbb{Z}^s} \exp(-i\mathbf{k}^T(\mathbf{v} - \mathbf{v}'))d\mathbf{v}' \qquad (22) \\ &= \frac{1}{(2\pi)^s} \sum_{\mathbf{k} \in \mathbb{Z}^s} \phi(\mathbf{k}) \exp(-i\mathbf{k}^T\mathbf{v}) \qquad (23) \end{aligned}$$

Here, $\phi(\mathbf{k}) = \mathbb{E}_p[\exp(i\mathbf{k}^\mathbf{v})]$ is the characteristic function of the distribution $p(\cdot)$ on the integer lattice. Therefore, any wrapped distribution can be written as a Fourier series with Fourier Coefficients being the characteristic function evaluated at the integer lattice.

Let $\phi_w(\cdot)$ be the characteristic function of the wrapped distribution. Further, $\phi_w(\mathbf{k}) = \phi(\mathbf{k}),\ \forall \mathbf{k} \in \mathbb{Z}^+$ while $\phi_w(\mathbf{s}) = 0$ when $\mathbf{s} \in \mathbb{R}^s\ \mathbb{Z}^s$ is not on the integer lattice. This is very clear from the Fourier serier representation of the wrapped distribution as in (22).

Lemma 6. *$\phi_w(\mathbf{k}) = \phi(\mathbf{k}) = 0,\ \forall \mathbf{k} : \mathbf{1}^T\mathbf{k} \neq 0$ when $p(\cdot)$ is the unwrapped joint distribution of $\left[\mathbf{w}^T\mathbf{x}_1 + b \ldots \mathbf{w}^T\mathbf{x}_{p_i} + b\right]$ where $\mathbf{w} \sim \mathcal{N}(0, 2\gamma\mathbf{I}_n)$ and $b \sim \text{Uniform}[0, 2\pi]$.*

Proof. Let $\mathbf{X} = \left[\mathbf{x}_1^T \dots \mathbf{x}_{p_r}^T\right]^T$. Therefore, variables $\mathbf{w}^T\mathbf{x}_j$ are jointly Gaussian with the covariance matrix $\Sigma = \mathbf{X}\mathbf{X}^T$. Given a fixed b, the conditional characteristic function over the integer lattice is given by:

$$\phi_{|b}(\mathbf{k}) = \exp(i(\mathbf{k}^T\mathbf{1})b)\exp(-\frac{1}{2}\mathbf{k}^T\Sigma\mathbf{k}) \tag{24}$$

This is the characteristic function of the standard multidimensional normal distribution.

$E_b[\exp(imb)] = 0$ for an integer m and $b \sim \text{Uniform}[0, 2\pi]$. Therefore, by (24), we have the desired result. □

We will show that the $\phi(\mathbf{k})$ is a function only of the pairwise distances between the points whenever $\mathbf{1}^T\mathbf{k} = 0$.

Lemma 7. *Let* $\mathbf{k}^{(i,j)} = [0 \dots \underset{\text{position } i}{1} \dots\dots \underset{\text{position } j}{-1} \dots]$). *Then,* $\phi\left(\mathbf{k}^{(i,j)}\right) = \exp(-2\gamma\|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$. *Further, whenever* $\mathbf{k}^T\mathbf{1} = 0$, $\phi_w(\mathbf{k}) = \phi(\mathbf{k}) = \prod_{i,j\in[1:p_r]} \left(\phi\left(\mathbf{k}^{(i,j)}\right)\right)^{m_{i,j}}$ *where* $m_{i,j}$ *are some integers that depend on the vector* $\mathbf{k}$ *alone.*

Proof of Lemma 7. Whenever $\mathbf{k}^T\mathbf{1} = 0$, by (24) $\phi(\mathbf{k})$ is a function of $\|\mathbf{k}^T\mathbf{X}\|^2$. Let $\sum|k_i| = 2t$, $t \in \mathbb{Z}^+$. The sum of absolute values is an even integer because $\sum k_i = 0$. Now, we can write $\|\mathbf{k}^T\mathbf{X}\|^2$ as follows:

$$\|\mathbf{k}^T\mathbf{X}\|^2 = \|\sum_{j=1}^{t}(\mathbf{g}_i - \mathbf{h}_i)\|_2^2 \tag{25}$$

where $\mathbf{g}_i = \mathbf{x}_j$ for some $j \in [1 : p_r]$ and $\mathbf{h}_i = \mathbf{x}_k$ for some $k \in [1 : p_r]$. Because any distinct data point $\mathbf{x}_j$ is multiplied only by either positive or negative integers, clearly $\{\mathbf{g}_i\}_{i=1}^t \bigcap \{\mathbf{h}_i\}_{i=}^t = \emptyset$.

Now, we have:

$$\begin{aligned}\|\sum_{j=1}^{t}(\mathbf{g}_i - \mathbf{h}_i)\|_2^2 = \sum_{j=1}^{t}\|(\mathbf{g}_j - \mathbf{h}_j)\|_2^2 + \\ 2 * \sum_{j,j'}(\mathbf{g}_j - \mathbf{h}_j)^T(\mathbf{g}_{j'} - \mathbf{h}_{j'})\end{aligned} \tag{26}$$

The first terms set of terms clearly are function of pairwise distances between points. Now we rewrite the cross terms as linear combination of pairwise distances in the following way.

$$\begin{aligned}2 * (\mathbf{g}_j - \mathbf{h}_j)^T(\mathbf{g}_{j'} - \mathbf{h}_{j'}) = \|\mathbf{g}_j - \mathbf{h}_{j'}\|_2^2 + \|\mathbf{g}_{j'} - \mathbf{h}_j\|_2^2 \\ - \|\mathbf{g}_j - \mathbf{g}_{j'}\|_2^2 - \|\mathbf{h}_j - \mathbf{h}_{j'}\|_2^2\end{aligned} \tag{27}$$

Hence, characteristic function can be written as pairwise distances between the data points.

Let $\mathbf{k}^{(i,j)} = [0 \dots \underset{\text{position } i}{1} \dots\dots \underset{\text{position } j}{-1} \dots]$). Then, $\phi\left(\mathbf{k}^{(i,j)}\right) = \exp(-2\gamma\|\mathbf{x}_i - \mathbf{x}_j\|_2^2)$. These are exactly the kernel values that the Aggregator is interested in. By (26) and (27), it is clear that the characteristic function can be written in terms of powers of $\phi\left(\mathbf{k}^{(i,j)}\right)$, i.e.

$$\phi(\mathbf{k}) = \prod_{i,j\in[1:p_r]}\left(\phi\left(\mathbf{k}^{(i,j)}\right)\right)^{m_{i,j}} \tag{28}$$

where $m_{i,j}$ are some integers that depend on the vector $\mathbf{k}$ alone. □

Proof of Theorem 9. The results of the two lemmas above prove the theorem. □

G Additional Experiments

As discussed before, we set the parameters of our algorithm as in Table 1.

$\gamma = 0.1$	$d = 140$
$T_{\text{init}}(= T, \ell = 1) = d^{1.5} = 1656$	$T_{\text{subs}}(= T, \ell \geq 2) = 5$
ε_v	$\varepsilon_{\ell,T}$
0.01	0.05 for $\ell = 1$, $\frac{0.01}{\sqrt{pT_{\text{subs}}}}$ for $\ell \geq 2$

Table 1: We describe the parameters for our experiments. Here γ is the RBF kernel parameter. d is the dimension of the Rahimi-Recht hash function $\mathfrak{h}_1(\cdot)$. We use two different T parameters for different epochs given by T_{init} (for the first epoch) and T_{subs} (for subsequent epochs). ϵ_v is the ϵ parameter for $\mathfrak{h}_2(\cdot)$ for the validation set and $\epsilon_{\ell,T}$ is set for $\mathfrak{h}_2(\cdot)$ on summaries D_s over epochs ℓ.

MNIST Dataset: We now demonstrate similar results on a standard hand-written digit recognition dataset namely MNIST. We start with a brief description of the setup.

Training: We distribute the MNIST training dataset among five data owners based on digit labels as follows. Splitting the digits into groups $[[0, 1], [3, 4], [5, 6], [7, 8], [9, 2]$, we allocate the training data corresponding to these digits to the corresponding data owners. *Testing:* The test set contains data corresponding to two labels $[3, 4]$ sampled with ratio $[0.7, 0.3]$. *Validation:* We sample (and remove) from the test set with probability 0.25 to construct the validation dataset.

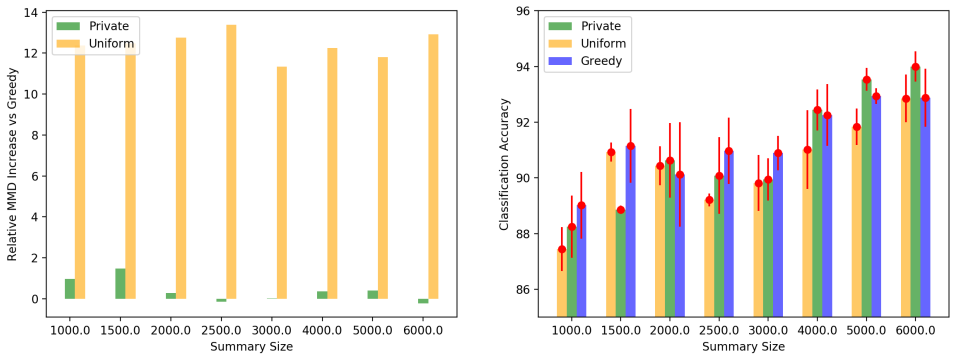

Figure 2: *MNIST Dataset* (Top): Comparison of the percentage increase in MMD^2 of both the private and uniform sampling algorithms with respect to baseline greedy algorithm. Lower values indicate better performance. Consistently there is **10-15**% performance difference from uniform sampling. (Bottom): Comparison of the classification accuracy of the three algorithms using a neural network with one hidden layer of 32 units. Higher numbers indicate better performance.

As before, we vary the number of samples and in Figure 2, compare the percentage increase in MMD^2 with respect to greedy, i.e., $\frac{MMD^2(ALGM) - MMD^2(GREEDY)}{MMD^2(GREEDY)} \times 100$. Recall from above that $ALGM$ is either our private greedy algorithm or the uniform sampling algorithm. Our results show that we consistently outperform the uniform sampling algorithm by at least 10-13%. In Figure 2, we compare the performance of these algorithms using a neural net with 32 neurons in a single hidden layer and drop out of 0.2. Note that since our goal is to demonstrate that the relative performance of these algorithms, we are not concerned with the actual performance numbers (prior works on this subject in fact use a much simple 1-Nearest Neighbor classifier). We again find that the private algorithm beats uniform sampling in most cases.