APPENDIX: Differentially Private Database Release via Kernel Mean Embeddings

A. Proofs

Here we provide proofs for the results stated in the main text, together with additional supporting lemmas required for these proofs.

A.1. Algorithm 1 (Synthetic Data Subspace): Consistency

Before proving Theorem 2, we obtain a Lemma showing that the "projection error" incurred due to projecting the KME $\hat{\mu}_X$ onto the finite-dimensional subspace $\mathcal{H}_M$ spanned by the synthetic data points, quantified by the RKHS distance between $\hat{\mu}_X$ and the projection $\overline{\mu}_X$, converges to 0 as $N \to \infty$:

Lemma 10. *Let $\mathcal{X}$ be a compact metric space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous kernel on $\mathcal{X}$. Suppose that the synthetic data points $z_1, z_2, \ldots$ are sampled i.i.d. from a probability distribution q on $\mathcal{X}$. If the support $\operatorname{supp}(X)$ of X is included in the support of q, then*

$$\|\overline{\mu}_X - \hat{\mu}_X\|_{\mathcal{H}} \xrightarrow{\mathbb{P}} 0 \text{ as } N \to \infty. \tag{6}$$

Proof. Let $\varepsilon > 0$. As k is continuous on $\mathcal{X} \times \mathcal{X}$, which as a product of compact spaces is itself compact by Tychonoff's theorem, the kernel k is uniformly continuous and in particular there exists $\delta > 0$ such that for all $x, x' \in \mathcal{X}$ we have $|k(x,x) - k(x,x')| < \varepsilon^2/2$ whenever $\|x - x'\|_{\mathcal{X}} < \delta$. As $\mathcal{X}$ is compact, it is totally bounded, and thus so is its subset $\operatorname{supp}(X)$. Therefore $\operatorname{supp}(X)$ can be covered with finitely many open balls $B_1, \ldots, B_K$ of radius $\delta/2$. Let the sequence $z_1, z_2, \ldots$ be sampled i.i.d. from q, and let E_M be the event that at least ones of these K balls contains no element of $z_1, \ldots, z_M$. Since $\operatorname{supp}(X) \subseteq \operatorname{supp}(q)$ by assumption, we have $q(B_k) > 0$ for all $k = 1, \ldots, K$ and therefore $\mathbb{P}[E_M] \to 0$ as $M \to \infty$.

Note that if all K balls contain an element of $z_1, \ldots, z_M$ (i.e., E_M^C holds), then for each $x \in \operatorname{supp}(X)$ one can find $1 \leq m(x) \leq M$ such that $\|x - z_{m(x)}\| < \delta/2 + \delta/2 = \delta$. In that case

$$\begin{aligned}
\|\overline{\mu}_X - \hat{\mu}_X\|_{\mathcal{H}} &= \inf_{h \in \mathcal{H}_M} \|h - \hat{\mu}_X\|_{\mathcal{H}} && \text{[property of projection]} \\
&\leq \left\| \frac{1}{N} \sum_{n=1}^{N} k(z_{m(x_n)}, \cdot) - \hat{\mu}_X \right\|_{\mathcal{H}} && \text{[as } \tfrac{1}{N}\textstyle\sum_{n=1}^{N} k(z_{m(x_n)}, \cdot) \in \mathcal{H}_M\text{]} \\
&\leq \frac{1}{N} \sum_{n=1}^{N} \left\| k(z_{m(x_n)}, \cdot) - k(x_n, \cdot) \right\|_{\mathcal{H}} && \text{[Triangle inequality]} \\
&< \frac{1}{N} \sum_{n=1}^{N} \varepsilon && \text{[see below]} \\
&= \varepsilon,
\end{aligned} \tag{7}$$

where we have used the reproducing property, the Triangle inequality and our choices of δ and $z_{m(x_n)}$ to see that for all $1 \leq n \leq N$,

$$\left\| k(z_{m(x_n)}, \cdot) - k(x_n, \cdot) \right\|_{\mathcal{H}} = \langle k(z_{m(x_n)},_{\cdot} - k(x_n, \cdot), k(z_{m(x_n)},_{\cdot} - k(x_n, \cdot) \rangle_{\mathcal{H}}^{1/2} \tag{8}$$

$$= \Big(k(z_{m(x_n)}, z_{m(x_n)}) - 2k(z_{m(x_n)}, x_n) + k(x_n, x_n) \Big)^{1/2} \tag{9}$$

$$\leq \big(|k(z_{m(x_n)}, z_{m(x_n)}) - k(z_{m(x_n)}, x_n)| + |k(x_n, x_n) - k(z_{m(x_n)}, x_n)| \big)^{1/2} \tag{10}$$

$$< \left(\frac{\varepsilon^2}{2} + \frac{\varepsilon^2}{2} \right)^{1/2} = \varepsilon. \tag{11}$$

Hence we have that $\mathbb{P}\left[\|\overline{\mu}_X - \hat{\mu}_X\|_{\mathcal{H}} > \varepsilon \right] \leq \mathbb{P}[E_M] \to 0$ as $M \to \infty$. But since $\varepsilon > 0$ was arbitrary and $M \to \infty$ as $N \to \infty$ by construction, the claimed convergence in probability result follows from definition. □

Theorem 2. Let $\mathcal{X}$ be a compact metric space and $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a continuous kernel on $\mathcal{X}$. Suppose that the synthetic data points $z_1, z_2, \ldots$ are sampled i.i.d. from a probability distribution q on $\mathcal{X}$. If the support of X is included in the support of q, then Algorithm 1 outputs a consistent estimator of the kernel mean embedding μ_X in the sense that

$$\sum_{m=1}^{M} w_m k(z_m, \cdot) \xrightarrow{\mathbb{P}} \mu_X \qquad \text{as } N \to \infty. \tag{12}$$

Proof. Using the Triangle inequality, we can upper bound the RKHS distance between the output $\tilde{\mu}_X$ of Algorithm 1 and the true kernel mean embedding μ_X as follows:

$$\|\tilde{\mu}_X - \mu_X\|_{\mathcal{H}} \leq \underbrace{\|\tilde{\mu}_X - \overline{\mu}_X\|_{\mathcal{H}}}_{\text{privacy error}} + \underbrace{\|\overline{\mu}_X - \hat{\mu}_X\|_{\mathcal{H}}}_{\text{projection error}} + \underbrace{\|\hat{\mu}_X - \mu_X\|_{\mathcal{H}}}_{\text{finite sample error}}. \tag{13}$$

The finite sample error tends to 0 as $N \to \infty$ by the law of large numbers, while the projection error tends to 0 as $N \to \infty$ by Lemma 10. For the privacy error, using orthonormality of the basis $\{b_1, \ldots, b_F\}$ we have

$$\|\tilde{\mu}_X - \overline{\mu}_X\|_{\mathcal{H}}^2 = \left\| \sum_{f=1}^{F} (\beta_f - \alpha_f) b_f \right\|_{\mathcal{H}}^2 = \sum_{f=1}^{F} (\beta_f - \alpha_f)^2 = \frac{8 \ln(1.25/\delta)}{N^2 \varepsilon^2} F \frac{1}{F} \sum_{f=1}^{F} \mathcal{N}(0,1)^2. \tag{14}$$

As a function of N, the size of the basis $F \in \mathbb{N}$ is a non-decreasing function, so it either converges to some $L \in \mathbb{N}$, in which case the obtained expression clearly tends to 0 as $N \to \infty$ with probability 1, or $F \to \infty$ as $N \to \infty$. In this latter case $\frac{1}{F} \sum_{f=1}^{F} \mathcal{N}(0,1)^2 \to 1$ as $N \to \infty$ a.s. by the strong law of large numbers, and $F/N^2 \to 0$ as $N \to \infty$ since $F \leq M = o(N^2)$. Hence the privacy error goes to 0 as $N \to \infty$ either way, as required to complete the proof. □

Theorem 11. *Suppose that the kernel k is c_0-universal (Sriperumbudur et al., 2011) and f is any continuous function mapping from $\mathcal{X}$ to some space $\mathcal{Y}$. Let $C \geq 1$ be any finite constant. If line 7 of Algorithm 1 is replaced with a regularised reduced set method solving the constrained minimisation problem*

$$\mathbf{w} = \operatorname*{argmin}_{\mathbf{u}:\|\mathbf{u}\|_1 \leq C} \left\| \tilde{\mu}_X - \sum_{m=1}^{M} u_m k(z_m, \cdot) \right\|_{\mathcal{H}}, \tag{15}$$

then the points output by Algorithm 1 yield a consistent estimator of the kernel mean embedding $\mathbb{E}[k(f(X), \cdot)]$ of $f(X)$ in the sense that

$$\sum_{m=1}^{M} w_m k(f(z_m), \cdot) \xrightarrow{\mathbb{P}} \mu_{f(X)} \qquad \text{as } N \to \infty. \tag{16}$$

Proof. Let $\mu_X^{\text{out}} := \sum_{m=1}^{M} w_m k(z_m, \cdot)$ be the RKHS element output by Algorithm 1 after adding the stated regularisation. First we show that despite the regularisation, μ_X^{out} remains a consistent estimator of the true kernel mean embedding μ_X as $N \to \infty$.

The modification introduces an additional regularisation error term $\|\mu_X^{\text{out}} - \tilde{\mu}_X\|_{\mathcal{H}}$ into the upper bound on $\|\mu_X^{\text{out}} - \mu_X\|$, compared to the corresponding bound (13) in the proof of Theorem 2. So to show the first desired consistency result, it remains to show that this extra regularisation error term converges to 0 in probability as $N \to \infty$. To this end, let $\varepsilon > 0$ be arbitrary. Define $\delta > 0$, the sequence $z_1, z_2, \ldots$ and $m(x)$ for $x \in \mathcal{X}$ as in the proof of Lemma 10. Note that the RKHS element $\frac{1}{N} \sum_{n=1}^{N} k(z_{m(x_n)}, \cdot)$ is in the feasible set of the regularised minimisation problem (15), because the sum of absolute values of expansions coefficients is

$$\sum_{m=1}^{M} \sum_{n:m(x_n)=n} \frac{1}{M} = \sum_{n=1}^{N} \frac{1}{N} = 1 \leq C \tag{17}$$

Therefore the regularisation error can be upper bounded as

$$\begin{aligned} \|\mu_X^{\text{out}} - \tilde{\mu}_X\|_{\mathcal{H}} &\leq \left\| \frac{1}{N} \sum_{n=1}^{N} k(z_{m(x_n)}, \cdot) - \tilde{\mu}_X \right\|_{\mathcal{H}} && \text{[property of min]} \\ &\leq \|\tilde{\mu}_X - \hat{\mu}_X\|_{\mathcal{H}} + \left\| \hat{\mu}_X - \frac{1}{N} \sum_{n=1}^{N} k(z_{m(x_n)}, \cdot) \right\|_{\mathcal{H}} && \text{[Triangle inequality]} \end{aligned}$$

The first term goes to 0 as $N \to \infty$ by the argument given in the proof of Theorem 2. The probability that the second term is larger than ε converges to 0 as $N \to \infty$ using the argument given in the proof of Lemma 10. Hence we have the desired convergence of the modified Algorithm 1's output μ_X^{out} to the true kernel mean embedding μ_X as $N \to \infty$, in probability.

This means that the modified algorithm still outputs a consistent estimator of the kernel mean embedding of μ_X. Moreover, the weights in the released finite expansion now have their L_1 norm $\sum_{m=1}^{M} |w_m|$ bounded by the constant C by construction, so Theorem 1 of (Simon-Gabriel et al., 2016) applies and gives the desired conclusion regarding consistency of the estimator for the kernel mean embedding $\mu_{f(X)}$ of $f(X)$. □

A.2. Algorithm 1 (Synthetic Data Subspace): Convergence Rates

Towards proving the convergence rate of Proposition 5, we will make use of the following Lemma 12, which is a refinement of the corresponding consistency result of Lemma 10 above. It uses the Lipschitz assumption on the kernel to establish a quantitative dependence between ε and δ, and the condition on q to establish a dependence between δ, K and $\mathbb{P}[E_M]$.

Lemma 12. *Suppose that $\mathcal{X}$ is a bounded subset of $\mathbb{R}^D$, the kernel k is Lipschitz with some Lipschitz constant $L \in \mathbb{R}^+$, and the synthetic data points $z_1, z_2, \ldots$ are sampled i.i.d. from a distribution q whose density is bounded away from 0 on any bounded subset of $\mathbb{R}^D$. Then*

$$\forall \gamma \in (0,1), a > 0 \quad \exists C \in \mathbb{R}, \varepsilon_0 > 0 \quad \forall \varepsilon \in (0, \varepsilon_0) \quad M \geq C\varepsilon^{-2D-a} \quad \Rightarrow \quad \mathbb{P}\left[\|\hat{\mu}_X - \bar{\mu}_X\|_{\mathcal{H}} \geq \varepsilon\right] \leq \gamma.$$

Proof. Let $\gamma \in (0,1)$ and $a > 0$. Suppose for the moment that C and ε_0 have already been chosen based on $\mathcal{X}, q, \gamma, a$ and based on the Lipschitz constant L of the kernel k. Let $\varepsilon \in (0, \varepsilon_0)$ and suppose that $M \geq C\varepsilon^{-2D-a}$.

Define $\delta = \frac{\varepsilon^2}{2L}$ and let $B_1, \ldots, B_K$ be a covering of $\operatorname{supp}(X)$ with K open balls of radii $\frac{\delta}{2}$. By the Lipschitz property

$$\|x - x'\|_{\mathcal{X}} < \delta \quad \Rightarrow \quad |k(x, x') - k(x, x)| \leq L\|x - x'\|_{\mathcal{X}} < L\delta = \frac{\varepsilon^2}{2}$$

and so by the argument appearing in the proof of Lemma 10, if each ball B_k contains at least one synthetic data point z_m, then $\|\hat{\mu}_X - \bar{\mu}_X\|_{\mathcal{H}} < \varepsilon$. Therefore it suffices to show that if $M \geq C\varepsilon^{-2D(1+a)}$, then the probability of some of the balls not containing any synthetic data point is at most γ.

To this end, let us look at the number of balls K, and the probability that a synthetic data point lands in a particular ball, as functions of ε (via the ball radius $\frac{\delta}{2}$). First, since $\mathcal{X}$ is a bounded subset of $\mathbb{R}^D$, there exists $C_1 \in \mathbb{R}$ such that for all $\delta > 0$, the space $\mathcal{X}$ can be covered with $\lfloor C_1 \delta^{-D} \rfloor$ open balls of radii $\delta/2$. Second, since the density of q is assumed to be bounded away from 0 on any bounded subset of $\mathbb{R}^D$, there exists $C_2 \in \mathbb{R}$ such that $q(B_k) \geq C_2 \delta^D$ for all k.

Let A_k^M be the event that the ball B_k remains without a synthetic data point after M of them have been sampled. Then the probability of the event E_M that *any* of the K balls remains empty can be upper bounded by a union bound as

$$\mathbb{P}[E_M] \leq \sum_{k=1}^{K} \mathbb{P}[A_k^M] = \sum_{k=1}^{K} (1 - q(B_k))^M \leq \sum_{k=1}^{K} (1 - C_2 \delta^D)^M \leq K \exp\left(-M C_2 \delta^D\right) \leq C_1 \delta^{-D} \exp\left(-M C_2 \delta^D\right).$$

Solving for M, we can easily verify that $\mathbb{P}[E_M] \leq \gamma$ is ensured whenever

$$M \geq \frac{1}{C_2 \delta^D} \left(D \ln \frac{1}{\delta} + \ln \frac{C_1}{\gamma} \right) = \frac{(2L)^D}{C_2} \frac{1}{\varepsilon^{2D}} \left(2 \ln \frac{1}{\varepsilon} + \ln \frac{C_1 (2L)^D}{\gamma} \right)$$

Since $\ln \frac{1}{\varepsilon} < \frac{1}{\varepsilon^a}$ for all sufficiently small ε, we see that we could have chosen $\varepsilon_0 > 0$ and $C \in \mathbb{R}$ such that the right-hand side is at most $C\varepsilon^{-2D-a}$ for all $\varepsilon \in (0, \varepsilon_0)$. But the condition $M \geq C\varepsilon^{-2D-a}$ is satisfied by supposition, and so we conclude that $\mathbb{P}\left[\|\hat{\mu}_X - \bar{\mu}_X\|_{\mathcal{H}}\right] \leq \mathbb{P}[E_M] \leq \gamma$. □

Proposition 5 Suppose that $\mathcal{X}$ is a bounded subset of $\mathbb{R}^D$, the kernel k is Lipschitz, and the synthetic data points $z_1, z_2, \dots$ are sampled i.i.d. from a distribution q whose density is bounded away from 0 on any bounded subset of $\mathbb{R}^D$. Then $M(N)$ can be chosen so that Algorithm 1 outputs an estimator that converges to the true kernel mean embedding μ_X in RKHS norm at a rate $\mathcal{O}_p(N^{-1/(D+1+c)})$, where c is any fixed positive number $c > 0$.

Proof. As in the proof of Theorem 2, we can decompose the error between the released element $\tilde{\mu}_X$ and the true μ_X as

$$\|\tilde{\mu}_X - \mu_X\|_{\mathcal{H}} \leq \underbrace{\|\hat{\mu}_X - \mu_X\|_{\mathcal{H}}}_{\text{finite sample error}} + \underbrace{\|\bar{\mu}_X - \hat{\mu}_X\|_{\mathcal{H}}}_{\text{projection error}} + \underbrace{\|\tilde{\mu}_X - \bar{\mu}_X\|_{\mathcal{H}}}_{\text{privacy error}}. \tag{18}$$

Using the standard empirical kernel mean embedding estimator, the finite sample error vanishes as $\mathcal{O}_p(N^{-1/2})$ (Muandet et al., 2016). From the proof of Theorem 2 we can see that the privacy error vanishes as $\mathcal{O}_p(\sqrt{F}/N) \subseteq \mathcal{O}_p(\sqrt{M}/N)$. Solving for ε in the statement of the preceding Lemma 12 we have that for all $\gamma \in (0, 1)$, $a > 0$ and all sufficiently large M,

$$\mathbb{P}\left[\|\hat{\mu}_X - \bar{\mu}_X\|_{\mathcal{H}} \geq \frac{1}{C} M^{-\frac{1}{2D+a}}\right] \leq \gamma.$$

The projection error thus vanishes at a rate $\mathcal{O}_p(M^{-1/(2D+a)})$, for any arbitrarily small $a > 0$. To achieve the claimed total rate $\mathcal{O}_p(N^{-1/(D+1+c)})$ we choose $M(N) = N^k$ with $k = 1 - 4/(2D + a + 2)$, and verify that

$$\mathcal{O}_p\left(\frac{1}{\sqrt{N}} + M^{\frac{-1}{2D+a}} + \frac{\sqrt{M}}{N}\right) = \mathcal{O}_p\left(\frac{1}{\sqrt{N}} + N^{\frac{-k}{2D+a}} + \frac{\sqrt{N^k}}{N}\right) = \mathcal{O}_p\left(\frac{1}{\sqrt{N}} + N^{-\frac{1}{D+1+a/2}}\right) = \mathcal{O}_p\left(N^{-\frac{1}{D+1+a/2}}\right)$$

and the claimed result follows by taking $a = 2c > 0$. □

Proposition 6 Suppose that a fixed proportion η of the private database can be published without modification. Using this part of the database as the synthetic data points, Algorithm 1 outputs a consistent estimator of μ_X that converges in RKHS norm at a rate $\mathcal{O}_p(N^{-1/2})$.

Proof. Let $\hat{\mu}^{\text{baseline}} := \frac{1}{M}\sum_{m=1}^{M} k(z_m, \cdot)$ be the baseline estimator that weights the M public points uniformly. Noting that $\hat{\mu}^{\text{baseline}} \in \mathcal{H}_M$ lies in the span of feature maps of synthetic data points, for the projection error as defined in equation (18) we have:

$$\begin{aligned}
\|\bar{\mu}_X - \hat{\mu}_X\|_{\mathcal{H}} &= \left\|\hat{\mu}^{\text{baseline}} - \hat{\mu}_X\right\|_{\mathcal{H}} && \text{[property of projection]}\\
&= \left\|\hat{\mu}^{\text{baseline}} - \mu_X\right\|_{\mathcal{H}} + \|\hat{\mu}_X - \mu_X\|_{\mathcal{H}} && \text{[Triangle inequality]}\\
&\in \mathcal{O}_p\left(M^{-1/2}\right) + \mathcal{O}_p\left(N^{-1/2}\right)
\end{aligned}$$

Using the error decomposition of equation (18) we thus have

$$\|\tilde{\mu}_X - \mu_X\|_{\mathcal{H}} \in \mathcal{O}_p\left(N^{-1/2} + (M^{-1/2} + N^{-1/2}) + \sqrt{M}/N\right)$$

and this is in $\mathcal{O}_p(N^{-1/2})$ when $M = \eta N$ is proportional to N. □

A.3. Algorithm 1 (Synthetic Data Subspace): Differential Privacy

The proof of Proposition 3 rests on the following simple calculation:

Lemma 13. *If $k(x, x) \leq 1$ for all $x \in \mathcal{X}$, then the RKHS norm sensitivity of the empirical kernel mean embedding $\hat{\mu}_X$ with respect to changing one data point is at most $\frac{2}{N}$.*

Proof. Let $D = \{x_1, \dots, x_N\}$ and $D' = \{x'_1, \dots, x'_N\}$ be two databases of the same cardinality N, differing in a single row. Without loss of generality $x_n = x'_n$ for $1 \leq n \leq N - 1$. Let $\hat{\mu}_X$ and $\hat{\mu}'_X$ be the empirical kernel mean embeddings computed using D and D', respectively. Then

$$\|\hat{\mu}_X - \hat{\mu}'_X\|_{\mathcal{H}} = \left\| \frac{1}{N}\sum_{n=1}^{N} k(x_n, \cdot) - \frac{1}{N}\sum_{n=1}^{N} k(x'_n, \cdot) \right\|_{\mathcal{H}} = \frac{1}{N} \|k(x_N, \cdot) - k(x'_N, \cdot)\|_{\mathcal{H}} \tag{19}$$

$$\leq \frac{1}{N} \left(\|k(x_N, \cdot)\|_{\mathcal{H}} + \|k(x_N, \cdot)\|_{\mathcal{H}}\right) = \frac{1}{N}\left(k(x_N, x_N)^{1/2} + k(x'_N, x'_N)^{1/2}\right) \leq \frac{2}{N}. \tag{20}$$

As D and D' were arbitrary neighbouring databases, the claimed result follows. □

Proposition 3. If $k(x, x) \leq 1$ for all $x \in \mathcal{X}$, then Algorithm 1 is (ε, δ)-differentially private.

Proof. As the synthetic data points $z_1, \dots, z_M$ do not depend on the private data, it suffices to show that the weights $w_1, \dots, w_M$ are (ε, δ)-differentially private. However, these weights result from data-independent post-processing of the coefficients $\boldsymbol{\beta}$, which are a perturbed version of the coefficients $\boldsymbol{\alpha}$, with the perturbation provided by the privacy-protecting *Gaussian mechanism* (Dwork & Roth, 2014). It remains to verify that the Gaussian mechanism employs sufficiently scaled noise; in particular we need to verify that $2/N \geq \Delta_2 := \sup_{D,D':D\sim D'} \|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_2$.

But indeed, since $b_1, \dots, b_F$ are orthonormal, for any $\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}'$ computed using neighbouring databases,

$$\|\boldsymbol{\alpha} - \boldsymbol{\alpha}'\|_2 = \left(\sum_{f=1}^{F} (\alpha_f - \alpha'_f)^2\right)^{1/2} = \left\|\overline{\hat{\mu}_N} - \overline{\hat{\mu}'_N}\right\|_{\mathcal{H}} \leq \|\hat{\mu}_N - \hat{\mu}'_N\|_{\mathcal{H}} \leq \frac{2}{N}, \tag{21}$$

(last inequality is Lemma 13) as required to verify the Gaussian mechanism. Then (ε, δ)-differential privacy for the entire algorithm follows. □

A.4. Algorithm 2 (Random Features RKHS Algorithm): Consistency

As a preliminary lemma, we first show that a uniform convergence result for the random features ϕ translates into a bound on the error incurred by Algorithm 2 due to using random features instead of the original kernel k.

Lemma 14. *Let $\hat{\mu}_X^{out} := \sum_{m=1}^{M} w_m k(z_m, \cdot) \in \mathcal{H}$ be the element in $\mathcal{H}$ represented by the output of Algorithm 2. Let $\hat{\mu}_X^{\phi,out} := \sum_{m=1}^{M} w_m \phi(z_m)$ be the corresponding element in the random features RKHS $\mathcal{H}_\phi$. If the random feature scheme ϕ is such that $\sup_{x,x'\in\mathcal{X}} |\phi(x)^T\phi(x') - k(x, x')| < \delta$, then the following bound on the "random features error" holds:*

$$\left| \left\| \hat{\mu}_X^{\phi,out} - \hat{\mu}_X^{\phi} \right\|_{\mathcal{H}_\phi} - \|\hat{\mu}_X^{out} - \hat{\mu}_X\|_{\mathcal{H}} \right| \leq 2\sqrt{\delta}.$$

Proof. Expanding the RKHS norms using bilinearity of inner products, we have

$$\begin{aligned}
&\left| \left\| \hat{\mu}_X^{\phi,\text{out}} - \hat{\mu}_X^{\phi} \right\|_{\mathcal{H}_\phi} - \left\|\hat{\mu}_X^{\text{out}} - \hat{\mu}_X\right\|_{\mathcal{H}} \right| \\
&= \Bigg| \left(\sum_{m_1=1}^{M}\sum_{m_2=1}^{M} w_{m_1} w_{m_2} \phi(z_{m_1})^T \phi(z_{m_2}) + \sum_{n_1=1}^{N}\sum_{n_2=1}^{N} \frac{1}{N}\frac{1}{N} \phi(x_{n_1})^T\phi(x_{n_2}) - 2\sum_{m=1}^{M}\sum_{n=1}^{N} w_m \frac{1}{N}\phi(z_m)^T\phi(x_n) \right)^{1/2} \\
&- \left(\sum_{m_1=1}^{M}\sum_{m_2=1}^{M} w_{m_1} w_{m_2} k(z_{m_1}, z_{m_2}) + \sum_{n_1=1}^{N}\sum_{n_2=1}^{N} \frac{1}{N}\frac{1}{N} k(x_{n_1}, x_{n_2}) - 2\sum_{m=1}^{M}\sum_{n=1}^{N} w_m \frac{1}{N} k(z_m, x_n) \right)^{1/2} \Bigg|
\end{aligned}$$

Since $\sum_{m=1}^{M} |w_m| \leq 1$ by construction and $\sum_{n=1}^{N} \frac{1}{N} = 1$, thanks to the assumption on ϕ this expression is of the form

$$\left|(a + b + 2c)^{1/2} - (A + B + 2C)^{1/2}\right|$$

for suitable $a, A, b, B, c, C \in \mathbb{R}$ with $|a - A|, |b - B|, |c - C| < \delta$. By monotonicity of the square root function, this expression is maximised when $A = a + \delta$, $B = b + \delta$, $C = c + \delta$. Writing $s := a + b + 2C$, we have

$$\left|(a + b + 2c)^{1/2} - (A + B + 2C)^{1/2}\right| \leq |s^{1/2} - (s + 4\delta)^{1/2}| = (s + 4\delta)^{1/2} - s^{1/2} \leq s^{1/2} + 2\delta^{1/2} - s^{1/2} = 2\delta^{1/2}. \quad □$$

Theorem 7 Suppose that the random features ϕ converge $\phi(\cdot)^T\phi(\cdot) \to k(\cdot,\cdot)$ uniformly in $\mathcal{X}$ as the number of random features $J \to \infty$. Assume also availability of an approximate Reduced set construction method that solves the minimisation (5) either up to a constant multiplicative error, or with an absolute error that can be made arbitrarily small. Then Algorithm 2 outputs a consistent estimator of the kernel mean embedding μ_X in the sense that

$$\sum_{m=1}^{M} w_m k(z_m,\cdot) \xrightarrow{\mathbb{P}} \mu_X \qquad \text{as } N \to \infty. \tag{22}$$

Proof. The output of Algorithm 2 specifies an element $\hat{\mu}_X^{\text{out}} := \sum_{m=1}^{M} w_m k(z_m,\cdot) \in \mathcal{H}$ in the RKHS $\mathcal{H}$ of k. Its RKHS distance to the true kernel mean embedding μ_X of X can be upper bounded by a decomposition using the Triangle inequality, where we write $\hat{\mu}_X^{\phi,\text{out}} := \sum_{m=1}^{M} w_m \phi(z_m)$ for the element of $\mathcal{H}_\phi$ that the Reduced set method constructs to approximate the privacy-protected $\tilde{\mu}_X^\phi$:

$$\begin{aligned}
\left\|\mu_X - \hat{\mu}_X^{\text{out}}\right\|_{\mathcal{H}} &\le \underbrace{\|\mu_X - \hat{\mu}_X\|_{\mathcal{H}}}_{\text{finite sample error}} + \underbrace{\left\|\hat{\mu}_X - \hat{\mu}_X^{\text{out}}\right\|_{\mathcal{H}}}_{\text{other errors}} \\
&\le \underbrace{\|\mu_X - \hat{\mu}_X\|_{\mathcal{H}}}_{\text{finite sample error}} + \underbrace{\left|\left\|\hat{\mu}_X^{\phi,\text{out}} - \hat{\mu}_X^{\phi}\right\|_{\mathcal{H}_\phi} - \left\|\hat{\mu}_N - \hat{\mu}_X^{\text{out}}\right\|_{\mathcal{H}}\right|}_{\text{random features error}} + \underbrace{\left\|\hat{\mu}_X^{\phi,\text{out}} - \hat{\mu}_X^{\phi}\right\|_{\mathcal{H}_\phi}}_{\text{other errors}} \\
&\le \underbrace{\|\mu_X - \hat{\mu}_X\|_{\mathcal{H}}}_{\text{finite sample error}} + \underbrace{\left|\left\|\hat{\mu}_X^{\phi,\text{out}} - \hat{\mu}_X^{\phi}\right\|_{\mathcal{H}_\phi} - \left\|\hat{\mu}_N - \hat{\mu}_X^{\text{out}}\right\|_{\mathcal{H}}\right|}_{\text{random features error}} \\
&\quad + \underbrace{\left\|\hat{\mu}_X^{\phi,\text{out}} - \tilde{\mu}_X^{\phi}\right\|_{\mathcal{H}_\phi}}_{\text{reduced set error}} + \underbrace{\left\|\tilde{\mu}_X^{\phi} - \hat{\mu}_X^{\phi}\right\|_{\mathcal{H}_\phi}}_{\text{privacy error}}.
\end{aligned} \tag{23}$$

The finite sample error tends to 0 as $N \to \infty$ in probability by consistency of the empirical kernel mean estimate. The random features error goes to 0 as $N \to \infty$ by Lemma 14, since $J \to \infty$ as $N \to \infty$ and $\phi(\cdot)^T\phi(\cdot) \to k(\cdot,\cdot)$ uniformly in $\mathcal{X}$ as $J \to \infty$. The privacy error goes to 0 as $N \to \infty$ by the same argument as in the proof of Theorem 2, with F replaced by J. So it remains to show that the reduced set error also goes to 0 as $N \to \infty$, in probability.

First, note that the private empirical kernel mean embedding $\hat{\mu}_X^\phi = \frac{1}{N}\sum_{n=1}^{N}\phi(x_n)$ is in the feasible set of the constrained minimisation problem solved by the reduced set method, as the sum of absolute values of weights in this expansion is $N|\frac{1}{N}| = 1 \le 1$. The RKHS $\mathcal{H}_\phi$ distance of $\hat{\mu}_X^\phi$ to the optimisation target $\tilde{\mu}_X^\phi$ equals the privacy error, so it follows that the reduced set error is upper bounded by the privacy error, and hence also goes to 0 as $N \to \infty$:

$$\underbrace{\left\|\hat{\mu}_X^{\phi,\text{out}} - \tilde{\mu}_X^{\phi}\right\|_{\mathcal{H}_\phi}}_{\text{reduced set error}} \le \underbrace{\left\|\tilde{\mu}_X^{\phi} - \hat{\mu}_X^{\phi}\right\|_{\mathcal{H}_\phi}}_{\text{privacy error}} \xrightarrow{\mathbb{P}} 0 \text{ as } N \to \infty, \tag{24}$$

as required to complete the proof. □

Corollary 15. *Let f be any continuous function. Then whenever k is a c_0-universal kernel, applying f to the points output by Algorithm 2 yields a consistent estimator of the kernel mean embedding $\mu_{f(X)}$ of $f(X)$.*

Proof. Noting that the sum of absolute values of weights w_m output by Algorithm 2 is at most C by construction, in light of Theorem 7 we see that Theorem 1 of (Simon-Gabriel et al., 2016) applies and gives the desired conclusion. □

A.5. Algorithm 2 (Random Features RKHS Algorithm): Convergence Rate

Proposition 7 Suppose that ϕ is a random feature scheme for the kernel k that converges uniformly on any compact set at a rate $\mathcal{O}_p(J^{-1/2})$ with the number J of random features. Then $J(N)$ can be chosen such that if the employed Reduced set method finds a global optimum of (5), Algorithm 2 outputs an element that converges to the true kernel mean embedding μ_X at a rate $\mathcal{O}_p(N^{-1/3})$.

Proof. Equation (23) shows that the error $\|\mu_X - \hat{\mu}_X^{\text{out}}\|_{\mathcal{H}}$ between the released element $\hat{\mu}_X^{\text{out}}$ and the true kernel mean embedding μ_X can be upper bounded by the sum of four terms: the finite sample error, the random features error, the reduced set error, and the privacy error. Arguing as in the proof of Proposition 5, the finite sample error vanishes at a rate $\mathcal{O}_p(N^{-1/2})$. The proof of Theorem 7 shows that the reduced set error is upper bounded by the privacy error, which itself vanishes at a rate of $\mathcal{O}_p(\sqrt{J}/N)$ by the argument given in the proof of Theorem 2, with F replaced by J. Lemma 14 implies that if the random features converge uniformly at a rate $\mathcal{O}_p(J^{-1/2})$, then the random features error vanishes at a rate $\mathcal{O}_p(J^{-1/4})$. The total convergence rate is thus

$$\mathcal{O}_p\left(N^{-1/2} + \frac{\sqrt{J}}{N} + J^{-1/4}\right)$$

and we can check that this becomes $\mathcal{O}_p(N^{-1/3})$ by setting $J = \lfloor N^{4/3} \rfloor$. □

A.6. Algorithm 2 (Random Features RKHS Algorithm): Differential Privacy

Proposition 9 Assume that the random feature vectors produced by ϕ are bounded by 1 in L_2 norm ($\|\phi(x)\|_2 \leq 1$ for all $x \in \mathcal{X}$). Then Algorithm 2 is (ε, δ)-differentially private.

Proof. The output of the algorithm is produced by a Reduced set method that is initialised blindly to the database and optimises RKHS distance to the element $\tilde{\mu}_X^{\phi} \in \mathcal{H}_{\phi}$, while only having access to the distance to it, rather than any representation of $\tilde{\mu}_X^{\phi}$. As $\tilde{\mu}_X^{\phi}$ can be seen as a vector in $\mathbb{R}^J$ obtained by perturbing $\hat{\mu}_X^{\phi}$ using the Gaussian mechanism with $\Delta_2 = \frac{2}{N}$, it suffices to show that the L_2-sensitivity of $\hat{\mu}_X^{\phi}$ is upper bounded by $\frac{2}{N}$. To this end, assume $D = \{x_1, \ldots, x_N\}$ and $D' = \{x'_1, \ldots, x'_N\}$ are two neighbouring databases of cardinality N, differing w.l.o.g. in their last element only. Then

$$\|\hat{\mu}_D^{\phi} - \hat{\mu}_{D'}^{\phi}\|_2 = \left\| \frac{1}{N}\sum_{n=1}^{N} \phi(x_n) - \frac{1}{N}\sum_{n=1}^{N} \phi(x'_n) \right\|_2 \tag{25}$$

$$= \frac{1}{N}\|\phi(x_N) - \phi(x'_N)\|_2 \tag{26}$$

$$\leq \frac{1}{N}\|\phi(x_N)\|_2 + \frac{1}{N}\|\phi(x'_N)\|_2 \leq \frac{2}{N}, \tag{27}$$

as required to complete the proof. □

B. Setup of Empirical Illustrations

We considered two scenarios in our basic empirical evaluations shown in Sections 4 and 5:

1. *No publishable subset*: No rows of the private database are, or can be made public without some privacy-ensuring modification.
2. *Publishable subset*: A small part of the private database is already public, or can be made public, perhaps for one of the several possible reasons outlined in Section 1.

To illustrate the impact of data dimensionality on the performance of the proposed algorithms, we provide results on datasets with data dimension $D = 2$ and $D = 5$. In both cases we constructed a synthetic private dataset by sampling $N = 100,000$ data points from a multivariate Gaussian mixture distribution. The mixture had 10 components, with mixing weights proportional to $1, \frac{1}{2}, \ldots, \frac{1}{10}$, and the means of the components were chosen randomly themselves from a spherical Gaussian distribution with mean $[100, \ldots, 100]$ and covariance $200 I_D$. Each of the N private data points was simulated by first sampling its mixture component using the mixing weights as probabilities, and then the point itself was sampled from a spherical Gaussian centered at the mean of the chosen mixture component and with covariance $30 I_D$.

We chose to work with the widely popular exponentiated quadratic kernel $k(x_1, x_2) = e^{-\gamma\|x_1 - x_2\|_2^2}$ for $\mathbb{R}^D$-valued data (also known as a Gaussian kernel, or a squared exponential kernel), with the parameter setting $\gamma = 10^{-4}/D$. This kernel is known to be *characteristic* (Fukumizu et al., 2008), and so as discussed in Section 2.2, no information about the data generating distribution p_X is lost by working with its kernel mean embedding μ_X.

We used our proposed algorithms to release an approximate version of the empirical KME of the private database, in such a way that the output satisfies the definition of (ε, δ)-differential privacy. We investigated the common privacy levels given by $\varepsilon \in \{0.01, 0.1, 1.0\}$, and used the fixed value of $\delta = 10^{-6}$, which satisfies the usual requirement that $\delta \ll \frac{1}{N}$.

B.1. Evaluation Metric

The geometry of the RKHS $\mathcal{H}$ allows comparing the performance of different algorithms by computing the RKHS distance Δ between the empirical KME $\hat{\mu}_X$ computed using all N private data points (and which could not have been released without violating differential privacy) and the element of the RKHS represented by the actually released weighted set of synthetic data points $(z_1, w_1), \ldots, (z_M, w_M)$:

$$\Delta := \left\| \hat{\mu}_X - \sum_{m=1}^{M} w_m k(z_m, \cdot) \right\|_{\mathcal{H}} .$$

Moreover, as the empirical KME $\hat{\mu}_X$ is based on a large sample size of $N = 100,000$ i.i.d. data points, it can be expected to be a good proxy for the true KME μ_X of the data-generating random variable X. In that case Δ is also a good proxy for the RKHS distance between the true KME μ_X and the RKHS element represented by the released dataset.

B.2. Scenario 1: No Publishable Subset

Algorithm 1 requires specifying the synthetic data points $z_1, \ldots, z_M$ in advance, before seeing the private data. If no part of the private data has already been published (which could then be used for the synthetic data points), one can construct the synthetic data points by sampling them randomly from a suitable probability distribution q. For the consistency result of Theorem 2 to apply, the support of q must include all possible private data points. In our case the private data takes values in $\mathbb{R}^D$, and so this requirement is satisfied by any distribution on $\mathbb{R}^D$ with full support. We used a spherical Gaussian distribution $q = \mathcal{N}(0, \sigma_q I_D)$ with $\sigma_q = 500$ for sampling the synthetic data points.

The implementation of Algorithm 2 used $J = 10,000$ random features for accurate approximation of the kernel, and an iterative gradient-based optimisation procedure to solve the reduced set problem (Equation (5) in Algorithm 2).

Figure 2 shows how the RKHS distance Δ changes with the number of synthetic data points M, for different requested privacy level ε for Algorithm 1 (solid lines) and Algorithm 2 (dashed lines), on datasets with dimensionality $D = 2$ (left subfigure) and $D = 5$ (right subfigure). We observe that the additional ability of Algorithm 2 to optimise the *locations* of the synthetic data points (rather than just the weights, as is the case for Algorithm 1) is more helpful in the higher-dimensional case $D = 5$, where the randomly sampled synthetic data points are less likely to land close to private data points.

B.3. Scenario 2: Publishable Subset

Here we explored the interesting scenario where one can exploit the fact that a small part of the private database is actually public, and use the public rows as the fixed synthetic data points in Algorithm 1. Specifically, we assume (without loss of generality) that the first M rows of the private database (where $M \ll N$) are public, and we take the synthetic data points to be $z_1 \leftarrow x_1, \ldots, z_M \leftarrow x_M$.

Observe that in this case $\hat{\mu}^{\text{baseline}} := \frac{1}{M} \sum_{m=1}^{M} k(z_m, \cdot)$, i.e., uniform weighting of the synthetic data points, is already expected to be a strong baseline since $\hat{\mu}^{\text{baseline}}$ is itself a consistent estimator of μ_X, (although based on a much smaller sample size $M \ll N$). The purpose of Algorithm 1 is to find (generally non-uniform) $w_1, \ldots, w_M$ that reweight the public data points using the information in the large private dataset, but respecting differential privacy. Figure 1 shows how the RKHS distance Δ changes with the number of public data points M, for different privacy levels ε.

For comparison, the figures also show the RKHS distances Δ achieved by the baseline that simply weights the public points uniformly. We can see that in both cases $D = 2$ and $D = 5$, if the ratio of public to private points is low enough, Algorithm 1 provides a substantial benefit over this baseline (note the logarithmic scaling). Since usually obtaining permission to publish a larger subset of the private data unchanged will come at an increased cost, the ability to instead reweight a smaller public dataset using Algorithm 1 in a differentially private manner is useful.