GENERALIZED KERNEL THINNING

Anonymous authors
Paper under double-blind review

ABSTRACT

The kernel thinning (KT) algorithm of Dwivedi and Mackey (2021) compresses an n point distributional summary into a $\sqrt{n}$ point summary with better-than-Monte-Carlo maximum mean discrepancy for a target kernel $\mathbf{k}$ by leveraging a less smooth square-root kernel. Here we provide four improvements. First, we show that KT applied directly to the target kernel yields a tighter $\mathcal{O}(\sqrt{\log n/n})$ integration error bound for each function f in the reproducing kernel Hilbert space. This modification extends the reach of KT to any kernel—even non-smooth kernels that do not admit a square-root, demonstrates that KT is suitable even for heavy-tailed target distributions, and eliminates the exponential dimension-dependence and $(\log n)^{d/2}$ factors of standard square-root KT. Second, we show that, for analytic kernels, like Gaussian and inverse multiquadric, target kernel KT admits maximum mean discrepancy (MMD) guarantees comparable to square-root KT without the need for an explicit square-root kernel. Third, we prove KT with a fractional α-power kernel $\mathbf{k}_\alpha$ for $\alpha > 1/2$ yields better-than-Monte-Carlo MMD guarantees for non-smooth kernels, like Laplace and Matérn, that do not have square-roots. Fourth, we establish that KT applied to a sum of $\mathbf{k}$ and $\mathbf{k}_\alpha$ (a procedure we call KT+) simultaneously inherits the improved MMD guarantees of power KT and the tighter individual function guarantees of KT on the target kernel. Finally, we illustrate the practical benefits of target KT and KT+ for compression after high-dimensional independent sampling and challenging Markov chain Monte Carlo posterior inference.

1 INTRODUCTION

A core task in probabilistic inference is learning a compact representation of a probability distribution $\mathbb{P}$. This problem is usually solved by sampling n points independently from $\mathbb{P}$ or, if direct sampling is intractable, generating n points from a Markov chain converging to $\mathbb{P}$. The benefit of these approaches is that they provide asymptotically exact sample estimates $\mathbb{P}_n f \triangleq \frac{1}{n}\sum_{i=1}^n f(x_i)$ for intractable expectations $\mathbb{P}f \triangleq \mathbb{E}_{X\sim\mathbb{P}}[f(X)]$. However, they also suffer from a serious drawback: the learned representations are unnecessarily large, requiring n points to achieve $|\mathbb{P}f - \mathbb{P}_n f| = \Theta(n^{-\frac{1}{2}})$ integration error. These inefficient representations quickly become prohibitive for expensive downstream tasks and function evaluations: for example, in computational cardiology, each function evaluation $f(x_i)$ initiates a heart or tissue simulation that consumes 1000s of CPU hours (Niederer et al., 2011; Augustin et al., 2016; Strocchi et al., 2020).

To reduce the downstream computational burden, a standard practice is to *thin* the initial sample by discarding every t-th sample point (Owen, 2017). Unfortunately, standard thinning often results in a substantial loss of accuracy: for example, thinning an i.i.d. or fast-mixing Markov chain sample from n points to $n^{\frac{1}{2}}$ points increases integration error from $\Theta(n^{-\frac{1}{2}})$ to $\Theta(n^{-\frac{1}{4}})$. Our goal is to produce more accurate compressed representations with provably better-than-i.i.d. integration error.

Problem setup Given a target distribution $\mathbb{P}$ on a domain $\mathcal{X}$, a sequence of input points $\mathcal{S}_{\text{in}} = (x_i)_{i=1}^n$ in $\mathcal{X}$, and a function space $\mathcal{H}$, our first goal is to identify a thinned $\mathcal{H}$ coreset, a point subsequence $\mathcal{S}_{\text{out}}$ of length $\sqrt{n}$ with better-than-i.i.d. integration error, $|\mathbb{P}f - \mathbb{P}_{\text{out}}f| = o(n^{-\frac{1}{4}})$, for each function $f \in \mathcal{H}$ and $\mathbb{P}_{\text{out}} \triangleq \frac{1}{\sqrt{n}}\sum_{x\in\mathcal{S}_{\text{out}}} \boldsymbol{\delta}_x$. We will focus our attention on the reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ associated with a given kernel $\mathbf{k}$ (Berlinet & Thomas-Agnan,

GAUSS(σ) $\sigma > 0$	LAPLACE(σ) $\sigma > 0$	MATÉRN(ν, γ) $\nu > \frac{d}{2}, \gamma > 0$	IMQ(ν, γ) $\nu > 0, \gamma > 0$	SINC-RAD(θ) $\theta \neq 0, d \leq 3$	SINC(θ) $\theta \neq 0$	B-SPLINE ($2\beta+1$) $\beta \in \mathbb{N}$
$\exp\left(-\frac{\lVert z\rVert_2^2}{2\sigma^2}\right)$	$\exp\left(-\frac{\lVert z\rVert_2}{\sigma}\right)$	$c_{\nu-\frac{d}{2}}(\gamma\lVert z\rVert_2)^{\nu-\frac{d}{2}} \cdot K_{\nu-\frac{d}{2}}(\gamma\lVert z\rVert_2)$	$\frac{1}{(1+\lVert z\rVert_2^2/\gamma^2)^\nu}$	$\frac{\sin(\theta\lVert z\rVert_2)}{\theta\lVert z\rVert_2}$	$\prod_{j=1}^d \frac{\sin(\theta z_j)}{\theta z_j}$	$\mathfrak{B}_{2\beta+2}^{-d}\prod_{j=1}^d h_\beta(z_j)$

Table 1: Common kernels $\mathbf{k}(x, y)$ with $z = x - y$. In each case, $\lVert\mathbf{k}\rVert_\infty = 1$. Here, $c_a \triangleq \frac{2^{1-a}}{\Gamma(a)}$, K_a is the modified Bessel function of the third kind of order a (Wendland, 2004, Def. 5.10), h_β is the recursive convolution of $2\beta + 2$ copies of $\mathbf{1}_{[-\frac{1}{2},\frac{1}{2}]}$, and $\mathfrak{B}_\beta \triangleq \frac{1}{(\beta-1)!}\sum_{j=0}^{\lfloor\beta/2\rfloor}(-1)^j\binom{\beta}{j}(\frac{\beta}{2}-j)^{\beta-1}$.

2011).[1] We additionally measure the quality of our thinned representations in terms of the kernel *maximum mean discrepancy* (MMD, Gretton et al., 2012), defined as the worst-case error between sample and target expectations over the unit ball of $\mathcal{H}$ equipped with RKHS norm $\lVert\cdot\rVert_{\mathbf{k}}$:

$$\mathrm{MMD}_{\mathbf{k}}(\mathbb{P}, \mathbb{P}_{\text{out}}) \triangleq \sup_{\lVert f\rVert_{\mathbf{k}}\leq 1}|\mathbb{P}f - \mathbb{P}_{\text{out}}f|.$$

Our second goal is to produce thinned MMD coresets $\mathcal{S}_{\text{out}}$ with $\mathrm{MMD}_{\mathbf{k}}(\mathbb{P}, \mathbb{P}_{\text{out}}) = o(n^{-\frac{1}{4}})$. For reference, we collect expressions for several commonly used kernels in Tab. 1.

Kernel thinning The starting point for our work is the recent *kernel thinning* (KT) algorithm of Dwivedi & Mackey (2021) which takes as input n points and returns $n^{\frac{1}{2}}$ points with $o(n^{-\frac{1}{4}})$-MMD error thereby beating the Monte Carlo rate. To effectively control $\mathrm{MMD}_{\mathbf{k}}$, KT uses non-uniform randomness and a less smooth *square-root kernel* $\mathbf{k}_{\mathrm{rt}}$ satisfying $\mathbf{k}(x, y) = (2\pi)^d \cdot \int_{\mathbb{R}^d}\mathbf{k}_{\mathrm{rt}}(x, z)\mathbf{k}_{\mathrm{rt}}(z, y)dz$, up to an immaterial constant rescaling. When the input points are sampled i.i.d. or from a fast-mixing Markov chain on $\mathbb{R}^d$, Dwivedi & Mackey prove that the KT output has, with high probability, $\mathcal{O}_d(n^{-\frac{1}{2}}\sqrt{\log n})$-$\mathrm{MMD}_{\mathbf{k}}$ error for $\mathbb{P}$ and $\mathbf{k}_{\mathrm{rt}}$ with bounded support, $\mathcal{O}_d(n^{-\frac{1}{2}}(\log^{d+1} n \log\log n)^{\frac{1}{2}})$-$\mathrm{MMD}_{\mathbf{k}}$ error for $\mathbb{P}$ and $\mathbf{k}_{\mathrm{rt}}$ with light tails, and $(n^{\frac{1}{2}}, \mathcal{O}_d(n^{-\frac{1}{2}+\frac{d}{2\rho}}\sqrt{\log n \log\log n}))$-$\mathrm{MMD}_{\mathbf{k}}$ error for $\mathbb{P}$ and $\mathbf{k}_{\mathrm{rt}}^2$ with $\rho > 2d$ moments.

Our contributions While KT provides a superior to Monte Carlo error rate for a range of settings, the algorithm in its original form, required the knowledge of a square-root kernel $\mathbf{k}_{\mathrm{rt}}$ to run the algorithm, and establish theoretical guarantees. Here we provide several generalizations. First, we directly analyze a variant of kernel thinning which uses only the target kernel $\mathbf{k}$. To distinguish the two cases, going forward, we use ROOT KT to denote the original kernel thinning algorithm using $\mathbf{k}_{\mathrm{rt}}$ and TARGET KT to denote kernel thinning just with $\mathbf{k}$. Furthermore, we introduce and analyze two other KT variants: (i) α-POWER KT, which uses a generalized power kernel $\mathbf{k}_\alpha$ of kernel $\mathbf{k}$ (see Def. 2), and (ii) **KT+**, which uses a sum of $\mathbf{k}_\alpha$ and $\mathbf{k}$. We note that the existence of generalized power $\mathbf{k}_\alpha$ is a weaker condition than the existence of $\mathbf{k}_{\mathrm{rt}}$.

We provide several theoretical and practical improvements with kernel thinning. First, we prove that the output of target KT provides $\mathcal{O}(n^{-\frac{1}{2}}\sqrt{\log n})$ integration error for any fixed function in the RKHS of **any kernel k on any domain**. In particular, when the domain is $\mathbb{R}^d$, our single function guarantee for target KT has explicit constants and is dimension independent. Second, we show that like root KT, even target KT provides $o(n^{-\frac{1}{4}})$-$\mathrm{MMD}_{\mathbf{k}}$ error with $\sqrt{n}$ points for a wide range of settings. Third, we show that α-power KT provides $o(n^{-1/4})$-MMD errors for non-smooth kernels which do not admit a square-root kernel. Finally, when $\mathbf{k}_{\mathrm{rt}}$ exists, we show that the KT+ inherits the uniform error guarantees of root KT and the tighter individual function guarantees of target KT.

Related work For bounded $\mathbf{k}$, both i.i.d. samples (Tolstikhin et al., 2017, Prop. A.1) and thinned geometrically ergodic Markov chains (Dwivedi & Mackey, 2021, Prop. 1) deliver $n^{\frac{1}{2}}$ points with $\mathcal{O}(n^{-\frac{1}{4}})$ MMD with high probability. The *online Haar strategy* of Dwivedi et al. (2019) and low discrepancy *quasi-Monte Carlo* methods (see, e.g., Hickernell, 1998; Novak & Wozniakowski, 2010; Dick et al., 2013) provide improved $\mathcal{O}_d(n^{-\frac{1}{2}}\log^d n)$ MMD guarantees but are tailored specifically to the uniform distribution on $[0, 1]^d$. Alternative coreset constructions for more general $\mathbb{P}$ include *kernel herding* (Chen et al., 2010), *discrepancy herding* (Harvey & Samadi, 2014), *super-sampling*

[1] Common kernel choices for generating and evaluating thinned coresets include Gaussian $\mathbf{k}$ (Chen et al., 2010), Sobolev space $\mathbf{k}$ on $[0, 1]^d$ underlying the QMC L^2 discrepancy (Hickernell, 1998), Stein kernels $\mathbf{k}$ (Riabiz et al., 2020a), and energy distance $\mathbf{k}(x, y) = \lVert x - y\rVert - \lVert x\rVert - \lVert y\rVert$ (Mak & Joseph, 2018).

with a reservoir Paige et al. (2016), *support points convex-concave procedures* (Mak & Joseph, 2018), *greedy sign selection* (Karnin & Liberty, 2019, Sec. 3.1), *Stein points MCMC* (Chen et al., 2019), and *Stein thinning* (Riabiz et al., 2020a). While some admit better-than-i.i.d. MMD guarantees for finite-dimensional kernels on $\mathbb{R}^d$ (Chen et al., 2010; Harvey & Samadi, 2014), apart from KT, none are known to provide better-than-i.i.d. MMD or integration error for infinite-dimensional kernels, like the convergence-determining Gaussian, Matérn, inverse multiquadric, B-spline, and sinc kernels studied in this work. The minimax lower bounds of Phillips & Tai (2020, Thm. 3.1) and Tolstikhin et al. (2017, Thm. 1) respectively establish that any procedure outputting $n^{\frac{1}{2}}$-sized coresets and any procedure estimating $\mathbb{P}$ based only on n i.i.d. sample points must incur $\Omega(n^{-\frac{1}{2}})$ MMD in the worst case. Our guarantees in Sec. 2 match these lower bounds up to logarithmic factors.

Notation We define the norm $\|\mathbf{k}\|_\infty = \sup_{x,y} |\mathbf{k}(x,y)|$ and the shorthand $[n] \triangleq \{1,\dots,n\}$, $\mathbb{P}_{\text{in}} \triangleq \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_{x_i}$, $\mathbb{R}_+ \triangleq \{x \in \mathbb{R} : x \geq 0\}$, $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$, $\mathcal{B}_{\mathbf{k}} \triangleq \{f \in \mathcal{H} : \|f\|_{\mathbf{k}} \leq 1\}$, and $\mathcal{B}_2(r) = \left\{y \in \mathbb{R}^d : \|y\|_2 \leq r\right\}$. We write $a \precsim b$ and $a \succsim b$ to mean $a = \mathcal{O}(b)$ and $a = \Omega(b)$ and use $\precsim_d$ when masking constants dependent on d. For point sequences $\mathcal{S}, \mathcal{S}'$ with empirical distributions $\mathbb{Q}_n, \mathbb{Q}'_n$, we define $\mathrm{MMD}_{\mathbf{k}}(\mathbb{P}, \mathcal{S}) \triangleq \mathrm{MMD}_{\mathbf{k}}(\mathbb{P}, \mathbb{Q}_n)$ and $\mathrm{MMD}_{\mathbf{k}}(\mathcal{S}, \mathcal{S}') \triangleq \mathrm{MMD}_{\mathbf{k}}(\mathbb{Q}_n, \mathbb{Q}'_n)$.

2 Generalized Kernel Thinning

Our generalization of kernel thinning in Alg. 1 proceeds in two steps: KT-SPLIT and KT-SWAP. Given a thinning parameter m and a target kernel $\mathbf{k}$, KT-SPLIT divides the input sequence $\mathcal{S}_{\text{in}} = (x_i)_{i=1}^n$ into 2^m candidate coresets of size $n/2^m$ in a non-uniform manner using an auxiliary kernel $\mathbf{k}_{\text{split}}$. Next, the KT-SWAP step adds a baseline coreset to the candidate list (for example, one produced by standard thinning or uniform subsampling), selects the candidate closest to $\mathcal{S}_{\text{in}}$ in terms of $\mathrm{MMD}_{\mathbf{k}}$, and refines the selected coreset by swapping each input point into the coreset if it offers an improvement in $\mathrm{MMD}_{\mathbf{k}}$. The original variant of kernel thinning uses $\mathbf{k}_{\text{split}} = \mathbf{k}_{\text{rt}}$ (and we thus refer to it as root KT), and here we establish theoretical guarantees for more general $\mathbf{k}_{\text{split}}$ with a notable example being $\mathbf{k}_{\text{split}} = \mathbf{k}$. Other than the choice of $\mathbf{k}_{\text{split}}$ being more general, the KT-SPLIT and KT-SWAP considered here are identical to Dwivedi & Mackey (2021, Alg 1a & 1b) and reproduced for reference in App. A. Overall, our generalization does not affect the runtime or storage of kernel thinning, which for any m has the time complexity dominated by $\mathcal{O}(n^2)$ evaluations of $\mathbf{k}_{\text{split}}$ and $\mathbf{k}$, while the space complexity is $\mathcal{O}(n \min(d, n))$, achieved by storing the smaller of the input sequence $(x_i)_{i=1}^n$ and the kernel matrices $(\mathbf{k}_{\text{split}}(x_i, x_j))_{i,j=1}^n$ and $(\mathbf{k}(x_i, x_j))_{i,j=1}^n$.

Algorithm 1: Generalized Kernel Thinning – Return coreset of size $\lfloor n/2^m \rfloor$ with small $\mathrm{MMD}_{\mathbf{k}}$

Input: split kernel $\mathbf{k}_{\text{split}}$, target kernel $\mathbf{k}$, point sequence $\mathcal{S}_{\text{in}} = (x_i)_{i=1}^n$, thinning parameter $m \in \mathbb{N}$, probabilities $(\delta_i)_{i=1}^{\lfloor n/2 \rfloor}$

$(\mathcal{S}^{(m,\ell)})_{\ell=1}^{2^m} \leftarrow$ KT-SPLIT $(\mathbf{k}_{\text{split}}, \mathcal{S}_{\text{in}}, m, (\delta_i)_{i=1}^{\lfloor n/2 \rfloor})$ // Split $\mathcal{S}_{\text{in}}$ into 2^m candidate coresets of size $\lfloor \frac{n}{2^m} \rfloor$

$\mathcal{S}_{\text{KT}} \leftarrow$ KT-SWAP $(\mathbf{k}, \mathcal{S}_{\text{in}}, (\mathcal{S}^{(m,\ell)})_{\ell=1}^{2^m})$ // Select best coreset and iteratively refine

return coreset $\mathcal{S}_{\text{KT}}$ of size $\lfloor n/2^m \rfloor$

2.1 Single function guarantees for KT-SPLIT

Our first main result, proved in App. B, provides a high probability bound for the integration error of any fixed function in the RKHS $\mathcal{H}_{\text{split}}$ of kernel $\mathbf{k}_{\text{split}}$ for the output of KT-SPLIT.

Theorem 1 (Single function guarantees for KT-SPLIT) *Consider* KT-SPLIT *(Alg. 1a) with oblivious*[2] $\mathcal{S}_{\text{in}}$ *and* $(\delta_i)_{i=1}^{n/2}$ *and* $\delta^\star \triangleq \min_i \delta_i$. *If* $\frac{n}{2^m} \in \mathbb{N}$, *then, for any fixed* $f \in \mathcal{H}_{split}$, *index* $\ell \in [2^m]$, *and scalar* $\delta' \in (0,1)$, *the output coreset* $\mathcal{S}^{(m,\ell)}$ *with* $\mathbb{P}_{split}^{(\ell)} \triangleq \frac{1}{n/2^m}\sum_{x \in \mathcal{S}^{(m,\ell)}} \boldsymbol{\delta}_x$ *satisfies*

$$|\mathbb{P}_{in} f - \mathbb{P}_{split}^{(\ell)} f| \leq \|f\|_{\mathbf{k}_{split}} \cdot \sigma_m \sqrt{2\log(\tfrac{2}{\delta'})} \quad \textit{for} \quad \sigma_m \triangleq \tfrac{2}{\sqrt{3}}\tfrac{2^m}{n}\sqrt{\|\mathbf{k}_{split}\|_{\infty,\text{in}} \cdot \log(\tfrac{4}{\delta^\star})} \tag{1}$$

with probability at least $p_{\text{sg}} \triangleq 1 - \sum_{j=1}^m \sum_{i=1}^{n/2^j} \delta_i - \delta'$. *Here,* $\|\mathbf{k}_{split}\|_{\infty,\text{in}} \triangleq \max_{x \in \mathcal{S}_{\text{in}}} \mathbf{k}_{split}(x,x)$.

[2] Throughout, *oblivious* indicates that a sequence is generated independently of any randomness in KT.

Remark 1 (Guarantees for known and oblivious stopping times) To ensure that the bound (1) holds with probability at least $1-\delta$, we can set $\delta' = \frac{\delta}{2}$ and $\delta_i = \frac{\delta}{2n}$ for a stopping time n known *a priori* or $\delta_i = \frac{\delta}{4(i+1)\log^2(i+1)}$ for an arbitrary oblivious stopping time n.

Remark 2 (Target KT-SPLIT yields better-than-i.i.d. thinning) For each $f \in \mathcal{H}$, KT-SPLIT with $\mathbf{k}_{\text{split}} = \mathbf{k}$, $m = \frac{1}{2}\log_2 n$, $\delta_i = \frac{\delta}{2n}$, and $\delta' = \frac{\delta}{2}$ provides $\mathcal{O}(n^{-\frac{1}{2}}\sqrt{\log n})$ integration error,

$$|\mathbb{P}_{\text{in}} f - \mathbb{P}^{(1)}_{\text{split}} f| \leq \|f\|_{\mathbf{k}} \sqrt{\tfrac{8}{3}\|\mathbf{k}\|_{\infty,\text{in}} \log(\tfrac{4}{\delta}) \tfrac{\log(8n/\delta)}{n}},$$

with probability at least $1-\delta$ using only $\sqrt{n}$ points. When $|\mathbb{P}f - \mathbb{P}_{\text{in}} f| = \mathcal{O}(n^{-\frac{1}{2}})$, the triangle inequality implies that $|\mathbb{P}f - \mathbb{P}^{(\ell)}_{\text{split}} f| = \mathcal{O}(n^{-\frac{1}{2}}\sqrt{\log n})$, a near-quadratic improvement over the $\Omega(n^{-\frac{1}{4}})$ integration error of $\sqrt{n}$ i.i.d. points.

Remark 3 (Guarantees for any target kernel $\mathbf{k}$) Thm. 1 applies to **any** kernel defined on any space including unbounded kernels on unbounded domains (e.g., energy distance and Stein kernels); kernels with slowly decaying square roots (e.g., sinc kernels); and non-smooth kernels without square roots (e.g., Laplace, Matérn with $\gamma \in (\frac{d}{2}, d]$), and the compactly supported kernels of Wendland (2004) with $s < \frac{1}{2}(d+1)$). In contrast, the MMD guarantees of Dwivedi & Mackey covered only bounded, smooth $\mathbf{k}$ on $\mathbb{R}^d$ with bounded, Lipschitz, and rapidly-decaying square-roots.

Remark 4 (Guarantees for any target distribution $\mathbb{P}$) Thm. 1 guarantees better-than-i.i.d. integration error for **any** target distribution with $|\mathbb{P}f - \mathbb{P}_{\text{in}} f| = o(n^{-\frac{1}{4}})$. In contrast, the MMD improvements of Dwivedi & Mackey (2021, cf. Tab. 2) applied only to $\mathbb{P}$ with at least d moments.

Remark 5 (Dimension-independent bounds) When $\|\mathbf{k}\|_\infty = 1$, the MMD bounds of Dwivedi & Mackey feature exponential dimension dependence of the form c^d or $(\log n)^{d/2}$ while the Thm. 1 guarantee is dimension-free and hence practically relevant even when d is large relative to n.

Remark 6 (Consequences for KT-SPLIT with square-root kernel $\mathbf{k}_{\text{rt}}$) When a bounded square-root kernel $\mathbf{k}_{\text{rt}}$ is available, we can apply Thm. 1 with $\mathbf{k}_{\text{split}} = \mathbf{k}_{\text{rt}}$ since any function f in the RKHS of $\mathbf{k}$ also lies in the RKHS of $\mathbf{k}_{\text{rt}}$: This bound with root KT-SPLIT, when compared to the bound (1) admits a multiplicative factor of $\sqrt{\frac{\|\mathbf{k}_{\text{rt}}\|_{\infty,\text{in}}}{\|\mathbf{k}\|_{\infty,\text{in}}}} \frac{\|f\|_{\mathbf{k}_{\text{rt}}}}{\|f\|_{\mathbf{k}}}$. For the kernels explicitly analyzed in Dwivedi & Mackey (2021), we show in App. G that this factor grows exponentially in dimension for Gaussian and Matérn kernels, and is 1 for B-spline kernels so that the single function guarantee from target KT is better than root KT for all the kernels explicitly considered in the prior work.

2.2 MMD GUARANTEES FOR TARGET KT

We now establish an MMD guarantee for *target KT*, our name for generalized KT with $\mathbf{k}_{\text{split}} = \mathbf{k}$.

Definition 1 (RKHS covering number) *Consider a kernel $\mathbf{k}$ defined on $\mathbb{R}^d \times \mathbb{R}^d$ with the corresponding RKHS $\mathcal{H}$, any set $\mathcal{X} \subset \mathbb{R}^d$, and a scalar $\varepsilon > 0$. We define $\mathcal{N}_{\mathbf{k}}(\mathcal{X}, \varepsilon)$ as the minimum cardinality over all possible covers $\mathcal{C} \subset \mathcal{B}_{\mathbf{k}}$ (the unit ball of RKHS) that satisfy*

$$\mathcal{B}_{\mathbf{k}} \subseteq \textstyle\bigcup_{h \in \mathcal{C}} \{g \in \mathcal{B}_{\mathbf{k}} : \sup_{x \in \mathcal{X}} |h(x) - g(x)| \leq \varepsilon\}, \quad \textit{and} \quad \mathcal{M}_{\mathbf{k}}(\mathcal{X}, \varepsilon) \triangleq \log \mathcal{N}_{\mathbf{k}}(\mathcal{X}, \varepsilon). \tag{2}$$

With the definitions in place, our next result provides an MMD guarantee for target KT in terms of the log-covering number $\mathcal{M}_{\mathbf{k}}$. Its proof based on Thm. 1 is provided in App. C.

Theorem 2 (MMD guarantee for target KT) *Consider generalized KT (Alg. 1) with $\mathbf{k}_{split} = \mathbf{k}$, oblivious $\mathcal{S}_{\text{in}}$ and $(\delta_i)_{i=1}^{\lfloor n/2 \rfloor}$, and $\delta^\star \triangleq \min_i \delta_i$. If $\frac{n}{2^m} \in \mathbb{N}$, then for any $\delta' \in (0,1)$, the output coreset $\mathcal{S}_{\text{tKT}}$ is of size $\frac{n}{2^m}$ and satisfies*

$$\text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}_{\text{tKT}}) \leq \inf_{\varepsilon \in (0,1),\ \mathcal{S}_{\text{in}} \subset \mathcal{A}} 2\varepsilon + \frac{2^m}{n} \cdot \sqrt{\tfrac{8}{3}\|\mathbf{k}\|_{\infty,\text{in}} \log(\tfrac{4}{\delta^\star}) \cdot \left[\log(\tfrac{4}{\delta'}) + \mathcal{M}_{\mathbf{k}}(\mathcal{A}, \varepsilon)\right]} \tag{3}$$

with probability at least p_{sg}, where $\|\mathbf{k}\|_{\infty,\text{in}}$ and p_{sg} were defined in Thm. 1.

Remark 7 (Guarantee for arbitrary target $\mathbb{P}$) By the triangle inequality, $\text{MMD}_{\mathbf{k}}(\mathbb{P}, \mathcal{S}_{\text{KT}}) \leq \text{MMD}_{\mathbf{k}}(\mathbb{P}, \mathcal{S}_{\text{in}}) + \text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}_{\text{KT}})$, yielding a guarantee for any target distribution $\mathbb{P}$.

Remark 8 (Improvement over baseline thinning) As in root KT, the target KT-SWAP step ensures that, deterministically, $\mathrm{MMD}_{\mathbf{k}}(\mathcal{S}_{\mathrm{in}}, \mathcal{S}_{\mathrm{KT}}) \leq \mathrm{MMD}_{\mathbf{k}}(\mathcal{S}_{\mathrm{in}}, \mathcal{S}_{\mathrm{base}})$ for $\mathcal{S}_{\mathrm{base}}$ a baseline thinned coreset of size $\frac{n}{2^m}$. Therefore, $\mathrm{MMD}_{\mathbf{k}}(\mathbb{P}, \mathcal{S}_{\mathrm{KT}}) \leq 2\,\mathrm{MMD}_{\mathbf{k}}(\mathbb{P}, \mathcal{S}_{\mathrm{in}}) + \mathrm{MMD}_{\mathbf{k}}(\mathbb{P}, \mathcal{S}_{\mathrm{base}})$.

Remark 9 (MMD guarantee for $\sqrt{n}$-thinning) Substituting $m = \frac{1}{2}\log_2 n$, $\delta_i = \frac{\delta}{2n}$, $\delta' = \frac{\delta}{2}$, $\varepsilon = n^{-\frac{1}{2}}$, $\mathcal{A} = \mathcal{B}_2(\mathfrak{R}_{\mathrm{in}})$ for $\mathfrak{R}_{\mathrm{in}} \triangleq \max_{x \in \mathcal{S}_{\mathrm{in}}} \|x\|_2$, we find that with probability at least $1-\delta$,

$$\mathrm{MMD}_{\mathbf{k}}(\mathcal{S}_{\mathrm{in}}, \mathcal{S}_{\mathrm{tKT}}) \precsim_{\mathbf{k},d,\delta} \sqrt{\tfrac{\log n}{n} \cdot \mathcal{M}_{\mathbf{k}}(\mathcal{B}_2(\mathfrak{R}_{\mathrm{in}}), \tfrac{1}{\sqrt{n}})}. \tag{4}$$

This bound yields an MMD guarantee for any kernel $\mathbf{k}$ with a covering number bound.

In Tab. 2, we summarize the MMD rates for the target KT output for a wide range of settings under different forms of growth conditions on $\mathcal{M}_{\mathbf{k}}$ and the input point radius $\mathfrak{R}_{\mathcal{S}_{\mathrm{in}}} \triangleq \max_{x \in \mathcal{S}_{\mathrm{in}}} \|x\|_2$. We consider two different growth conditions on $\mathcal{M}_{\mathbf{k}}$ (i) **LogGrowth** ($\mathcal{M}_{\mathbf{k}}(\mathcal{B}_2(r), \varepsilon) \precsim_d r^d \log^{\omega}(\frac{1}{\varepsilon})$), and (ii) **PolyGrowth** ($\mathcal{M}_{\mathbf{k}}(\mathcal{B}_2(r), \varepsilon) \precsim_d r^d \varepsilon^{-\omega}$). We show in Props. 2 and 3 that the LogGrowth condition applies to infinitely differentiable analytic kernels, including Gaussian, inverse multiquadric, and sinc kernels, and the PolyGrowth condition applies to kernels differentiable finitely many times, including Matérn and B-spline.

Our conditions on $\mathfrak{R}_{\mathcal{S}_{\mathrm{in}}}$, arise from four forms of target distribution tail decay: (1) **Compact** ($\mathfrak{R}_{\mathcal{S}_{\mathrm{in}}} \precsim_d 1$), (2) **SubGauss** ($\mathfrak{R}_{\mathcal{S}_{\mathrm{in}}} \precsim_d \sqrt{\log n}$), (3) **SubExp** ($\mathfrak{R}_{\mathcal{S}_{\mathrm{in}}} \precsim_d \log n$), and (4) **HeavyTail** ($\mathfrak{R}_{\mathcal{S}_{\mathrm{in}}} \precsim_d n^{1/\rho}$). The first setting arises with a compactly supported $\mathbb{P}$ (e.g., on the unit cube $[0,1]^d$), and the other three settings arise in expectation and with high probability when $\mathbb{P}$ has, respectively, sub-Gaussian tails, sub-exponential tails, or ρ moments and $\mathcal{S}_{\mathrm{in}}$ is generated i.i.d. from $\mathbb{P}$.

Substituting these conditions in (4) yields the eight entries of Tab. 2. We, thus, conclude that for a LogGrowth kernel, target KT provides an $o(n^{-1/4})$ MMD error for all four settings of $\mathbb{P}$; and for a PolyGrowth kernel, it provides an $o(n^{-1/4})$ MMD error if $\omega < \frac{1}{2}$ for Compact, SubGauss, and SubExp $\mathbb{P}$, and if $\omega + 2d/\rho < 1/2$ for HeavyTail $\mathbb{P}$ with tail exponent ρ.

	Compact $\mathbb{P}$ $\mathfrak{R}_{\mathrm{in}} \precsim_d 1$	SubGauss $\mathbb{P}$ $\mathfrak{R}_{\mathrm{in}} \precsim_d \sqrt{\log n}$	SubExp $\mathbb{P}$ $\mathfrak{R}_{\mathrm{in}} \precsim_d \log n$	HeavyTail $\mathbb{P}$ $\mathfrak{R}_{\mathrm{in}} \precsim_d n^{1/\rho}$
LogGrowth $\mathbf{k}$ $\mathcal{M}_{\mathbf{k}}(\mathcal{B}_2(r), \varepsilon) \precsim_d r^d \log^{\omega}(\frac{1}{\varepsilon})$	$\sqrt{\frac{(\log n)^{\omega+1}}{n}}$	$\sqrt{\frac{(\log n)^{d+\omega+1}}{n}}$	$\sqrt{\frac{(\log n)^{2d+\omega+1}}{n}}$	$\sqrt{\frac{(\log n)^{\omega+1}}{n^{1-2d/\rho}}}$
PolyGrowth $\mathbf{k}$ $\mathcal{M}_{\mathbf{k}}(\mathcal{B}_2(r), \varepsilon) \precsim_d r^d \varepsilon^{-\omega}$	$\sqrt{\frac{\log n}{n^{1-\omega}}}$	$\sqrt{\frac{(\log n)^{d+1}}{n^{1-\omega}}}$	$\sqrt{\frac{(\log n)^{2d+1}}{n^{1-\omega}}}$	$\sqrt{\frac{\log n}{n^{1-\omega-2d/\rho}}}$

Table 2: MMD guarantees for target KT under $\mathcal{M}_{\mathbf{k}}$ growth and $\mathbb{P}$ tail decay. We report the $\mathrm{MMD}_{\mathbf{k}}(\mathcal{S}_{\mathrm{in}}, \mathcal{S}_{\mathrm{KT}})$ bound of Thm. 2 for target KT with n input points and $\sqrt{n}$ output points, up to constants depending on d, δ, δ', and $\|\mathbf{k}\|_{\infty,\mathrm{in}}$. Here $\mathfrak{R}_{\mathrm{in}} \triangleq \max_{x \in \mathcal{S}_{\mathrm{in}}} \|x\|_2$.

Next in Tab. 3, we summarize the MMD rates for target KT and root KT for all kernels displayed in Tab. 1. The results for root KT follow from Dwivedi & Mackey (2021), while the results for target KT are obtained by combining Thm. 2 with kernel-specific covering number bounds derived in Apps. I and J. Notably, Thm. 2 provides better-than-i.i.d. and better-than-root KT guarantees for a variety of kernels with slowly decaying or non-existent square-roots (e.g., IMQ with $\nu < \frac{d}{2}$, sinc, sinc-rad, and B-spline) and nearly matches known root KT guarantees for analytic kernels like Gauss and IMQ with $\nu \geq \frac{d}{2}$, even though target KT makes no explicit use of a square-root kernel.

3 Kernel Thinning+

We next introduce two new instantiations of generalized KT: (i) power KT which leverages a power kernel $\mathbf{k}_{\alpha}$ that interpolates between $\mathbf{k}$ and $\mathbf{k}_{\mathrm{rt}}$ to improve upon the MMD guarantees of target KT even when $\mathbf{k}_{\mathrm{rt}}$ is not available and (ii) KT+ which uses a sum of $\mathbf{k}$ and $\mathbf{k}_{\alpha}$ to retain both the improved MMD guarantee of $\mathbf{k}_{\alpha}$ and the superior single function guarantees of $\mathbf{k}$.

Kernel	**Target KT** $\mathrm{MMD}(\mathcal{S}_{\mathrm{in}}, \mathcal{S}_{\mathbf{tKT}}) \precsim$	**Root KT** $\mathrm{MMD}(\mathcal{S}_{\mathrm{in}}, \mathcal{S}_{\mathbf{rKT}}) \precsim$	**KT+** $\mathrm{MMD}(\mathcal{S}_{\mathrm{in}}, \mathcal{S}_{\mathbf{KT+}}) \precsim$
GAUSS(σ)	$\frac{(\log n)^{\frac{3d}{4}+1}}{\sqrt{n \cdot c_n^d}}$	$\frac{(\log n)^{\frac{d}{4}+\frac{1}{2}}\sqrt{c_n}}{\sqrt{n}}$	$\frac{(\log n)^{\frac{d}{4}+\frac{1}{2}}\sqrt{c_n}}{\sqrt{n}}$
LAPLACE(σ)	$n^{-\frac{1}{4}}$	N/A	$\left(\frac{c_n (\log n)^{1+2d(1-\alpha)}}{n}\right)^{\frac{1}{4\alpha}}$
MATÉRN(ν, γ) $\nu \in (\frac{d}{2}, d]$	$n^{-\frac{1}{4}}$	N/A	$\left(\frac{c_n (\log n)^{1+2d(1-\alpha)}}{n}\right)^{\frac{1}{4\alpha}}$
MATÉRN(ν, γ) $\nu > d$	$\sqrt{\frac{\log n}{n^{1-d/(2\nu)}}}$	$\frac{(\log n)^{\frac{d+1}{2}}\sqrt{c_n}}{\sqrt{n}}$	$\frac{(\log n)^{\frac{d+1}{2}}\sqrt{c_n}}{\sqrt{n}}$
IMQ(ν, γ) $\nu < \frac{d}{2}$	$\frac{(\log n)^{d+1}}{\sqrt{n}}$	$\min(n^{-\frac{1}{4}}, \frac{\log n}{\sqrt{n^{1-d/(4\nu)}}})$	$\frac{(\log n)^{d+1}}{\sqrt{n}}$
IMQ(ν, γ) $\nu \geq \frac{d}{2}$	$\frac{(\log n)^{d+1}}{\sqrt{n}}$	$\frac{(\log n)^{\frac{d+1}{2}}\sqrt{c_n}}{\sqrt{n}}$	$\frac{(\log n)^{\frac{d+1}{2}}\sqrt{c_n}}{\sqrt{n}}$
SINC(θ)	$\frac{(\log n)^2}{\sqrt{n}}$	$n^{-\frac{1}{4}}$	$\frac{(\log n)^2}{\sqrt{n}}$
SINC-RAD(θ) $d \leq 3$	$\frac{(\log n)^{d+1}}{\sqrt{n}}$	N/A	N/A
B-SPLINE$(2\beta+1)$ $\beta \in 2\mathbb{N}_0$	$\sqrt{\frac{\log n}{n^{2\beta/(2\beta+1)}}}$	N/A	$\sqrt{\frac{\log n}{n}}$ $(\beta \geq 2)$
B-SPLINE$(2\beta+1)$ $\beta \in 2\mathbb{N}+1$	$\sqrt{\frac{\log n}{n^{2\beta/(2\beta+1)}}}$	$\sqrt{\frac{\log n}{n}}$	$\sqrt{\frac{\log n}{n}}$

Table 3: MMD guarantees for target KT, root KT, and KT+ for commonly used kernels. For n input points and $\sqrt{n}$ thinned points, we report the MMD bounds from Thm. 2 for target KT, Thm. 4 for KT+ (with $\alpha = \frac{1}{2}$ wherever feasible), and Dwivedi & Mackey (2021, Tab. 3) for root KT. We report the rates for a SUBGAUSS $\mathbb{P}$ for the Gaussian kernel, a COMPACT $\mathbb{P}$ for the B-SPLINE kernel, and a SUBEXP $\mathbb{P}$ for all other kernels (see Tab. 2 for a definition of the various $\mathbb{P}$). Here, $c_n \triangleq \log\log n$, and the MMD error is reported up to constants depending on $\mathbf{k}, d, \delta, \alpha$. The KT+ guarantee for LAPLACE applies with $\alpha > \frac{d}{d+1}$, and for MATÉRN with $\alpha > \frac{d}{2\nu}$. The best rate is highlighted in **blue**.

3.1 POWER KERNEL THINNING

First, we define a generalization of the square-root kernel for a shift-invariant kernel. Let $\widehat{f}$ denote the generalized Fourier transform (GFT) (Wendland, 2004, Def. 8.9) of $f : \mathbb{R}^d \to \mathbb{R}^d$.

Definition 2 (α-power kernel) *For $\alpha > 0$, a kernel $\mathbf{k}_\alpha(x, y) = \kappa_\alpha(x - y)$ is an α-power kernel for $\mathbf{k}(x, y) = \kappa(x - y)$ if $\widehat{\kappa_\alpha} = \widehat{\kappa}^\alpha$.*

By Dwivedi & Mackey (2021, Prop. 2), $\mathbf{k}_{\frac{1}{2}} = \mathbf{k}_{\mathrm{rt}}$ for any shift-invariant $\mathbf{k}$. Given a power kernel $\mathbf{k}_\alpha$, we instantiate generalized KT with $\mathbf{k}_{\mathrm{split}} = \mathbf{k}_\alpha$ and call it *power KT*. To state an MMD guarantee for power KT, we need additional notation. Define the $\mathbf{k}_\alpha$ and $\mathcal{S}_{\mathrm{in}}$ tail radii:

$$\begin{aligned}
&\mathfrak{R}^\dagger_{\mathbf{k}_\alpha, n} \triangleq \min\{r : \tau_{\mathbf{k}_\alpha}(r) \leq \tfrac{\|\mathbf{k}_\alpha\|_\infty}{n}\}, \quad \text{where} \quad \tau_{\mathbf{k}_\alpha}(R) \triangleq (\sup_x \textstyle\int_{\|y\|_2 \geq R} \mathbf{k}_\alpha^2(x, x-y)dy)^{\frac{1}{2}}, \\
&\mathfrak{R}_{\mathbf{k}_\alpha, n} \triangleq \min\{r : \sup_{\|x-y\|_2 \geq r} |\mathbf{k}_\alpha(x, y)| \leq \tfrac{\|\mathbf{k}_\alpha\|_\infty}{n}\}, \\
&\mathfrak{R}_{\mathcal{S}_{\mathrm{in}}} \triangleq \max_{x \in \mathcal{S}_{\mathrm{in}}} \|x\|_2, \quad \text{and} \quad \mathfrak{R}_{\mathcal{S}_{\mathrm{in}}, \mathbf{k}_\alpha, n} \triangleq \min\left(\mathfrak{R}_{\mathcal{S}_{\mathrm{in}}}, n^{1+\frac{1}{d}}\mathfrak{R}_{\mathbf{k}_\alpha, n} + n^{\frac{1}{d}}\tfrac{\|\mathbf{k}_\alpha\|_\infty}{L_{\mathbf{k}_\alpha}}\right)
\end{aligned} \tag{5}$$

and the inflation factor

$$\mathfrak{M}_{\mathbf{k}_\alpha}(n, d, \delta, \delta', R) \triangleq 37\sqrt{\log\left(\tfrac{4}{\delta}\right)}\left[\sqrt{\log\left(\tfrac{4}{\delta'}\right)} + 5\sqrt{d\log(2 + 2\tfrac{L_{\mathbf{k}_\alpha}}{\|\mathbf{k}_\alpha\|_\infty}(\mathfrak{R}_{\mathbf{k}_\alpha, n} + R))}\right],$$

where $L_{\mathbf{k}_\alpha}$ denotes the Lipschitz-constant of $\mathbf{k}_\alpha$, i.e., $|\mathbf{k}_\alpha(x, y) - \mathbf{k}_\alpha(x, z)| \leq L_{\mathbf{k}_\alpha}\|y - z\|_2$ for all $x, y, z \in \mathbb{R}^d$. Our next result (with proof in App. D) provides an MMD guarantee for power KT.

Theorem 3 (MMD guarantee for power KT) *Consider generalizd KT (Alg. 1) with $\mathbf{k}_{split} = \mathbf{k}_\alpha$ for some $\alpha \in [\frac{1}{2}, 1]$, oblivious sequences $\mathcal{S}_{\mathrm{in}}$ and $(\delta_i)_{i=1}^{\lfloor n/2 \rfloor}$, and $\delta^\star \triangleq \min_i \delta_i$. If $\frac{n}{2^m} \in \mathbb{N}$, then for*

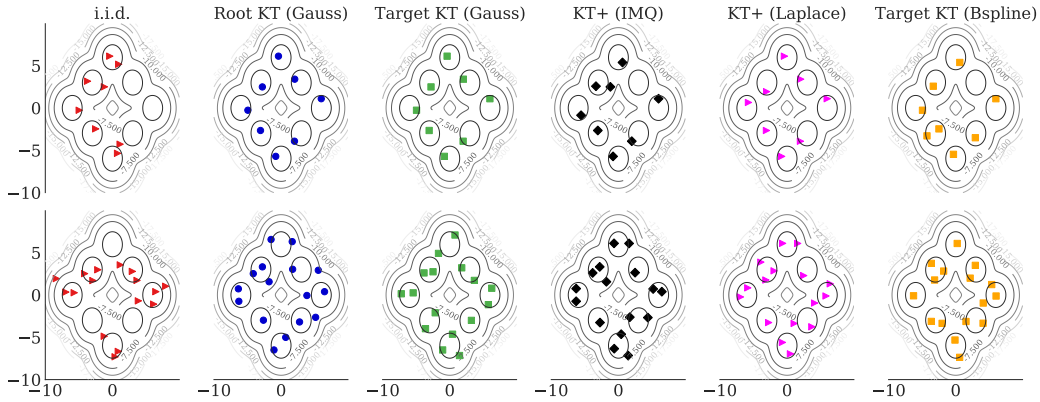

Figure 1: Generalized kernel thinning (KT) and i.i.d. coresets for various kernels $\mathbf{k}$ (in parentheses) and an 8-component mixture of Gaussian target $\mathbb{P}$ with equidensity contours underlaid.

any $\delta' \in (0,1)$, the output coreset $\mathcal{S}_{\alpha\mathrm{KT}}$ is of size $\frac{n}{2^m}$ and satisfies

$$\mathrm{MMD}_{\mathbf{k}}(\mathcal{S}_{\mathrm{in}}, \mathcal{S}_{\alpha\mathrm{KT}}) \leq \left(\tfrac{2^m}{n}\|\mathbf{k}_\alpha\|_\infty\right)^{\frac{1}{2\alpha}} (2 \cdot \widetilde{\mathfrak{M}}_\alpha)^{1-\frac{1}{2\alpha}} \left(2 + \sqrt{\tfrac{(4\pi)^{d/2}}{\Gamma(\frac{d}{2}+1)}} \cdot \mathfrak{R}_{\max}^{\frac{d}{2}} \cdot \widetilde{\mathfrak{M}}_\alpha\right)^{\frac{1}{\alpha}-1}, \tag{6}$$

with probability at least p_{sg} (defined in Thm. 1) where

$$\widetilde{\mathfrak{M}}_\alpha \triangleq \mathfrak{M}_{\mathbf{k}_\alpha}(n, d, \delta^\star, \delta', \mathfrak{R}_{\mathcal{S}_{\mathrm{in}}, \mathbf{k}_\alpha, n}) \quad \textit{and} \quad \mathfrak{R}_{\max} \triangleq \max(\mathfrak{R}_{\mathcal{S}_{\mathrm{in}}}, \mathfrak{R}^\dagger_{\mathbf{k}_\alpha, n/2^m}). \tag{7}$$

This result generalizes the root KT guarantees of Dwivedi & Mackey (2021) to power KT, and is proven via a novel interpolation result (of independent interest) for MMDs, that relates $\mathrm{MMD}_{\mathbf{k}}$ in terms of $\mathrm{MMD}_{\mathbf{k}_\alpha}$ and $\mathrm{MMD}_{\mathbf{k}_{2\alpha}}$ (Prop. 1). In fact, we can recover Dwivedi & Mackey (2021, Thm. 1) from Thm. 3 by simply substituting $\alpha = \frac{1}{2}$. For less-smooth kernels like LAPLACE and MATÉRN(ν, γ) with $\nu \in (\frac{d}{2}, d]$, $\mathbf{k}_{\mathrm{rt}}$ does not exist and our target KT guarantees are no better than the Monte Carlo rate. However, as shown in App. J, these kernels have MATÉRN kernels as their α-power kernels, which when substituted appropriately in Thm. 3 yields $o(n^{-\frac{1}{4}})$ MMD error.

3.2 KERNEL THINNING+

Our next result shows that choosing the split kernel as the scaled sum of target and power kernels, $\mathbf{k}^\dagger \triangleq \mathbf{k}/\|\mathbf{k}\|_\infty + \mathbf{k}_\alpha/\|\mathbf{k}_\alpha\|_\infty$.[3], simultaneously provides improved MMD error guarantees and dimension-free single function guarantees. See App. E for the proof.

Theorem 4 (Single function & MMD guarantees for KT+) *Consider kernel thinning (Alg. 1) with $\mathbf{k}_{split} = \mathbf{k}^\dagger$, oblivious $\mathcal{S}_{\mathrm{in}}$ and $(\delta_i)_{i=1}^{\lfloor n/2 \rfloor}$, $\delta^\star \triangleq \min_i \delta_i$, and $\frac{n}{2^m} \in \mathbb{N}$. For any fixed function $f \in \mathcal{H}$, index $\ell \in [2^m]$, and scalar $\delta' \in (0,1)$, the* KT-SPLIT *coreset $\mathcal{S}^{(m,\ell)}$ satisfies*

$$\left|\mathbb{P}_{in} f - \mathbb{P}^{(\ell)}_{split} f\right| \leq \tfrac{2^m}{n} \cdot \sqrt{\tfrac{16}{3}\log(\tfrac{4}{\delta^\star})\log(\tfrac{2}{\delta'})}\|f\|_{\mathbf{k}}\sqrt{\|\mathbf{k}\|_\infty}, \tag{8}$$

with probability at least p_{sg} (for p_{sg} and $\mathbb{P}^{(\ell)}_{split}$ defined in Thm. 1). Moreover,

$$\mathrm{MMD}_{\mathbf{k}}(\mathcal{S}_{\mathrm{in}}, \mathcal{S}_{\mathrm{KT+}}) \leq \min\left[\sqrt{2} \cdot \overline{\mathbf{M}}_{\mathrm{targetKT}}(\mathbf{k}), \quad 2^{\frac{1}{2\alpha}} \cdot \overline{\mathbf{M}}_{\mathrm{powerKT}}(\mathbf{k}_\alpha)\right] \tag{9}$$

with probability at least p_{sg} where $\overline{\mathbf{M}}_{\mathrm{targetKT}}(\mathbf{k})$ denotes the right hand side of (3) with $\|\mathbf{k}\|_{\infty,\mathrm{in}}$ replaced by $\|\mathbf{k}\|_\infty$, and $\overline{\mathbf{M}}_{\mathrm{powerKT}}(\mathbf{k}_\alpha)$ denotes the right hand side of (6).

[3]When $\mathcal{S}_{\mathrm{in}}$ is known in advance, one can alternatively choose $\mathbf{k}^\dagger \triangleq \mathbf{k}/\|\mathbf{k}\|_{\infty,\mathrm{in}} + \mathbf{k}_\alpha/\|\mathbf{k}_\alpha\|_{\infty,\mathrm{in}}$.

4 EXPERIMENTS

We now evaluate the empirical performance of generalized KT when producing a coreset of size $\sqrt{n}$ from n input points. We consider three classes of target distributions: (i) mixture of Gaussians (MoG) $\mathbb{P} = \frac{1}{M}\sum_{j=1}^{M}\mathcal{N}(\mu_j, \mathbf{I}_2)$ with M component means $\mu_j \in \mathbb{R}^2$ defined in App. H, (ii) Gaussian $\mathbb{P} = \mathcal{N}(0, \mathbf{I}_d)$ for $d \in \{2, 10, 20, 50, 100\}$, and (iii) MCMC targets based on 12 posterior inference experiments from Riabiz et al. (2020a) targeting the posteriors of four distinct coupled ordinary differential equation models: the *Goodwin (1965) model* of oscillatory enzymatic control ($d = 4$), the *Lotka (1925) model* of oscillatory predator-prey evolution ($d = 4$), the *Hinch et al. (2004) model* of calcium signalling in cardiac cells ($d = 38$), and a tempered Hinch posterior.

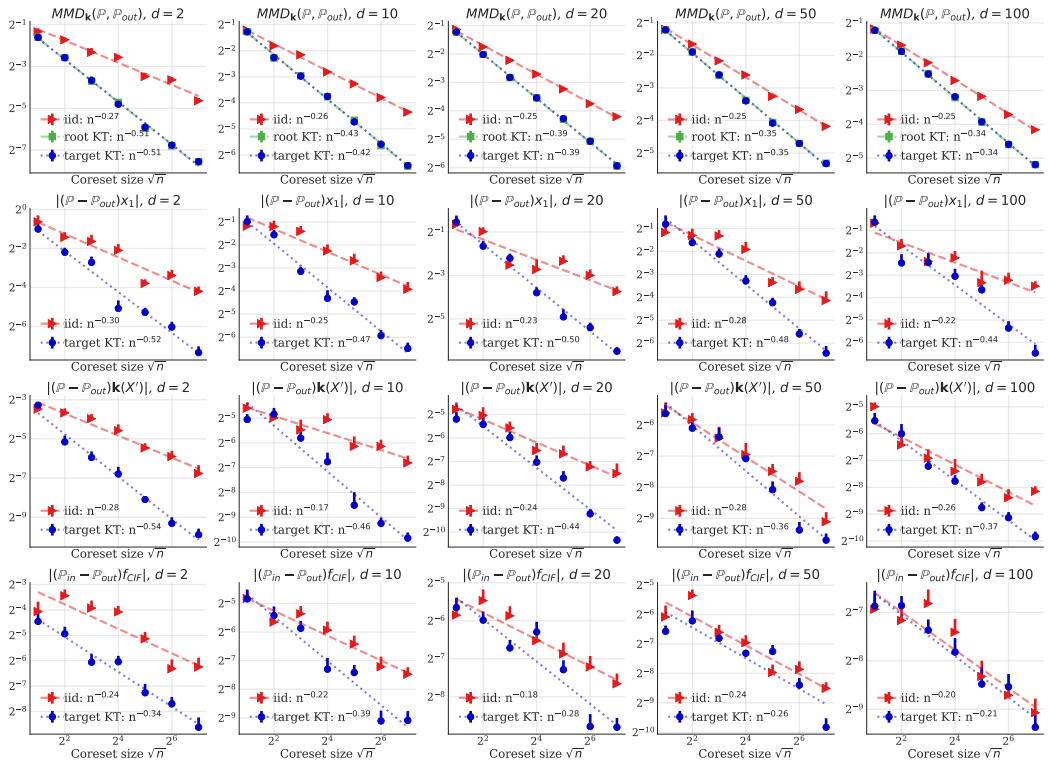

Figure 2: MMD and single-function integration error for standard Gaussian $\mathbb{P}$ in $\mathbb{R}^d$. Without using a square-root kernel, target KT matches the MMD performance of root KT and improves upon i.i.d. MMD and single-function integration error, even in $d = 100$ dimensions.

To test the ability of generalized KT to improve integration both inside and outside of $\mathcal{H}$, we evaluate integration error for (a) a random element of the target kernel RKHS ($f(x) = \mathbf{k}(X', x)$ described in App. H), (b) moments ($f(x) = x_1$ and $f(x) = x_1^2$), and (c) a standard numerical integration benchmark test function from the *continuous integrand family* (CIF, Genz, 1984), $f_{\text{CIF}}(x) = \exp(-\sum_{i=1}^{d} a_i|x_i - u_i|)$ for u_i drawn i.i.d. from uniform distribution on $[0, 1]$, and $a_i = 1/d$. We let $\mathbb{P}_{\text{out}}$ denote the empirical distribution of the output points.

For the MoG $\mathbb{P}$ we visualize in Fig. 1 the coresets from root KT with GAUSS(4), target KT with GAUSS(4) and B-SPLINE (1), KT+ for LAPLACE(2) $\mathbf{k}$ and $\mathbf{k}_{0.7}$, and KT+ with IMQ(0.5, 2) $\mathbf{k}$ and $\mathbf{k}_{0.5}$. Notably, the B-SPLINE (1) and LAPLACE kernels do not admit a square-root. In all cases, the spatial distribution of the KT output is superior to i.i.d. points and provides a better approximation of the underlying target (also see Fig. 4). This observation can be quantified via the MMD error plots in Fig. 5 in App. H, where we notice that the MMD error for all variants of KT decays more rapidly than i.i.d. with a rate close to $n^{-\frac{1}{2}}$ across all settings.

For the Gaussian targets, we use a Gaussian kernel with squared bandwidth $\sigma^2 = 2d$ and an i.i.d. input sequence $\mathcal{S}_{\text{in}}$ from $\mathbb{P}$. In line with our theory, we observe from Fig. 2 that the target and root

KT MMD errors are nearly identical and that both variants are significantly superior to i.i.d. points even in 50 and 100 dimensions with as few as 8 points. Moreover, for single function integration error, target KT provides a significant improvement over i.i.d. points in almost all settings.

Since the exact posterior $\mathbb{P}$ is unknown for the MCMC targets, we measure error with respect to the input distribution $\mathbb{P}_{\text{in}}$ (consistent with our guarantees). We employ KT+ with LAPLACE(σ) $\mathbf{k}$ and $\mathbf{k}_{0.81}$ for Goodwin and Lotka-Volterra and IMQ($0.5, \frac{1}{\sigma}$) $\mathbf{k}$ and $\mathbf{k}_{0.5}$ for Hinch. Here the kernel bandwidth σ is chosen via the popular median heuristic (see, e.g., Garreau et al., 2017): for all n, σ is set to the median distance between all pairs of points in the largest $\mathcal{S}_{\text{in}}$ tested. In Fig. 3, we plot one posterior setting from each of the Goodwin, Lotka-Volterra, and Hinch models and observe that that KT+ uniformly improves the MMD error over standard thinning in all settings. Furthermore, it provides a significantly smaller integration error for all four test functions in almost all settings. For the other 9 MCMC settings, we provide the results in Fig. 6 in App. H.‘

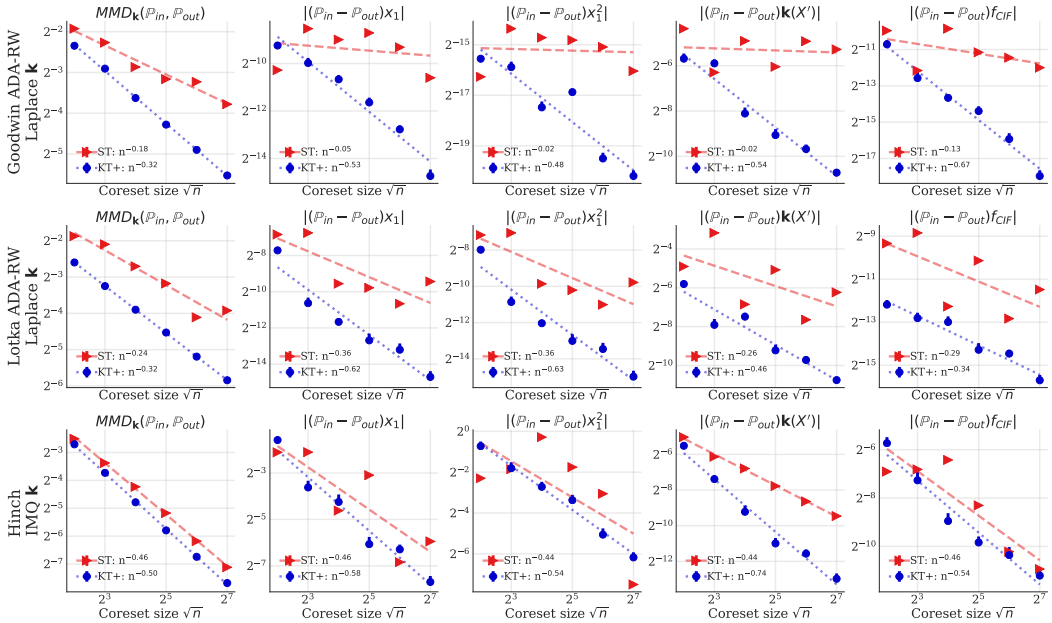

Figure 3: Kernel thinning+ (KT+) vs. standard MCMC thinning (ST). For kernels without fast-decaying square-roots, KT+ improves MMD and integration error decay rates in each posterior inference task.

5 CONCLUSION

We introduced and analyzed generalized kernel thinning, a generalization of the kernel thinning algorithm of Dwivedi & Mackey (2021) for generating better-than-i.i.d. distributional summaries. We provided the first analysis of KT applied directly on the target kernel and thereby extended the reach of KT to kernels without fast-decaying square-roots. Moreover, we established dimension-independent single function error guarantees for any kernels. We also introduced two new variants of KT, namely power KT and KT+ that provide improved MMD guarantees whenever a less-smooth power kernel if available, while simultaneously providing single function guarantees like target KT.

We highlight that while generalized KT provides unweighted coresets with better than Monte Carlo MMD, its guarantees also extend to any improved reweighting of the output points. In particular, if the downstream inferential task supports weighted coresets, one can optimally reweight $\mathbb{P}_{\text{out}}$ to approximate $\mathbb{P}_{\text{in}}$ in $\mathcal{O}(n^{\frac{3}{2}})$ time by directly minimizing the MMD. Finally, we note that like the original root KT algorithm, target KT and KT+ have a runtime of n^2 that might be prohibitive for large values of n. An important open question is whether one can substantially improve the runtime of KT without significant degradation of the MMD or single function guarantees.

REPRODUCIBILITY STATEMENT

Python code reproducing all experiments is available at **redacted**, and supplementary experimental details can be found in App. H.

REFERENCES

Christoph M Augustin, Aurel Neic, Manfred Liebmann, Anton J Prassl, Steven A Niederer, Gundolf Haase, and Gernot Plank. Anatomically accurate high resolution modeling of human whole heart electromechanics: A strongly scalable algebraic multigrid solver method for nonlinear deformation. *Journal of computational physics*, 305:622–646, 2016.

Necdet Batir. Bounds for the gamma function. *Results in Mathematics*, 72(1):865–874, 2017. doi: 10.1007/s00025-017-0698-0. URL https://doi.org/10.1007/s00025-017-0698-0.

Alain Berlinet and Christine Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Springer Science & Business Media, 2011.

Wilson Ye Chen, Alessandro Barp, François-Xavier Briol, Jackson Gorham, Mark Girolami, Lester Mackey, and Chris Oates. Stein point Markov chain Monte Carlo. In *International Conference on Machine Learning*, pp. 1011–1021. PMLR, 2019.

Yutian Chen, Max Welling, and Alex Smola. Super-samples from kernel herding. In *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, pp. 109–116, Arlington, Virginia, USA, 2010. AUAI Press. ISBN 9780974903965.

Josef Dick, Frances Y Kuo, and Ian H Sloan. High-dimensional integration: the quasi-Monte Carlo way. *Acta Numerica*, 22:133–288, 2013.

Raaz Dwivedi and Lester Mackey. Kernel thinning. *arXiv preprint arXiv:2105.05842*, 2021.

Raaz Dwivedi, Ohad N Feldheim, Ori Gurel-Gurevich, and Aaditya Ramdas. The power of online thinning in reducing discrepancy. *Probability Theory and Related Fields*, 174(1):103–131, 2019.

Damien Garreau, Wittawat Jitkrittum, and Motonobu Kanagawa. Large sample analysis of the median heuristic. *arXiv preprint arXiv:1707.07269*, 2017.

Alan Genz. Testing multidimensional integration routines. In *Proc. of international conference on Tools, methods and languages for scientific and engineering computation*, pp. 81–94, 1984.

Mark Girolami and Ben Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123–214, 2011.

Brian C Goodwin. Oscillatory behavior in enzymatic control process. *Advances in Enzyme Regulation*, 3: 318–356, 1965.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Heikki Haario, Eero Saksman, and Johanna Tamminen. Adaptive proposal distribution for random walk Metropolis algorithm. *Computational Statistics*, 14(3):375–395, 1999.

Nick Harvey and Samira Samadi. Near-optimal herding. In *Conference on Learning Theory*, pp. 1165–1182, 2014.

Fred Hickernell. A generalized discrepancy and quadrature error bound. *Mathematics of computation*, 67(221): 299–322, 1998.

Robert Hinch, JL Greenstein, AJ Tanskanen, L Xu, and RL Winslow. A simplified local control model of calcium-induced calcium release in cardiac ventricular myocytes. *Biophysical journal*, 87(6):3723–3736, 2004.

Zohar Karnin and Edo Liberty. Discrepancy, coresets, and sketches in machine learning. In *Conference on Learning Theory*, pp. 1975–1993. PMLR, 2019.

Alfred James Lotka. *Elements of physical biology*. Williams & Wilkins, 1925.

Simon Mak and V Roshan Joseph. Support points. *The Annals of Statistics*, 46(6A):2562–2592, 2018.

Steven A Niederer, Lawrence Mitchell, Nicolas Smith, and Gernot Plank. Simulating human cardiac electrophysiology on clinical time-scales. *Frontiers in Physiology*, 2:14, 2011.

E Novak and H Wozniakowski. Tractability of multivariate problems, volume ii: Standard information for functionals, european math. *Soc. Publ. House, Zürich*, 3, 2010.

Art B Owen. Statistically efficient thinning of a Markov chain sampler. *Journal of Computational and Graphical Statistics*, 26(3):738–744, 2017.

Brooks Paige, Dino Sejdinovic, and Frank Wood. Super-sampling with a reservoir. In *Proceedings of the Thirty-Second Conference on Uncertainty in Artificial Intelligence*, pp. 567–576, 2016.

Jeff M Phillips and Wai Ming Tai. Near-optimal coresets of kernel density estimates. *Discrete & Computational Geometry*, 63(4):867–887, 2020.

Marina Riabiz, Wilson Chen, Jon Cockayne, Pawel Swietach, Steven A Niederer, Lester Mackey, and Chris Oates. Optimal thinning of MCMC output. *arXiv preprint arXiv:2005.03952*, 2020a.

Marina Riabiz, Wilson Ye Chen, Jon Cockayne, Pawel Swietach, Steven A. Niederer, Lester Mackey, and Chris J. Oates. Replication Data for: Optimal Thinning of MCMC Output, 2020b. URL `https://doi.org/10.7910/DVN/MDKNWM`. Accessed on Mar 23, 2021.

Gareth O Roberts and Richard L Tweedie. Exponential convergence of Langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363, 1996.

Alessandro Rudi, Ulysse Marteau-Ferey, and Francis Bach. Finding global minima via kernel approximations. *arXiv preprint arXiv:2012.11978*, 2020.

Isaac J Schoenberg. Metric spaces and completely monotone functions. *Annals of Mathematics*, pp. 811–841, 1938.

Michael L Stein. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

Ingo Steinwart and Simon Fischer. A closer look at covering number bounds for Gaussian kernels. *Journal of Complexity*, 62:101513, 2021.

Marina Strocchi, Matthias AF Gsell, Christoph M Augustin, Orod Razeghi, Caroline H Roney, Anton J Prassl, Edward J Vigmond, Jonathan M Behar, Justin S Gould, Christopher A Rinaldi, Martin J Bishop, Gernot Plank, and Steven A Niederer. Simulating ventricular systolic motion in a four-chamber heart model with spatially varying robin boundary conditions to model the effect of the pericardium. *Journal of Biomechanics*, 101:109645, 2020.

Hong-Wei Sun and Ding-Xuan Zhou. Reproducing kernel Hilbert spaces associated with analytic translation-invariant Mercer kernels. *Journal of Fourier Analysis and Applications*, 14(1):89–101, 2008.

Ilya Tolstikhin, Bharath K Sriperumbudur, and Krikamol Muandet. Minimax estimation of kernel mean embeddings. *The Journal of Machine Learning Research*, 18(1):3002–3048, 2017.

Vito Volterra. Variazioni e fluttuazioni del numero d'individui in specie animali conviventi. 1926.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.

Holger Wendland. *Scattered data approximation*, volume 17. Cambridge university press, 2004.

Haizhang Zhang and Liang Zhao. On the inclusion relation of reproducing kernel hilbert spaces. *Analysis and Applications*, 11(02):1350014, 2013.

Ding-Xuan Zhou. Capacity of reproducing kernel spaces in learning theory. *IEEE Transactions on Information Theory*, 49(7):1743–1752, 2003.

APPENDIX

A Details of KT-SPLIT and KT-SWAP 12

B Proof of Thm. 1: Single function guarantees for KT-SPLIT 13

C Proof of Thm. 2: MMD guarantee for target KT 13

D Proof of Thm. 3: MMD guarantee for power KT 14

E Proof of Thm. 4: Single function & MMD guarantees for KT+ 15

F Proof of Prop. 1: An interpolation result for MMD 16

G Sub-optimality of single function guarantees with root KT 16

H Additional experimental results 17

I Upper bounds on RKHS covering numbers 20

J Proof of Tab. 3 results 25

A DETAILS OF KT-SPLIT AND KT-SWAP

Algorithm 1a: KT-SPLIT – Divide points into candidate coresets of size $\lfloor n/2^m \rfloor$

Input: kernel $\mathbf{k}_{\text{split}}$, point sequence $\mathcal{S}_{\text{in}} = (x_i)_{i=1}^n$, thinning parameter $m \in \mathbb{N}$, probabilities $(\delta_i)_{i=1}^{\lfloor n/2 \rfloor}$

$\mathcal{S}^{(j,\ell)} \leftarrow \{\}$ for $j = 0, \dots, m$ and $\ell = 1, \dots, 2^j$ // Empty coresets: $\mathcal{S}^{(j,\ell)}$ has size $\lfloor i/2^j \rfloor$ after round i

$\sigma_{j,\ell} \leftarrow 0$ for $j = 1, \dots, m$ and $\ell = 1, \dots, 2^{j-1}$ // Swapping parameters

for $i = 1, \dots, n$ **do**

  $\mathcal{S}^{(0,1)}$. append(x_i)

  // Every 2^j iterations, add one point from each parent coreset $\mathcal{S}^{(j-1,\ell)}$ to each child $\mathcal{S}^{(j,2\ell-1)}$, $\mathcal{S}^{(j,2\ell)}$

  for ($j = 1;\ j \leq m$ **and** $i/2^j \in \mathbb{N};\ j = j+1$) **do**

    for $\ell = 1, \dots, 2^{j-1}$ **do**

      $(\mathcal{S}, \mathcal{S}') \leftarrow (\mathcal{S}^{(j-1,\ell)}, \mathcal{S}^{(j,2\ell-1)})$; $(x, x') \leftarrow$ get_last_two_points($\mathcal{S}$)

      $(p_{\text{swap}}, \sigma_{j,\ell}) \leftarrow$ **Update_params** ($\mathbf{k}_{split}, \mathcal{S}, \mathcal{S}', x, x', \sigma_{j,\ell}, \delta_{i/2^j}$)

      $(x, x') \leftarrow (x', x)$ *with probability* p_{swap}

      $\mathcal{S}^{(j,2\ell-1)}$. append(x); $\mathcal{S}^{(j,2\ell)}$. append(x')

    end

  end

end

return $(\mathcal{S}^{(m,\ell)})_{\ell=1}^{2^m}$, candidate coresets of size $\lfloor n/2^m \rfloor$

Update_params ($\mathbf{k}_{split}, \mathcal{S}, \mathcal{S}', x, x', \sigma, \delta$) *:*

  $\mathfrak{b}^2 \leftarrow \mathbf{k}_{\text{split}}(x,x) + \mathbf{k}_{\text{split}}(x',x') - 2\mathbf{k}_{\text{split}}(x,x')$; $\mathfrak{a} \leftarrow \max(\mathfrak{b}\sigma\sqrt{2\log(4/\delta)}, \mathfrak{b}^2)$

  $\sigma \leftarrow \sigma + \mathfrak{b}^2(1 + (\mathfrak{b}^2 - 2\mathfrak{a})\sigma^2/\mathfrak{a}^2)_+$

  $\alpha \leftarrow \mathbf{k}(x',x') - \mathbf{k}(x,x) + \Sigma_{y \in \mathcal{S}}(\mathbf{k}_{\text{split}}(y,x) - \mathbf{k}(y,x')) - 2\Sigma_{z \in \mathcal{S}'}(\mathbf{k}(z,x) - \mathbf{k}(z,x'))$

  $p \leftarrow \min(1, \frac{1}{2}(1 - \alpha/\mathfrak{a})_+)$

  return (p, σ)

Algorithm 1b: KT-SWAP − Identify and refine the best candidate coreset

Input: kernel $\mathbf{k}$, point sequence $\mathcal{S}_{\text{in}} = (x_i)_{i=1}^n$, candidate coresets $(\mathcal{S}^{(m,\ell)})_{\ell=1}^{2^m}$

$\mathcal{S}^{(m,0)} \leftarrow \texttt{baseline_thinning}(\mathcal{S}_{\text{in}}, \texttt{size} = \lfloor n/2^m \rfloor)$ // Compare to baseline

$\mathcal{S}_{\text{KT}} \leftarrow \mathcal{S}^{(m,\ell^\star)}$ for $\ell^\star \leftarrow \text{argmin}_{\ell \in \{0,1,\ldots,2^m\}} \text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}^{(m,\ell)})$ // Select best candidate coreset

for $i = 1, \ldots, \lfloor n/2^m \rfloor$ **do**

  $\mathcal{S}_{\text{KT}}[i] \leftarrow \text{argmin}_{z \in \mathcal{S}_{\text{in}}} \text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}_{\text{KT}} \text{ with } \mathcal{S}_{\text{KT}}[i] = z)$

  // Swap out each point in $\mathcal{S}_{\text{KT}}$ for best alternative in $\mathcal{S}_{\text{in}}$

end

return $\mathcal{S}_{\text{KT}}$, refined coreset of size $\lfloor n/2^m \rfloor$

B Proof of Thm. 1: Single function guarantees for KT-SPLIT

We prove the theorem for the case $\ell = 1$, and for other indices, the proof can be derived by repeating the arguments. Define

$$\widetilde{\mathcal{W}}_m \triangleq \mathcal{W}_{1,m} = \mathbb{P}_{\text{in}}\mathbf{k} - \mathbb{P}_{\text{out}}^{(1)}\mathbf{k} = \tfrac{1}{n}\sum_{x \in \mathcal{S}_{\text{in}}} \mathbf{k}(x, \cdot) - \tfrac{1}{n/2^m}\sum_{x \in \mathcal{S}^{(m,1)}} \mathbf{k}(x, \cdot).$$

Next, we use the results about an intermediate algorithm, kernel halving (Dwivedi & Mackey, 2021, Alg. 3) that was introduced for the analysis of kernel thinning. Using the arguments from Dwivedi & Mackey (2021, Sec. 5.2), we conclude that KT-SPLIT with $\mathbf{k}_{\text{split}}$ set as $\mathbf{k}$ and thinning parameter m, is equivalent to repeated kernel halving with kernel $\mathbf{k}$ for m rounds (with no `Failure` in any rounds of kernel halving). On this event of equivalence, denoted by $\mathcal{E}_{\text{equi}}$, Dwivedi & Mackey (2021, Eqns. (50, 51)) imply that the function $\widetilde{\mathcal{W}}_m \in \mathcal{H}$ is equal in distribution to another random function $\mathcal{W}_m$, where $\mathcal{W}_m$ is unconditionally sub-Gaussian with parameter

$$\sigma_m = \tfrac{2}{\sqrt{3}}\tfrac{2^m}{n}\sqrt{\|\mathbf{k}\|_\infty \log(\tfrac{4}{\delta^\star})},$$

that is,

$$\mathbb{E}[\exp(\langle \mathcal{W}_m, f \rangle_{\mathbf{k}})] \leq \exp(\tfrac{1}{2}\sigma_m^2 \|f\|_{\mathbf{k}}^2) \quad \text{for all} \quad f \in \mathcal{H}, \tag{10}$$

where we note that the analysis of Dwivedi & Mackey (2021) remains unaffected when we replace $\|\mathbf{k}\|_\infty$ by $\|\mathbf{k}\|_{\infty,\text{in}}$ in all the arguments. Applying the sub-Gaussian Hoeffding inequality (Wainwright, 2019, Prop. 2.5) along with (10), we obtain that

$$\mathbb{P}[|\langle \mathcal{W}_m, f \rangle_{\mathbf{k}}| > t] \leq 2\exp(-\tfrac{1}{2}t^2/(\sigma_m^2\|f\|_{\mathbf{k}}^2)) \leq \delta' \text{ for } t \triangleq \sigma_m \|f\|_{\mathbf{k}}\sqrt{2\log(\tfrac{2}{\delta'})}.$$

Call this event $\mathcal{E}_{\text{sg}}$. As noted above, conditional to the event $\mathcal{E}_{\text{equi}}$, we also have

$$\mathcal{W}_m \overset{d}{=} \widetilde{\mathcal{W}}_m \quad \Longrightarrow \quad \langle \mathcal{W}_m, f \rangle_{\mathbf{k}} \overset{d}{=} \mathbb{P}_{\text{in}} f - \mathbb{P}_{\text{out}}^{(1)} f,$$

where $\overset{d}{=}$ denotes equality in distribution. Furthermore, Dwivedi & Mackey (2021, Eqn. 48) implies that

$$\mathbb{P}(\mathcal{E}_{\text{equi}}) \geq 1 - \textstyle\sum_{j=1}^m \sum_{i=1}^{n/2^j} \delta_i.$$

Putting the pieces together, we have

$$\mathbb{P}[|\mathbb{P}_{\text{in}} f - \mathbb{P}_{\text{out}}^{(1)} f| \leq t] \geq \mathbb{P}(\mathcal{E}_{\text{equi}} \cap \mathcal{E}_{\text{sg}}^c) \geq \mathbb{P}(\mathcal{E}_{\text{equi}}) - \mathbb{P}(\mathcal{E}_{\text{sg}}) \geq 1 - \textstyle\sum_{j=1}^m \sum_{i=1}^{n/2^j} \delta_i - \delta' = p_{\text{sg}},$$

as claimed. The proof is now complete.

C Proof of Thm. 2: MMD guarantee for target KT

First, we note that by design, KT-SWAP ensures

$$\text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}_{\text{KT}}) \leq \text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}^{(m,1)}),$$

where $\mathcal{S}^{(m,1)}$ denotes the first coreset returned by KT-SPLIT. Thus it suffices to show that $\text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}^{(m,1)})$ is bounded by the term stated on the right hand side of (3). Let $\mathbb{P}_{\text{out}}^{(1)} \triangleq \frac{1}{n/2^m}\sum_{x \in \mathcal{S}^{(m,1)}} \boldsymbol{\delta}_x$. By design of KT-SPLIT, $\text{supp}(\mathbb{P}_{\text{out}}^{(1)}) \subseteq \text{supp}(\mathbb{P}_{\text{in}})$. Recall the set $\mathcal{A}$ is such that $\text{supp}(\mathbb{P}_{\text{in}}) \subseteq \mathcal{A}$.

Proof of (3) Let $\mathcal{C} \triangleq \mathcal{C}_{\mathbf{k},\varepsilon}(\mathcal{A})$ denote the cover of minimum cardinality satisfying (2). Fix any $f \in \mathcal{B}_{\mathbf{k}}$. By the triangle inequality and the covering property (2) of $\mathcal{C}$, we have

$$\begin{aligned}\left|(\mathbb{P}_{\text{in}} - \mathbb{P}_{\text{out}}^{(1)})f\right| &\leq \inf_{g\in\mathcal{C}}\left|(\mathbb{P}_{\text{in}} - \mathbb{P}_{\text{out}}^{(1)})(f-g)\right| + \left|(\mathbb{P}_{\text{in}} - \mathbb{P}_{\text{out}}^{(1)})(g)\right| \\ &\leq \inf_{g\in\mathcal{C}}|\mathbb{P}_{\text{in}}(f-g)| + \left|\mathbb{P}_{\text{out}}^{(1)}(f-g)\right| + \sup_{g\in\mathcal{C}}\left|(\mathbb{P}_{\text{in}} - \mathbb{P}_{\text{out}}^{(1)})(g)\right| \\ &\leq \inf_{g\in\mathcal{C}} 2\sup_{x\in\mathcal{A}}|f(x)-g(x)| + \sup_{g\in\mathcal{C}}\left|(\mathbb{P}_{\text{in}} - \mathbb{P}_{\text{out}}^{(1)})(g)\right| \\ &\leq 2\varepsilon + \sup_{g\in\mathcal{C}}\left|(\mathbb{P}_{\text{in}} - \mathbb{P}_{\text{out}}^{(1)})(g)\right|. \end{aligned} \tag{11}$$

Applying Thm. 1, we have

$$\left|(\mathbb{P}_{\text{in}} - \mathbb{P}_{\text{out}}^{(1)})(g)\right| \leq \tfrac{2^m}{n}\|g\|_{\mathbf{k}}\sqrt{\tfrac{8}{3}\|\mathbf{k}\|_{\infty,\text{in}}\cdot\log(\tfrac{4}{\delta^\star})\log(\tfrac{4}{\delta'})} \tag{12}$$

with probability at least $1-\delta'-\sum_{j=1}^{m}\sum_{i=1}^{n/2^j}\delta_i = p_{\text{sg}} - \delta'$. A standard union bound then yields that

$$\sup_{g\in\mathcal{C}}\left|(\mathbb{P}_{\text{in}} - \mathbb{P}_{\text{out}}^{(1)})(g)\right| \leq \tfrac{2^m}{n}\sup_{g\in\mathcal{C}}\|g\|_{\mathbf{k}}\sqrt{\tfrac{8}{3}\|\mathbf{k}\|_{\infty,\text{in}}\cdot\log(\tfrac{4}{\delta^\star})\left[\log|\mathcal{C}| + \log(\tfrac{4}{\delta'})\right]}$$

probability at least $p_{\text{sg}} - \delta'$. Since $f \in \mathcal{B}_{\mathbf{k}}$ was arbitrary, and $\mathcal{C} \subset \mathcal{B}_{\mathbf{k}}$ and thus $\sup_{g\in\mathcal{C}}\|g\|_{\mathbf{k}} \leq 1$, we therefore have

$$\begin{aligned}\text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}^{(m,1)}) &= \sup_{\|f\|_{\mathbf{k}}\leq 1}\left|(\mathbb{P}_{\text{in}} - \mathbb{P}_{\text{out}}^{(1)})f\right| \overset{(11)}{\leq} 2\varepsilon + \sup_{g\in\mathcal{C}}\left|(\mathbb{P}_{\text{in}} - \mathbb{P}_{\text{out}}^{(1)})(g)\right| \\ &\leq 2\varepsilon + \sqrt{\tfrac{8\|\mathbf{k}\|_\infty}{3}}\cdot\tfrac{2^m}{n}\sqrt{\log(\tfrac{4}{\delta^\star})\left[\log|\mathcal{C}| + \log(\tfrac{4}{\delta'})\right]},\end{aligned}$$

with probability at least $p_{\text{sg}} - \delta'$ as claimed.

D Proof of Thm. 3: MMD guarantee for power KT

KT-SWAP step ensures that

$$\text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}_{\alpha\text{KT}}) \leq \text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}_{\alpha}^{(m,1)}),$$

where $\mathcal{S}_{\alpha}^{(m,1)}$ denotes the first coreset output by KT-SPLIT with $\mathbf{k}_{\text{split}} = \mathbf{k}_\alpha$. Next, we state a key interpolation result for $\text{MMD}_{\mathbf{k}}$ that relates it to the MMD of its power kernels (Def. 2) (see App. F for the proof).

Proposition 1 (An interpolation result for MMD) *Consider a shift-invariant kernel* $\mathbf{k}$ *that admits valid* α *and* 2α*-power kernels* $\mathbf{k}_\alpha$ *and* $\mathbf{k}_{2\alpha}$ *respectively for some* $\alpha \in [\frac{1}{2}, 1]$. *Then for any two discrete measures* $\mathbb{P}$ *and* $\mathbb{Q}$ *supported on finitely many points, we have*

$$\text{MMD}_{\mathbf{k}}(\mathbb{P},\mathbb{Q}) \leq (\text{MMD}_{\mathbf{k}_\alpha}(\mathbb{P},\mathbb{Q}))^{2-\frac{1}{\alpha}}\cdot(\text{MMD}_{\mathbf{k}_{2\alpha}}(\mathbb{P},\mathbb{Q}))^{\frac{1}{\alpha}-1}. \tag{13}$$

Given Prop. 1, it remains to establish suitable upper bounds on MMDs of $\mathbf{k}_\alpha$ and $\mathbf{k}_{2\alpha}$. To this end, first we note that for any reproducing kernel $\mathbf{k}$ and any two distributions $\mathbb{P}$ and $\mathbb{Q}$, Hölder's inequality implies that

$$\begin{aligned}\text{MMD}_{\mathbf{k}}^2(\mathbb{P},\mathbb{Q}) = \|(\mathbb{P}-\mathbb{Q})\mathbf{k}\|_{\mathbf{k}}^2 = (\mathbb{P}-\mathbb{Q})(\mathbb{P}-\mathbb{Q})\mathbf{k} &\leq \|\mathbb{P}-\mathbb{Q}\|_1\|(\mathbb{P}-\mathbb{Q})\mathbf{k}\|_\infty \\ &\leq 2\|(\mathbb{P}-\mathbb{Q})\mathbf{k}\|_\infty.\end{aligned}$$

Now, let $\mathbb{P}_{\text{in}}$ and $\mathbb{P}_{\alpha}^{(m,1)}$ denote the empirical distributions of $\mathcal{S}_{\text{in}}$ and $\mathcal{S}_{\alpha}^{(m,1)}$. Now applying Dwivedi & Mackey (2021, Thm. 4(b)), we find that

$$\text{MMD}_{\mathbf{k}_\alpha}(\mathcal{S}_{\text{in}}, \mathcal{S}_{\alpha}^{(m,1)}) \leq \sqrt{2\|(\mathbb{P}_{\text{in}} - \mathbb{P}_{\alpha}^{(m,1)})\mathbf{k}_\alpha\|_{\infty,\text{in}}} \leq \sqrt{2\cdot\tfrac{2^m}{n}\|\mathbf{k}_\alpha\|_{\infty,\text{in}}\widetilde{\mathfrak{M}}_{\mathbf{k}_\alpha}} \tag{14}$$

with probability $p_{\text{sg}} - \delta'$, where $\widetilde{\mathfrak{M}}_{\mathbf{k}_\alpha}$ was defined in (7). We note that while Dwivedi & Mackey (2021, Thm. 4(b)) uses $\|\mathbf{k}_\alpha\|_\infty$ in their bounds, we can replace it by $\|\mathbf{k}_\alpha\|_{\infty,\text{in}}$, and verifying that

all the steps of the proof continue to be valid (noting that $\|\mathbf{k}_\alpha\|_{\infty,\text{in}}$ is deterministic given $\mathcal{S}_{\text{in}}$). Furthermore, Dwivedi & Mackey (2021, Thm. 4(b)) yields that

$$\text{MMD}_{\mathbf{k}_{2\alpha}}(\mathcal{S}_{\text{in}}, \mathcal{S}_\alpha^{(m,1)}) \leq \tfrac{2^m}{n}\|\mathbf{k}_\alpha\|_{\infty,\text{in}}\left(2+\sqrt{\tfrac{(4\pi)^{d/2}}{\Gamma(\frac{d}{2}+1)}}\cdot \mathfrak{R}_{\max}^{\frac{d}{2}}\cdot \widetilde{\mathfrak{M}}_\alpha\right), \tag{15}$$

with probability $p_{\text{sg}} - \delta'$, where we have once again replaced the term $\|\mathbf{k}_\alpha\|_\infty$ with $\|\mathbf{k}_\alpha\|_{\infty,\text{in}}$ for the same reasons as stated above. We note that the two bounds (14) and (15) apply under the same high probability event as noted in Dwivedi & Mackey (2021, proof of Thm. 1, eqn. (18)). Putting together the pieces, we find that

$$\begin{aligned}
\text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}_\alpha^{(m,1)}) &\overset{(13)}{\leq} (\text{MMD}_{\mathbf{k}_\alpha}(\mathcal{S}_{\text{in}}, \mathcal{S}_\alpha^{(m,1)})^{2-\frac{1}{\alpha}}\cdot(\text{MMD}_{\mathbf{k}_{2\alpha}}(\mathcal{S}_{\text{in}}, \mathcal{S}_\alpha^{(m,1)}))^{\frac{1}{\alpha}-1} \\
&\overset{(14,15)}{\leq} \left[2\cdot\tfrac{2^m}{n}\|\mathbf{k}_\alpha\|_{\infty,\text{in}}\widetilde{\mathfrak{M}}_\alpha\right]^{1-\frac{1}{2\alpha}}\left[\tfrac{2^m}{n}\|\mathbf{k}_\alpha\|_{\infty,\text{in}}\left(2+\sqrt{\tfrac{(4\pi)^{d/2}}{\Gamma(\frac{d}{2}+1)}}\cdot\mathfrak{R}_{\max}^{\frac{d}{2}}\cdot\widetilde{\mathfrak{M}}_\alpha\right)\right]^{\frac{1}{\alpha}-1} \\
&= \left(\tfrac{2^m}{n}\|\mathbf{k}_\alpha\|_{\infty,\text{in}}\right)^{\frac{1}{2\alpha}}(2\cdot\widetilde{\mathfrak{M}}_\alpha)^{1-\frac{1}{2\alpha}}\left(2+\sqrt{\tfrac{(4\pi)^{d/2}}{\Gamma(\frac{d}{2}+1)}}\cdot\mathfrak{R}_{\max}^{\frac{d}{2}}\cdot\widetilde{\mathfrak{M}}_\alpha\right)^{\frac{1}{\alpha}-1},
\end{aligned}$$

as claimed. The proof is now complete.

E Proof of Thm. 4: Single function & MMD guarantees for KT+

Proof of part (8) First, we note that the RKHS $\mathcal{H}$ of $\mathbf{k}$ is contained in the RKHS $\mathcal{H}^\dagger$ of $\mathbf{k}^\dagger$ Berlinet & Thomas-Agnan (2011, Thm. 5). Now, applying Thm. 1 with $\mathbf{k}_{\text{split}} = \mathbf{k}^\dagger$ for any fixed function $f \in \mathcal{H} \subset \mathcal{H}^\dagger$ and $\delta' \in (0,1)$, we obtain that

$$\begin{aligned}
\left|\mathbb{P}_{\text{in}} f - \mathbb{P}_{\text{split}}^{(\ell)} f\right| &\leq \|f\|_{\mathbf{k}^\dagger}\cdot\tfrac{2}{\sqrt{3}}\tfrac{2^m}{n}\sqrt{\|\mathbf{k}^\dagger\|_{\infty,\text{in}}\cdot\log(\tfrac{4}{\delta^\star})}\sqrt{2\log(\tfrac{2}{\delta'})} \\
&\overset{(i)}{\leq} \|f\|_{\mathbf{k}^\dagger}\cdot\tfrac{2^m}{n}\sqrt{\tfrac{16}{3}\log(\tfrac{4}{\delta^\star})\log(\tfrac{2}{\delta'})}, \\
&\overset{(ii)}{\leq} \|f\|_{\mathbf{k}}\cdot\tfrac{2^m}{n}\sqrt{\tfrac{16}{3}\|\mathbf{k}\|_\infty\log(\tfrac{4}{\delta^\star})\log(\tfrac{2}{\delta'})},
\end{aligned}$$

with probability at least $p_{\text{sg}} - \delta'$. Here step (i) follows from the inequality $\|\mathbf{k}^\dagger\|_\infty \leq 2$, and step (ii) follows from the inequality $\|f\|_{\mathbf{k}^\dagger} \leq \sqrt{\|\mathbf{k}\|_\infty}\|f\|_{\mathbf{k}}$, which in turn follows from the standard facts that

$$\|f\|_{\lambda\mathbf{k}} \overset{(i)}{=} \tfrac{\|f\|_{\mathbf{k}}}{\sqrt{\lambda}}, \quad \text{and} \quad \|f\|_{\lambda\mathbf{k}+\mathbf{k}_\alpha} \overset{(ii)}{\leq} \|f\|_{\lambda\mathbf{k}} \quad \text{for} \quad \lambda > 0, f \in \mathcal{H},$$

see, e.g., Zhang & Zhao (2013, Proof of Prop. 2.5) for a proof of step (i), Berlinet & Thomas-Agnan (2011, Thm. 5) for step (ii). The proof for the bound (8) is now complete.

Proof of part (9) Repeating the proof of Thm. 2 with the bound (12) replaced by (8) yields that

$$\begin{aligned}
\text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}_{\text{KT+}}) &\leq \inf_{\varepsilon, \mathcal{S}_{\text{in}}\subset\mathcal{A}} 2\varepsilon + \tfrac{2^m}{n}\sqrt{\tfrac{16}{3}\|\mathbf{k}\|_\infty\log(\tfrac{4}{\delta^\star})\cdot\left[\log(\tfrac{4}{\delta'})+\mathcal{M}_{\mathbf{k}}(\mathcal{A},\varepsilon)\right]} \\
&\leq \sqrt{2}\cdot\overline{\mathbf{M}}_{\text{targetKT}}(\mathbf{k})
\end{aligned} \tag{16}$$

with probability at least p_{sg}. Let us denote this event by $\mathcal{E}_1$.

To establish the other bound, first we note that KT-SWAP step ensures that

$$\text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}_{\text{KT+}}) \leq \text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}_{\text{KT+}}^{(m,1)}), \tag{17}$$

where $\mathcal{S}_{\text{KT+}}^{(m,1)}$ denotes the first coreset output by KT-SPLIT with $\mathbf{k}_{\text{split}} = \mathbf{k}^\dagger$. We can now repeat the proof of Thm. 3, using the sub-Gaussian tail bound (8), and with a minor substitution, namely, $\|\mathbf{k}_\alpha\|_{\infty,\text{in}}$ replaced by $2\|\mathbf{k}_\alpha\|_\infty$. Putting it together with (17), we conclude that

$$\begin{aligned}
\text{MMD}_{\mathbf{k}}(\mathcal{S}_{\text{in}}, \mathcal{S}_{\text{KT+}}) &\leq \left(\tfrac{2^m}{n}2\|\mathbf{k}_\alpha\|_{\infty,\text{in}}\right)^{\frac{1}{2\alpha}}(2\widetilde{\mathfrak{M}}_\alpha)^{1-\frac{1}{2\alpha}}\left(2+\sqrt{\tfrac{(4\pi)^{d/2}}{\Gamma(\frac{d}{2}+1)}}\cdot\mathfrak{R}_{\max}^{\frac{d}{2}}\cdot\widetilde{\mathfrak{M}}_\alpha\right)^{\frac{1}{\alpha}-1} \\
&= 2^{\frac{1}{2\alpha}}\cdot\overline{\mathbf{M}}_{\text{powerKT}}(\mathbf{k}_\alpha),
\end{aligned} \tag{18}$$

with probability at least p_{sg}. Let us denote this event by $\mathcal{E}_2$.

Note that the quantities on the right hand side of the bounds (16) and (18) are deterministic given $\mathcal{S}_{\mathrm{in}}$, and thus can be computed apriori. Consequently, we apply the high probability bound only for one of the two events $\mathcal{E}_1$ or $\mathcal{E}_2$ depending on which of the two quantities (deterministically) attains the minimum. Thus, the bound (9) holds with probability at least p_{sg} as claimed, and we are done.

F Proof of Prop. 1: An interpolation result for MMD

For two arbitrary distributions $\mathbb{P}$ and $\mathbb{Q}$, and any reproducing kernel $\mathbf{k}$, Gretton et al. (2012, Lem. 4) yields that

$$\mathrm{MMD}^2_{\mathbf{k}}(\mathbb{P},\mathbb{Q}) = \|(\mathbb{P}-\mathbb{Q})\mathbf{k}\|^2_{\mathbf{k}}. \tag{19}$$

Let $\mathcal{F}$ denote the generalized Fourier transform (GFT) operator (Wendland (2004, Def. 8.9)). Since $\mathbf{k}(x,y)=\kappa(x-y)$, Wendland (2004, Thm. 10.21) yields that

$$\|f\|^2_{\mathbf{k}} = \tfrac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\tfrac{(\mathcal{F}(f)(\omega))^2}{\mathcal{F}(\kappa)(\omega)}d\omega, \quad \text{for} \quad f\in\mathcal{H}. \tag{20}$$

Let $\widehat{\kappa}\triangleq\mathcal{F}(\kappa)$, and consider a discrete measure $\mathbb{D}=\sum_{i=1}^n w_i\boldsymbol{\delta}_{x_i}$ supported on finitely many points, and let $\mathbb{D}\mathbf{k}(x)\triangleq\sum w_i\mathbf{k}(x_i,x)=\sum w_i\kappa(x_i-x)$. Now using the linearity of the GFT operator $\mathcal{F}$, we find that for any $\omega\in\mathbb{R}^d$,

$$\begin{aligned}\mathcal{F}(\mathbb{D}\mathbf{k})(\omega) = \mathcal{F}(\textstyle\sum_{i=1}^n w_i\kappa(x_i-\cdot)) = \sum_{i=1}^n w_i\mathcal{F}(\kappa(x_i-\cdot)) &= (\textstyle\sum_{i=1}^n w_i e^{\langle\omega,x_i\rangle})\cdot\widehat{\kappa}(\omega)\\ &= \widehat{D}(\omega)\widehat{\kappa}(\omega)\end{aligned} \tag{21}$$

where we used the fact that $\mathcal{F}(\kappa(x_i-\cdot))(\omega)=e^{\langle\omega,x_i\rangle}\widehat{\kappa}(\omega)$,[4] and used the shorthand $\widehat{D}(\omega)\triangleq(\sum_{i=1}^n w_i e^{\langle\omega,x_i\rangle})$ in the last step. Putting together (19) to (21) with $\mathbb{D}=\mathbb{P}-\mathbb{Q}$, we find that

$$\begin{aligned}\mathrm{MMD}^2_{\mathbf{k}}(\mathbb{P},\mathbb{Q}) &= \tfrac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\widehat{D}^2(\omega)\widehat{\kappa}(\omega)d\omega \qquad (22)\\
&= \tfrac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\widehat{D}^2(\omega)\widehat{\kappa}^\alpha(\omega)(\widehat{\kappa}^\alpha(\omega))^{\frac{1-\alpha}{\alpha}}d\omega\\
&= \tfrac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\widehat{D}^2(\omega')\widehat{\kappa}^\alpha(\omega')d\omega'\int_{\mathbb{R}^d}\tfrac{\widehat{D}^2(\omega)\widehat{\kappa}^\alpha(\omega)}{\int_{\mathbb{R}^d}\widehat{D}^2(\omega')\widehat{\kappa}^\alpha(\omega')d\omega'}(\widehat{\kappa}^\alpha(\omega))^{\frac{1-\alpha}{\alpha}}d\omega\\
&\overset{(i)}{\le} \tfrac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\widehat{D}^2(\omega')\widehat{\kappa}^\alpha(\omega')d\omega'\left(\int_{\mathbb{R}^d}\tfrac{\widehat{D}^2(\omega)\widehat{\kappa}^\alpha(\omega)}{\int_{\mathbb{R}^d}\widehat{D}^2(\omega')\widehat{\kappa}^\alpha(\omega')d\omega'}\widehat{\kappa}^\alpha(\omega)d\omega\right)^{\frac{1-\alpha}{\alpha}}\\
&= \tfrac{1}{(2\pi)^{d/2}}\left(\int_{\mathbb{R}^d}\widehat{D}^2(\omega')\widehat{\kappa}^\alpha(\omega')d\omega'\right)^{2-\frac{1}{\alpha}}\left(\int_{\mathbb{R}^d}\tfrac{\widehat{D}^2(\omega)\widehat{\kappa}^{2\alpha}(\omega)}{d}\omega\right)^{\frac{1-\alpha}{\alpha}}\\
&= \left(\tfrac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\widehat{D}^2(\omega')\widehat{\kappa}^\alpha(\omega')d\omega'\right)^{2-\frac{1}{\alpha}}\left(\tfrac{1}{(2\pi)^{d/2}}\int_{\mathbb{R}^d}\tfrac{\widehat{D}^2(\omega)\widehat{\kappa}^{2\alpha}(\omega)}{d}\omega\right)^{\frac{1-\alpha}{\alpha}}\\
&\overset{(ii)}{=} (\mathrm{MMD}^2_{\mathbf{k}_\alpha}(\mathbb{P},\mathbb{Q}))^{2-\frac{1}{\alpha}}\cdot(\mathrm{MMD}^2_{\mathbf{k}_{2\alpha}}(\mathbb{P},\mathbb{Q}))^{\frac{1}{\alpha}-1},\end{aligned}$$

where step (i) makes use of Jensen's inequality and the fact that the function $t\mapsto t^{\frac{1-\alpha}{\alpha}}$ for $t\ge 0$ is concave for $\alpha\in[\frac{1}{2},1]$, and step (ii) follows by applying (22) for kernels $\mathbf{k}_\alpha$ and $\mathbf{k}_{2\alpha}$ and noting that by definition $\mathcal{F}(\mathbf{k}_\alpha)=\widehat{\kappa}^\alpha$, and $\mathcal{F}(\mathbf{k}_{2\alpha})=\widehat{\kappa}^{2\alpha}$. Noting MMD is a non-negative quantity, and taking square-root establishes the claim (13).

G Sub-optimality of single function guarantees with root KT

Define $\widetilde{\mathbf{k}}_{\mathrm{rt}}$ as the scaled version of $\mathbf{k}_{\mathrm{rt}}$, i.e., $\widetilde{\mathbf{k}}_{\mathrm{rt}}\triangleq\mathbf{k}_{\mathrm{rt}}/\|\mathbf{k}_{\mathrm{rt}}\|_\infty$ that is bounded by 1. Then Zhang & Zhao (2013, Proof of Prop. 2.3) implies that

$$\|f\|_{\mathbf{k}_{\mathrm{rt}}} = \tfrac{1}{\sqrt{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}}\|f\|_{\widetilde{\mathbf{k}}_{\mathrm{rt}}}. \tag{23}$$

[4]The property $\mathcal{F}(\kappa(x_i-\cdot))(\omega)=e^{\langle\omega,x_i\rangle}\widehat{\kappa}(\omega)$ often called the *shifting property*, can be verified directly for GFT using Wendland (2004, Def. 8.9), a change of variable, and noting that $\widehat{\gamma}$ denotes a Fourier transform (in Wendland's notation) for a Schwarz function; and hence $\widehat{\gamma}$ itself satisfies the shifting property (see Wendland (2004, Thm. 5.16(4)).

And thus we also have $\mathcal{H}_{\mathrm{rt}} = \widetilde{\mathcal{H}}_{\mathrm{rt}}$ where $\mathcal{H}_{\mathrm{rt}}$ and $\widetilde{\mathcal{H}}_{\mathrm{rt}}$ respectively denote the RKHSs of $\mathbf{k}_{\mathrm{rt}}$ and $\widetilde{\mathbf{k}}_{\mathrm{rt}}$.

Next, we note that for any two kernels $\mathbf{k}_1$ and $\mathbf{k}_2$ with corresponding RKHSs $\mathcal{H}_1$ and $\mathcal{H}_2$ with $\mathcal{H}_1 \subset \mathcal{H}_2$, in the convention of Zhang & Zhao (2013, Lem. 2.2, Prop. 2.3), we have

$$\frac{\|f\|_{\mathbf{k}_2}}{\|f\|_{\mathbf{k}_1}} \leq \beta(\mathcal{H}_1, \mathcal{H}_2) \leq \sqrt{\lambda(\mathcal{H}_1, \mathcal{H}_2)} \quad \text{for} \quad f \in \mathcal{H}. \tag{24}$$

Consequently, we have

$$\sqrt{\max_{x \in \mathcal{S}_{\mathrm{in}}} \mathbf{k}_{\mathrm{rt}}(x,x)} \frac{\|f\|_{\mathbf{k}_{\mathrm{rt}}}}{\|f\|_{\mathbf{k}}} \leq \sqrt{\|\mathbf{k}_{\mathrm{rt}}\|_\infty} \frac{\|f\|_{\mathbf{k}_{\mathrm{rt}}}}{\|f\|_{\mathbf{k}}} \overset{(23)}{=} \frac{\|f\|_{\widetilde{\mathbf{k}}_{\mathrm{rt}}}}{\|f\|_{\mathbf{k}}} \leq \sqrt{\lambda(\mathcal{H}, \widetilde{\mathcal{H}}_{\mathrm{rt}})}, \tag{25}$$

where in the last step, we have applied the bound (24) with $(\mathbf{k}_1, \mathcal{H}_1) \leftarrow (\mathbf{k}, \mathcal{H})$ and $(\mathbf{k}_2, \mathcal{H}_2) \leftarrow (\widetilde{\mathbf{k}}_{\mathrm{rt}}, \widetilde{\mathbf{k}}_{\mathrm{rt}})$ since $\mathcal{H} \subset \mathcal{H}_{\mathrm{rt}} = \widetilde{\mathbf{k}}_{\mathrm{rt}}$.

Next, we use (25) to the kernels studied in Dwivedi & Mackey (2021) where we note that all the kernels in that work were scaled to ensure $\|\mathbf{k}\|_\infty = 1$ and in fact satisfied $\mathbf{k}(x,x) = 1$. Consequently, the multiplicative factor stated in the discussion after Thm. 1, namely, $\sqrt{\frac{\|\mathbf{k}_{\mathrm{rt}}\|_{\infty,\mathrm{in}}}{\|\mathbf{k}\|_{\infty,\mathrm{in}}}} \frac{\|f\|_{\mathbf{k}_{\mathrm{rt}}}}{\|f\|_{\mathbf{k}}}$ can be bounded by $\sqrt{\lambda(\mathcal{H}, \widetilde{\mathcal{H}}_{\mathrm{rt}})}$ given the arguments above.

For $\mathbf{k} = \textbf{Gauss}(\sigma)$ kernels, Zhang & Zhao (2013, Prop. 3.5(1)) yields that

$$\lambda(\mathcal{H}, \widetilde{\mathcal{H}}_{\mathrm{rt}}) = 2^{d/2}.$$

For $\mathbf{k} = \textbf{B-spline}(2\beta + 1)$ with $\beta \in 2\mathbb{N} + 1$, Zhang & Zhao (2013, Prop. 3.5(1)) yields that

$$\lambda(\mathcal{H}, \widetilde{\mathcal{H}}_{\mathrm{rt}}) = 1.$$

For $\mathbf{k} =$**Matérn**(ν, γ) with $\nu > d$, some algebra along with Zhang & Zhao (2013, Prop 3.1) yields that

$$\lambda(\mathcal{H}, \widetilde{\mathcal{H}}_{\mathrm{rt}}) = \frac{\Gamma(\nu)\Gamma((\nu-d)/2)}{\Gamma(\nu-d/2)\Gamma(\nu/2)} \geq 1.$$

H Additional Experimental Results

This section provides additional experimental details and results deferred from Sec. 4.

Common settings and error computation To obtain an output coreset of size $n^{\frac{1}{2}}$ with n input points, we (a) take every $n^{\frac{1}{2}}$-th point for standard thinning (ST) and (b) run KT with $m = \frac{1}{2}\log_2 n$ using an ST coreset as the base coreset in KT-SWAP. For Gaussian and MoG target we use i.i.d. points as input, and for MCMC targets we use an ST coreset after burn-in as the input (see App. H for more details). We compute errors with respect to $\mathbb{P}$ whenever available in closed form and otherwise use $\mathbb{P}_{\mathrm{in}}$. For each input sample size $n \in \{2^4, 2^6, \ldots, 2^{14}\}$ with $\delta_i = \frac{1}{2n}$, we report the mean MMD or function integration error ± 1 standard error across 10 independent replications of the experiment (the standard errors are too small to be visible in all experiments). We also plot the ordinary least squares fit (for log mean error vs log coreset size), with the slope of the fit denoted as the empirical decay rate, e.g., for an OLS fit with slope -0.25, we display the decay rate of $n^{-0.25}$.

Details of test functions We note the following: (a) For Gaussian targets, the error with CIF function and i.i.d. input is measured across the sample mean over the n input points and $\sqrt{n}$ output points obtained by standard thinning the input sequence, since $\mathbb{P}f_{\mathrm{CIF}}$ does not admit a closed form. (b) To define the function $f : x \mapsto \mathbf{k}(X', x)$, first we draw a sample $X \sim \mathbb{P}$, independent of the input, and then set $X' = 2X$. For the MCMC targets, we draw a point uniformly from a held out data point not used as input for KT. For each target, the sample is drawn exactly once and then fixed throughout all sample sizes and repetions.

H.1 Mixture of Gaussians Experiments

Our mixture of Gaussians target is given by $\mathbb{P} = \frac{1}{M}\sum_{j=1}^{M} \mathcal{N}(\mu_j, \mathbf{I}_d)$ for $M \in \{4, 6, 8\}$ where

$$\mu_1 = [-3, 3]^\top, \quad \mu_2 = [-3, 3]^\top, \quad \mu_3 = [-3, -3]^\top, \quad \mu_4 = [3, -3]^\top,$$
$$\mu_5 = [0, 6]^\top, \quad \mu_6 = [-6, 0]^\top, \quad \mu_7 = [6, 0]^\top, \quad \mu_8 = [0, -6]^\top.$$

We provide visualizations similar to Fig. 1 from two other independent runs of the experiment in Fig. 4; where we once again notice that the KT output visually provides a significantly improved approximation of the underlying 8-MoG target compared to i.i.d. points. Finally, we provide the MMD errors for target KT and root KT with GAUSS(4) $\mathbf{k}$, target KT with GAUSS(4) and B-SPLINE (1), KT+ for LAPLACE(2) $\mathbf{k}$ and $\mathbf{k}_{0.7}$, and KT+ with IMQ(0.5, 2) $\mathbf{k}$ and $\mathbf{k}_{0.5}$ in Fig. 5

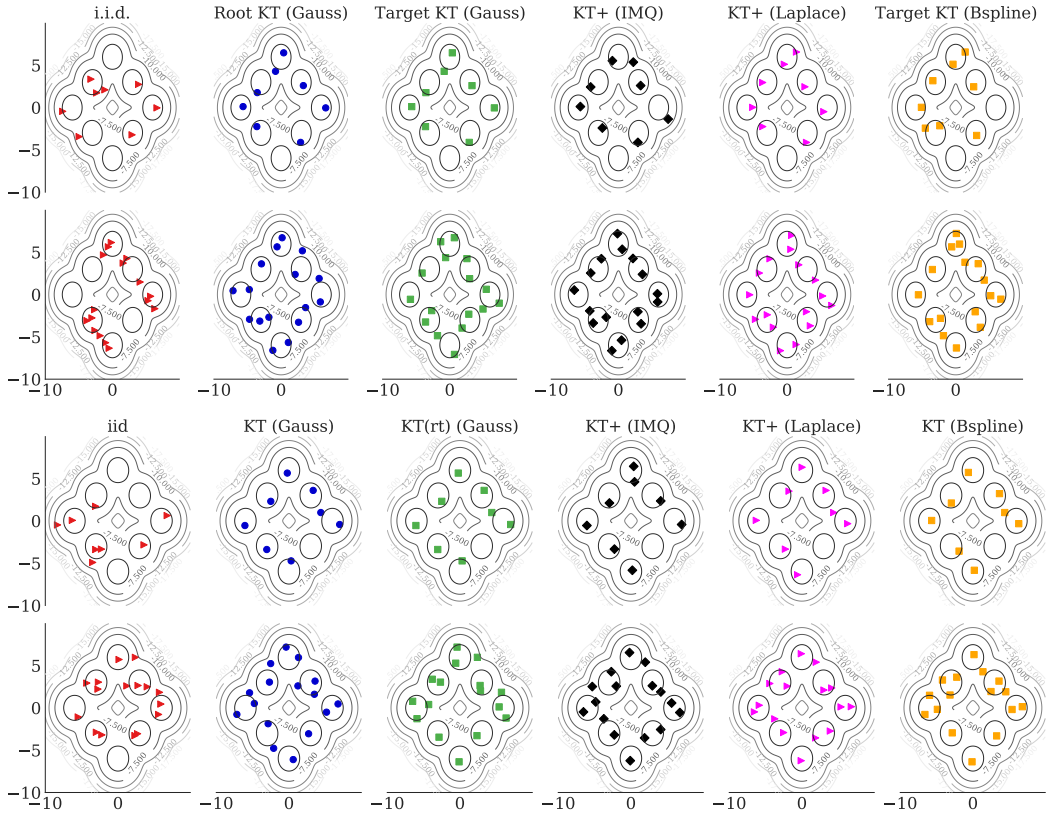

Figure 4: Scatter plots of i.i.d. coresets, and coresets obtained from target KT, root KT, and KT+ with various kernels for 8-component mixture of Gaussian targets with equidensity contours of the target underlaid. These plots are from two independent runs different from Fig. 1. See Sec. 4 for more details.

H.2 MCMC EXPERIMENTS

Our set-up for MCMC experiments follows closely that of Dwivedi & Mackey (2021). For complete details on the targets and sampling algorithms we refer the reader to Riabiz et al. (2020a, Sec. 4).

Goodwin and Lotka-Volterra experiments From Riabiz et al. (2020b), we use the output of four distinct MCMC procedures targeting each of two $d = 4$-dimensional posterior distributions $\mathbb{P}$: (1) a posterior over the parameters of the *Goodwin model* of oscillatory enzymatic control (Goodwin, 1965) and (2) a posterior over the parameters of the *Lotka-Volterra model* of oscillatory predator-prey evolution (Lotka, 1925; Volterra, 1926). For each of these targets, Riabiz et al. (2020b) provide 2×10^6 sample points from the following four MCMC algorithms: Gaussian random walk (RW), adaptive Gaussian random walk (adaRW, Haario et al., 1999), Metropolis-adjusted Langevin algorithm (MALA, Roberts & Tweedie, 1996), and pre-conditioned MALA (pMALA, Girolami & Calderhead, 2011).

Hinch experiments Riabiz et al. (2020b) also provide the output of two independent Gaussian random walk MCMC chains targeting each of two $d = 38$-dimensional posterior distributions $\mathbb{P}$: (1) a posterior over the parameters of the Hinch model of calcium signalling in cardiac cells (Hinch

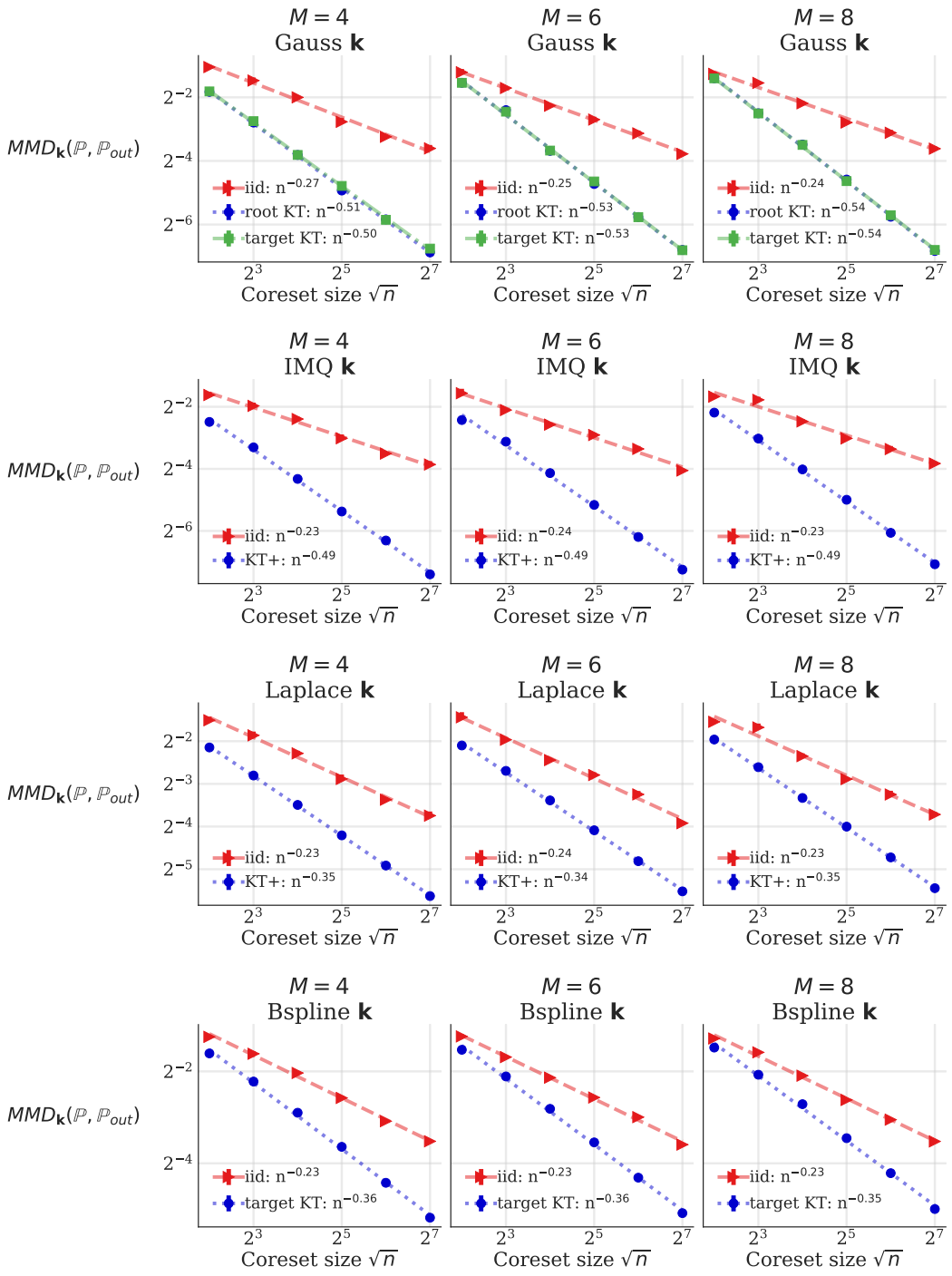

Figure 5: Kernel thinning vs i.i.d. sampling for MoG target for $M \in \{4, 6, 8\}$ with Gaussian, Laplace, IMQ, and Bspline kernels as discussed in Sec. 4.

et al., 2004) and (2) a tempered version of the same posterior, as defined by Riabiz et al. (2020a, App. S5.4).

Burn-in and standard thinning We discard the initial burn-in points of each chain using the maximum burn-in period reported in Riabiz et al. (2020a, Tabs. S4 & S6, App. S5.4). Furthermore, we also normalize each Hinch chain by subtracting the post-burn-in sample mean and dividing each

coordinate by its post-burn-in sample standard deviation. To obtain an input sequence $\mathcal{S}_{\text{in}}$ of length n to be fed into a thinning algorithm, we downsample the remaining even indices of points using standard thinning (odd indices are held out). When applying standard thinning to any Markov chain output, we adopt the convention of keeping the final sample point.

The selected burn-in periods for the Goodwin task were 820,000 for RW; 824,000 for adaRW; 1,615,000 for MALA; and 1,475,000 for pMALA. The respective numbers for the Lotka-Volterra task were 1,512,000 for RW; 1,797,000 for adaRW; 1,573,000 for MALA; and 1,251,000 for pMALA.

I UPPER BOUNDS ON RKHS COVERING NUMBERS

In this section, we state several results on covering bounds for RKHSs for generic and some specific kernels. These bounds can then be used with Thm. 2 (or Tab. 2) to establish MMD guarantees for the output of kernel thinning as summarized in Tab. 3.

We first state covering number bounds for RKHS associated with generic kernels, that are either (a) analytic, or (b) finitely many times differentiable. These results follow essentially from Sun & Zhou (2008); Steinwart & Christmann (2008), but we provide a proof in App. I.2 for completeness.

Proposition 2 (Covering numbers for analytic and differentiable kernels) *The following results hold true.*

(a) ***Analytic kernels:*** *Suppose that* $\mathbf{k}(x,y) = \kappa(\|x-y\|_2^2)$ *for* $\kappa : \mathbb{R}_+ \to \mathbb{R}$ *real-analytic with convergence radius* R_κ*, that is,*

$$\left|\tfrac{1}{j!}\kappa_+^{(j)}(0)\right| \leq C_\kappa (2/R_\kappa)^j \quad \textit{for all} \quad j \in \mathbb{N}_0 \tag{26}$$

for some constant C_κ*, where* $\kappa_+^{(j)}$ *denotes the right-sided* j*-th derivative of* κ*. Then for any set* $\mathcal{A} \subset \mathbb{R}^d$ *and any* $\varepsilon \in (0, \frac{1}{2})$*, we have*

$$\mathcal{M}_{\mathbf{k}}(\mathcal{A}, \varepsilon) \leq \mathcal{N}_2(\mathcal{A}, r^\dagger/2) \cdot \left(4\log(1/\varepsilon) + 2 + 4\log(16\sqrt{C_\kappa}+1)\right)^{d+1}, \tag{27}$$

$$\textit{where } r^\dagger \triangleq \min\left(\tfrac{\sqrt{R_\kappa}}{2d}, \sqrt{R_\kappa + D_{\mathcal{A}}^2} - D_{\mathcal{A}}\right), \textit{ and } D_{\mathcal{A}} \triangleq \max_{x,y\in\mathcal{A}} \|x-y\|_2. \tag{28}$$

(b) ***Differentiable kernels:*** *Suppose that for* $\mathcal{X} \subset \mathbb{R}^d$*, the kernel* $\mathbf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *is* s*-times continuously differentiable, i.e., all partial derivatives* $\partial^{\alpha,\alpha}\mathbf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ *exist and are continuous for all multi-indices* $\alpha \in \mathbb{N}_0^d$ *with* $|\alpha| \leq s$*. Then, for any closed Euclidean ball* $\bar{\mathcal{B}}_2(r)$ *contained in* $\mathcal{X}$ *and any* $\varepsilon > 0$*, we have*

$$\mathcal{M}_{\mathbf{k}}(\bar{\mathcal{B}}_2(r), \varepsilon) \leq c_{s,d,\mathbf{k}} \cdot r^d \cdot (1/\varepsilon)^{d/s}, \tag{29}$$

for some constant $c_{s,d,\mathbf{k}}$ *that depends only on on* s, d *and* $\mathbf{k}$*.*

Next, we state several explicit bounds on covering numbers for several popular kernels. See App. I.3 for the proof.

Proposition 3 (Covering numbers for specific kernels) *The following statements hold true.*

(a) When $\mathbf{k} = \textsc{Gauss}(\sigma)$*, we have*

$$\mathcal{M}_{\mathbf{k}}(\mathcal{B}_2(r), \varepsilon) \leq C_{\text{Gauss},d} \cdot \left(\tfrac{\log(4/\varepsilon)}{\log\log(4/\varepsilon)}\right)^d \log(1/\varepsilon) \cdot \begin{cases} 1 & \textit{when } r \leq \frac{1}{\sqrt{2}\sigma}, \\ (3\sqrt{2}r\sigma)^d & \textit{otherwise,} \end{cases} \tag{30}$$

$$\textit{where } C_{\textsc{Gauss},d} \triangleq \tbinom{4e+d}{d} e^{-d} \leq \begin{cases} 4.3679 & \textit{for } d = 1 \\ 0.05 \cdot d^{4e} e^{-d} & \textit{for } d \geq 2 \end{cases} \leq 30 \textit{ for all } d \geq 1. \tag{31}$$

(b) When $\mathbf{k} = \textsc{Matérn}(\nu, \gamma)$*, then for some constant* $C_{\textsc{Matérn},\nu,\gamma,d}$*, we have*

$$\mathcal{M}_{\mathbf{k}}(\mathcal{B}_2(r), \varepsilon) \leq C_{\textsc{Matérn},\nu,\gamma,d} \cdot r^d \cdot (1/\varepsilon)^{d/\nu}. \tag{32}$$

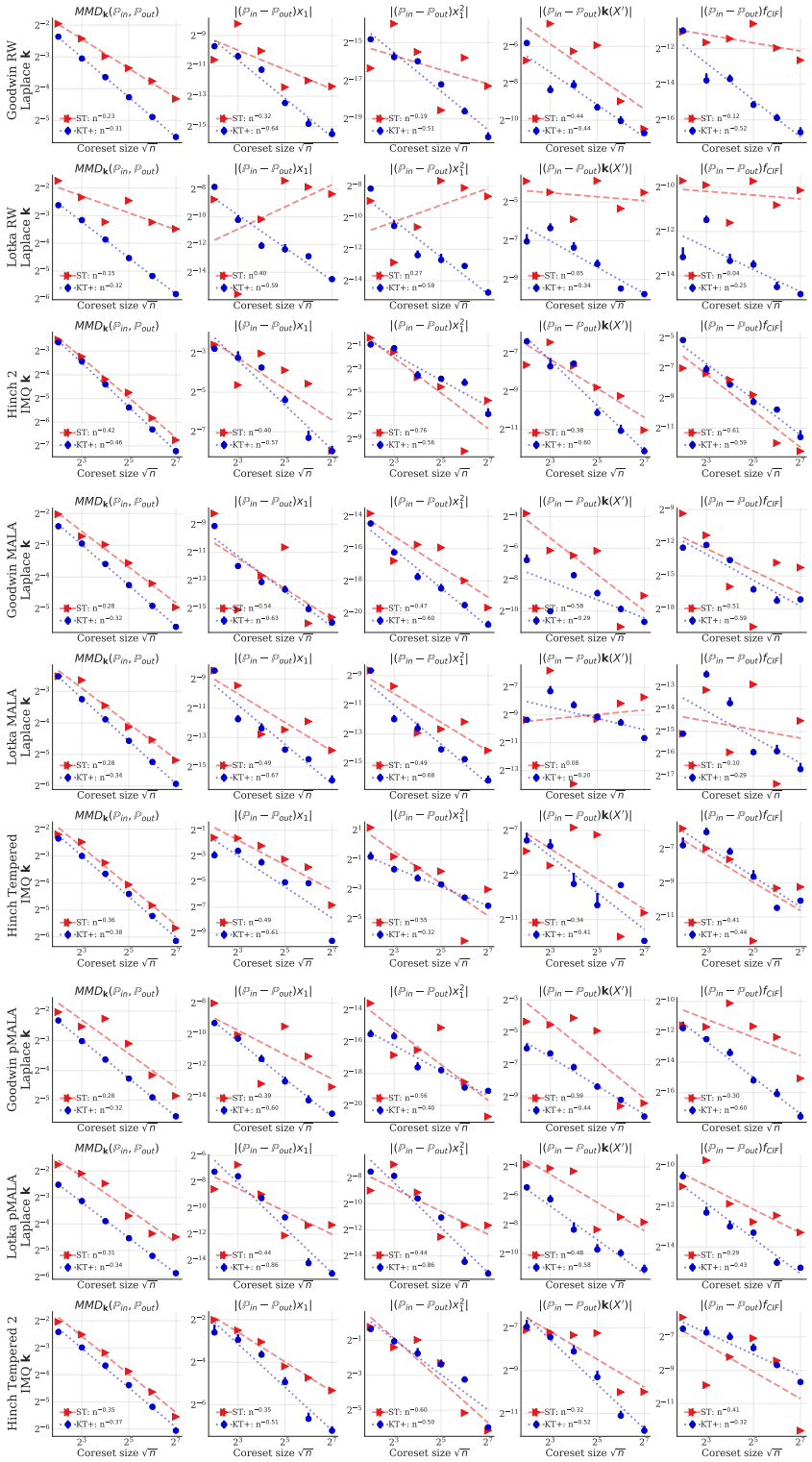

Figure 6: KT+ vs standard thinning (ST) for MCMC targets. Here we present the results for nine out of twelve MCMC experiments deferred from Sec. 4.

(c) When $\mathbf{k} = \mathrm{IMQ}(\nu, \gamma)$*, we have*

$$\mathcal{M}_{\mathbf{k}}(\mathcal{B}_2(r), \varepsilon) \leq (1 + \tfrac{4r}{\widetilde{r}})^d \cdot (4\log(1/\varepsilon) + 2 + C_{\mathrm{IMQ},\nu,\gamma})^{d+1}, \tag{33}$$

$$\text{where } C_{\mathrm{IMQ},\nu,\gamma} \triangleq 4\log\Big(16\tfrac{(2\nu+1)^{\nu+1}}{\gamma^{2\nu}} + 1\Big), \text{ and } \widetilde{r} \triangleq \min\Big(\tfrac{\gamma}{2d}, \sqrt{\gamma^2 + 4r^2} - 2r\Big). \tag{34}$$

(d) When $\mathbf{k} = \textsc{sinc-rad}(\theta)$*, then for* $d \in \{1, 2, 3\}$ *and* $\varepsilon \in (0, \frac{1}{2})$*, we have*

$$\mathcal{M}_{\mathbf{k}}(\mathcal{B}_2(r), \varepsilon) \leq (1 + \tfrac{4r}{\widetilde{r}_{\theta,d}})^d \cdot (4\log(1/\varepsilon) + 2 + 4\log 17)^{d+1}, \tag{35}$$

$$\text{where } \widetilde{r}_{\theta,d} \triangleq \min\Big(\tfrac{\sqrt{3}}{|\theta|d}, \sqrt{\tfrac{12}{\theta^2} + 4r^2} - 2r\Big). \tag{36}$$

(e) When $\mathbf{k} = \textsc{sinc}(\theta)$*, then for* $\varepsilon \in (0, \frac{1}{2})$ *and* $\widetilde{r}_{\theta,1}$ *as in* (36)*, we have*

$$\mathcal{M}_{\mathbf{k}}([-r, r]^d, \varepsilon) \leq d \cdot (1 + \tfrac{4r}{\widetilde{r}_{\theta,1}}) \cdot (4\log(d/\varepsilon) + 2 + 4\log 17)^2. \tag{37}$$

(f) When $\mathbf{k} = \textsc{b-spline}(2\beta + 1)$*, then for some universal constant* $C_{\textsc{b-spline}}$*, we have*

$$\mathcal{M}_{\mathbf{k}}([-\tfrac{1}{2}, \tfrac{1}{2}]^d, \varepsilon) \leq d \cdot C_{\textsc{b-spline}} \cdot (d/\varepsilon)^{\frac{1}{\beta+\frac{1}{2}}}. \tag{38}$$

I.1 Auxiliary results about RKHS and Euclidean covering numbers

In this section, we collect several results regarding the covering numbers of Euclidean and RKHS spaces that come in handy for our proofs. These results can also be of independent interest.

We start by defining the notion of restricted kernel and its unit ball (Rudi et al. (2020, Prop. 8)). For $\mathcal{X} \subset \mathbb{R}^d$, let $|_{\mathcal{X}}$ denotes the restriction operator. That is, for any function $f : \mathbb{R}^d \to \mathbb{R}$, we have $f|_{\mathcal{X}} : \mathcal{X} \to \mathbb{R}$ such that $f|_{\mathcal{A}}(x) = f(x)$ for $x \in \mathcal{X}$.

Definition 3 (Restricted kernel and its RKHS) *Consider a kernel* $\mathbf{k}$ *defined on* $\mathbb{R}^d \times \mathbb{R}^d$ *with the corresponding RKHS* $\mathcal{H}$*, any set* $\mathcal{X} \subset \mathbb{R}^d$*. The restricted kernel* $\mathbf{k}_{|\mathcal{X}}$ *is defined as*

$$\mathbf{k}|_{\mathcal{X}} : \mathcal{X} \times \mathcal{X} \to \mathbb{R} \quad \text{such that} \quad \mathbf{k}|_{\mathcal{X}}(x, y) \triangleq \mathbf{k}|_{\mathcal{X}\times\mathcal{X}}(x, y) = \mathbf{k}(x, y) \quad \text{for all} \quad x, y \in \mathcal{X},$$

and $\mathcal{H}|_{\mathcal{X}}$ *denotes its RKHS. For* $f \in \mathcal{H}|_{\mathcal{X}}$*, the restricted RKHS norm is defined as follows:*

$$\|f\|_{\mathbf{k}|_{\mathcal{X}}} = \inf_{h\in\mathcal{H}} \|h\|_{\mathbf{k}} \quad \text{such that} \quad h|_{\mathcal{X}} = f.$$

Furthermore, we use $\mathcal{B}_{\mathbf{k}_{|\mathcal{X}}} \triangleq \{f \in \mathcal{H}|_{\mathcal{X}} : \|f\|_{\mathbf{k}|_{\mathcal{X}}} \leq 1\}$ *to denote the unit ball of the RKHS corresponding to this restricted kernel.*

In this notation, the unit ball of unrestricted kernel satisfies $\mathcal{B}_{\mathbf{k}} \triangleq \mathcal{B}_{\mathbf{k}_{|\mathbb{R}^d}}$. Now, recall the RKHS covering number definition from Def. 1. In the sequel, we also use the covering number of the restricted kernel defined as follows:

$$\mathcal{N}^{\dagger}_{\mathbf{k}}(\mathcal{X}, \varepsilon) = \mathcal{N}_{\mathbf{k}|_{\mathcal{X}}}(\mathcal{X}, \varepsilon), \tag{39}$$

that is $\mathcal{N}^{\dagger}_{\mathbf{k}}(\mathcal{X}, \varepsilon)$ denotes the minimum cardinality over all possible covers $\mathcal{C} \subset \mathcal{B}_{\mathbf{k}_{|\mathcal{X}}}$ that satisfy

$$\mathcal{B}_{\mathbf{k}_{|\mathcal{X}}} \subset \textstyle\bigcup_{h\in\mathcal{C}} \big\{g \in \mathcal{B}_{\mathbf{k}_{|\mathcal{X}}} : \sup_{x\in\mathcal{X}} |h(x) - g(x)| \leq \varepsilon\big\}.$$

With this notation in place, we now state a result that relates the covering numbers $\mathcal{N}^{\dagger}$ (39) and $\mathcal{N}$ Def. 1.

Lemma 1 (Relation between restricted and unrestricted RKHS covering numbers) *We have*

$$\mathcal{N}_{\mathbf{k},\varepsilon}(\mathcal{X}) \leq \mathcal{N}^{\dagger}_{\mathbf{k},\varepsilon}(\mathcal{X})$$

Proof of Lem. 1 Rudi et al. (2020, Prop. 8(d,f)) imply that there exists a bounded linear extension operator $E : \mathcal{H}|_{\mathcal{X}} \to \mathcal{H}$ with operator norm bounded by 1, which when combined with Steinwart & Christmann (2008, eqns. (A.38), (A.39)) yields the claim.

Next, we state results that relate RKHS covering numbers for a change of domain for a shift-invariant kernel. We use $\mathcal{B}_{\|\cdot\|}(x; r) \triangleq \big\{y \in \mathbb{R}^d : \|x - y\| \leq r\big\}$ to denote the r radius ball in $\mathbb{R}^d$ defined by the metric induced by a norm $\|\cdot\|$.

Definition 4 (Euclidean covering numbers) *Given a set $\mathcal{X} \subset \mathbb{R}^d$, a norm $\|\cdot\|$, and a scalar $\varepsilon > 0$, we use $\mathcal{N}_{\|\cdot\|}(\mathcal{X}, \varepsilon)$ to denote the ε-covering number of $\mathcal{X}$ with respect to $\|\cdot\|$-norm. That is, $\mathcal{N}_{\|\cdot\|}(\mathcal{X}, \varepsilon)$ denotes the minimum cardinality over all possible covers $\mathcal{C} \subset \mathcal{X}$ that satisfy*

$$\mathcal{X} \subset \cup_{z \in \mathcal{C}} \mathcal{B}_{\|\cdot\|}(z; \varepsilon).$$

When $\|\cdot\| = \|\cdot\|_q$ for some $q \in [1, \infty]$, we use the shorthand $\mathcal{N}_q \triangleq \mathcal{N}_{\|\cdot\|_q}$.

Lemma 2 (Relation between RKHS covering numbers on different domains) *Given a shift-invariant kernel $\mathbf{k}$, a norm $\|\cdot\|$ on $\mathbb{R}^d$, and any set $\mathcal{X} \subset \mathbb{R}^d$, we have*

$$\mathcal{N}^{\dagger}_{\mathbf{k}}(\mathcal{X}, \varepsilon) \leq \left[\mathcal{N}^{\dagger}_{\mathbf{k}}(\mathcal{B}_{\|\cdot\|}, \varepsilon)\right]^{\mathcal{N}_{\|\cdot\|}(\mathcal{X}, 1)}.$$

Proof of Lem. 2 Let $\mathcal{C} \subset \mathcal{X}$ denote the cover of minimum cardinality such that

$$\mathcal{X} \subseteq \bigcup_{z \in \mathcal{C}} \mathcal{B}_{\|\cdot\|}(z, 1).$$

We then have

$$\mathcal{N}^{\dagger}_{\mathbf{k}}(\mathcal{X}, \varepsilon) \overset{(i)}{\leq} \prod_{z \in \mathcal{C}} \mathcal{N}^{\dagger}_{\mathbf{k}}(\mathcal{B}_{\|\cdot\|}(z, 1), \varepsilon) \overset{(ii)}{\leq} \prod_{z \in \mathcal{C}} \mathcal{N}^{\dagger}_{\mathbf{k}}(\mathcal{B}_{\|\cdot\|}, \varepsilon) \leq \left[\mathcal{N}^{\dagger}_{\mathbf{k}}(\mathcal{B}_{\|\cdot\|}, \varepsilon)\right]^{|\mathcal{C}|},$$

where step (i) follows by applying Steinwart & Fischer (2021, Lem. 3.11),[5] and step (ii) follows by applying Steinwart & Fischer (2021, Lem. 3.10). The claim follows by noting that $\mathcal{C}$ denotes a cover of minimum cardinality, and hence by definition $|\mathcal{C}| = \mathcal{N}_{\|\cdot\|}(\mathcal{X}, 1)$.

Lemma 3 (Covering number for product kernel) *Given $\mathcal{X} \subset \mathbb{R}$ and a reproducing kernel $\kappa : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, consider the product kernel $\mathbf{k} \triangleq \kappa^{\otimes d} : \mathcal{X}^{\otimes 2d} \to \mathbb{R}$ defined as*

$$\mathbf{k}(x, y) = \prod_{i=1}^{d} \kappa(x_i, y_i) \quad \textit{for} \quad x, y \in \mathcal{X}^{\otimes d} \triangleq \underbrace{\mathcal{X} \times \mathcal{X} \ldots \times \mathcal{X}}_{d \text{ times}} \subset \mathbb{R}^d.$$

Then the covering numbers of the two kernels are related as follows:

$$\mathcal{N}^{\dagger}_{\mathbf{k}}(\mathcal{X}^{\otimes d}, \varepsilon) \leq \left[\mathcal{N}^{\dagger}_{\kappa}(\mathcal{X}, \varepsilon / (d \|\kappa\|_{\infty}^{\frac{d-1}{2}}))\right]^{d}. \tag{40}$$

Proof of Lem. 3 Let $\mathcal{H}$ denote the RKHS corresponding to κ. Then the RKHS corresponding to the kernel $\mathbf{k}$ is given by the tensor product $\mathcal{H}_{\mathbf{k}} \triangleq \mathcal{H} \times \mathcal{H} \times \ldots \times \mathcal{H}$ Berlinet & Thomas-Agnan (2011, Sec. 4.6), i.e., for any $f \in \mathcal{H}_{\mathbf{k}}$, there exists $f_1, f_2, \ldots, f_d \in \mathcal{H}$ such that

$$f(x) = \prod_{i=1}^{d} f_i(x_i) \quad \text{for all} \quad x \in \mathcal{X}^{\otimes d}. \tag{41}$$

Let $\mathcal{C}_{\kappa}(\mathcal{X}, \varepsilon) \subset \mathcal{B}_{\kappa}$ denote an ε-cover of $\mathcal{B}_{\kappa}$ in L^{∞}-norm (Def. 1). Then for each $f_i \in \mathcal{H}$, we have $\widetilde{f}_i \in \mathcal{C}_{\kappa}(\mathcal{X}, \varepsilon)$ such that

$$\sup_{z \in \mathcal{X}} \left| f_i(z) - \widetilde{f}_i(z) \right| \leq \varepsilon. \tag{42}$$

Now, we claim that for every $f \in \mathcal{B}_{\mathbf{k}}$, there exists $g \in \mathcal{C}_{\mathbf{k}} \triangleq (\mathcal{C}_{\kappa}(\mathcal{X}, \varepsilon))^{\otimes d}$ such that

$$\sup_{x \in \mathcal{X}^{\otimes d}} |f(x) - g(x)| \leq d\varepsilon \|\kappa\|_{\infty}^{\frac{d-1}{2}}, \tag{43}$$

which immediately implies the claimed bound (40) on the covering number. We now prove the claim (43). For any fixed $f \in \mathcal{H}_{\mathbf{k}}$, let $f_i, \widetilde{f}_i$ denote the functions satisfying (41) and (42) respectively. Then, we prove our claim (43) with $g = \prod_{i=1}^{d} \widetilde{f}_i \in \mathcal{C}_{\mathbf{k}}$. Using the convention $\prod_{k=1}^{0} \widetilde{f}_k(x_k) = 1$, we find that

$$\begin{aligned} |f(x) - g(x)| &= \left| \prod_{i=1}^{d} f_i(x_i) - \prod_{i=1}^{d} \widetilde{f}_i(x_i) \right| \\ &\leq \sum_{i=1}^{d} \left| f_i(x_i) - \widetilde{f}_i(x_i) \right| \left| \prod_{j=i+1}^{d} f_j(x_j) \prod_{k=1}^{i-1} \widetilde{f}_k(x_k) \right| \\ &\overset{(42)}{\leq} d\varepsilon \cdot \sup_{h \in \mathcal{B}_{\kappa}} \|h\|_{\infty}^{d-1} \leq d\varepsilon \|\kappa\|_{\infty}^{\frac{d-1}{2}}, \end{aligned}$$

where in the last step we have used the following argument:

$$\sup_{z \in \mathcal{X}} h(x) = \sup_{z \in \mathcal{X}} h, \kappa(z, \cdot)_{\kappa} \leq \|h\|_{\kappa} \sqrt{\kappa(z, z)} \leq \sqrt{\|\kappa\|_{\infty}} \quad \text{for any} \quad h \in \mathcal{B}_{\kappa}.$$

The proof is now complete.

[5] Steinwart & Fischer (2021, Lem. 3.11) is stated for disjoint partition of $\mathcal{X}$ in two sets, but the argument can be repeated for any finite cover of $\mathcal{X}$.

Lemma 4 (Relation between Euclidean covering numbers) *We have*

$$\mathcal{N}_\infty(\mathcal{B}_2(r), 1) \leq \frac{1}{\sqrt{\pi d}} \cdot \left[(1 + \frac{2r}{\sqrt{d}})\sqrt{2\pi e}\right]^d \quad \textit{for all} \quad d \geq 1.$$

Proof of Lem. 4 We apply Wainwright (2019, Lem. 5.7) with $\mathcal{B} = \mathcal{B}_2(r)$ and $\mathcal{B}' = \mathcal{B}_\infty(1)$ to conclude that

$$\mathcal{N}_\infty(\mathcal{B}_2(r), 1) \leq \frac{\mathrm{Vol}(2\mathcal{B}_2(r)+\mathcal{B}_\infty(1))}{\mathrm{Vol}(\mathcal{B}_\infty(1))} \leq \mathrm{Vol}(\mathcal{B}_2(2r + \sqrt{d})) \leq \frac{\pi^{d/2}}{\Gamma(\frac{d}{2}+1)} \cdot (2r + \sqrt{d})^d,$$

where $\mathrm{Vol}(\mathcal{X})$ denotes the d-dimensional Euclidean volume of $\mathcal{X} \subset \mathbb{R}^d$, and $\Gamma(a)$ denotes the Gamma function. Next, we apply the following bounds on the Gamma function from Batir (2017, Thm. 2.2):

$$\Gamma(b+1) \geq (b/e)^b\sqrt{2\pi b} \text{ for any } b \geq 1, \quad \text{and} \quad \Gamma(b+1) \leq (b/e)^b\sqrt{e^2 b} \text{ for any } b \geq 1.1.$$

Thus, we have

$$\mathcal{N}_\infty(\mathcal{B}_2(r), 1) \leq \frac{\pi^{d/2}}{\sqrt{2\pi d}(\frac{d}{2e})^{d/2}} \cdot (2r + \sqrt{d})^d \leq \frac{1}{\sqrt{\pi d}} \cdot \left[(1 + \frac{2r}{\sqrt{d}})\sqrt{2e\pi}\right]^d,$$

as claimed, and we are done.

I.2 Proof of Prop. 2: Covering numbers for analytic and differentiable kernels

First we apply Lem. 1 so that it remains to establish the stated bounds simply on $\log \mathcal{N}^\dagger_{\mathbf{k}}(\mathcal{X}, \varepsilon)$.

Proof of bound (27) in part (a) The bound (27) for the real-analytic kernel is a restatement of Sun & Zhou (2008, Thm. 2) in our notation (in particular, after making the following substitutions in their notation: $R \leftarrow 1, C_0 \leftarrow C_\kappa, r \leftarrow R_\kappa, \mathcal{X} \leftarrow \mathcal{A}, \widetilde{r} \leftarrow r^\dagger, \eta \leftarrow \varepsilon, D \leftarrow D^2_{\mathcal{A}}, n \leftarrow d$).

Proof of bound (29) for part (b): Under these assumptions, Steinwart & Christmann (2008, Thm. 6.26) states that the i-th dyadic entropy number Steinwart & Christmann (2008, Def. 6.20) of the identity inclusion mapping from $\mathcal{H}|_{\bar{\mathcal{B}}_2(r)}$ to $L^\infty_{\bar{\mathcal{B}}_2(r)}$ is bounded by $c'_{s,d,\mathbf{k}} \cdot r^s i^{-s/d}$ for some constant $c'_{s,d,\mathbf{k}}$ independent of ε and r. Given this bound on the entropy number, and applying Steinwart & Christmann (2008, Lem. 6.21), we conclude that the log-covering number $\log \mathcal{N}^\dagger_{\mathbf{k}}(\bar{\mathcal{B}}_2(r), \varepsilon)$ is bounded by $\ln 4 \cdot (c'_{s,d,\mathbf{k}} r^s/\varepsilon)^{d/s} = c_{s,d,\mathbf{k}} r^d \cdot (1/\varepsilon)^{d/s}$ as claimed.

I.3 Proof of Prop. 3: Covering numbers for specific kernels

First we apply Lem. 1 so that it remains to establish the stated bounds in each part on the corresponding $\log \mathcal{N}_{\mathbf{k}}$.

Proof for Gauss kernel: Part (a) The bound (30) for the Gaussian kernel follows directly from Steinwart & Fischer (2021, Eqn. 11) along with the discussion stated just before it. Furthermore, the bound (31) for $C_{\mathrm{Gauss},d}$ are established in Steinwart & Fischer (2021, Eqn. 6), and in the discussion around it.

Proof for Matérn kernel: Part (b) Since $\textsc{Matérn}(\nu, \gamma)$ is ν-times continuously differentiable Stein (2012), the bound (32) immediately follows from Prop. 3(b).

Proof for IMQ kernel: Part (c) The bounds (33) and (34) follow from Sun & Zhou (2008, Ex. 3), and noting that $\mathcal{N}_2(\mathcal{B}_2(r), \widetilde{r}/2)$ is bounded by $(1 + \frac{4r}{\widetilde{r}})^d$ Wainwright (2019, Lem. 5.7).

Proof for sinc-rad kernel: Part (d) First, we note that the sinc-rad kernel is positive definite in $d = \{1, 2, 3\}$ Schoenberg (1938), and thus is a reproducing kernel. Next, we claim that it is an analytic kernel that satisfies the condition (26) of Prop. 2(a) with $\kappa(t) = \textsc{sinc}(\theta\sqrt{t})$, $R_\kappa = \frac{12}{\theta^2}$, and

$C_\kappa = 1$; and thus applying the bounds (27) and (28) from Prop. 2(a) with $\mathcal{A} = \mathcal{B}_2^d(r)$ yields the claimed bound (35) and (36). To verify the condition (26) with the stated parameters, we note that

$$\kappa(t) = \textsc{sinc}(\theta\sqrt{t}) = \tfrac{1}{|\theta|\sqrt{t}} \sum_{j=0}^{\infty} \tfrac{1}{(2j+1)!} \cdot (\theta\sqrt{t})^{2j+1} = \sum_{j=0}^{\infty} \tfrac{1}{(2j+1)!} \cdot (\theta\sqrt{t})^{2j}$$
$$= \sum_{j=0}^{\infty} \tfrac{1}{(2j+1)!} \cdot \theta^{2j} \cdot t^j$$

which implies

$$\left|\kappa_+^{(j)}(0)\right| = \tfrac{1}{(2j+1)!} \cdot \theta^{2j} j! \leq (2/R_\kappa)^j j! \quad \text{for} \quad R_\kappa \triangleq \tfrac{2}{\theta^2} \cdot \inf_{j\geq 1}((2j+1)!)^{1/j} = \tfrac{12}{\theta^2},$$

and we are done.

Proof for SINC kernel: Part (e) For $\mathbf{k} = \textsc{sinc}(\theta)$, we can write $\mathbf{k}(x,y) = \prod_{i=1}^d \kappa_\theta(x_i - y_i)$ for $\kappa_\theta : \mathbb{R} \to \mathbb{R}$ defined as $\kappa_\theta(t) = \frac{\sin(\theta t)}{\theta t} \overset{(i)}{=} \frac{\sin(|\theta t|)}{|\theta t|}$, where step (i) follows from the fact that $t \mapsto \sin t/t$ is an even function. Thus, we can apply Lem. 3. Given the bound (40), and noting that $\|\kappa_\theta\|_\infty = 1$, it suffices to establish the univariate version of the bound (37), namely,

$$\mathcal{M}_{\mathbf{k}}([-r,r],\varepsilon) \leq (1 + \tfrac{4r}{\widetilde{r}_{\theta,1}}) \cdot (4\log(1/\varepsilon) + 2 + 4\log 17)^2,$$

which follows directly by substituting $d = 1$ in part(d).

Proof for B-SPLINE kernel: Part (f) For $\mathbf{k} = \text{B-SPLINE}(2\beta+1)$, we can write $\mathbf{k}(x,y) = \prod_{i=1}^d \kappa_\beta(x_i - y_i)$ for $\kappa_\beta : \mathbb{R} \to \mathbb{R}$ defined as $\kappa_\beta(t) = \mathfrak{B}_{2\beta+2}^{-1} \circledast^{2\beta+2} \mathbf{1}_{[-\frac{1}{2},\frac{1}{2}]}(t)$, and thus we can apply Lem. 3. Given the bound (40), and noting that $\|\kappa_\beta\|_\infty = 1$ (Dwivedi & Mackey (2021, Eqn. 107)), it suffices to establish the univariate version of the bound (38), namely,

$$\log \mathcal{N}_{\kappa_\beta}^{\dagger}([-\tfrac{1}{2},\tfrac{1}{2}],\varepsilon) \leq C_{\text{B-SPLINE}} \cdot (1/\varepsilon)^{\frac{1}{\beta+\frac{1}{2}}}, \tag{44}$$

where we abuse notation and use κ_β to denote the univariate B-SPLINE kernel. The bound (44) follows from Zhou (2003, Ex. 4) and using the fact that $\log \mathcal{N}_{\kappa_\beta}^{\dagger}([-\frac{1}{2},\frac{1}{2}],\varepsilon) = \log \mathcal{N}_{\kappa_\beta}^{\dagger}([0,1],\varepsilon)$ since κ_β is shift-invariant.

.

J PROOF OF TAB. 3 RESULTS

In Tab. 2, the stated results for all the entries in target KT column follow directly by substituting the covering number bounds from Prop. 3 in the corresponding entry along with the stated radii growth conditions for the target $\mathbb{P}$. For the KT+ column, results follow from the first column, and noting the following root or power KT guarantees:

For IMQ(ν,γ) kernel, one can establish a guarantee for root KT using a square-root dominating kernel $\widetilde{\mathbf{k}}_{\text{rt}}$ IMQ(ν',γ') Dwivedi & Mackey (2021, Def.2), and Dwivedi & Mackey (2021, Eqn.(117)) shows that they are always defined for appropriate choices of ν',γ'. The best root KT guarantees are obtained by choosing largest possible ν' (to allow the most rapid decay of tails), and Dwivedi & Mackey (2021, Eqn.(117)) implies with $\nu < \frac{d}{2}$, the best possible parameter satisfies $\nu' \leq \frac{d}{4} + \frac{\nu}{2}$. For this parameter, some algebra shows that $\max(\mathfrak{R}^{\dagger}_{\widetilde{\mathbf{k}}_{\text{rt}},n} \mathfrak{R}_{\widetilde{\mathbf{k}}_{\text{rt}},n}) \precsim_{d,\nu,\gamma} n^{1/2\nu}$, leading to the guarantee stated for root KT in the corresponding row. When $\nu \geq \frac{d}{2}$, we can use a MATÉRN kernel as a square-root dominating kernel from Dwivedi & Mackey (2021, Prop. 3), and then applying the bounds for the kernel radii (5), and the inflation factor (7) for a generic Matérn kernel from Dwivedi & Mackey (2021, Tab. 3) yields the entry for root KT stated in Tab. 2.

For MATÉRN(ν,γ) kernel, the α-power kernel is given by MATÉRN$(a\nu,\gamma)$ if $a\nu > \frac{d}{2}$ (a proof of this follows from Def. 2 and Dwivedi & Mackey (2021, Eqns (71-72))). Since LAPLACE(σ) = MATÉRN$(\frac{d+1}{2},\sigma^{-1})$, we conclude that its α-power kernel is defined for $a > \frac{d}{d+1}$. And using the various tail radii (5), and the inflation factor (7) for a generic Matérn kernel from Dwivedi & Mackey (2021, Tab. 3), we conclude that the MMD bounds remain same as root KT (upto factors depending on kernel parameters and d), leading to the entries in the KT+ column for MATÉRN row.

For SINC(θ) kernel, the square-root kernel is SINC(θ) itself (as can be seen by Dwivedi & Mackey (2021, Prop.2) and the fact that Fourier transform of SINC is a rectangle function on bounded domain). However, one can directly verify that the tail of SINC kernel does not decay fast enough leading to a vacuous bound of $n^{-\frac{1}{4}}$ for the root KT (with baseline thinning as i.i.d. thinning as noted in Dwivedi & Mackey (2021, Rem. 2)).

For B-SPLINE $(2\beta+1)$ kernel, using arguments similar to that in Dwivedi & Mackey (2021, Tab.4), we conclude that the α-power kernel is defined to be B-SPLINE$(A+1)$ whenever $A \triangleq 2\alpha\beta + 2\alpha - 1 \in 2\mathbb{N}_0+1$. For odd β we can always take $\alpha = \frac{1}{2}$ and the square-root kernel is well defined Dwivedi & Mackey (2021). For even β, we have to choose $\alpha \triangleq \frac{p+1}{\beta+1} \in (\frac{1}{2}, 1)$ by taking $p \in \mathbb{N}$ suitably as long as $\beta \geq 2$. Whenever the α-power kernel is defined, we can then apply the various tail radii (5), and the inflation factor (7) for the power B-SPLINE kernel from Dwivedi & Mackey (2021, Tab. 3) to obtain the MMD rates for power KT from Dwivedi & Mackey (2021, Tab. 2) (which remains the same as root KT upto factors depending on α and β), thereby providing the corresponding entry for KT+ in our Tab. 3.