Supplementary Material

This supplement is structured as follows: In Appendix A we present proofs for all novel theoretical results stated in Section 5 of the main text. In Appendices B and C we provide additional experimental results to support the discussion in Section 4 of the main text.

A Proof of Theoretical Results

In what follows we let $\mathcal{H}$ denote the reproducing kernel Hilbert space $\mathcal{H}(k)$ reproduced by the kernel k and let $\|\cdot\|_{\mathcal{H}}$ denote the induced norm in $\mathcal{H}$.

A.1 Proof of Theorem 1

To start the proof, define

$$\begin{aligned} a_m &:= (ms)^2 \, \mathrm{MMD}_{\mu,k}\left(\frac{1}{ms}\sum_{i=1}^{m}\sum_{j=1}^{s}\delta(x_{\pi(i,j)})\right)^2 \\ &= \sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{s}\sum_{j'=1}^{s} k(x_{\pi(i,j)}, x_{\pi(i',j')}) - 2ms\sum_{i=1}^{m}\sum_{j=1}^{s}\int k(x_{\pi(i,j)}, x)\,\mathrm{d}\mu(x) + (ms)^2 \iint k(x,x')\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x') \\ f_m(\cdot) &:= \sum_{i=1}^{m}\sum_{j=1}^{s} k(x_{\pi(i,j)}, \cdot) - ms\int k(\cdot, x)\,\mathrm{d}\mu(x) \end{aligned}$$

and note immediately that $a_m = \|f_m\|_{\mathcal{H}}^2$. Then we can write a recursive relation

$$\begin{aligned} a_m = a_{m-1} + &\underbrace{\sum_{j=1}^{s}\sum_{j'=1}^{s} k(x_{\pi(m,j)}, x_{\pi(m,j')}) + 2\sum_{i=1}^{m-1}\sum_{j=1}^{s}\sum_{j'=1}^{s} k(x_{\pi(m,j)}, x_{\pi(i,j')}) - 2ms\sum_{j=1}^{s}\int k(x_{\pi(m,j)}, x)\,\mathrm{d}\mu(x)}_{(*)} \\ &\underbrace{- 2s\sum_{i=1}^{m-1}\sum_{j=1}^{s}\int k(x_{\pi(i,j)}, x)\,\mathrm{d}\mu(x) \; + \; s^2(2m-1)\iint k(x,x')\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x')}_{(**)} \end{aligned}$$

We will first derive an upper bound for $(*)$, then one for $(**)$.

Bounding $(*)$: Noting that the algorithm chooses the $S \in \{1,\dots,n\}^s$ that minimises

$$\begin{aligned} &\sum_{j\in S}\sum_{j'\in S} k(x_j, x_{j'}) + 2\sum_{j\in S}\sum_{j'=1}^{s}\sum_{i=1}^{m-1} k(x_j, x_{\pi(i,j')}) - 2ms\sum_{j\in S}\int k(x_j, x)\,\mathrm{d}\mu(x) \\ &\qquad = \sum_{j\in S}\sum_{j'\in S} k(x_j, x_{j'}) - 2s\sum_{j\in S}\int k(x_j, x)\,\mathrm{d}\mu(x) + 2\sum_{j\in S} f_{m-1}(x_j), \end{aligned}$$

we therefore have that

$$\begin{aligned} (*) &= \min_{S\in\{1,\dots,n\}^s}\left[\sum_{j\in S}\sum_{j'\in S} k(x_j, x_{j'}) - 2s\sum_{j\in S}\int k(x_j, x)\,\mathrm{d}\mu(x) + 2\sum_{j\in S} f_{m-1}(x_j)\right] \\ &\le \max_{S\in\{1,\dots,n\}^s}\left[\sum_{j\in S}\sum_{j'\in S} k(x_j, x_{j'}) - 2s\sum_{j\in S}\int k(x_j, x)\,\mathrm{d}\mu(x)\right] + 2\min_{S\in\{1,\dots,n\}^s}\sum_{j\in S} f_{m-1}(x_j) \\ &= \max_{S\in\{1,\dots,n\}^s}\left[\sum_{j\in S}\sum_{j'\in S} k(x_j, x_{j'}) - 2s\sum_{j\in S}\int \langle k(x_j,\cdot), k(x,\cdot)\rangle_{\mathcal{H}}\,\mathrm{d}\mu(x)\right] + 2\min_{S\in\{1,\dots,n\}^s}\sum_{j\in S} f_{m-1}(x_j) \qquad (8) \end{aligned}$$

$$\leq \max_{S\in\{1,\ldots,n\}^s}\left[\sum_{j\in S}\sum_{j'\in S}k(x_j,x_{j'})+2s\sum_{j\in S}\big\|k(x_j,\cdot)\big\|_{\mathcal{H}}\cdot\int\big\|k(x,\cdot)\big\|_{\mathcal{H}}\,\mathrm{d}\mu(x)\right]+2\min_{S\in\{1,\ldots,n\}^s}\sum_{j\in S}f_{m-1}(x_j) \quad (9)$$

$$\leq s^2\max_{j\in\{1,\ldots,n\}}k(x_j,x_j)+2s^2\max_{j\in\{1,\ldots,n\}}\sqrt{k(x_j,x_j)}\cdot\int\sqrt{k(x,x)}\,\mathrm{d}\mu(x)+2\min_{S\in\{1,\ldots,n\}^s}\sum_{j\in S}f_{m-1}(x_j)$$

$$\leq s^2C_{n,k}^2+2s^2C_{n,k}\left(\int k(x,x)\mathrm{d}\mu(x)\right)^{1/2}+2\min_{S\in\{1,\ldots,n\}^s}\sum_{j\in S}f_{m-1}(x_j) \quad (10)$$

$$= s^2C_{n,k}^2+2s^2C_{n,k}C_{\mu,k}+2\min_{S\in\{1,\ldots,n\}^s}\sum_{j\in S}f_{m-1}(x_j) \quad (11)$$

In (8) we used the reproducing property, while in (9) we used the Cauchy–Schwarz inequality and in (10) we used Jensen's inequality. To bound the third term in (11), we write

$$\min_{S\in\{1,\ldots,n\}^s}\sum_{j\in S}f_{m-1}(x_j)=\min_{S\in\{1,\ldots,n\}^s}\left\langle f_{m-1},\sum_{j\in S}k(\cdot,x_j)\right\rangle_{\mathcal{H}}$$

Define $\mathcal{M}$ as the convex hull in $\mathcal{H}$ of $\left\{s^{-1}\sum_{j\in S}k(\cdot,x_j), S\in\{1,\ldots,n\}^s\right\}$. Since the extreme points of $\mathcal{M}$ correspond to the vertices $(x_i,\ldots,x_i)$ we have that

$$\mathcal{M}=\left\{\sum_{i=1}^n c_ik(\cdot,x_i):c_i\geq 0,\sum_{i=1}^n c_i=1\right\}.$$

Then we have, for any $h\in\mathcal{M}$,

$$\langle f_{m-1},h\rangle_{\mathcal{H}}=\left\langle f_{m-1},\sum_{i=1}^n c_ik(\cdot,x_i)\right\rangle_{\mathcal{H}}=\sum_{i=1}^n c_if_{m-1}(x_i).$$

This linear combination is clearly minimised by taking each of the x_i equal to a candidate point x_j that minimises $f_{m-1}(x_j)$, and taking the corresponding $c_j=1$, and all other $c_i=0$. Now consider an element $h_w=\sum_{i=1}^n w_ik(\cdot,x_i)$ for which the weights $w=(w_1,\ldots,w_n)^\top$ minimise $\mathrm{MMD}_{\mu,k}(\sum_{i=1}^n w_i\delta(x_i))$ subject to $1^\top w=1$ and $w_i\geq 0$. Clearly $h_w\in\mathcal{M}$. Thus

$$\min_{S\in\{1,\ldots,n\}^s}\sum_{j\in S}f_{m-1}(x_j)=s\cdot\inf_{h\in\mathcal{M}}\langle f_{m-1},h\rangle_{\mathcal{H}}\leq s\cdot\langle f_{m-1},h_w\rangle_{\mathcal{H}}.$$

Combining this with (11) provides an overall bound on $(*)$.

Bounding $(**)$: To upper bound $(**)$ we can in fact just use an equality;

$$\begin{aligned}(**)&=-2s\left[\sum_{i=1}^{m-1}\sum_{j=1}^{s}\int k(x_{\pi(i,j)},x)\,\mathrm{d}\mu(x)\;+\;s(m-1)\iint k(x,x')\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x')\right]\\&\qquad+s^2\iint k(x,x')\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x')\\&=-2s\langle f_{m-1},h_\mu\rangle_{\mathcal{H}}+s^2\|h_\mu\|_{\mathcal{H}}^2\end{aligned}$$

where $h_\mu=\int k(\cdot,x)\,\mathrm{d}\mu(x)$.

Bound on the Iterates: Combining our bounds on $(*)$ and $(**)$, we obtain

$$\begin{aligned}a_m&\leq a_{m-1}+s^2C_{n,k}^2+2s^2C_{n,k}C_{\mu,k}+2s\langle f_{m-1},h_w\rangle_{\mathcal{H}}-2s\langle f_{m-1},h_\mu\rangle_{\mathcal{H}}+s^2\|h_\mu\|_{\mathcal{H}}^2\\&=a_{m-1}+s^2C_{n,k}^2+2s^2C_{n,k}C_{\mu,k}+2s\langle f_{m-1},h_w-h_\mu\rangle_{\mathcal{H}}+s^2\|h_\mu\|_{\mathcal{H}}^2\\&\leq a_{m-1}+s^2C_{n,k}^2+2s^2C_{n,k}C_{\mu,k}+2s\|f_{m-1}\|_{\mathcal{H}}\cdot\|h_w-h_\mu\|_{\mathcal{H}}+s^2\|h_\mu\|_{\mathcal{H}}^2\\&\leq a_{m-1}+\left(s^2C_{n,k}^2+2s^2C_{n,k}C_{\mu,k}+s^2C_{\mu,k}^2\right)+2s\sqrt{a_{m-1}}\cdot\|h_w-h_\mu\|_{\mathcal{H}}\end{aligned}$$

The last line arises because

$$\begin{aligned}
\|h_\mu\|_{\mathcal{H}}^2 = \iint k(x,x')\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x') &= \iint \langle k(x,\cdot), k(x',\cdot)\rangle\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x') && (12)\\
&\le \iint |\langle k(x,\cdot), k(x',\cdot)\rangle|\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x')\\
&\le \iint \|k(x,\cdot)\|_{\mathcal{H}}\|k(x',\cdot)\|_{\mathcal{H}}\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x') && (13)\\
&= \left(\int \sqrt{k(x,x)}\,\mathrm{d}\mu(x)\right)^2\\
&\le \int k(x,x)\,\mathrm{d}\mu(x) = C_{\mu,k}^2. && (14)
\end{aligned}$$

In (12) we used the reproducing property, while in (13) we used the Cauchy–Schwarz inequality and in (14) we used Jensen's inequality.

We now note that

$$\begin{aligned}
\|h_w - h_\mu\|_{\mathcal{H}}^2 &= \langle h_w - h_\mu, h_w - h_\mu\rangle_{\mathcal{H}}\\
&= \left\langle \sum_{i=1}^n w_i k(\cdot, x_i) - \int k(\cdot, x)\,\mathrm{d}\mu(x), \sum_{i'=1}^n w_{i'} k(\cdot, x_{i'}) - \int k(\cdot, x')\,\mathrm{d}\mu(x') \right\rangle_{\mathcal{H}}\\
&= \sum_{i=1}^n \sum_{i'=1}^n w_i w_{i'} k(x_i, x_{i'}) - 2\sum_{i=1}^n w_i \int k(x_i, x)\,\mathrm{d}\mu(x) + \iint k(x,x')\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x')\\
&= \mathrm{MMD}_{\mu,k}\left(\sum_{i=1}^n w_i \delta(x_i)\right)^2 \quad =: \Phi^2,
\end{aligned}$$

which gives

$$a_m \le a_{m-1} + s^2 (C_{n,k} + C_{\mu,k})^2 + 2s\sqrt{a_{m-1}} \cdot \Phi$$

as an overall bound on the iterates a_m.

Inductive Argument: Next we follow a similar argument to Theorem 1 in Riabiz et al. (2020) to establish an induction in a_m. Defining $C^2 := (C_{n,k} + C_{\mu,k})^2$ for brevity and noting that C^2 is a constant satisfying $C^2 \ge 0$, we assert

$$a_m \le (sm)^2(\Phi^2 + K_m), \qquad \text{with} \qquad K_m := \frac{1}{m}(C^2 - \Phi^2)\sum_{j=1}^m \frac{1}{j}$$

For $m = 1$, we have $a_1 \le s^2(C_{n,k}^2 + 2C_{n,k}C_{\mu,k} + C_{\mu,k}^2) = s^2C^2$, so the root of the induction holds. We now assume that $a_{m-1} \le s^2(m-1)^2(\Phi^2 + K_{m-1})$. Then

$$\begin{aligned}
a_m &\le a_{m-1} + s^2C^2 + 2s\sqrt{a_{m-1}} \cdot \Phi\\
&\le s^2(m-1)^2(\Phi^2 + K_{m-1}) + s^2C^2 + 2s^2(m-1)\Phi\sqrt{\Phi^2 + K_{m-1}}\\
&\le s^2\left[(m-1)^2(\Phi^2 + K_{m-1}) + C^2 + (m-1)(2\Phi^2 + K_{m-1})\right] && (15)\\
&= s^2\left[(m^2-1)\Phi^2 + m(m-1)K_{m-1} + C^2\right]\\
&= s^2\left[(m^2-1)\Phi^2 + m(C^2 - \Phi^2)\sum_{j=1}^{m-1}\frac{1}{j} + C^2\right]\\
&= s^2\left[(m^2-1)\Phi^2 + m(C^2 - \Phi^2)\sum_{j=1}^{m}\frac{1}{j} - m(C^2 - \Phi^2)\frac{1}{m} + C^2\right]\\
&= s^2\left[m^2\Phi^2 + m(C^2 - \Phi^2)\sum_{j=1}^{m}\frac{1}{j}\right]\\
&= (sm)^2(\Phi^2 + K_m),
\end{aligned}$$

which proves the induction. Here (15) follows from the fact that for any $a, b > 0$, it holds that $2a\sqrt{a^2 + b} \le 2a^2 + b$.

Overall Bound: To complete the proof, we first show that $\Phi^2 \leq C^2$ by writing

$$\Phi^2 = \|h_w - h_\mu\|_{\mathcal{H}}^2 \leq \|h_w\|_{\mathcal{H}}^2 + 2\|h_w\|_{\mathcal{H}} \cdot \|h_\mu\|_{\mathcal{H}} + \|h_\mu\|_{\mathcal{H}}^2$$

and noting that, since $k(x_i, x_{i'}) \leq \sqrt{k(x_i, x_i)}\sqrt{k(x_{i'}, x_{i'})}$ and $\sum_{i=1}^n w_i = 1$, it holds that

$$\|h_w\|_{\mathcal{H}}^2 = \sum_{i=1}^n \sum_{i'=1}^n w_i w_{i'} k(x_i, x_{i'}) \leq C_{n,k}^2.$$

We have already shown that $\|h_\mu\|^2 \leq C_{\mu,k}^2$, thus it follows that $\Phi^2 \leq C_{n,k}^2 + 2C_{n,k}C_{\mu,k} + C_{\mu,k}^2 \equiv C^2$ as required. Using this bound in conjunction with the elementary series inequality $\sum_{j=1}^m j^{-1} \leq (1 + \log m)$, we have $K_m \geq 0$ and

$$K_m = \frac{1}{m}(C^2 - \Phi^2)\sum_{j=1}^m \frac{1}{j} \;\leq\; \frac{1}{m}C^2 \sum_{j=1}^m \frac{1}{j} \;\leq \left(\frac{1 + \log m}{m}\right) C^2$$

Finally, the theorem follows by noting

$$\mathrm{MMD}_{\mu,k}\left(\frac{1}{ms}\sum_{i=1}^m \sum_{j=1}^s \delta(x_{\pi(i,j)})\right)^2 = \frac{a_m}{(sm)^2} \leq \Phi^2 + K_m = \Phi^2 + \left(\frac{1 + \log m}{m}\right) C^2,$$

as claimed. □

Remark: We observe that, in the myopic case only ($s = 1$), one can alternatively recover Theorem 1 as a consequence of Theorem 1 in Riabiz et al. (2020) (refer also to Theorem 5 of Chen et al., 2019). This can be seen by noting that $\mathrm{MMD}_{\mu,k_0}(\nu) = \mathrm{MMD}_{\mu,k}(\nu)$ for all $\nu \in \mathcal{P}(\mathcal{X})$, where k_0 is the kernel

$$k_0(x, y) := k(x, y) - \int k(x, x')\mathrm{d}\mu(x') - \int k(y, y')\mathrm{d}\mu(y') + \iint k(x', y')\mathrm{d}\mu(x')\mathrm{d}\mu(y'), \tag{16}$$

which satisfies the precondition $\int k_0(x, y')\mathrm{d}\mu(y') = 0$ for all $x \in \mathcal{X}$ in Theorem 1 of Riabiz et al. (2020). Indeed,

$$\begin{aligned}
\mathrm{MMD}_{\mu,k_0}(\nu)^2 &= \left\| \int k_0(\cdot, y')\mathrm{d}\nu(y') - \int k_0(\cdot, y')\mathrm{d}\mu(y') \right\|_{\mathcal{H}(k_0)}^2 \\
&= \left\| \int k_0(\cdot, y')\mathrm{d}\nu(y') \right\|_{\mathcal{H}(k_0)}^2 \\
&= \iint \left[k(x, y) - \int k(x, y')\mathrm{d}\mu(y') - \int k(x', y)\mathrm{d}\mu(x') + \iint k(x', y')\mathrm{d}\mu(x')\mathrm{d}\mu(y') \right] \mathrm{d}\nu(x)\mathrm{d}\nu(y) \\
&= \iint k(x, y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) - \iint k(x, y)\mathrm{d}\mu(x)\mathrm{d}\nu(y) - \iint k(x, y)\mathrm{d}\nu(x)\mathrm{d}\nu(y) \\
&\qquad + \iint k(x, y)\mathrm{d}\nu(x)\mathrm{d}\nu(y) \\
&= \mathrm{MMD}_{\mu,k}(\nu)^2.
\end{aligned}$$

A.2 Proof of Theorem 2

First note that the preconditions of Theorem 1 are satisfied. We may therefore take expectations of the bound obtained in Theorem 1, to obtain that:

$$\mathbb{E}\left[\mathrm{MMD}_{\mu,k}\left(\frac{1}{ms}\sum_{i=1}^m \sum_{j=1}^s \delta(x_{\pi(i,j)})\right)^2\right] \leq \mathbb{E}\left[\min_{\substack{1^T w = 1 \\ w_i \geq 0}} \mathrm{MMD}_{\mu,k}\left(\sum_{i=1}^n w_i \delta(x_i)\right)^2\right] + \mathbb{E}[C^2]\left(\frac{1 + \log m}{m}\right), \tag{17}$$

To bound the first expectation we proceed as follows:

$$\begin{aligned}
\mathbb{E}\left[\min_{\substack{1^T w=1\\ w_i\geq 0}} \mathrm{MMD}_{\mu,k}\left(\sum_{i=1}^n w_i\delta(x_i)\right)^2\right] &\leq \mathbb{E}\left[\mathrm{MMD}_{\mu,k}\left(\frac{1}{n}\sum_{i=1}^n \delta(x_i)\right)^2\right] \qquad (18)\\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(x_i,x_j) - \frac{2}{n}\sum_{i=1}^n \int k(x,x_i)\,\mathrm{d}\mu(x) + \iint k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\right]\\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^n\sum_{j=1}^n k(x_i,x_j)\right] - \iint k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \qquad (\text{since } x_i\sim\mu)\\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^n k(x_i,x_i) + \frac{1}{n^2}\sum_{i=1}^n\sum_{j\neq i} k(x_i,x_j)\right] - \iint k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^n k(x_i,x_i)\right] - \frac{1}{n}\iint k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y) \qquad (\text{since } x_i\sim\mu)\\
&= \frac{1}{n}\mathbb{E}\left[k(x_1,x_1)\right] - \frac{C_{\mu,k}^2}{n}\\
&= \frac{1}{n\gamma}\mathbb{E}\left[\log e^{\gamma k(x_i,x_i)}\right] - \frac{C_{\mu,k}^2}{n}\\
&\leq \frac{1}{n\gamma}\log\mathbb{E}\left[e^{\gamma k(x_i,x_i)}\right] - \frac{C_{\mu,k}^2}{n}\\
&\leq \frac{1}{n\gamma}\log(C_1) - \frac{C_{\mu,k}^2}{n}\\
&\leq \frac{1}{n\gamma}\log(C_1). \qquad (19)
\end{aligned}$$

To bound the second expectation we use the fact that $C^2 = (C_{\mu,k}+C_{n,k})^2 \leq 2C_{\mu,k}^2 + 2C_{n,k}^2$ where $C_{\mu,k}$ is independent of the set $\{x_i\}_{i=1}^n$ to focus only on the term $C_{n,k}$. Here we have that

$$\begin{aligned}
\mathbb{E}[C_{n,k}^2] := \mathbb{E}\left[\max_{i=1,\dots,n} k(x_i,x_i)\right] &= \mathbb{E}\left[\frac{1}{\gamma}\log\max_{i=1,\dots,n} e^{\gamma k(x_i,x_i)}\right] \qquad (20)\\
&\leq \mathbb{E}\left[\frac{1}{\gamma}\log\sum_{i=1}^n e^{\gamma k(x_i,x_i)}\right]\\
&\leq \frac{1}{\gamma}\log\left(\sum_{i=1}^n \mathbb{E}\left[e^{\gamma k(x_i,x_i)}\right]\right) = \frac{\log(nC_1)}{\gamma}. \qquad (21)
\end{aligned}$$

Thus we arrive at the overall bound

$$\mathbb{E}\left[\mathrm{MMD}_{\mu,k}\left(\frac{1}{ms}\sum_{i=1}^m\sum_{j=1}^s \delta(x_{\pi(i,j)})\right)^2\right] \leq \frac{\log(C_1)}{n\gamma} + 2\left(C_{\mu,k}^2 + \frac{\log(nC_1)}{\gamma}\right)\left(\frac{1+\log m}{m}\right),$$

as claimed. □

Remark: We observe that, in the myopic case only ($s=1$), one can alternatively recover Theorem 2 as a consequence of Theorem 2 in Riabiz et al. (2020), once again using the observation that the kernel in (16) satisfies the preconditions of Theorem 2 in Riabiz et al. (2020).

A.3 Proof of Theorem 3

The following proof combines parts of the arguments used to establish Theorem 1 and Theorem 2, with additional notation required to deal with the mini-batching involved.

In a natural extension to the proof of Theorem 1, we define

$$
\begin{aligned}
a_m &:= (ms)^2 \,\mathrm{MMD}_{\mu,k}\left(\frac{1}{ms}\sum_{i=1}^{m}\sum_{j=1}^{s}\delta(x^i_{\pi(i,j)})\right)^2 \\
&= \sum_{i=1}^{m}\sum_{i'=1}^{m}\sum_{j=1}^{s}\sum_{j'=1}^{s} k(x^i_{\pi(i,j)}, x^{i'}_{\pi(i',j')}) - 2ms\sum_{i=1}^{m}\sum_{j=1}^{s}\int k(x^i_{\pi(i,j)}, x)\,\mathrm{d}\mu(x) + (ms)^2 \iint k(x,x')\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x') \\
f_m(\cdot) &:= \sum_{i=1}^{m}\sum_{j=1}^{s} k(x^i_{\pi(i,j)},\cdot) - ms\int k(\cdot,x)\,\mathrm{d}\mu(x)
\end{aligned}
$$

and note immediately that $a_m = \|f_m\|^2_{\mathcal{H}}$. Then, similarly to Theorem 1, we write a recursive relation

$$
\begin{aligned}
a_m = a_{m-1} + &\underbrace{\sum_{j=1}^{s}\sum_{j'=1}^{s} k(x^m_{\pi(m,j)}, x^m_{\pi(m,j')}) + 2\sum_{i=1}^{m-1}\sum_{j=1}^{s}\sum_{j'=1}^{s} k(x^m_{\pi(m,j)}, x^i_{\pi(i,j')}) - 2ms\sum_{j=1}^{s}\int k(x^m_{\pi(m,j)}, x)\,\mathrm{d}\mu(x)}_{(*)} \\
&\underbrace{- 2s\sum_{i=1}^{m-1}\sum_{j=1}^{s}\int k(x^i_{\pi(i,j)}, x)\,\mathrm{d}\mu(x) \;+\; s^2(2m-1)\iint k(x,x')\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x')}_{(**)}\,.
\end{aligned}
$$

We will first derive an upper bound for $(*)$, then one for $(**)$.

Bounding $(*)$: Noting that at iteration m the algorithm chooses the $S \in \{1,\dots,b\}^s$ that minimises

$$
\begin{aligned}
&\sum_{j\in S}\sum_{j'\in S} k(x^m_j, x^m_{j'}) + 2\sum_{j\in S}\sum_{j'=1}^{s}\sum_{i=1}^{m-1} k(x^m_j, x^i_{\pi(i,j')}) - 2ms\sum_{j\in S}\int k(x^m_j, x)\,\mathrm{d}\mu(x) \\
&\qquad = \sum_{j\in S}\sum_{j'\in S} k(x^m_j, x^m_{j'}) - 2s\sum_{j\in S}\int k(x^m_j, x)\,\mathrm{d}\mu(x) + 2\sum_{j\in S} f_{m-1}(x^m_j),
\end{aligned}
$$

we have that

$$
\begin{aligned}
(*) &= \min_{S\in\{1,\dots,b\}^s}\left[\sum_{j\in S}\sum_{j'\in S} k(x^m_j, x^m_{j'}) - 2s\sum_{j\in S}\int k(x^m_j, x)\,\mathrm{d}\mu(x) + 2\sum_{j\in S} f_{m-1}(x^m_j)\right] \\
&\leq \max_{S\in\{1,\dots,b\}^s}\left[\sum_{j\in S}\sum_{j'\in S} k(x^m_j, x^m_{j'}) - 2s\sum_{j\in S}\int k(x^m_j, x)\,\mathrm{d}\mu(x)\right] + 2\min_{S\in\{1,\dots,b\}^s}\sum_{j\in S} f_{m-1}(x^m_j) \\
&= \max_{S\in\{1,\dots,b\}^s}\left[\sum_{j\in S}\sum_{j'\in S} k(x^m_j, x^m_{j'}) - 2s\sum_{j\in S}\int \left\langle k(x^m_j,\cdot), k(x,\cdot)\right\rangle_{\mathcal{H}}\,\mathrm{d}\mu(x)\right] + 2\min_{S\in\{1,\dots,b\}^s}\sum_{j\in S} f_{m-1}(x^m_j) \qquad (22) \\
&\leq \max_{S\in\{1,\dots,n\}^b}\left[\sum_{j\in S}\sum_{j'\in S} k(x^m_j, x^m_{j'}) + 2s\sum_{j\in S}\left\|k(x^m_j,\cdot)\right\|_{\mathcal{H}}\cdot\int \|k(x,\cdot)\|_{\mathcal{H}}\,\mathrm{d}\mu(x)\right] + 2\min_{S\in\{1,\dots,b\}^s}\sum_{j\in S} f_{m-1}(x^m_j) \\
&\qquad (23) \\
&\leq s^2\max_{j\in\{1,\dots,b\}} k(x^m_j, x^m_j) + 2s^2\max_{j\in\{1,\dots,b\}}\sqrt{k(x^m_j, x^m_j)}\cdot\int\sqrt{k(x,x)}\,\mathrm{d}\mu(x) + 2\min_{S\in\{1,\dots,b\}^s}\sum_{j\in S} f_{m-1}(x^m_j)
\end{aligned}
$$

$$\leq s^2 C_{b,m,k}^2 + 2s^2 C_{b,m,k} \left(\int k(x,x) \mathrm{d}\mu(x) \right)^{1/2} + 2 \min_{S \in \{1,\ldots,b\}^s} \sum_{j \in S} f_{m-1}(x_j^m) \tag{24}$$
$$= s^2 C_{b,m,k}^2 + 2s^2 C_{b.m,k} C_{\mu,k} + 2 \min_{S \in \{1,\ldots,b\}^s} \sum_{j \in S} f_{m-1}(x_j^m)$$

In (22) we used the reproducing property. In (23) we used the Cauchy–Schwarz inequality. In (24) we used Jensen's inequality.

To bound the third term, we write

$$\min_{S \in \{1,\ldots,b\}^s} \sum_{j \in S} f_{m-1}(x_j^m) = \min_{S \in \{1,\ldots,b\}^s} \left\langle f_{m-1}, \sum_{j \in S} k(\cdot, x_j^m) \right\rangle_{\mathcal{H}}$$

Define $\mathcal{M}_m$ as the convex hull in $\mathcal{H}$ of $\left\{ s^{-1} \sum_{j \in S} k(\cdot, x_j^m), S \in \{1,\ldots,b\}^s \right\}$. Since the extreme points of $\mathcal{M}_m$ correspond to the vertices $(x_i^m, \ldots, x_i^m)$ we have that

$$\mathcal{M}_m = \left\{ \sum_{i=1}^n c_i k(\cdot, x_i^m) : c_i \geq 0, \sum_{i=1}^n c_i = 1 \right\}$$

Then we have for any $h \in \mathcal{M}_m$

$$\langle f_{m-1}, h \rangle_{\mathcal{H}} = \left\langle f_{m-1}, \sum_{i=1}^n c_i k(\cdot, x_i^m) \right\rangle_{\mathcal{H}} = \sum_{i=1}^n c_i f_{m-1}(x_i^m)$$

This linear combination is clearly minimised by taking the $x_j^m \in \{x_i^m\}_{i=1}^b$ that minimises $f_{m-1}(x_j^m)$, and taking the corresponding $c_j = 1$, and all other $c_i = 0$. Now consider the element $h_w^m = \sum_{i=1}^b w_i^m k(\cdot, x_i^m)$ for which the weights are equal to the optimal weight vector w^m. Clearly $h_w^m \in \mathcal{M}_m$. Thus

$$\min_{S \in \{1,\ldots,b\}^s} \sum_{j \in S} f_{m-1}(x_j^m) = s \cdot \inf_{h \in \mathcal{M}_m} \langle f_{m-1}, h \rangle_{\mathcal{H}} \leq s \cdot \langle f_{m-1}, h_w^m \rangle_{\mathcal{H}}.$$

Bounding $(**)$: Our bound on $(**)$ is actually just an equality:

$$(**) = -2s \left[\sum_{i=1}^{m-1} \sum_{j=1}^{s} \int k(x_{\pi(i,j)}^i, x) \, \mathrm{d}\mu(x) \; + \; s(m-1) \iint k(x,x') \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(x') \right] + s^2 \iint k(x,x') \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(x')$$
$$= -2s \langle f_{m-1}, h_\mu \rangle_{\mathcal{H}} + s^2 \|h_\mu\|_{\mathcal{H}}^2$$

where $h_\mu = \int k(\cdot, x) \, \mathrm{d}\mu(x)$.

Bound on the Iterates: Combining our bounds on $(*)$ and $(**)$ leads to the following bound on the iterates:

$$\begin{aligned} a_m &\leq a_{m-1} + s^2 C_{b,m,k}^2 + 2s^2 C_{b,m,k} C_{\mu,k} + 2s \langle f_{m-1}, h_w^m \rangle_{\mathcal{H}} - 2s \langle f_{m-1}, h_\mu \rangle_{\mathcal{H}} + s^2 \|h_\mu\|_{\mathcal{H}}^2 \\ &= a_{m-1} + s^2 C_{b,m,k}^2 + 2s^2 C_{b,m,k} C_{\mu,k} + 2s \langle f_{m-1}, h_w^m - h_\mu \rangle_{\mathcal{H}} + s^2 \|h_\mu\|_{\mathcal{H}}^2 \\ &\leq a_{m-1} + s^2 C_{b,m,k}^2 + 2s^2 C_{b,m,k} C_{\mu,k} + 2s \|f_{m-1}\|_{\mathcal{H}} \cdot \|h_w^m - h_\mu\|_{\mathcal{H}} + s^2 \|h_\mu\|_{\mathcal{H}}^2 \\ &\leq a_{m-1} + \left(s^2 C_{b,m,k}^2 + 2s^2 C_{b,m,k} C_{\mu,k} + s^2 C_{\mu,k}^2 \right) + 2s \sqrt{a_{m-1}} \cdot \|h_w^m - h_\mu\|_{\mathcal{H}} \end{aligned}$$

The last line arises because

$$\|h_\mu\|_{\mathcal{H}}^2 = \iint k(x,x') \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(x') = \iint \langle k(x,\cdot), k(x',\cdot) \rangle \, \mathrm{d}\mu(x) \, \mathrm{d}\mu(x') \tag{25}$$

$$
\begin{aligned}
&\leq \iint |\langle k(x,\cdot), k(x',\cdot)\rangle| \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x') \\
&\leq \iint \|k(x,\cdot)\|_{\mathcal{H}} \|k(x',\cdot)\|_{\mathcal{H}} \,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x') \qquad (26) \\
&= \left(\int \sqrt{k(x,x)} \,\mathrm{d}\mu(x) \right)^2 \\
&\leq \int k(x,x)\,\mathrm{d}\mu(x) = C_{\mu,k}^2 \qquad (27)
\end{aligned}
$$

In (25) we used the reproducing property. In (26) we used the Cauchy–Schwarz inequality. In (27) we used Jensen's inequality.

We now note that

$$
\begin{aligned}
\|h_w^m - h_\mu\|_{\mathcal{H}}^2 &= \langle h_w^m - h_\mu, h_w^m - h_\mu \rangle_{\mathcal{H}} \\
&= \left\langle \sum_{i=1}^b w_i^m k(\cdot, x_i^m) - \int k(\cdot, x)\,\mathrm{d}\mu(x), \sum_{i'=1}^b w_{i'}^m k(\cdot, x_{i'}^m) - \int k(\cdot, x')\,\mathrm{d}\mu(x') \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^b \sum_{i'=1}^b w_i^m w_{i'}^m k(x_i^m, x_{i'}^m) - 2\sum_{i=1}^b w_i^m \int k(x_i^m, x)\,\mathrm{d}\mu(x) + \iint k(x,x')\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(x') \\
&= \mathrm{MMD}_{\mu,k}\left(\sum_{i=1}^b w_i^m \delta(x_i^m) \right)^2 \quad =: \Phi_m^2,
\end{aligned}
$$

which gives

$$
a_m \leq a_{m-1} + s^2 (C_{b,m,k} + C_{\mu,k})^2 + 2s\sqrt{a_{m-1}} \cdot \Phi_m.
$$

We then follow a similar argument to Theorem 1 in Riabiz et al. (2020) to establish an induction in a_m.

Inductive Argument: Let $c_m^2 := (C_{b,m,k} + C_{\mu,k})^2$. We assert

$$
\mathbb{E}[a_m] \leq (sm)^2 \mathbb{E}[\Phi_m^2 + K_m], \qquad \text{with} \qquad K_m := \frac{1}{m}(c_m^2 - \Phi_m^2) \sum_{j=1}^m \frac{1}{j}
$$

For $m = 1$, the induction holds since $a_1 \leq s^2 c_1$. We now assume that $\mathbb{E}[a_{m-1}] \leq s^2(m-1)^2 \mathbb{E}[\Phi_{m-1}^2 + K_{m-1}]$. Then

$$
\begin{aligned}
\mathbb{E}[a_m] &\leq \mathbb{E}[a_{m-1}] + s^2 \mathbb{E}[c_m^2] + 2s\mathbb{E}[\sqrt{a_{m-1}} \cdot \Phi_m] \\
&= \mathbb{E}[a_{m-1}] + s^2 \mathbb{E}[c_m^2] + 2s\mathbb{E}[\sqrt{a_{m-1}}] \cdot \mathbb{E}[\Phi_m] \qquad \text{(independence of } a_{m-1} \text{ and } \Phi_m) \\
&\leq \mathbb{E}[a_{m-1}] + s^2 \mathbb{E}[c_m^2] + 2s\sqrt{\mathbb{E}[a_{m-1}]} \cdot \mathbb{E}[\Phi_m] \qquad \text{(Jensen's inequality)} \\
&\leq s^2(m-1)^2 \mathbb{E}[\Phi_{m-1}^2 + K_{m-1}] + s^2 \mathbb{E}[c_m^2] + 2s^2(m-1)\mathbb{E}[\Phi_m]\sqrt{\mathbb{E}[\Phi_{m-1}^2 + K_{m-1}]} \\
&\leq s^2(m-1)^2 \mathbb{E}[\Phi_m^2 + K_{m-1}] + s^2 \mathbb{E}[c_m^2] + 2s^2(m-1)\mathbb{E}[\Phi_m]\sqrt{\mathbb{E}[\Phi_m^2 + K_{m-1}]} \qquad (\text{since } \Phi_{m-1} \overset{d}{=} \Phi_m) \\
&\leq s^2(m-1)^2 \mathbb{E}[\Phi_m^2 + K_{m-1}] + s^2 \mathbb{E}[c_m^2] + 2s^2(m-1)\mathbb{E}[\Phi_m^2]^{1/2}\sqrt{\mathbb{E}[\Phi_m^2 + K_{m-1}]} \qquad \text{(Jensen's inequality)} \\
&\leq s^2 \left[(m-1)^2 \mathbb{E}[\Phi_m^2 + K_{m-1}] + \mathbb{E}[c_m^2] + (m-1)(2\mathbb{E}[\Phi_m^2] + \mathbb{E}[K_{m-1}]) \right] \qquad (28) \\
&= s^2 \mathbb{E}\left[(m^2-1)\Phi_m^2 + m(m-1)K_{m-1} + c_m^2 \right] \\
&= s^2 \mathbb{E}\left[(m^2-1)\Phi_m^2 + m(c_{m-1}^2 - \Phi_{m-1}^2) \sum_{j=1}^{m-1} \frac{1}{j} + c_m^2 \right] \\
&= s^2 \mathbb{E}\left[(m^2-1)\Phi_m^2 + m(c_{m-1}^2 - \Phi_{m-1}^2) \sum_{j=1}^{m} \frac{1}{j} - m(c_{m-1}^2 - \Phi_{m-1}^2)\frac{1}{m} + c_m^2 \right] \\
&= s^2 \mathbb{E}\left[(m^2-1)\Phi_m^2 + m(c_{m-1}^2 - \Phi_{m-1}^2) \sum_{j=1}^{m} \frac{1}{j} - m(c_m^2 - \Phi_m^2)\frac{1}{m} + c_m^2 \right] \qquad (\text{since } c_{m-1} \overset{d}{=} c_m,\ \Phi_{m-1} \overset{d}{=} \Phi_m)
\end{aligned}
$$

$$= s^2\mathbb{E}\left[m^2\Phi_m^2 + m(c_{m-1}^2 - \Phi_{m-1}^2)\sum_{j=1}^{m}\frac{1}{j}\right]$$
$$= (sm)^2\mathbb{E}[\Phi_m^2 + K_m]$$

which proves the induction. The line (28) follows from the second by the fact that for any $a, b > 0$, it holds that $2a\sqrt{a^2+b} \leq 2a^2 + b$.

Overall Bound: We now show that $\Phi_m^2 \leq c_m^2$, by writing

$$\Phi_m^2 = \|h_w^m - h_\mu\|_{\mathcal{H}}^2 \leq \|h_w^m\|_{\mathcal{H}}^2 + 2\|h_w^m\|_{\mathcal{H}} \cdot \|h_\mu\|_{\mathcal{H}} + \|h_\mu\|_{\mathcal{H}}^2$$

and noting that since $\sum_{i=1}^n w_i^m = 1$, it holds that

$$\|h_w^m\|_{\mathcal{H}}^2 = \sum_{i=1}^{b}\sum_{i'=1}^{b} w_i^m w_{i'}^m k(x_i^m, x_{i'}^m) \leq C_{b,m,k}^2.$$

We have already shown that $\|h_\mu\|^2 \leq C_{\mu,k}^2$, thus it follows that $\Phi_m^2 \leq C_{b,m,k}^2 + 2C_{b,m,k}C_{\mu,k} + C_{\mu,k}^2 = c_m^2$ as required. Using this bound in conjunction with the elementary series inequality $\sum_{j=1}^m j^{-1} \leq (1+\log m)$, we have $K_m \geq 0$ and

$$K_m = \frac{1}{m}(c_m^2 - \Phi_m^2)\sum_{j=1}^{m}\frac{1}{j} \;\leq\; \frac{1}{m}c_m^2\sum_{j=1}^{m}\frac{1}{j} \;\leq\; \left(\frac{1+\log m}{m}\right)c_m^2$$

An identical argument to that used between (20) and (21) shows that

$$\mathbb{E}[C_{b,m,k}^2] = \frac{\log(nC_1)}{\gamma}$$

and therefore

$$\mathbb{E}[c_m^2] \leq 2C_{\mu,k}^2 + 2\mathbb{E}[C_{b,m,k}^2] \leq 2C_{\mu,k}^2 + \frac{2\log(bC_1)}{\gamma}.$$

An identical argument to (18)-(19) gives that

$$\mathbb{E}[\Phi_m^2] \leq \frac{\log(C_1)}{b\gamma}$$

From this the theorem follows by noting

$$\mathbb{E}\left[\mathrm{MMD}_{\mu,k}\left(\frac{1}{ms}\sum_{i=1}^{m}\sum_{j=1}^{s}\delta(x_{\pi(i,j)}^i)\right)^2\right] = \frac{\mathbb{E}[a_m]}{(sm)^2} \leq \mathbb{E}[\Phi_m^2] + \left(\frac{1+\log m}{m}\right)\mathbb{E}[c_m^2]$$
$$\leq \frac{\log(C_1)}{b\gamma} + 2\left(C_{\mu,k}^2 + \frac{\log(bC_1)}{\gamma}\right)\left(\frac{1+\log m}{m}\right).$$

□

This argument relied on independence between mini-batches and therefore it may not easily generalise to the MCMC context.

Remarks: We observe that, in the myopic case only ($s = 1$), one can alternatively recover Theorem 3 as a consequence of Theorem 6 in Chen et al. (2019), once again using the observation that the kernel in (16) satisfies the preconditions of Theorem 6 in Chen et al. (2019).

The argument used to prove Theorem 3 relies on independence between mini-batches and therefore it may not easily generalise to the MCMC context, in which this is unlikely to be true. Theorem 7 in Chen et al. (2019) considered a particular form of dependence between mini-batches (once again, only for the case $s = 1$), but this result does not directly apply to mini-batches sampled from MCMC output.

A.4 Proof of Theorem 4

The argument below is almost identical to that used in Theorem 2 of Riabiz et al. (2020), with most of the effort required to handle the non-myopic optimisation having already been performed in Theorem 1. In particular, it relies on the following technical result:

Lemma 1 (Lemma 3 in Riabiz et al. (2020))**.** *Let $\mathcal{X}$ be a measurable space and let μ be a probability distribution on $\mathcal{X}$. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a reproducing kernel with $\int k(x, \cdot)\mathrm{d}\mu(x) = 0$ for all $x \in \mathcal{X}$. Consider a μ-invariant, time-homogeneous reversible Markov chain $(x_i)_{i\in\mathbb{N}} \subset \mathcal{X}$ generated using a V-uniformly ergodic transition kernel, such that $V(x) \geq \sqrt{k(x,x)}$ for all $x \in \mathcal{X}$, with parameters $R \in [0, \infty)$ and $\rho \in (0, 1)$ as in (7). Then we have that*

$$\sum_{i=1}^{n} \sum_{r\in\{1,\ldots,n\}\setminus\{i\}} \mathbb{E}\left[k(x_i, x_r)\right] \;\leq\; C_3 \sum_{i=1}^{n-1} \mathbb{E}\left[\sqrt{k(x_i, x_i)}V(X_i)\right]$$

with $C_3 := \frac{2R\rho}{1-\rho}$. □

The proof starts in a similar manner to the proof of Theorem 2, taking expectations of the bound obtained in Theorem 1 to arrive at (17).

An identical argument to that used in the proof of Theorem 2 allows us to bound

$$\mathbb{E}[C^2] \leq 2\left(C_{\mu,k}^2 + \frac{\log(nC_1)}{\gamma}\right).$$

Thus it remains to bound the first term in (17) under the assumptions that we have made on the Markov chain $(x_i)_{i\in\mathbb{N}}$. To this end, we have that

$$\begin{aligned}
\mathbb{E}\left[\min_{\substack{1^T w=1\\ w_i\geq 0}} \mathrm{MMD}_{\mu,k}\left(\sum_{i=1}^{n} w_i\delta(x_i)\right)^2\right] &\leq \mathbb{E}\left[\mathrm{MMD}_{\mu,k}\left(\frac{1}{n}\sum_{i=1}^{n}\delta(x_i)\right)^2\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(x_i,x_j) - \frac{2}{n}\sum_{i=1}^{n}\int k(x,x_i)\,\mathrm{d}\mu(x) + \iint k(x,y)\,\mathrm{d}\mu(x)\,\mathrm{d}\mu(y)\right] \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}k(x_i,x_j)\right] \qquad \left(\text{since } \textstyle\int k(x,\cdot)\mathrm{d}\mu(x) = 0\right) \\
&= \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}k(x_i,x_i)\right] + \mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}k(x_i,x_j)\right]. \qquad (29)
\end{aligned}$$

The first term in (29) is handled as follows:

$$\begin{aligned}
\frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[k(x_i,x_i)\right] &= \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}\left[\frac{1}{\gamma}\log e^{\gamma k(x_i,x_i)}\right] \\
&\leq \frac{1}{\gamma n^2}\sum_{i=1}^{n}\log\left(\mathbb{E}\left[e^{\gamma k(x_i,x_i)}\right]\right) \;\leq\; \frac{\log(C_1)}{\gamma n}
\end{aligned}$$

The second term in (29) can be controlled using Lemma 1:

$$\mathbb{E}\left[\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j\neq i}k(x_i,x_j)\right] \leq \frac{C}{n^2}\sum_{i=1}^{n-1}\mathbb{E}\left[\sqrt{k(x_i,x_i)}V(X_i)\right] \leq \frac{C_3}{n^2}(n-1)C_2 \leq \frac{C_2C_3}{n}.$$

Thus we arrive at the overall bound

$$\mathbb{E}\left[\mathrm{MMD}_{\mu,k}\left(\frac{1}{ms}\sum_{i=1}^{m}\sum_{j=1}^{s}\delta(x_{\pi(i,j)})\right)^2\right] \leq \frac{\log(C_1)}{n\gamma} + \frac{C_2C_3}{n} + 2\left(C_{\mu,k}^2 + \frac{\log(nC_1)}{\gamma}\right)\left(\frac{1+\log m}{m}\right),$$

as claimed. □

B Semidefinite Relaxation

In this supplement we briefly explain how to construct a relaxation of the discrete optimisation problem (5). The standard technique for relaxation of a quadratic programme of this form is to construct an approximating semidefinite programme (SDP). This not only convexifies the problem but also replaces a quadratic problem in v with a linear problem in a semidefinite matrix M. To simplify the presentation we consider[5] the BQP setting of Remark 1, so that $v \in \{0,1\}^n$. We also employ a change of variable $\tilde{v}_j := 2v_j - 1$, so that $\tilde{v} \in \{-1,1\}^n$. By analogy with (4) we recast an optimal subset π as the solution to the following BQP.

$$\underset{\tilde{v}\in\{-1,1\}^n}{\operatorname{argmin}} \ \tilde{v}^\top K \tilde{v} + 2(\mathbf{1}^\top K + c_j^{i\top})\tilde{v}, \ \text{s.t. } \mathbf{1}^\top \tilde{v} = 2s - n. \tag{30}$$

The relaxation treats $\tilde{v}$ as a continuous variable whose feasible set is the entire convex hull of $\{-1,1\}^n$. Define $\tilde{V} = \tilde{v}\tilde{v}^\top$ and then relax this non-convex equality, so that $\tilde{V} - \tilde{v}\tilde{v}^\top \succeq 0$ rather than the $\tilde{V} - \tilde{v}\tilde{v}^\top = 0$. Then rewrite this as a Schur complement, using the relation:

$$M := \begin{pmatrix} 1 & \tilde{v}^\top \\ \tilde{v} & \tilde{V} \end{pmatrix} \succeq 0 \iff \tilde{V} - \tilde{v}\tilde{v}^\top \succeq 0$$

Consider now the two $(n+1) \times (n+1)$ matrices constructed as follows

$$A = \begin{pmatrix} \mathbf{1}^\top K \,\mathbf{1} + 2c_j^{i\top} & \mathbf{1}^\top K + c_j^{i\top} \\ K\,\mathbf{1} + c_j^i & K \end{pmatrix} \qquad B = \begin{pmatrix} 0 & \frac{1}{2}\mathbf{1}^\top \\ \frac{1}{2}\mathbf{1} & \mathbf{0}\mathbf{0}^\top \end{pmatrix}$$

The SDP relaxation of (30) is then

$$\begin{aligned} \text{minimise } M \bullet A \quad \text{s.t. } & \operatorname{diag}(M) = \mathbf{1} \\ & B \bullet M = 2s - n \\ & M \succeq 0 \end{aligned} \tag{31}$$

($X \bullet Y \equiv \sum\sum_{i,j=1}^n X_{ij}Y_{ij}$). Note that (31) collapses to (30) when $\tilde{V} = \tilde{v}\tilde{v}^\top$ and $\tilde{v} \in \{-1,1\}^n$ are enforced. Note that if the cardinality constraint $B \bullet M = 2s - n$ is omitted, then (31) is equivalent to the classical graph partitioning problem *MAX-CUT* (Goemans and Williamson, 1995).

The SDP (31) is linear in M and is soluble to within any $\varepsilon > 0$ of the true optimum in polynomial time. Its solution M^*, however, only solves the BQP (30) if $\tilde{V}^* = \tilde{v}^*\tilde{v}^{*\top}$, or equivalently $\operatorname{rank}(M^*) = 1$. This will not be true in general and the second part of a relaxation procedure is to round the output $\tilde{v}^* \in [-1,1]^n$ to a feasible vector $\tilde{v} \in \{-1,1\}^n$ for the BQP. Goemans and Williamson (1995) introduced a popular randomised rounding approach for *MAX-CUT*, and for the following exploratory simulations we adopted a similar approach. This starts by performing an incomplete Cholesky decomposition $\tilde{V}^* = UU^\top$ with $\operatorname{rank}(U) = r$. Since $\operatorname{diag}(\tilde{V}^*) = 1$, the columns of U all lie on the unit r-sphere.

To select exactly m points we draw a random hyperplane through the origin of this sphere and translate it affinely until exactly m points are separated from the rest (it is this translation that is a modification of the original approach for non-cardinality constrained problems, and which means the analysis of Goemans and Williamson (1995) is not directly applicable). The resulting approximations are presented only as an empirical benchmark for Algorithms 1-3 and the detailed analysis of rounding procedures is well beyond the scope of this work.

We also find improved output by drawing $R > 1$ points on the r-sphere and choosing the one for which the points separated off are best, in the sense of lowest cumulative KSD. This process imposes trivial additional computational cost. The semi-definite optimisations are performed using the `Python` optimisation package `MOSEK`.

Figure 5 shows that the semi-definite relaxation approach can be competitive in time-adjusted KSD. Each line in left pane represents the drawing of 1000 samples. The non-relaxed and best-of-50 SDR approaches closely mirror each other in time-adjusted KSD, though the non-relaxed approach is more efficient in that it achieves the same KSD in the same time with fewer samples chosen. Choosing $R > 1$ imposes little additional computation time, leading to a performance improvement for $R = 50$ over $R = 10$, though past a certain point (visible here for $R = 200$) this additional computation does become significant and harms performance.

[5] The more general IQP setting, in which candidate points can be repeatedly selected, can similarly be cast as an SDP by proceeding with s copies of the candidate set and $v \in \{0,1\}^{ns}$.

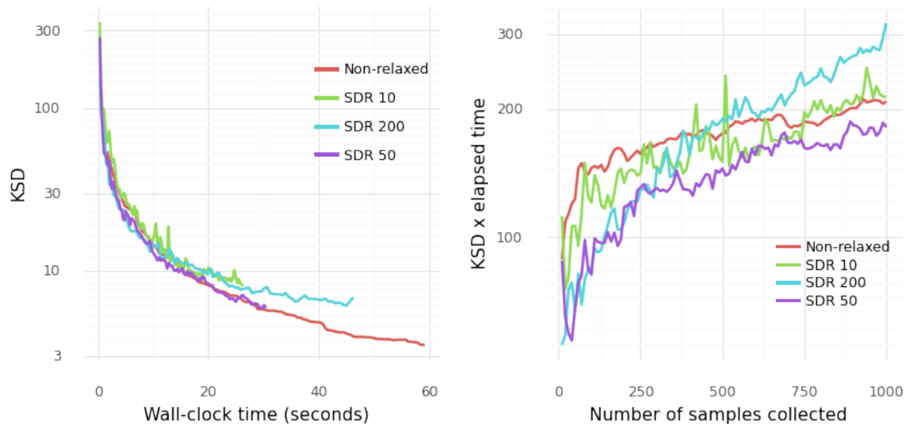

Figure 5: KSD vs. wall-clock time, and time-adjusted KSD vs. number of selected samples, for the 4-dim Lotka–Volterra model also used in Section 4, and with the same kernel specification. We draw 1000 samples using batch-size $b = 100$ and choosing $s = 10$ points simultaneously at each iteration. The four lines refer to the non-relaxed method (generated using the same code as in Figure 3), as well as the approach employing semi-definite relaxation (taking the best of 10, 50 and 200 point selections, determined by drawing that many points on the sphere).

C Choice of Kernel

As with all kernel-based methods, the specification of the kernel itself is of key importance. For the MMD experiments in Section 4.1, we employed the squared-exponential kernel $k(x, y; \ell) = \exp(-\frac{1}{2}\ell^{-2}\|x - y\|^2)$, and for the KSD experiments in Section 4.2 we followed Chen et al. (2018, 2019) and Riabiz et al. (2020) and used the inverse multi-quadric kernel $k(x, y; \ell) = (1 + \ell^{-2}\|x - y\|^2)^{-1/2}$ as the 'base kernel' k in (3) from which the compound Stein kernel k_μ is built up. The latter choice ensures that, under suitable conditions on μ, KSD controls weak convergence to μ in $\mathcal{P}(\mathbb{R}^d)$, meaning that if $\mathrm{MMD}_{\mu,k_\mu}(\nu) \to 0$ then $\nu \Rightarrow \mu$ (Gorham and Mackey, 2017, Thm. 8).

The next consideration is the length scale ℓ. There are several possible approaches. For the simulations in Sections 4.1 and 4.2, we use the median heuristic (Garreau et al., 2017). The length-scale $\hat{\ell}$ is calculated from the dataset themselves, using the formula $\hat{\ell} = \sqrt{\frac{1}{2}\mathrm{Med}\{\|x_i - x_j\|^2\}}$. The indices i, j can run over the entire dataset, or more commonly in practice, a uniformly-sampled subset of it. For the large datasets in Section 4, we use 1000 points to calculate $\hat{\ell}$.

To explore the impact of the choice of length scale on the approximations that our methods produce, in Figure 6 we start with $\tilde{\ell} = 0.25$ (the value used to produce Figure 1 in the main text) and now vary this parameter, considering $0.1\tilde{\ell}$ and $10\tilde{\ell}$. The difference in the quality of the approximation of ν to μ is immediately visually evident, even for such a simple model. It appears that, at least in this instance, the median heuristic is helpful in avoiding pathologies that can occur when an inappropriate length-scale is used.

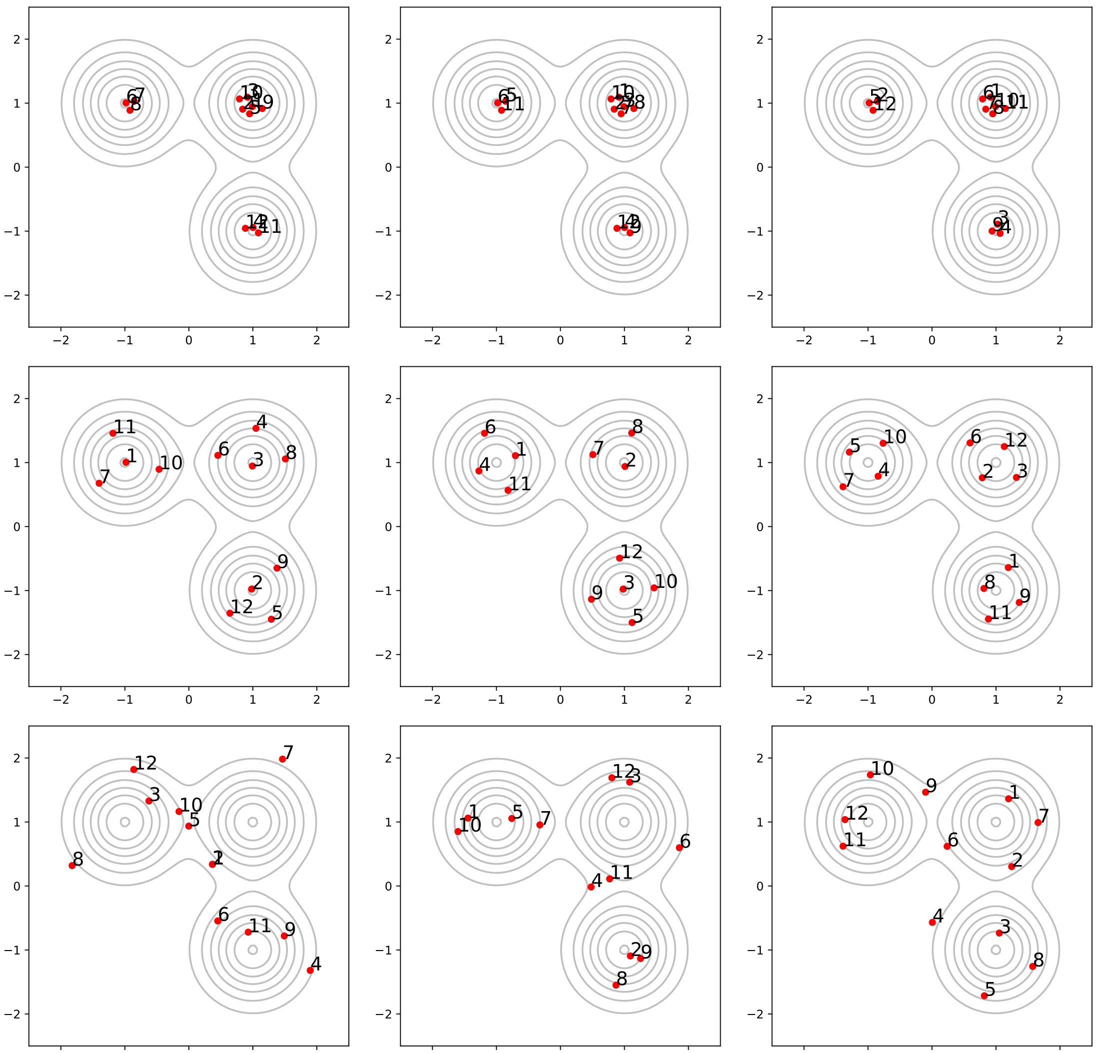

Figure 6: Investigating the role of the length-scale parameter ℓ in the squared-exponential kernel $k(x, y; \ell) = \exp(-\frac{1}{2}\ell^{-2}\|x - y\|^2)$. Model and simulation set-up as in Figure 1. Here 12 representative points were selected using the myopic method (left column), a non-myopic method (centre column), and by simultaneous selection of all 12 points (right column). The kernel length-scale parameter $\tilde{\ell}$ set to 0.025 (top row), 0.25 (middle row; as Figure 1) and 2.5 (bottom row).