Supplementary Material

A Proofs

We will conduct our analysis in terms of general noise covariance Σ for the added noise, $\mathbf{n} \sim \mathcal{N}(0, \Sigma)$. The results will depend on various norms of Σ, as well as $\|\Sigma^{1/2}\mathbf{a}\|$, where $\mathbf{a} = \boldsymbol{\mu}_\phi(\mathcal{D}) - \boldsymbol{\mu}_\phi(\tilde{\mathcal{D}})$ is the difference between empirical mean embeddings $\boldsymbol{\mu}_\phi(\mathcal{D}) = \frac{1}{|\mathcal{D}|}\sum_{\mathbf{x}\in\mathcal{D}} \phi(\mathbf{x})$. (Recall that $\mathrm{MMD}(\mathcal{D}, \tilde{\mathcal{D}}) = \|\mathbf{a}\|$.)

When we use only normalized first-moment features, the quantities appearing in the bounds are

$$\begin{gathered} \Sigma = \frac{4\sigma^2}{m^2} I_D \\ \|\Sigma\|_{op} = \frac{4\sigma^2}{m^2} \qquad \|\Sigma\|_F = \frac{4\sigma^2}{m^2}\sqrt{D} \qquad \mathrm{Tr}(\Sigma) = \frac{4\sigma^2}{m^2} D \\ \|\Sigma^{1/2}\mathbf{a}\|_2 = \sqrt{\mathbf{a}^\top \Sigma \mathbf{a}} = \frac{2\sigma}{m}\,\mathrm{MMD}_{k_{\boldsymbol{\phi}}}(\mathcal{D}, \tilde{\mathcal{D}}). \end{gathered} \tag{8}$$

When we use first- and second-moment features with respective scales C_1 and C_2 (both 1 in our experiments here), we have

$$\begin{gathered} \Sigma = \begin{bmatrix} \sigma^2 \left(\frac{2C_1}{m}\right)^2 I_D & 0 \\ 0 & \sigma^2 \left(\frac{2C_2}{m}\right)^2 I_D \end{bmatrix} = \frac{4\sigma^2}{m^2} \begin{bmatrix} C_1^2 I_D & 0 \\ 0 & C_2^2 I_D \end{bmatrix} \\ \|\Sigma\|_{op} = \frac{4\sigma^2}{m^2}\max(C_1^2, C_2^2) \quad \|\Sigma\|_F = \frac{4\sigma^2}{m^2}(C_1^2 + C_2^2)\sqrt{D} \quad \mathrm{Tr}(\Sigma) = \frac{4\sigma^2}{m^2}(C_1^2 + C_2^2)\, D \\ \|\Sigma^{1/2}\mathbf{a}\|_2 = \sqrt{\mathbf{a}^\top \Sigma \mathbf{a}} = \frac{2\sigma}{m}\sqrt{C_1^2\,\mathrm{MMD}\, k_{\boldsymbol{\phi}_1}(\mathcal{D}, \tilde{\mathcal{D}})^2 + C_2^2\,\mathrm{MMD}\, k_{\boldsymbol{\phi}_2}(\mathcal{D}, \tilde{\mathcal{D}})^2}. \end{gathered} \tag{9}$$

Note that if $C_1 = C_2 = C$, then

$$\sqrt{C_1^2\,\mathrm{MMD}\, k_{\boldsymbol{\phi}_1}(\mathcal{D}, \tilde{\mathcal{D}})^2 + C_2^2\,\mathrm{MMD}\, k_{\boldsymbol{\phi}_2}(\mathcal{D}, \tilde{\mathcal{D}})^2} = C\,\mathrm{MMD}_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}).$$

A.1 Mean absolute error of loss function

Proposition A.1. *Given datasets $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^m$ and $\tilde{\mathcal{D}} = \{\tilde{\mathbf{x}}_j\}_{j=1}^n$ and a kernel k_ϕ with a D-dimensional embedding ϕ, let $\mathbf{a} = \mu_\phi(\mathcal{D}) - \mu_\phi(\tilde{\mathcal{D}})$. Define $\widetilde{\mathrm{MMD}}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}) = \|\mathbf{a} + \mathbf{n}\|^2$ for a noise vector $\mathbf{n} \sim \mathcal{N}(0, \Sigma)$. Introducing the noise $\mathbf{n}$ affects the expected absolute error as*

$$\mathbb{E}_{\mathbf{n}}\left[\left|\widetilde{\mathrm{MMD}}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}) - \mathrm{MMD}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}})\right|\right] \leq \mathrm{Tr}(\Sigma) + 2\sqrt{\frac{2}{\pi}}\|\Sigma^{1/2}\mathbf{a}\|. \tag{10}$$

Proof. We have that

$$\begin{aligned} &\mathbb{E}_{\mathbf{n}}\left[\left|\,\widetilde{\mathrm{MMD}}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}) - \mathrm{MMD}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}})\right|\right] \\ &\qquad = \mathbb{E}_{\mathbf{n}}\left[\left|\,\|\mathbf{a} + \mathbf{n}\|^2 - \|\mathbf{a}\|^2\,\right|\right] = \mathbb{E}_{\mathbf{n}}\left[\left|\mathbf{n}^\top\mathbf{n} + 2\mathbf{n}^\top\mathbf{a}\right|\right] \leq \mathbb{E}_{\mathbf{n}}\left[\mathbf{n}^\top\mathbf{n}\right] + 2\,\mathbb{E}_{\mathbf{n}}\left[\left|\mathbf{n}^\top\mathbf{a}\right|\right]. \end{aligned} \tag{11}$$

The first term is standard:

$$\mathbb{E}\,\mathbf{n}^\top\mathbf{n} = \mathbb{E}\,\mathrm{Tr}(\mathbf{n}^\top\mathbf{n}) = \mathbb{E}\,\mathrm{Tr}(\mathbf{n}\mathbf{n}^\top) = \mathrm{Tr}(\mathbb{E}\,\mathbf{n}\mathbf{n}^\top) = \mathrm{Tr}(\Sigma).$$

For the second, note that

$$\mathbf{a}^\top\mathbf{n} \sim \mathcal{N}(0, \mathbf{a}^\top\Sigma\mathbf{a}),$$

and so its absolute value is $\sqrt{\mathbf{a}^\top\Sigma\mathbf{a}}$ times a $\chi(1)$ random variable. Since the mean of a $\chi(1)$ distribution is $\frac{\sqrt{2}\,\Gamma(1)}{\Gamma(1/2)} = \sqrt{\frac{2}{\pi}}$, we obtain the desired bound. □

A.2 High-probability bound on the error

Proposition A.2. *Given datasets $\mathcal{D} = \{\mathbf{x}_i\}_{i=1}^m$ and $\tilde{\mathcal{D}} = \{\tilde{\mathbf{x}}_j\}_{j=1}^n$, let $\mathbf{a} = \mu_\phi(\mathcal{D}) - \mu_\phi(\tilde{\mathcal{D}})$, and define $\widetilde{\mathrm{MMD}}_{k_\Phi}^2(\mathcal{D}, \tilde{\mathcal{D}}) = \|\mathbf{a} + \mathbf{n}\|^2$ for a noise vector $\mathbf{n} \sim \mathcal{N}(0, \Sigma)$. Then for any $\rho \in (0,1)$, it holds with probability at least $1-\rho$ over the choice of $\mathbf{n}$ that*

$$\begin{aligned}&\big|\widetilde{\mathrm{MMD}}_{k_\Phi}^2(\mathcal{D}, \tilde{\mathcal{D}}) - \mathrm{MMD}_{k_\Phi}^2(\mathcal{D}, \tilde{\mathcal{D}})\big| \\ &\qquad\le \mathrm{Tr}(\Sigma) + \sqrt{\tfrac{2}{\pi}}\|\Sigma^{\frac{1}{2}}\mathbf{a}\|_2 + 2\left(\|\Sigma\|_F + \sqrt{2}\|\Sigma^{\frac{1}{2}}\mathbf{a}\|_2\right)\sqrt{\log(\tfrac{2}{\rho})} + 2\|\Sigma\|_{op}\log(\tfrac{2}{\rho}). \quad (12)\end{aligned}$$

This implies that

$$\big|\widetilde{\mathrm{MMD}}_{k_\Phi}^2(\mathcal{D}, \tilde{\mathcal{D}}) - \mathrm{MMD}_{k_\Phi}^2(\mathcal{D}, \tilde{\mathcal{D}})\big| = \mathcal{O}_P\left(\mathrm{Tr}(\Sigma) + \|\Sigma^{1/2}\mathbf{a}\|_2\right).$$

Proof. Introduce $\mathbf{z} \sim \mathcal{N}(0, I)$ such that $\mathbf{n} = \Sigma^{\frac{1}{2}}\mathbf{z}$ into Equation 11:

$$\big|\widetilde{\mathrm{MMD}}_{k_\Phi}^2(\mathcal{D}, \tilde{\mathcal{D}}) - \mathrm{MMD}_{k_\Phi}^2(\mathcal{D}, \tilde{\mathcal{D}})\big| \le \mathbf{n}^\top\mathbf{n} + 2\left|\mathbf{n}^\top\mathbf{a}\right| = \mathbf{z}^\top\Sigma\mathbf{z} + 2\left|\mathbf{a}^\top\Sigma^{1/2}\mathbf{z}\right|. \quad (13)$$

For the first term, denoting the eigendecomposition of Σ as $\mathbf{Q}\mathbf{\Lambda}\mathbf{Q}^\top$, we can write

$$\mathbf{z}^\top\Sigma\mathbf{z} = (\mathbf{Q}^\top\mathbf{z})^\top\mathbf{\Lambda}(\mathbf{Q}^\top\mathbf{z}),$$

in which $\mathbf{Q}^\top\mathbf{z} \sim \mathcal{N}(0, I)$ and $\mathbf{\Lambda}$ is diagonal. Thus, applying Lemma 1 of Laurent & Massart (2000), we obtain that with probability at least $1 - \frac{\rho}{2}$,

$$\mathbf{z}^\top\Sigma\mathbf{z} \le \mathrm{Tr}(\Sigma) + 2\|\Sigma\|_F\sqrt{\log(\tfrac{2}{\rho})} + 2\|\Sigma\|_{op}\log(\tfrac{2}{\rho}). \quad (14)$$

In the second term, $\left|\mathbf{a}^\top\Sigma^{\frac{1}{2}}\mathbf{z}\right|$, can be viewed as a function of a standard normal variable $\mathbf{z}$ with Lipschitz constant at most $\|\Sigma^{\frac{1}{2}}\mathbf{a}\|_2$. Thus, applying the standard Gaussian Lipschitz concentration inequality (Boucheron et al., 2013, Theorem 5.6), we obtain that with probability at least $1 - \frac{\rho}{2}$,

$$\left|\mathbf{z}^\top\Sigma^{\frac{1}{2}}\mathbf{a}\right| \le \mathbb{E}\left|\mathbf{z}^\top\Sigma^{\frac{1}{2}}\mathbf{a}\right| + \|\Sigma^{\frac{1}{2}}\mathbf{a}\|_2\sqrt{2\log(\tfrac{2}{\rho})} = \|\Sigma^{\frac{1}{2}}\mathbf{a}\|_2\left(\sqrt{\tfrac{2}{\pi}} + \sqrt{2\log(\tfrac{2}{\rho})}\right).$$

The first statement in the theorem follows by a union bound. The $\mathcal{O}_P$ form follows by Lemma A.1 and the fact that $\mathrm{Tr}(A) \ge \|A\|_F \ge \|A\|_{op}$ for positive semi-definite matrices A. □

The following lemma shows how to convert high-probability bounds with both sub-exponential and sub-Gaussian tails into a $\mathcal{O}_P$ statement.

Lemma A.1. *If a sequence of random variables X_n satisfies*

$$X_n \le A_n + B_n\sqrt{\log\frac{b_n}{\rho}} + C_n\log\frac{c_n}{\rho} \qquad \textit{with probability at least } 1-\rho,$$

then the sequence of variables X_n is

$$\mathcal{O}_P\left(\max\left(A_n, B_n\max(\sqrt{\log b_n}, 1), C_n\max(\log c_n, 1)\right)\right).$$

Proof. The definition of a sequence of random variables X_n being $\mathcal{O}_P(Q_n)$, where Q_n is a sequence of scalars, means that the sequence $\frac{X_n}{Q_n}$ is stochastically bounded: for each ρ, there is some constant R_ρ such that $\Pr(X_n/Q_n \ge R_\rho) \le \rho$.

Here, we have for all n with probability at least $1 - \rho$ that

$$\begin{aligned}
&\frac{X_n}{\max\left(A_n, B_n \max(\sqrt{\log b_n}, 1), C_n \max(\log c_n, 1)\right)} \\
&\qquad \leq \frac{A_n + B_n \sqrt{\log \frac{b_n}{\rho}} + C_n \log \frac{c_n}{\rho}}{\max\left(A_n, B_n \max(\sqrt{\log b_n}, 1), C_n \max(\log c_n, 1)\right)} \\
&\qquad = \frac{A_n + B_n \sqrt{\log b_n + \log \frac{1}{\rho}} + C_n \left[\log c_n + \log \frac{1}{\rho}\right]}{\max\left(A_n, B_n \max(\sqrt{\log b_n}, 1), C_n \max(\log c_n, 1)\right)} \\
&\qquad \leq \frac{A_n + B_n \sqrt{\log b_n} + B_n \sqrt{\log \frac{1}{\rho}} + C_n \log c_n + C_n \log \frac{1}{\rho}}{\max\left(A_n, B_n \max(\sqrt{\log b_n}, 1), C_n \max(\log c_n, 1)\right)} \\
&\qquad \leq 1 + 1 + \sqrt{\log \frac{1}{\rho}} + 1 + \log \frac{1}{\rho}.
\end{aligned}$$

Thus the desired bound holds with $R_\rho = 3 + \sqrt{\log \frac{1}{\rho}} + \log \frac{1}{\rho}$. □

A.3 Quality of the private minimizer: worst-case analysis

We first show uniform convergence of the privatized MMD to the non-private MMD.

Proposition A.3. *Suppose that $\Phi : \mathcal{X} \to \mathbb{R}^D$ is such that $\sup_x \|\Phi(x)\| \leq B$, and let $\widetilde{\mathrm{MMD}}_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}) = \|\mu_\Phi(\mathcal{D}) - \mu_\Phi(\tilde{\mathcal{D}}) + \mathbf{n}\|$ for $\mathbf{n} \sim \mathcal{N}(0, \Sigma)$. Then, with probability at least $1 - \rho$ over the choice of $\mathbf{n}$,*

$$\begin{aligned}
&\sup_{\mathcal{D}, \tilde{\mathcal{D}}} \left| \widetilde{\mathrm{MMD}}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}) - \mathrm{MMD}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}) \right| \\
&\qquad \leq \mathrm{Tr}(\Sigma) + 4B\sqrt{\mathrm{Tr}(\Sigma)} + 2\left(\|\Sigma\|_F + 2B\|\Sigma\|_{op}^{\frac{1}{2}}\right)\sqrt{\log(\tfrac{2}{\rho})} + 2\|\Sigma\|_{op} \log(\tfrac{2}{\rho}) \\
&\qquad = \mathcal{O}_P\left(\mathrm{Tr}(\Sigma) + B\sqrt{\mathrm{Tr}(\Sigma)}\right),
\end{aligned}$$

where the supremum is taken over all probability distributions, including the empirical distribution of datasets $\mathcal{D}, \tilde{\mathcal{D}}$ of any size.

Proof. Introducing $\mathbf{z} \sim \mathcal{N}(0, I_D)$ such that $\mathbf{n} = \Sigma^{1/2}\mathbf{z}$, we have that

$$\begin{aligned}
\sup_{\mathcal{D}, \tilde{\mathcal{D}}} \left| \widetilde{\mathrm{MMD}}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}) - \mathrm{MMD}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}) \right| &\leq \sup_{\mathcal{D}, \tilde{\mathcal{D}}} \mathbf{z}^\top \Sigma \mathbf{z} + 2\left|\mathbf{a}^\top \Sigma^{1/2}\mathbf{z}\right| \\
&\leq \mathbf{z}^\top \Sigma \mathbf{z} + 2 \sup_{\mathbf{a} : \|\mathbf{a}\| \leq 2B} \left|\mathbf{a}^\top \Sigma^{1/2}\mathbf{z}\right| \\
&\leq \mathbf{z}^\top \Sigma \mathbf{z} + 2 \sup_{\mathbf{a} : \|\mathbf{a}\| \leq 2B} \|\mathbf{a}\| \|\Sigma^{1/2}\mathbf{z}\| \\
&= \mathbf{z}^\top \Sigma \mathbf{z} + 4B\|\Sigma^{1/2}\mathbf{z}\|.
\end{aligned}$$

To apply Gaussian Lipschitz concentration, we also need to know that

$$\mathbb{E}\|\Sigma^{1/2}\mathbf{z}\| \leq \sqrt{\mathbb{E}\|\Sigma^{1/2}\mathbf{z}\|^2} = \sqrt{\mathrm{Tr}(\Sigma)};$$

the exact expectation of a χ variable with more than one degree of freedom is inconvenient, but the gap is generally not asymptotically significant. Then we get that, with probability at least $1 - \frac{\rho}{2}$,

$$\|\Sigma^{1/2}\mathbf{z}\| \leq \sqrt{\mathrm{Tr}(\Sigma)} + \|\Sigma\|_{op}^{1/2}\sqrt{2 \log \tfrac{2}{\rho}}.$$

Again combining with the bound of Equation 14, we get the stated bound. □

This bound is looser than in Proposition A.2, since the term depending on $\mathbf{a}$ is now "looking at" $\mathbf{z}$ in many directions rather than just one: we end up with a $\chi(\dim(\Sigma)$ random variable instead of $\chi(1)$.

We can use this uniform convergence bound to show that the minimizer of the private loss approximately minimizes the non-private loss:

Proposition A.4. *Fix a target dataset $\mathcal{D}$. For each $\boldsymbol{\theta}$ in some set Θ, fix a corresponding $\tilde{\mathcal{D}}_{\boldsymbol{\theta}}$; in particular, $\Theta = \mathbb{R}^p$ could be the set of all generator parameters, and $\tilde{\mathcal{D}}_{\boldsymbol{\theta}}$ either the outcome of running a generator $g_{\boldsymbol{\theta}}$ on a fixed set of "seeds," $\tilde{\mathcal{D}}_{\boldsymbol{\theta}} = \{g_{\boldsymbol{\theta}}(\mathbf{z}_i)\}_{i=1}^n$, or the full output distribution of the generator $Q_{g_{\boldsymbol{\theta}}}$. Suppose that $\Phi : \mathcal{X} \to \mathbb{R}^D$ is such that $\sup_x \|\Phi(x)\| \le B$, and let $\widetilde{\mathrm{MMD}}_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}) = \|\mu_\Phi(\mathcal{D}) - \mu_\Phi(\tilde{\mathcal{D}}) + \mathbf{n}\|$ for $\mathbf{n} \sim \mathcal{N}(0, \Sigma)$. Let $\widetilde{\boldsymbol{\theta}} \in \arg\min_{\theta\in\Theta} \widetilde{\mathrm{MMD}}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_\theta)$ be the private minimizer, and $\widehat{\boldsymbol{\theta}} \in \arg\min_{\theta\in\Theta} \widetilde{\mathrm{MMD}}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_\theta)$ the non-private minimizer. For any $\rho \in (0, 1)$, with probability at least $1 - \rho$ over the choice of $\mathbf{n}$,*

$$\begin{aligned}
&\mathrm{MMD}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_{\widetilde{\boldsymbol{\theta}}}) - \mathrm{MMD}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_{\widehat{\boldsymbol{\theta}}}) \\
&\qquad \le 2\mathrm{Tr}(\Sigma) + 8B\sqrt{\mathrm{Tr}(\Sigma)} + 4\left(\|\Sigma\|_F + 2B\|\Sigma\|_{op}^{\frac{1}{2}}\right)\sqrt{\log(\tfrac{2}{\rho})} + 4\|\Sigma\|_{op}\log(\tfrac{2}{\rho}) \\
&\qquad = \mathcal{O}_P\left(\mathrm{Tr}(\Sigma) + B\sqrt{\mathrm{Tr}(\Sigma)}\right).
\end{aligned}$$

Proof. Let α represent the uniform error bound of Proposition A.2. Applying Proposition A.2, the definition of $\widetilde{\boldsymbol{\theta}}$, then Proposition A.2 again:

$$\mathrm{MMD}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_{\widetilde{\boldsymbol{\theta}}}) \le \widetilde{\mathrm{MMD}}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_{\widetilde{\boldsymbol{\theta}}}) + \alpha \le \widetilde{\mathrm{MMD}}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_{\widehat{\boldsymbol{\theta}}}) + \alpha \le \mathrm{MMD}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_{\widehat{\boldsymbol{\theta}}}) + 2\alpha. \qquad \square$$

A.4 Quality of the private minimizer: "optimistic" analysis

The preceding analysis is quite "worst-case," since we upper-bounded the MMD by the maximum possible value everywhere. Noticing that the approximation in Proposition A.2 is tighter when $\|\Sigma^{1/2}\mathbf{a}\|$ is smaller, we can instead show an "optimistic" rate which takes advantage of this fact to show tighter approximation for the minimizer of the noised loss. In the "interpolating" case where the generator can achieve zero empirical MMD, the convergence rate substantially improves (generally improving the squared MMD from $\mathcal{O}_P(1/m)$ to $\mathcal{O}_P(1/m^2)$).

Proposition A.5. *In the setup of Proposition A.4, we have with probability at least $1 - \rho$ over $\mathbf{n}$ that*

$$\begin{aligned}
&\mathrm{MMD}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_{\widetilde{\boldsymbol{\theta}}}) - \mathrm{MMD}^2_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_{\widehat{\boldsymbol{\theta}}}) \\
&\quad \le 9\mathrm{Tr}(\Sigma) + 4\sqrt{\mathrm{Tr}(\Sigma)}\,\mathrm{MMD}_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_{\widehat{\boldsymbol{\theta}}}) \\
&\qquad + 2\left(9\|\Sigma\|_F + 2\sqrt{2\|\Sigma\|_{op}}\,\mathrm{MMD}_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_{\widehat{\boldsymbol{\theta}}})\right)\sqrt{\log\frac{2}{\rho}} + 18\|\Sigma\|_{op}\log\frac{2}{\rho} \\
&\quad = \mathcal{O}_P\left(\mathrm{Tr}(\Sigma) + \sqrt{\mathrm{Tr}(\Sigma)}\,\mathrm{MMD}_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_{\widehat{\boldsymbol{\theta}}})\right).
\end{aligned}$$

Proof. Let's use $\widehat{\mathrm{MMD}}(\boldsymbol{\theta})$ to denote $\mathrm{MMD}_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_{\boldsymbol{\theta}})$, and $\widetilde{\mathrm{MMD}}(\boldsymbol{\theta})$ for $\widetilde{\mathrm{MMD}}_{k_\Phi}(\mathcal{D}, \tilde{\mathcal{D}}_{\boldsymbol{\theta}})$.

For all $\boldsymbol{\theta}$, we have that

$$\begin{aligned}
\left|\widetilde{\mathrm{MMD}}^2(\boldsymbol{\theta}) - \widehat{\mathrm{MMD}}^2(\boldsymbol{\theta})\right| &\le \mathbf{z}^\top\Sigma\mathbf{z} + 2\left|(\mu^\Phi(\mathcal{D}) - \mu^\Phi(\tilde{\mathcal{D}}))^\top\Sigma^{1/2}\mathbf{z}\right| \\
&\le \mathbf{z}^\top\Sigma\mathbf{z} + 2\,\widehat{\mathrm{MMD}}(\boldsymbol{\theta})\|\Sigma^{1/2}\mathbf{z}\|.
\end{aligned}$$

Thus, applying this inequality in both the first and third lines,

$$\begin{aligned}
\widehat{\mathrm{MMD}}^2(\widetilde{\boldsymbol{\theta}}) &\le \widetilde{\mathrm{MMD}}^2(\widetilde{\boldsymbol{\theta}}) + \mathbf{z}^\top\Sigma\mathbf{z} + 2\,\widehat{\mathrm{MMD}}(\widetilde{\boldsymbol{\theta}})\|\Sigma^{1/2}\mathbf{z}\| \\
&\le \widetilde{\mathrm{MMD}}^2(\widehat{\boldsymbol{\theta}}) + \mathbf{z}^\top\Sigma\mathbf{z} + 2\,\widehat{\mathrm{MMD}}(\widetilde{\boldsymbol{\theta}})\|\Sigma^{1/2}\mathbf{z}\| \\
&\le \widehat{\mathrm{MMD}}^2(\widehat{\boldsymbol{\theta}}) + 2\mathbf{z}^\top\Sigma\mathbf{z} + 2\left(\widehat{\mathrm{MMD}}(\widetilde{\boldsymbol{\theta}}) + \widehat{\mathrm{MMD}}(\widehat{\boldsymbol{\theta}})\right)\|\Sigma^{1/2}\mathbf{z}\|;
\end{aligned}$$

in the second line we used that $\widetilde{\mathrm{MMD}}(\widetilde{\boldsymbol{\theta}}) \leq \widetilde{\mathrm{MMD}}(\widehat{\boldsymbol{\theta}})$. Rearranging, we get that

$$\widehat{\mathrm{MMD}}^2(\widetilde{\boldsymbol{\theta}}) - \beta\,\widehat{\mathrm{MMD}}(\widetilde{\boldsymbol{\theta}}) - \gamma \leq 0, \tag{15}$$

where

$$\begin{gathered}
\beta = 2\|\Sigma^{1/2}\mathbf{z}\| \geq 0 \\
\gamma = \widehat{\mathrm{MMD}}^2(\widehat{\boldsymbol{\theta}}) + 2\mathbf{z}^\top\Sigma\mathbf{z} + 2\,\widehat{\mathrm{MMD}}(\widehat{\boldsymbol{\theta}})\|\Sigma^{1/2}\mathbf{z}\| \geq 0.
\end{gathered}$$

The left-hand side of Equation 15 is a quadratic in $\widehat{\mathrm{MMD}}(\tilde{\boldsymbol{\theta}})$ with positive curvature; it has two roots, at

$$\frac{\beta}{2} \pm \sqrt{\left(\frac{\beta}{2}\right)^2 + \gamma}.$$

Thus the inequality Equation 15 can only hold in between the roots; the root with a minus sign is negative, and so does not concern us since we know that $\widehat{\mathrm{MMD}}(\boldsymbol{\theta}) \geq 0$. Thus, for Equation 15 to hold, we must have

$$\begin{aligned}
\widehat{\mathrm{MMD}}(\widetilde{\boldsymbol{\theta}}) &\leq \tfrac{\beta}{2} + \sqrt{\left(\tfrac{\beta}{2}\right)^2 + \gamma} \\
\widehat{\mathrm{MMD}}^2(\widetilde{\boldsymbol{\theta}}) &\leq \tfrac{\beta^2}{4} + \left(\tfrac{\beta}{2}\right)^2 + \gamma + \beta\sqrt{\left(\tfrac{\beta}{2}\right)^2 + \gamma} \\
&\leq \gamma + \beta^2 + \beta\sqrt{\gamma}.
\end{aligned}$$

Also note that

$$\gamma = \widehat{\mathrm{MMD}}^2(\widehat{\boldsymbol{\theta}}) + 2\mathbf{z}^\top\Sigma\mathbf{z} + 2\,\widehat{\mathrm{MMD}}(\widehat{\boldsymbol{\theta}})\|\Sigma^{1/2}\mathbf{z}\| \leq \left(\widehat{\mathrm{MMD}}(\widehat{\boldsymbol{\theta}}) + \sqrt{2}\|\Sigma^{1/2}\mathbf{z}\|\right)^2.$$

Thus, substituting in for β and γ then simplifying, we have that

$$\widehat{\mathrm{MMD}}^2(\widetilde{\boldsymbol{\theta}}) \leq \widehat{\mathrm{MMD}}^2(\widehat{\boldsymbol{\theta}}) + (6 + 2\sqrt{2})\mathbf{z}^\top\Sigma\mathbf{z} + 4\|\Sigma^{1/2}\mathbf{z}\|\,\widehat{\mathrm{MMD}}(\widehat{\boldsymbol{\theta}}).$$

Using the same bounds on $\mathbf{z}^\top\Sigma\mathbf{z}$ and $\|\Sigma^{1/2}\mathbf{z}\|$ as in Proposition A.3, and $6\sqrt{2} < 9$, gives the claimed bound. □

B Extended Implementation details

Repository. Our anonymized code is available at `https://anonymous.4open.science/r/dp-gfmn`; the readme files contain further instructions on how to run the code.

B.1 Hyperparameter settings

For each dataset, we tune the generator learning rate (LR$_{gen}$) and moving average learning rate (LR$_{mavg}$) from choices 10^{-k} and $3 \cdot 10^{-k}$ with $k \in \{3, 4, 5\}$ once for the non-private setting and once at $\epsilon = 2$. The latter is used in all private experiments for that dataset, as shown in 7. After some initial unstructured experimentation, hyperparameters are chosen with identical values across dataset shown in 8

For the Cifar10 DP-MERF baseline we tested random tuned random features dimension $d \in \{10000, 50000\}$, random features sampling distribution $\sigma \in \{100, 300, 1000\}$, learning rate decay by 10% every $e \in \{1000, 10000\}$ iterations and learning rate 10^{-k} with $k \in \{2, 3, 4, 5, 6\}$. Results presented use $d = 500000, \sigma = 1000, e = 10000, k = 3$.

The DP-GAN baseline for Cifar10 and CelebA uses the same generator as DP-MEPF with 3 residual blocks and a total of 8 convolutional layers and is paired with a ResNet9 discriminator which uses Groupnorm instead of Batchnorm to allow for per-sample gradient computation. We pre-train the model non-privately to convergence on downsampled imagenet in order to maintain the same resolution of 32×32 and then fine-tune the model for a smaller number of epochs. In case of the CelebA 64×64 data we add another residual block to discriminator and generator to account for the doubling in resolution. The base multiplier for number of feature maps is reduced from 64 to 50 to lessen the increase in number of weights. Results are the best scores of a grid-search over the following parameters at $\epsilon = 2$, which is then used in all settings: number of epochs $\{1, 10, 30, 50\}$ generator and discriminator learning rate separately for 10^{-k} and $3 \cdot 10^{-k}$ with $k \in \{3, 4, 5\}$, clip-norm $\{10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$, batch size $\{128, 256, 512\}$ and, as advised in Anonymous (2023), number of discriminator updates per generator $\{1, 10, 30, 50\}$. The chosen values are given in table 9.

Table 7: Learning rate hyperparameters across datasets

	LR$_{gen}$	LR$_{mavg}$
MNIST-nonDP	10^{-5}	10^{-3}
MNIST-DP	10^{-5}	10^{-4}
FashionMNIST-nonDP	10^{-5}	10^{-3}
FashionMNIST-DP	10^{-4}	10^{-3}
CelebA32-nonDP	10^{-3}	$\cdot 10^{-4}$
CelebA32-DP	10^{-3}	$\cdot 10^{-4}$
CelebA64-nonDP	10^{-4}	$3 \cdot 10^{-4}$
CelebA64-DP	10^{-4}	$3 \cdot 10^{-4}$
Cifar10-nonDP labeled	10^{-3}	10^{-2}
Cifar10-DP labeled	10^{-3}	10^{-2}
Cifar10-nonDP unlabeled	10^{-3}	$3 \cdot 10^{-4}$
Cifar10-DP unlabeled	10^{-3}	$3 \cdot 10^{-4}$

Table 8: Hyperparameters fixed across datasets

Parameter	Value
(ϕ_1)-bound	1
(ϕ_2)-bound	1
iterations (MNIST & FashionMNIST)	100,000
batch size (MNIST and FashionMNIST)	100
iterations (Cifar10 & CelebA)	200,000
batch size (Cifar10 and CelebA)	128
seeds	1,2,3,4,5

Table 9: Hyperparameters of DP-GAN for Cifar10 and CelebA

	Cifar10	CelebA 32×32	CelebA 64×64			
			$\epsilon \in \{0.2, 0.5\}$	$\epsilon = 1$	$\epsilon = 2$	$\epsilon \in \{5, 10\}$
LR$_{gen}$	10^{-4}	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$	$3 \cdot 10^{-4}$
LR$_{dis}$	10^{-3}	$3 \cdot 10^{-4}$	10^{-3}	$3 \cdot 10^{-4}$	10^{-3}	10^{-3}
batch size	512	512	512	512	512	512
epochs	10	10	10	10	10	10
discriminator frequency	10	10	30	30	10	10
clip norm	10^{-5}	10^{-4}	10^{-5}	10^{-5}	10^{-4}	10^{-5}

C Detailed Tables

Below we present the results from the main paper with added $a \pm b$ notation, where a is the mean and b is the standard deviation of the score distribution across three independent runs for MNIST and FashionMNIST and 5 independent runs for Cifar10 and CelebA.

Table 10: Downstream accuracies of our method for MNIST at varying values of ϵ

		$\epsilon = \infty$	$\epsilon = 10$	$\epsilon = 5$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.2$
MLP	DP-MEPF (ϕ_1, ϕ_2)	91.4 ± 0.3	89.8 ± 0.5	89.9 ± 0.2	89.3 ± 0.3	89.3 ± 0.6	79.9 ± 1.3
	DP-MEPF (ϕ_1)	88.2 ± 0.6	88.8 ± 0.1	88.4 ± 0.5	88.0 ± 0.2	87.5 ± 0.6	77.1 ± 0.4
LogReg	DP-MEPF (ϕ_1, ϕ_2)	84.6 ± 0.5	83.4 ± 0.6	83.3 ± 0.7	82.9 ± 0.7	82.5 ± 0.5	75.8 ± 1.1
	DP-MEPF (ϕ_1)	81.4 ± 0.4	80.8 ± 0.9	80.8 ± 0.8	80.5 ± 0.6	79.0 ± 0.6	72.1 ± 1.4

Table 11: Downstream accuracies of our method for FashionMNIST at varying values of ϵ

		$\epsilon = \infty$	$\epsilon = 10$	$\epsilon = 5$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.2$
MLP	DP-MEPF (ϕ_1, ϕ_2)	74.4 ± 0.3	76.0 ± 0.4	75.8 ± 0.6	75.1 ± 0.3	74.7 ± 1.1	70.4 ± 1.9
	DP-MEPF (ϕ_1)	73.8 ± 0.5	75.5 ± 0.6	75.1 ± 0.8	75.8 ± 0.7	75.0 ± 1.8	69.0 ± 1.5
LogReg	DP-MEPF (ϕ_1, ϕ_2)	74.3 ± 0.1	75.7 ± 1.0	75.2 ± 0.4	75.8 ± 0.4	75.4 ± 1.1	72.5 ± 1.2
	DP-MEPF (ϕ_1)	72.8 ± 0.5	75.5 ± 0.1	75.5 ± 0.8	76.4 ± 0.8	76.2 ± 0.8	71.7 ± 0.4

Table 12: CelebA FID scores 32×32 (lower is better)

	$\epsilon = \infty$	$\epsilon = 10$	$\epsilon = 5$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.5$	$\epsilon = 0.2$
DP-MEPF (ϕ_1, ϕ_2)	13.9 ± 1.6	15.1 ± 4.7	14.3 ± 2.3	13.9 ± 1.1	14.9 ± 2.5	14.4 ± 1.7	19.3 ± 3.0
DP-MEPF (ϕ_1)	12.8 ±	11.7 ± 0.6	12.1 ± 1.1	12.6 ± 1.0	13.2 ± 1.6	14.4 ± 1.1	18.1 ± 2.3

Table 13: CelebA FID scores 64×64 (lower is better)

	$\epsilon = \infty$	$\epsilon = 10$	$\epsilon = 5$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.5$	$\epsilon = 0.2$
DP-MEPF (ϕ_1, ϕ_2)	12.8 ± 0.6	13.0 ± 0.7	13.1 ± 0.9	13.2 ± 0.3	13.5 ± 1.1	15.5 ± 1.0	24.8 ± 1.6
DP-MEPF (ϕ_1)	11.2 ± 0.5	11.7 ± 0.7	11.7 ± 0.6	11.6 ± 0.4	13.0 ± 0.7	16.2 ± 0.7	27.3 ± 2.3

Table 14: FID scores for synthetic *labelled* CIFAR-10 data (generating both labels and input images)

	$\epsilon = \infty$	$\epsilon = 10$	$\epsilon = 5$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.5$	$\epsilon = 0.2$
DP-MEPF (ϕ_1, ϕ_2)	27.3 ± 1.5	26.6 ± 2.2	27.6 ± 2.4	27.6 ± 0.3	38.6 ± 1.9	64.4 ± 5.6	325.0 ± 15.9
DP-MEPF (ϕ_1)	25.8 ± 2.3	27.1 ± 1.0	27.7 ± 2.2	28.7 ± 1.1	39.0 ± 0.5	78.4 ± 8.1	469.3 ± 8.8
DP-MERF	127.4 ± 1.8	124.4 ± 2.3	124.0 ± 0.8	126.5 ± 2.8	122.7 ± 1.1		412.8 ± 0.8

Table 15: Test accuracies (higher better) of ResNet9 trained on CIFAR-10 synthetic data with varying privacy guarantees. When trained on real data, test accuracy is 88.3%

	$\epsilon = \infty$	$\epsilon = 10$	$\epsilon = 5$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.5$	$\epsilon = 0.2$
DP-MEPF (ϕ_1, ϕ_2)	47.2 ± 1.3	48.9 ± 3.5	47.9 ± 1.4	38.7 ± 2.3	28.9 ± 1.5	19.7 ± 3.6	12.4 ± 1.9
DP-MEPF (ϕ_1)	50.8 ± 1.7	51.0 ± 2.1	48.5 ± 2.6	42.5 ± 0.8	29.4 ± 2.9	19.4 ± 2.9	13.8 ± 1.9
DP-MERF	13.2 ± 0.4	13.4 ± 0.4	13.5 ± 0.5	13.8 ± 1.4	13.1 ± 0.7		10.4 ± 0.5

D Encoder architecture comparison

We are testing a large collection of classifiers of different sizes from the torchvision library including VGG, ResNet, ConvNext and EfficientNet. For each we look at unlabelled Cifar10 generation quality in the non-DP setting and at $\epsilon = 0.2$. In each architecture, we use all activations from convolutional layers with a kernel size greater than 1x1. We list the number of extracted features along with the achieved FID score in table 17, where each result is the best result obtained by tuning learning rates. As already observed in dos Santos et al. (2019), we find that VGG architectures appear to learn particularly useful features for feature matching. We hypothesized that in the private setting other architectures with fewer features might outperform the VGG model, but have found this to not be the case.

E Public dataset comparison

We pretrained a ResNet18 using ImageNet, CIFAR10, and SVHN as our public data, respectively. We then used the perceptual features to train a generator using CelebA dataset as our private data at a privacy budget of $\epsilon = 0.2$ and obtained the scores shown in 18. These numbers reflect our intuition that as long as the public data is sufficiently similar and contains more complex patterns than private data, e.g., transferring the knowledge learned from ImageNet as public data to generate CelebA images as private data, the learned features from public data are useful enough to generate good synthetic data. In addition, as the public data become more simplistic (from CIFAR10 to SVHN), the usefulness of such features reduces in producing good CelebA synthetic samples.

Table 16: FID scores for synthetic *unlabelled* CIFAR-10 data

	$\epsilon = \infty$	$\epsilon = 10$	$\epsilon = 5$	$\epsilon = 2$	$\epsilon = 1$	$\epsilon = 0.5$	$\epsilon = 0.2$
DP-MEPF (ϕ_1, ϕ_2)	24.3 ± 1.2	27.1 ± 2.7	24.9 ± 1.0	26.0 ± 0.7	27.2 ± 3.0	34.8 ± 2.5	56.6 ± 7.9
DP-MEPF (ϕ_1)	26.6 ± 1.6	26.8 ± 1.6	25.9 ± 0.9	28.9 ± 2.8	32.0 ± 1.7	38.6 ± 4.7	53.9 ± 2.1

Table 17: Unlabeled Cifar10 FID scores achieved with different feature extractors. VGG models yield the best results both in non-DP and high DP settings.

Encoder model	#features	$\epsilon = \infty$		$\epsilon = 0.2$	
		(ϕ_1, ϕ_2)	(ϕ_1)	(ϕ_1, ϕ_2)	(ϕ_1)
VGG19	303104	24.7	25.5	46.5	52.5
VGG16	276480	25.4	27.3	52.1	56.5
VGG13	249856	24.4	25.7	45.7	58.0
VGG11	151552	25.0	25.1	53.9	48.9
ResNet152	429568	46.6	67.7	77.7	80.0
ResNet101	300544	59.3	104.7	64.7	73.8
ResNet50	196096	58.8	65.8	80.0	91.2
ResNet34	72704	59.8	70.3	65.8	66.8
ResNet18	47104	71.9	82.1	90.4	83.8
ConvNext large	161280	110.4	242.4	130.3	236.9
ConvNext base	107520	119.9	241.5	128.9	240.3
ConvNext small	80640	103.0	227.5	151.4	216.4
ConvNext tiny	52992	94.2	227.9	124.5	223.5
EfficientNet L	119168	126.1	126.1	210.1	216.2
EfficientNet M	68704	109.8	121.6	196.1	174.3
EfficientNet S	47488	99.6	120.3	155.9	154.8

Table 18: FID scores achieved for CelebA 32×32 using a ResNet encoder with different public training sets

	ImageNet	Cifar10	SVHN
FID	37	135	172

F Training DP-MEPF without auxiliary data

While DP-MEPF is explicitly designed to take advantage of available public data, one might wonder how the method performs if no such data is available. The following experiment on CIFAR10 explores this scenario. We assume that a privacy budget of $\epsilon = 10$ is given. We use some part of the budget for feature extractor (i.e. the classifier) training and the rest of the budget for the generator training.

For a feature extractor, we have trained ResNet-20 classifiers with DP-SGD at three different levels of $\epsilon \in \{2, 5, 8\}$ for classifying the CIFAR10 dataset. We set the clipping norm to 0.01 and trained the classifiers for $7, 49$ and 98 epochs, respectively. Their test accuracies are $38.4\%, 49.5\%$ and 54.0% respectively. We also include scores for DP-MEPF applied to the untrained Classifier, denoted as $\epsilon = 0$.

Then, we train the generator using these four sets of features to generate CIFAR10 images, where each generator training uses the rest of the budget, i.e., $\epsilon \in \{8, 5, 2\}$ and $\epsilon = 10$ for the untrained classifier. We tune the learning rate in each of the four settings and keep other hyperparameters at default values.

Table 19: DP-MEPF results in CIFAR10 when using a DP feature extractor ($\epsilon = 0$ is an untrained extractor)

ϵ for feature extractor training	for generator training	FID
0	10	97.6
2	8	124.1
5	5	82.0
8	2	99.1

As expected, in Table 19 we see a considerable increase in the FID score, compared to DP-MEPF with public data. A balanced allocation of privacy budget with $\epsilon = 5$ each for classifier and generator training yields the best result at an FID score of 82 and performs significantly better than just using a randomly initialized feature extractor, which only achieves a score of 97.6. For comparison: with public data DP-MEPF achieves an FID score of 24.9 at $\varepsilon = 5$, highlighting the importance of such data to our method.

G Additional Plots

Below we show samples from our generated MNIST and FashionMNIST data in Figure 8 and Figure 9 respectively.

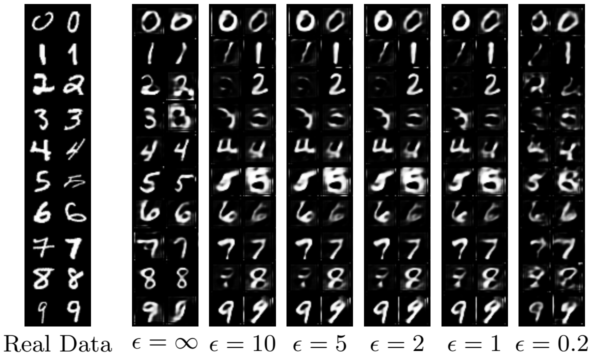

Figure 8: MNIST samples produced with DP-MEPF (ϕ_1, ϕ_2) at various levels of privacy

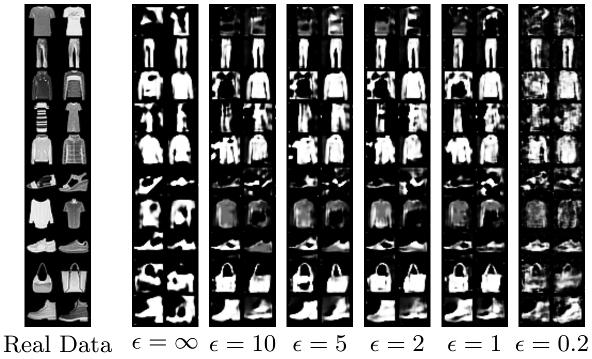

Figure 9: Fashion-MNIST samples produced with DP-MEPF (ϕ_1, ϕ_2) at various levels of privacy