Journal of Machine Learning Research 25 (2024) 1-77 Submitted 11/21; Revised 3/24; Published 4/24

Kernel Thinning

Raaz Dwivedi DWIVEDI@CORNELL.EDU
Cornell Tech

Lester Mackey LMACKEY@MICROSOFT.COM
Microsoft Research New England

Editor: Ingo Steinwart

Abstract

We introduce kernel thinning, a new procedure for compressing a distribution $\mathbb{P}$ more effectively than i.i.d. sampling or standard thinning. Given a suitable reproducing kernel $\mathbf{k}_\star$ and $\mathcal{O}(n^2)$ time, kernel thinning compresses an n-point approximation to $\mathbb{P}$ into a $\sqrt{n}$-point approximation with comparable worst-case integration error across the associated reproducing kernel Hilbert space. The maximum discrepancy in integration error is $\mathcal{O}_d(n^{-1/2}\sqrt{\log n})$ in probability for compactly supported $\mathbb{P}$ and $\mathcal{O}_d(n^{-\frac{1}{2}}(\log n)^{(d+1)/2}\sqrt{\log\log n})$ for sub-exponential $\mathbb{P}$ on $\mathbb{R}^d$. In contrast, an equal-sized i.i.d. sample from $\mathbb{P}$ suffers $\Omega(n^{-1/4})$ integration error. Our sub-exponential guarantees resemble the classical quasi-Monte Carlo error rates for uniform $\mathbb{P}$ on $[0,1]^d$ but apply to general distributions on $\mathbb{R}^d$ and a wide range of common kernels. Moreover, the same construction delivers near-optimal L^∞ coresets in $\mathcal{O}(n^2)$ time. We use our results to derive explicit non-asymptotic maximum mean discrepancy bounds for Gaussian, Matérn, and B-spline kernels and present two vignettes illustrating the practical benefits of kernel thinning over i.i.d. sampling and standard Markov chain Monte Carlo thinning, in dimensions $d = 2$ through 100.

Keywords: coresets, distribution compression, Markov chain Monte Carlo, maximum mean discrepancy, reproducing kernel Hilbert space, thinning

1. Introduction

Monte Carlo and Markov chain Monte Carlo (MCMC) methods (Brooks et al., 2011) are commonly used to approximate intractable target expectations $\mathbb{P}f \triangleq \mathbb{E}_{X\sim\mathbb{P}}[f(X)]$ of $\mathbb{P}$-integrable functions f with asymptotically exact averages $\mathbb{P}_n f \triangleq \frac{1}{n}\sum_{i=1}^n f(x_i)$ based on points $(x_i)_{i=1}^n$ generated from a Markov chain. A standard practice, to minimize the expense of downstream function evaluation, is to *thin* the Markov chain output down to a smaller size n_{out} by keeping only every (n/n_{out})-th sample point (Owen, 2017). We call this approach *standard thinning*, and such sample compression is critical in fields like computational cardiology in which each function evaluation triggers an organ or tissue simulation consuming thousands of CPU hours (Niederer et al., 2011; Augustin et al., 2016; Strocchi et al., 2020). Unfortunately, standard thinning also leads to a significant reduction in accuracy. For example, thinning one's chain down to $n_{\mathrm{out}} = \sqrt{n}$ sample points increases integration error from $\mathcal{O}(n^{-\frac{1}{2}})$ in probability to $\Omega(n^{-\frac{1}{4}})$ by the Markov chain central limit theorem (Roberts and Rosenthal, 2004, Prop. 29). Our primary contribution is a more effective thinning strategy, which provides $o_p(n^{-\frac{1}{4}})$-integration error when $n^{\frac{1}{2}}$ points are returned.

1.1 Thinned MMD coresets

We focus on integration error in a reproducing kernel Hilbert space (RKHS, Steinwart and Christmann, 2008, Def. 4.18) of bounded, measurable functions with a target kernel $\mathbf{k}_\star : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ for $d \in \mathbb{N}$ and RKHS norm $\|\cdot\|_{\mathbf{k}_\star}$.

Assumption 1 (RKHS of bounded, measurable functions)**.** *The RKHS $\mathcal{H}_\mathbf{k}$ of a kernel $\mathbf{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ contains only bounded measurable functions. Equivalently, $\mathbf{k}$ is bounded with $\mathbf{k}(x, \cdot)$ measurable for all $x \in \mathbb{R}^d$ (Steinwart and Christmann, 2008, Lems. 4.23, 4.24).*[1]

The worst-case integration error over the RKHS unit ball is given by the kernel *maximum mean discrepancy* (MMD, Gretton et al., 2012).

Definition 1 (Maximum mean discrepancy (Gretton et al., 2012))**.** *For a kernel $\mathbf{k}$ satisfying Assump. 1, we define the kernel* maximum mean discrepancy*,*

$$\mathrm{MMD}_\mathbf{k}(\mu, \nu) \triangleq \sup_{f \in \mathcal{H}_\mathbf{k} : \|f\|_\mathbf{k} \le 1} |\mu f - \nu f| \quad \textit{for all probability measures } \mu, \nu \textit{ on } \mathbb{R}^d. \tag{1}$$

For sequences of points $\mathcal{S}$ and $\mathcal{S}'$ in $\mathbb{R}^d$ with empirical distributions $\mathbb{Q}$ and $\mathbb{Q}'$, we overload this notation to write $\mathrm{MMD}_\mathbf{k}(\mathbb{P}, \mathcal{S}) \triangleq \mathrm{MMD}_\mathbf{k}(\mathbb{P}, \mathbb{Q})$ and $\mathrm{MMD}_\mathbf{k}(\mathcal{S}, \mathcal{S}') \triangleq \mathrm{MMD}_\mathbf{k}(\mathbb{Q}, \mathbb{Q}')$.

Given $\mathbf{k}_\star$ satisfying Assump. 1, a target distribution $\mathbb{P}$ on $\mathbb{R}^d$, and a sequence of $\mathbb{R}^d$-valued points $(x_i)_{i=1}^n$ generated to approximate $\mathbb{P}$, our aim is to identify a *thinned MMD coreset*, a shorter subsequence that continues to approximate $\mathbb{P}$ well in $\mathrm{MMD}_{\mathbf{k}_\star}$.

Definition 2 (MMD coreset)**.** *We call a sequence of n_{out} points in $\mathbb{R}^d$ with empirical measure $\mathbb{Q}$ an $(n_{\mathrm{out}}, \varepsilon)$-*MMD coreset *for $(\mathbf{k}_\star, \mathbb{P})$ if $\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathbb{Q}) \le \varepsilon$.*

Notably, when the initial sequence is drawn i.i.d. or from a fast-mixing Markov chain targeting $\mathbb{P}$, standard thinning down to size $n_{\mathrm{out}} = n^{\frac{1}{2}}$ yields an order $(n^{\frac{1}{2}}, n^{-\frac{1}{4}})$-MMD coreset in probability (see Prop. 1). A benchmark for improvement is provided by the online Haar strategy of Dwivedi et al. (2019), which generates an $(n^{\frac{1}{2}}, \mathcal{O}_d(n^{-\frac{1}{2}} \log^{2d} n))$-MMD coreset in probability from $2n^{\frac{1}{2}}$ i.i.d. sample points when $\mathbb{P}$ is specifically the uniform distribution on the unit cube $[0, 1]^d$.[2] Our goal is to develop thinned coresets of improved quality for any target $\mathbb{P}$ with sufficiently fast tail decay.

1.2 Our contributions

To this end, we introduce *kernel thinning* (Alg. 1), a new, practical solution to the thinned MMD coreset problem that takes as input an $(n, \mathcal{O}_p(n^{-\frac{1}{2}}))$-MMD coreset and outputs an $(n^{\frac{1}{2}}, o_p(n^{-\frac{1}{4}}))$-MMD coreset for a wide-range of $(\mathbf{k}_\star, \mathbb{P})$. Kernel thinning uses non-uniform randomness and evaluations of a less smooth *square-root kernel* $\mathbf{k}_{\mathrm{rt}}$ (see Def. 5) to partition the input into subsets of comparable quality and then greedily refines the best of these subsets using $\mathbf{k}_\star$. Our primary contributions include:

1. Throughout, we use $\mathbf{k}$ for statements involving a generic kernel that is potentially distinct from the target kernel $\mathbf{k}_\star$.
2. Dwivedi et al. (2019) specifically control the *star discrepancy*, a quantity which in turn upper bounds a Sobolev space MMD called the L^2 *discrepancy* (Hickernell, 1998; Novak and Wozniakowski, 2010).

1. **Better-than-i.i.d. MMD coresets:** Given n input points sampled i.i.d. or from a fast-mixing Markov chain, kernel thinning yields, in probability, an $(n^{\frac{1}{2}}, \mathcal{O}_d(n^{-\frac{1}{2}}\sqrt{\log n}))$-MMD coreset for $\mathbb{P}$ and $\mathbf{k}_{\mathrm{rt}}$ with bounded support, an $(n^{\frac{1}{2}}, \mathcal{O}_d(n^{-\frac{1}{2}}\sqrt{\log^{d+1} n \log\log n}))$-MMD coreset for $\mathbb{P}$ and $\mathbf{k}_{\mathrm{rt}}$ with light tails, and an $(n^{\frac{1}{2}}, \mathcal{O}_d(n^{-\frac{1}{2}+\frac{d}{2\rho}}\sqrt{\log n \log\log n}))$-MMD coreset for $\mathbb{P}$ and $\mathbf{k}_{\mathrm{rt}}^2$ with $\rho > 2d$ moments (Thm. 1 and Cor. 1). For compactly supported or light-tailed $\mathbb{P}$ and $\mathbf{k}_{\mathrm{rt}}$, these results compare favorably with known $\Omega_d(n^{-\frac{1}{2}})$ lower bounds (see Sec. 8.1). Our guarantees extend to more general input point sequences, including deterministic sequences based on quadrature or kernel herding (Chen et al., 2010), and give rise to explicit, non-asymptotic error bounds for a wide variety of popular kernels including Gaussian, Matérn, and B-spline kernels. While $(n^{\frac{1}{2}}, \mathcal{O}_d(n^{-\frac{1}{2}}\log^{\frac{d-1}{2}} n))$-MMD coresets have been developed for specific $(\mathbf{k}_\star, \mathbb{P})$ pairings like the uniform distribution on $[0,1]^d$ and an L^2 discrepancy kernel $\mathbf{k}_\star$ (see Sec. 8.1), to the best of our knowledge, no prior $(n^{\frac{1}{2}}, o_p(n^{-\frac{1}{4}}))$-MMD coreset constructions were known for the range of $\mathbb{P}$ and $\mathbf{k}_\star$ studied in this work.

2. **MMD error from square-root L^∞ error:** To derive our MMD guarantees for kernel thinning, we first establish an important link between MMD coresets for $\mathbf{k}_\star$ and L^∞ *coresets* for $\mathbf{k}_{\mathrm{rt}}$.

   Definition 3 (L^∞ coreset)**.** *For any kernel* $\mathbf{k}$ *satisfying Assump. 1, probability measure* μ *on* $\mathbb{R}^d$*, and* $z \in \mathbb{R}^d$*, let* $\mu\mathbf{k}(z) \triangleq \mathbb{E}_{X\sim\mu}[\mathbf{k}(X,z)]$*. We call a sequence of* n_{out} *points in* $\mathbb{R}^d$ *with empirical measure* $\mathbb{Q}$ *an* $(n_{\mathrm{out}}, \varepsilon)$*-*$L^\infty$ coreset *for* $(\mathbf{k}, \mathbb{P})$ *if* $\|\mathbb{P}\mathbf{k} - \mathbb{Q}\mathbf{k}\|_\infty \le \varepsilon$*.*

   Thm. 2 and Cor. 2 show that *any* L^∞ coreset for $(\mathbf{k}_{\mathrm{rt}}, \mathbb{P})$ is also an MMD coreset for $(\mathbf{k}_\star, \mathbb{P})$ with quality depending on the tail decay of $\mathbf{k}_{\mathrm{rt}}$ and $\mathbb{P}$.

3. **Online vector balancing in Hilbert spaces:** As a building block for constructing high-quality coresets, we introduce and analyze a Hilbert space generalization of the self-balancing walk of Alweiss et al. (2021) to partition a sequence of functions (like $(\mathbf{k}_{\mathrm{rt}}(x_i, \cdot))_{i=1}^n$) into nearly equal halves. Our analysis of this *self-balancing Hilbert walk* (SBHW, Alg. 3) in Thm. 3 may be of independent interest for solving the online vector balancing problem of Spencer (1977) in Hilbert spaces (Cor. 4).

4. **Efficient, near-optimal L^∞ coresets:** We then design a symmetrized version of SBHW for RKHSes—kernel halving—that delivers 2-thinned coresets with small L^∞ error (Alg. 2 and Thm. 4). The first stage of kernel thinning, KT-SPLIT, recursively applies kernel halving to $\mathbf{k}_{\mathrm{rt}}$ to obtain near-minimax-optimal L^∞ coresets in $\mathcal{O}(n^2)$ time with $\mathcal{O}(n\min(d,n))$ space (Cors. 5 and 6).

After describing our kernel and input point requirements in Sec. 2, we detail the kernel thinning and kernel halving algorithms in Sec. 3. Sec. 4 houses our main MMD guarantees, both for kernel thinning and for generic L^∞ square-root kernel coresets. We introduce and analyze the self-balancing Hilbert walk in Sec. 5 and present our main L^∞ guarantees for kernel halving and KT-SPLIT in Sec. 6. Sec. 7 complements our theoretical contributions with two vignettes illustrating the practical benefits of kernel thinning over (a) i.i.d. sampling in dimensions $d = 2$ through 100 and (b) standard MCMC thinning across twelve experiments targeting challenging differential equation posterior distributions. We conclude with a discussion of our results, related work, and future directions in Sec. 8 and defer all proofs to the appendices.

Notation We define the shorthand $[n] \triangleq \{1, \dots, n\}$ for $n \in \mathbb{N}$, $a \wedge b \triangleq \min(a, b)$ for $a, b \in \mathbb{R}$, $\mathbb{R}_+ \triangleq \{x \in \mathbb{R} : x \geq 0\}$, and $\mathcal{B}(x; r) \triangleq \left\{y \in \mathbb{R}^d \mid \|x - y\|_2 < r\right\}$ for $r \in \mathbb{R}$. We use $\mathrm{Vol}(\mathcal{B})$ to denote the volume of a compact set $\mathcal{B} \subset \mathbb{R}^d$. We let $\mathcal{A}^c$ denote the complement of a set $\mathcal{A} \subset \mathbb{R}^d$ and $\mathbb{I}_{\mathcal{A}}(x) = 1$ if $x \in \mathcal{A}$ and 0 otherwise. We use $\Pr(\mathcal{E})$ to denote the probability of an event $\mathcal{E}$. For real-valued kernels $\mathbf{k}$ and functions f on $\mathbb{R}^d$, we make frequent use of the norms $\|\mathbf{k}\|_\infty = \sup_{x,y\in\mathbb{R}^d} |\mathbf{k}(x, y)|$ and $\|f\|_\infty = \sup_{x\in\mathbb{R}^d} |f(x)|$. For $x > 0$, we use $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}\, dt$ to denote the Gamma function (with $\Gamma(n) = (n-1)!$ for $n \in \mathbb{N}$).

For two sequences of real numbers $(a_n)_{n\in\mathbb{N}}$ and $(b_n)_{n\in\mathbb{N}}$, we say that a_n is of order b_n and write $a_n = \mathcal{O}(b_n)$ or $a_n \precsim b_n$ to denote that $a_n \leq cb_n$ for all $n \in \mathbb{N}$ and some constant $c > 0$. We write $a_n = \Omega(b_n)$ if $b_n = \mathcal{O}(a_n)$ and $a_n = \Theta(b_n)$ when $a = \Omega(b_n)$ and $a = \mathcal{O}(b_n)$. Moreover, we use $a_n = \mathcal{O}_d(b_n), a_n \precsim_d b_n, a_n = \Omega_d(b_n), a_n \succsim_d b_n$ to indicate dependency of underlying universal constant on d. We say $a_n = o(b_n)$ if $\lim_{n\to\infty} a_n/b_n = 0$. For a sequence of real-valued random variables $(X_n)_{n\in\mathbb{N}}$, we write $X_n = \mathcal{O}_P(a_n)$ or $X_n = \mathcal{O}(a_n)$ in probability, when $\frac{X_n}{a_n}$ is stochastically bounded, i.e., for all $\delta > 0$, there exists finite c_δ and n_δ such that $\Pr(|\frac{X_n}{a_n}| > c_\delta) < \delta$, for all $n > n_\delta$. We write $X_n = \Omega_P(a_n)$ if $1/X_n = \mathcal{O}_P(1/b_n)$ and $X_n = o_p(a_n)$ when $\frac{X_n}{a_n} \to 0$ in probability, i.e., for all $\varepsilon > 0$, $\Pr(|\frac{X_n}{a_n}| \geq \varepsilon) \to 0$.

We write *order* (n, ε)*-MMD (or* L^∞*) coreset* to mean an $(n, \mathcal{O}(\epsilon))$-MMD (or L^∞) coreset and append *in probability* to mean an $(n, \mathcal{O}_p(\epsilon))$-MMD (or L^∞) coreset.

2. Input Point and Kernel Requirements

Given a target distribution $\mathbb{P}$ on $\mathbb{R}^d$, a kernel $\mathbf{k}_\star$ satisfying Assump. 1, and a sequence of $\mathbb{R}^d$-valued input points $\mathcal{S}_n = (x_i)_{i=1}^n$ generated either randomly or deterministically, our goal is to identify a better-than-i.i.d. thinned MMD coreset, that is, a subsequence $\mathcal{S}_{\mathrm{out}}$ of size $n^{\frac{1}{2}}$ satisfying $\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathcal{S}_{\mathrm{out}}) = o_p(n^{-\frac{1}{4}})$. When drawing asymptotic conclusions, we will view d as fixed and $\mathcal{S}_n$ as a prefix of an infinite sequence of points $\mathcal{S}_\infty \triangleq (x_i)_{i=1}^\infty$.

2.1 Input point requirements

Our algorithms are designed to return high quality MMD coresets for the *input* $\mathcal{S}_n$. To translate these into high quality coresets for the target $\mathbb{P}$, it suffices, by the triangle inequality, for the input points to have quality $\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathcal{S}_n) = \mathcal{O}_p(n^{-\frac{1}{2}})$. As we discuss in Sec. 8.1, input sequences generated by i.i.d. sampling, kernel herding (Chen et al., 2010), Stein Point MCMC (Chen et al., 2019), and greedy sign selection (Karnin and Liberty, 2019) all satisfy this property. Moreover, we prove in App. B that an analogous guarantee holds for the iterates of a fast-mixing Markov chain.

Proposition 1 (MMD guarantee for MCMC)**.** *Consider a homogeneous* ϕ*-irreducible geometrically ergodic Markov chain (Gallegos-Herrada et al., 2023, Thm. 1xi) with initial state* x_0*, subsequent iterates* $\mathcal{S}_\infty$*, and stationary distribution* $\mathbb{P}$*. If* $\mathbf{k}_\star$ *satisfies Assump. 1, then there exists a* $\mathbb{P}$*-almost everywhere finite function* $c : \mathbb{R}^d \to (0, \infty]$ *such that, for any given* $n \in \mathbb{N}$ *and* $\delta \in (0, 1)$*,* $\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathcal{S}_n) \leq \sqrt{\frac{c(x_0)\|\mathbf{k}_\star\|_\infty \log(e/\delta)}{n}}$ *with probability* $1 - \delta$ *given* x_0*.*

The *input radius*,

$$R_{\mathcal{S}_n} \triangleq \max_{x\in\mathcal{S}_n} \|x\|_2, \tag{2}$$

will also play an important role in our results. In particular, the growth rate of this radius as a function of n impacts the growth rate of our MMD bounds. Our next definition assigns familiar names to the most commonly encountered growth rates.

Definition 4 (Input radius growth rates). *We say the point sequence $\mathcal{S}_\infty$ with prefixes $\mathcal{S}_n$ for $n \in \mathbb{N}$ is* Compact *if $R_{\mathcal{S}_n} = \mathcal{O}_d(1)$,* SubGauss *if $R_{\mathcal{S}_n} = \mathcal{O}_d(\sqrt{\log n})$,* SubExp *if $R_{\mathcal{S}_n} = \mathcal{O}_d(\log n)$, and* HeavyTail$(\rho)$ *with $\rho > 0$ if $R_{\mathcal{S}_n} = \mathcal{O}_d(n^{1/\rho})$.*

These growth rates are exactly those which arise with probability 1 when an input sequence is generated identically from $\mathbb{P}$ with corresponding tail behavior or from a fast-mixing Markov chain targeting $\mathbb{P}$. Our proof of this result is given in App. C.

Proposition 2 (Almost sure radius growth). *Consider either (i) points $\mathcal{S}_\infty$ sampled identically (but not necessarily independently) from $\mathbb{P}$ with x_0 independent or (ii) a homogeneous ϕ-irreducible geometrically ergodic Markov chain with initial state x_0, subsequent iterates $\mathcal{S}_\infty$, and stationary distribution $\mathbb{P}$. Then the following statements hold true for any non-negative c and ρ and $\mathbb{P}$-almost every x_0.*

(a) If $\mathbb{P}$ is compactly supported, then, with probability 1 conditional on x_0, $\mathcal{S}_\infty$ is Compact.

(b) If $\mathbb{E}_{X\sim\mathbb{P}}[e^{c\|X\|_2^2}] < \infty$, then, with probability 1 conditional on x_0, $\mathcal{S}_\infty$ is SubGauss.

(c) If $\mathbb{E}_{X\sim\mathbb{P}}[e^{c\|X\|_2}] < \infty$, then, with probability 1 conditional on x_0, $\mathcal{S}_\infty$ is SubExp.

(d) If $\mathbb{E}_{X\sim\mathbb{P}}[\|X\|_2^\rho] < \infty$, then, with probability 1 conditional on x_0, $\mathcal{S}_\infty$ is HeavyTail(ρ).

Finally, we will also require $\mathcal{S}_\infty$ to be *oblivious*, that is, generated independently of any randomness in the thinning algorithm. To capture this assumption, we treat $\mathcal{S}_\infty$ as fixed and deterministic hereafter. This treatment is without loss of generality since our results hold conditional on the observed values of $(x_i)_{i=1}^\infty$ when the points are random and oblivious.

2.2 Kernel requirements

We use the terms *reproducing kernel* and *kernel* interchangeably to indicate that $\mathbf{k}_\star$ is symmetric and positive definite, i.e., that the kernel matrix $(\mathbf{k}_\star(z_i, z_j))_{i,j=1}^l$ is symmetric and positive semidefinite for any evaluation points $(z_i)_{i=1}^l$ in $\mathbb{R}^d$. In addition to $\mathbf{k}_\star$, our algorithm takes as input a *square-root kernel* for $\mathbf{k}_\star$.

Definition 5 (Square-root kernel). *We say a kernel $\mathbf{k}_{\mathrm{rt}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is a* square-root kernel *for $\mathbf{k}_\star : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ if $\mathbf{k}_{\mathrm{rt}}(x, \cdot)$ is square integrable for all $x \in \mathbb{R}^d$ with*

$$\mathbf{k}_\star(x, y) = \textstyle\int_{\mathbb{R}^d} \mathbf{k}_{\mathrm{rt}}(x, z)\mathbf{k}_{\mathrm{rt}}(y, z) dz \quad \textit{for all} \quad x, y \in \mathbb{R}^d. \tag{3}$$

We highlight that a square-root kernel need not be unique and that its existence is an indication of a certain degree of smoothness in the target kernel $\mathbf{k}_\star$. One convenient tool for deriving square-root kernels is the notion of a *spectral density*.

Name of kernel $\mathbf{k}_\star(x,y)=\kappa(x-y)$	Expression for $\kappa(z)$	Fourier transform $\widehat{\kappa}(\omega)$	Square-root kernel $\mathbf{k}_{\mathrm{rt}}$
$\mathbf{Gaussian}(\sigma)$: $\sigma>0$	$\exp\Big(-\frac{\lVert z\rVert_2^2}{2\sigma^2}\Big)$	$\sigma^d\exp\Big(-\frac{\sigma^2\lVert\omega\rVert_2^2}{2}\Big)$	$\big(\frac{2}{\pi\sigma^2}\big)^{\frac{d}{4}}\mathbf{Gaussian}\Big(\frac{\sigma}{\sqrt{2}}\Big)$
$\mathbf{Mat\acute{e}rn}(\nu,\gamma)$: $\nu>d,\gamma>0$	$c_{\nu-\frac{d}{2}}(\gamma\lVert z\rVert_2)^{\nu-\frac{d}{2}}K_{\nu-\frac{d}{2}}(\gamma\lVert z\rVert_2)$	$\phi_{d,\nu,\gamma}\,(\gamma^2+\lVert\omega\rVert_2^2)^{-\nu}$	$A_{\nu,\gamma,d}\mathbf{Mat\acute{e}rn}(\frac{\nu}{2},\gamma)$
$\mathbf{B\text{-}spline}(2\beta+1)$: $\beta\in 2\mathbb{N}+1$	$S_{2\beta+2,d}\prod_{j=1}^{d}\circledast^{2\beta+2}\mathbb{I}_{[-\frac{1}{2},\frac{1}{2}]}(z_j)$	$S'_{2\beta+2,d}\prod_{j=1}^{d}\frac{\sin^{2\beta+2}(\frac{\omega_j}{2})}{\omega_j^{2\beta+2}}$	$\widetilde{S}_{\beta,d}\mathbf{B\text{-}spline}(\beta)$

Table 1: Square-root kernels $\mathbf{k}_{\mathrm{rt}}$ for common target kernels $\mathbf{k}_\star$. Each $\mathbf{k}_\star$ satisfies $\lVert\mathbf{k}_\star\rVert_\infty=1$, and the parameter range ensures the existence of $\mathbf{k}_{\mathrm{rt}}$. Above, $\circledast^\ell$ denotes recursive convolution with ℓ function copies, K_a denotes the modified Bessel function of the third kind (Wendland, 2004, Def. 5.10), $c_b\triangleq\frac{2^{1-b}}{\Gamma(b)}$, $\phi_{d,\nu,\gamma}=\frac{c_{\nu-d/2}}{c_\nu}\gamma^{2\nu-d}$, $A_{\nu,\gamma,d}\triangleq\big(\frac{1}{4\pi}\gamma^2\big)^{d/4}\sqrt{\frac{\Gamma(\nu)}{\Gamma(\nu-d/2)}}\cdot\frac{\Gamma((\nu-d)/2)}{\Gamma(\nu/2)}$, $S'_{2\beta+2,d}\triangleq S_{2\beta+2,d}\cdot(\frac{4^{\beta+1}}{\sqrt{2\pi}})^d$, and $\widetilde{S}_{\beta,d}\triangleq\frac{\sqrt{S_{2\beta+2,d}}}{S_{\beta+1,d}}$ where $S_{\beta,d}$ is defined in (89). See App. N for our derivation.

Definition 6 (Shift invariance and spectral density)**.** *We call a kernel of the form* $\mathbf{k}(x,y)=\kappa(x-y)$ *for* $\kappa:\mathbb{R}^d\to\mathbb{R}$ shift-invariant *and say* $\mathbf{k}$ *has* spectral density $\widehat{\kappa}$ *if* κ *is the Fourier transform of a finite measure with Lebesgue density* $\widehat{\kappa}$*, i.e.,* $\kappa(z)=\frac{1}{(2\pi)^{d/2}}\int e^{-i\langle\omega,z\rangle}\widehat{\kappa}(\omega)d\omega$.

As we show in Apps. N and Q, many familiar kernels admit spectral densities, including Gaussian, Matérn, B-spline, inverse multiquadric, sech, and Wendland's compactly supported kernels. Moreover, by Bochner's theorem (Bochner, 1933; Wendland, 2004, Thm. 6.6) and the Fourier inversion theorem (Wendland, 2004, Cor. 5.24), any continuous $\mathbf{k}_\star(x,y)=\kappa(x-y)$ with absolutely integrable κ has a spectral density equal to the Fourier transform of κ. Our next result (proved in App. D) derives a square-root kernel for any shift-invariant $\mathbf{k}_\star$ with a square-root integrable spectral density.

Proposition 3 (Shift-invariant square-root kernels)**.** *If a kernel* $\mathbf{k}_\star(x,y)=\kappa(x-y)$ *admits a spectral density (Def. 6)* $\widehat{\kappa}$ *with* $\int\sqrt{\widehat{\kappa}(\omega)}d\omega<\infty$*, then* $\mathbf{k}_{\mathrm{rt}}(x,y)=\frac{\kappa_{\mathrm{rt}}(x-y)}{(2\pi)^{d/4}}$ *is a square-root kernel of* $\mathbf{k}_\star$ *for* κ_{rt} *the Fourier transform of* $\sqrt{\widehat{\kappa}}$.

Tab. 1 gives several examples of common kernels satisfying the conditions of Prop. 3 along with their associated square-root kernels. For example, if $\mathbf{k}_\star$ is Gaussian with bandwidth σ, then a rescaled Gaussian kernel with bandwidth $\frac{\sigma}{\sqrt{2}}$ is a valid choice for $\mathbf{k}_{\mathrm{rt}}$. For simplicity, our results in the sequel assume the use of an exact square-root kernel $\mathbf{k}_{\mathrm{rt}}$, but, as we detail in App. Q, it suffices to use the square-root of any kernel that dominates $\mathbf{k}_\star$ in the positive-definite order (see Def. 8). For example, we show in Prop. 4 of App. Q that a standard Matérn kernel is a suitable square-root dominating kernel for any sufficiently-smooth shift-invariant $\mathbf{k}_\star$ with absolutely integrable κ. In Tab. 5 of App. Q, we also derive convenient tailored square-root dominating kernels for inverse multiquadric, sech, and Wendland's compactly supported kernels.

Finally, we define several kernel growth and decay properties that will be explicitly assumed in some of our results.

Assumption 2 (Lipschitz kernel)**.** *The kernel* $\mathbf{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *admits a Lipschitz constant*

$$L_{\mathbf{k}} \triangleq \sup_{x,y,z} \frac{|\mathbf{k}(x,y)-\mathbf{k}(x,z)|}{\|y-z\|_2} < \infty. \tag{4}$$

Assumption 3 (Kernel tail decay)**.** *The kernel* $\mathbf{k}$ *satisfies Assump. 1 and, for each* $\varepsilon > 0$*,*

$$\begin{aligned} &\max\{ \ \inf\{r : \sup_{\substack{x,y:\\ \|x-y\|_2 \geq r}} |\mathbf{k}(x,y)| \leq \varepsilon\}, \quad \inf\{r : \tau_{\mathbf{k}}(r) \leq \varepsilon\} \ \} < \infty, \\ &\textit{where} \quad \tau_{\mathbf{k}}(r) \triangleq (\sup_x \textstyle\int_{\|y\|_2 \geq r} \mathbf{k}^2(x, x-y)dy)^{\frac{1}{2}} \quad \textit{for} \quad r \geq 0. \end{aligned} \tag{5}$$

The following definition gives familiar names to commonly encountered tail decay rates.

Definition 7 (Kernel tail decay rate)**.** *For a kernel* $\mathbf{k}$ *satisfying Assump. 3, define*

$$\begin{aligned} &R_{\mathbf{k},n} \triangleq \inf\{r : \sup_{\substack{x,y:\\ \|x-y\|_2 \geq r}} |\mathbf{k}(x,y)| \leq \tfrac{\|\mathbf{k}\|_\infty}{n}\}, \quad R'_{\mathbf{k},n} \triangleq \inf\{r : \tau_{\mathbf{k}}(r) \leq \tfrac{\|\mathbf{k}\|_\infty}{\sqrt{n}}\}, \\ \textit{and} \quad &R^\dagger_{\mathbf{k},n} \triangleq \max\{R_{\mathbf{k},n}, R'_{\mathbf{k},n}\} \end{aligned} \tag{6}$$

for $\tau_{\mathbf{k}}$ *defined in* (5)*. We say* $\mathbf{k}$ *is (a)* Compact *if* $R^\dagger_{\mathbf{k},n} = \mathcal{O}_d(1)$*, (b)* SubGauss *if* $R^\dagger_{\mathbf{k},n} = \mathcal{O}_d(\sqrt{\log n})$*, (c)* SubExp *if* $R^\dagger_{\mathbf{k},n} = \mathcal{O}_d(\log n)$*, (d)* HeavyTail$(\rho)$ *for* $\rho > 0$ *if* $R^\dagger_{\mathbf{k},n} = \mathcal{O}_d(n^{1/\rho})$*, and (e)* SubPoly *if* $\log R^\dagger_{\mathbf{k},n} = \mathcal{O}_d(\log n)$*. Notably, any* Compact*,* SubGauss*,* SubExp*, or* HeavyTail(ρ) $\mathbf{k}$ *is also* SubPoly*.*

Remark 1 ($\mathbf{k}_{\mathrm{rt}}$ tail decay implies $\mathbf{k}_\star$ boundedness)**.** *If* $\mathbf{k}_{\mathrm{rt}}$ *satisfying Assump. 3 is a square-root kernel of* $\mathbf{k}_\star$*, then there exists a finite* r *for which* $\mathbf{k}_\star(x,x) = \int \mathbf{k}^2_{\mathrm{rt}}(x, x-y)dy \leq \|\mathbf{k}_{\mathrm{rt}}\|^2_\infty \mathrm{Vol}(\mathcal{B}(0;r)) + \tau^2_{\mathbf{k}_{\mathrm{rt}}}(r) < \infty$*.*

Popular examples of Compact, SubGauss, and SubExp $\mathbf{k}$ are B-spline, Gaussian, and Matérn kernels respectively (see Tab. 2). Moreover, one can directly verify that an inverse multiquadric $\mathbf{k}(x,y) = (\gamma^2 + \|x-y\|_2^2)^{-\nu}$ with $\gamma > 0$ and $\nu > \frac{d}{4}$ is HeavyTail$(2\nu \wedge (4\nu - d))$.

3. Kernel Thinning

Our solution to the thinned coreset problem is *kernel thinning*, described in Alg. 1. Given a thinning parameter $m \in \mathbb{N}$, kernel thinning proceeds in two stages: kt-split and kt-swap.

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

Input: kernels $(\mathbf{k}_\star, \mathbf{k}_{\mathrm{rt}})$, input points $\mathcal{S}_n = (x_i)_{i=1}^n$, thinning parameter $m \in \mathbb{N}$, probabilities $(\delta_i)_{i=1}^{\lfloor \frac{n}{2} \rfloor}$

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

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

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

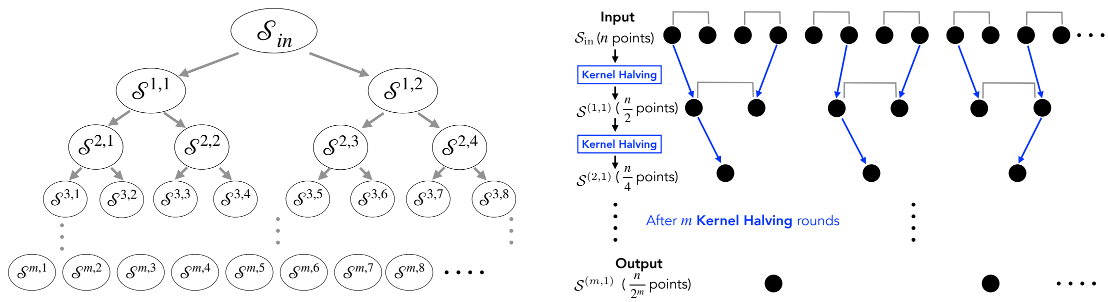


Figure 1: Overview of KT-SPLIT. (Left) KT-SPLIT recursively partitions its input $\mathcal{S}_n$ into 2^m balanced coresets $\mathcal{S}^{(m,\ell)}$ of size $\lfloor \frac{n}{2^m} \rfloor$. (Right) In Sec. 6, we interpret each coreset $\mathcal{S}^{(m,\ell)}$ as the output of repeated *kernel halving*: on each halving round, remaining points are paired, and one point from each pair is selected using non-uniform randomness.

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

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

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

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

for $i = 1, \ldots, \lfloor n/2 \rfloor$ **do**
  $\mathcal{S}^{(0,1)}$`.append`(x_{2i-1}); $\mathcal{S}^{(0,1)}$`.append`(x_{2i})
  // Every 2^{j-1} rounds add point from parent coreset $\mathcal{S}^{(j-1,\ell)}$ to each child $\mathcal{S}^{(j,2\ell-1)}$, $\mathcal{S}^{(j,2\ell)}$
  for ($j = 1$; $j \le m$ **and** $i/2^{j-1} \in \mathbb{N}$; $j = j+1$) **do**
    for $\ell = 1, \ldots, 2^{j-1}$ **do**
      $(\mathcal{S}, \mathcal{S}') \leftarrow (\mathcal{S}^{(j-1,\ell)}, \mathcal{S}^{(j,2\ell-1)})$; $(x, x') \leftarrow$ `get_last_two_points`$(\mathcal{S})$
      // Compute swapping threshold $\mathfrak{a}$
      $\mathfrak{a}, \sigma_{j,\ell} \leftarrow$`get_swap_params`$(\sigma_{j,\ell}, \mathfrak{b}, \delta_{|\mathcal{S}|/2} \cdot \frac{2^{j-1}}{m})$ for $\mathfrak{b}^2 = \mathbf{k}_{\mathrm{rt}}(x,x) + \mathbf{k}_{\mathrm{rt}}(x',x') - 2\mathbf{k}_{\mathrm{rt}}(x,x')$
      // Assign one point to each child after probabilistic swapping
      $\alpha \leftarrow \mathbf{k}_{\mathrm{rt}}(x',x') - \mathbf{k}_{\mathrm{rt}}(x,x) + \Sigma_{y \in \mathcal{S}}(\mathbf{k}_{\mathrm{rt}}(y,x) - \mathbf{k}_{\mathrm{rt}}(y,x')) - 2\Sigma_{z \in \mathcal{S}'}(\mathbf{k}_{\mathrm{rt}}(z,x) - \mathbf{k}_{\mathrm{rt}}(z,x'))$
      $(x, x') \leftarrow (x', x)$ *with probability* $\min(1, \frac{1}{2}(1 - \frac{\alpha}{\mathfrak{a}})_+)$
      $\mathcal{S}^{(j,2\ell-1)}$`.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 $\lfloor n/2^m \rfloor$

---

function `get_swap_params`$(\sigma, \mathfrak{b}, \delta)$:
  $\mathfrak{a} \leftarrow \max(\mathfrak{b}\sigma\sqrt{2\log(2/\delta)}, \mathfrak{b}^2)$
  $\sigma^2 \leftarrow \sigma^2 + \mathfrak{b}^2(1 + (\mathfrak{b}^2 - 2\mathfrak{a})\sigma^2/\mathfrak{a}^2)_+$
return $(\mathfrak{a}, \sigma)$

KT-SPLIT The first stage, KT-SPLIT, is an initialization stage that partitions the input sequence $\mathcal{S}_n = (x_i)_{i=1}^n$ into 2^m balanced candidate coresets, each of size $\lfloor \frac{n}{2^m} \rfloor$.[3] As depicted in Fig. 1, this partitioning is carried out recursively in m rounds, first dividing the input sequence in half, then halving those halves into quarters, and so on until coresets of size $\lfloor \frac{n}{2^m} \rfloor$ are produced. The details of KT-SPLIT can appear a bit complicated as, in practice, all m halving rounds are carried out concurrently in an online manner. However, under the hood, each candidate coreset is generated by recursively applying a new simple subroutine called *kernel halving*.

Kernel halving Kernel halving (KH, Alg. 2) is a simple randomized procedure for dividing an input sequence $\mathcal{S}_n$ into two balanced, equal-sized coresets $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ using a kernel $\mathbf{k}$. KH begins with two empty coresets $\mathcal{S}^{(1)}$ and $\mathcal{S}^{(2)}$ and adds points (x, x') from the input sequence two at a time, assigning one point from each pair to each coreset. To encourage balance between the coresets during generation, KH effectively computes which assignment of (x, x') leads to a smaller $\mathrm{MMD}_{\mathbf{k}}(\mathcal{S}^{(1)}, \mathcal{S}^{(2)})$ and then favors that assignment using non-uniform randomness. More precisely, on the i-th step with $(x, x') = (x_{2i-1}, x_{2i})$, KH computes the imbalance contrast

$$\alpha_i = i^2\big(\mathrm{MMD}_{\mathbf{k}}^2(\mathcal{S}^{(1)} \cup \{x\}, \mathcal{S}^{(2)} \cup \{x'\}) - \mathrm{MMD}_{\mathbf{k}}^2(\mathcal{S}^{(1)} \cup \{x'\}, \mathcal{S}^{(2)} \cup \{x\})\big)$$

and then adds x to $\mathcal{S}^{(1)}$ and x' to $\mathcal{S}^{(2)}$ or x' to $\mathcal{S}^{(1)}$ and x to $\mathcal{S}^{(2)}$ with probability biased toward the more balanced outcome. We refer to this step as *probabilistic swapping* in the algorithm statements. The exact value of this probability depends on a *swapping threshold* $\mathfrak{a}_i$ that is produced automatically each round based on the user-supplied inputs $(\delta_i)_{i=1}^{\lfloor n/2 \rfloor}$. In Sec. 4.1, we will learn how to set these inputs to achieve better balance than standard thinning or uniform subsampling, and in Sec. 6.1 we will discuss the role of $\mathfrak{a}_i$ and its generation procedure `get_swap_params` in achieving this balance.

Notably, in the context of KT-SPLIT, KH is run specifically with the square-root kernel $\mathbf{k}_{\mathrm{rt}}$ rather than the target kernel $\mathbf{k}_\star$. This choice enables us to take advantage of the strong L^∞ balance properties established for KH in Sec. 6 and the close connection between square-root L^∞ error and target MMD error revealed in Sec. 4.2.

KT-SWAP The second stage, KT-SWAP, refines the candidate coresets produced by KT-SPLIT in three steps. First, KT-SWAP adds a baseline coreset of size $\lfloor \frac{n}{2^m} \rfloor$ to the candidate list (for example, one produced by standard thinning or uniform subsampling) to ensure that the KT-SWAP output is never worse than that of the baseline. Next, it selects the candidate coreset closest to $\mathcal{S}_n$ in terms of $\mathrm{MMD}_{\mathbf{k}_\star}$. Finally, it refines the selected coreset by replacing each coreset point in turn with the best alternative in $\mathcal{S}_n$, as measured by $\mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \cdot)$. This stage serves to greedily improve upon the MMD of the initial KT-SPLIT candidates, and, when computable, $\mathrm{MMD}_{\mathbf{k}_\star}$ to the target distribution $\mathbb{P}$ can be substituted for the surrogate $\mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \cdot)$ throughout.

Complexity For any m, the time complexity of kernel thinning is dominated by $\mathcal{O}(n^2)$ kernel evaluations, 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 matrix $(\mathbf{k}_{\mathrm{rt}}(x_i, x_j))_{i,j=1}^n$. In addition,

3. When 2^m does not evenly divide n, the final $n - 2^m \lfloor \frac{n}{2^m} \rfloor$ points are discarded.

Algorithm 2: Kernel Halving

Input: kernel $\mathbf{k}$, point sequence $\mathcal{S}_n = (x_i)_{i=1}^n$, probability sequence $(\delta_i)_{i=1}^{\lfloor n/2 \rfloor}$

$\mathcal{S}^{(1)}, \mathcal{S}^{(2)} \leftarrow \{\}$; $\psi_0 \leftarrow \mathbf{0} \in \mathcal{H}$ // Initialize empty coresets: $\mathcal{S}^{(1)}, \mathcal{S}^{(2)}$ have size i after round i

$\sigma_0 \leftarrow 0$ // Swapping parameter

for $i = 1, 2, \dots, \lfloor n/2 \rfloor$ **do**

// Construct kernel difference function using next two points

$(x, x') \leftarrow (x_{2i-1}, x_{2i})$; $f_i \leftarrow \mathbf{k}(x_{2i-1}, \cdot) - \mathbf{k}(x_{2i}, \cdot)$; $\eta_i \leftarrow -1$

// Compute swapping threshold $\mathfrak{a}_i$

$\mathfrak{a}_i, \sigma_i \leftarrow \texttt{get\_swap\_params}(\sigma_{i-1}, \mathfrak{b}, \delta_i)$ with $\mathfrak{b}^2 = \|f_i\|_{\mathbf{k}}^2 = \mathbf{k}(x,x) + \mathbf{k}(x',x') - 2\mathbf{k}(x,x')$

// Compute RKHS inner product $\langle \psi_{i-1}, f_i \rangle_{\mathbf{k}}$, which has a simple form

$\alpha_i \leftarrow \sum_{j=1}^{2i-2} (\mathbf{k}(x_j, x) - \mathbf{k}(x_j, x')) - 2 \sum_{z \in \mathcal{S}^{(1)}} (\mathbf{k}(z, x) - \mathbf{k}(z, x'))$

// Assign one point to each coreset after probabilistic swapping

$(x, x') \leftarrow (x', x)$ and $\eta_i \leftarrow 1$ *with probability* $\min(1, \frac{1}{2}(1 - \frac{\alpha_i}{\mathfrak{a}_i})_+)$

$\mathcal{S}^{(1)}.\texttt{append}(x)$; $\mathcal{S}^{(2)}.\texttt{append}(x')$; $\psi_i \leftarrow \psi_{i-1} + \eta_i f_i$ // $\psi_i = \sum_{x' \in \mathcal{S}^{(2)}} \mathbf{k}(x', \cdot) - \sum_{x \in \mathcal{S}^{(1)}} \mathbf{k}(x, \cdot)$

end

return $\mathcal{S}^{(1)}$, coreset of size $\lfloor n/2 \rfloor$

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

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

$\mathcal{S}^{(m,0)} \leftarrow \texttt{baseline\_coreset}(\mathcal{S}_n, \texttt{size} = \lfloor n/2^m \rfloor)$ // Compare to baseline (e.g., standard thinning)

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

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

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

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

end

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

scaling either $\mathbf{k}_\star$ or $\mathbf{k}_{\mathrm{rt}}$ by a positive multiplier has no impact on Alg. 1, so the kernels need only be specified up to arbitrary rescalings.

4. MMD Guarantees

We are now prepared to present our main MMD guarantees.

4.1 MMD guarantees for kernel thinning

Our first main result, proved in App. E, bounds the MMD of a kernel thinning coreset in terms of the input (2) and kernel (6) radii, the combined radii

$$R_{\mathcal{S}_n, \mathbf{k}, n} \triangleq \min\left(R_{\mathcal{S}_n}, n^{1+\frac{1}{d}} R_{\mathbf{k},n} + n^{\frac{1}{d}} \frac{\|\mathbf{k}\|_\infty}{L_{\mathbf{k}}}\right) \quad \text{and} \quad R'_{\mathcal{S}_n, \mathbf{k}, n} \triangleq \max\left(R_{\mathcal{S}_n}, R'_{\mathbf{k},n}\right), \tag{7}$$

and the kernel thinning inflation factor

$$\mathfrak{M}_{\mathbf{k}}(n,m,d,\delta,\delta',R) \triangleq 2\,\mathbb{I}(\tfrac{n}{2^m} \notin \mathbb{N}) + 37\sqrt{\log(\tfrac{6m}{2^m\delta})}\Big[\sqrt{\log(\tfrac{4}{\delta'})} + 5\sqrt{d\log\big(2+\tfrac{2L_{\mathbf{k}}}{\|\mathbf{k}\|_\infty}(R_{\mathbf{k},n}+R)\big)}\Big], \quad (8)$$

defined for any kernel $\mathbf{k}$ satisfying Assumps. 2 and 3, $n, m, d \in \mathbb{N}$, $\delta \in (0, \frac{6m}{2^m}]$, $\delta' \in (0,1]$, and $R \geq 0$.

Theorem 1 (MMD guarantee for kernel thinning)**.** *Consider kernel thinning (Alg. 1) with $\mathbf{k}_\star$ satisfying Assump. 1, $\mathbf{k}_{\mathrm{rt}}$ a square-root kernel of $\mathbf{k}_\star$, $\delta^\star \triangleq \min_i \delta_i$, and $n_{\mathrm{out}} \triangleq \lfloor \frac{n}{2^m} \rfloor$ for $m \leq \lfloor \log_2 n \rfloor$. If $\mathbf{k}_{\mathrm{rt}}$ satisfies Assumps. 2 and 3, then, for any fixed $\delta' \in (0,1)$, we have*

$$\mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S}_{\mathrm{KT}}) \leq \frac{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{n_{\mathrm{out}}}\Big[2 + \sqrt{\tfrac{(4\pi)^{d/2}}{\Gamma(\frac{d}{2}+1)}}(R'_{\mathcal{S}_n,\mathbf{k}_{\mathrm{rt}},n_{\mathrm{out}}})^{\frac{d}{2}}\,\mathfrak{M}_{\mathbf{k}_{\mathrm{rt}}}(n, m, d, \delta^\star, \delta', R_{\mathcal{S}_n,\mathbf{k}_{\mathrm{rt}},n})\Big], \quad (9)$$

with probability at least $1-\delta'-\sum_{j=1}^m \frac{2^{j-1}}{m}\sum_{i=1}^{2^{m-j}n_{\mathrm{out}}} \delta_i$.

Remark 2 (Guarantee for target $\mathbb{P}$)**.** *A guarantee for any target distribution $\mathbb{P}$ follows directly from the triangle inequality,* $\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathcal{S}_{\mathrm{KT}}) \leq \mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathcal{S}_n) + \mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S}_{\mathrm{KT}})$*.*

Remark 3 (Comparison with baseline thinning)**.** *The* KT-SWAP *step ensures that, deterministically,* $\mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S}_{\mathrm{KT}}) \leq \mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S}_{\mathrm{base}})$ *for $\mathcal{S}_{\mathrm{base}}$ a baseline thinned coreset of size n_{out}. Therefore, we additionally have* $\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathcal{S}_{\mathrm{KT}}) \leq 2\,\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathcal{S}_n) + \mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathcal{S}_{\mathrm{base}})$.

Remark 4 (Finite-time and anytime guarantees)**.** *To obtain a success probability of at least $1-\delta$ with $\delta' = \frac{\delta}{2}$, it suffices to choose $\delta_i = \frac{\delta}{n}$ when the input size n is known in advance and $\delta_i = \frac{m\delta}{2^{m+2}(i+1)\log^2(i+1)}$ when the input size n is not known in advance (but is chosen independently of the randomness used in kernel thinning). In either case, $\delta^\star \leq \frac{6m}{2^m}$ is a valid argument to $\mathfrak{M}_{\mathbf{k}_{\mathrm{rt}}}$ (8). See App. F for our proof.*

Our next corollary, proved in App. G, translates Thm. 1 into specific rates of MMD decay depending on the radius growth of $\mathcal{S}_\infty$ and the tail decay of $\mathbf{k}_{\mathrm{rt}}$.

Corollary 1 (MMD rates for kernel thinning)**.** *Under the notation and assumptions of Thm. 1, consider a sequence of kernel thinning runs (Alg. 1), indexed by $n \in \mathbb{N}$, with $m = \lfloor \frac{1}{2}\log_2 n \rfloor$, $\log(1/\delta^\star) = \mathcal{O}(\log n)$, and $\sum_{j=1}^m \frac{2^{j-1}}{m}\sum_{i=1}^{\lfloor n/2^j \rfloor} \delta_i = o(1)$ as $n \to \infty$. If $\mathcal{S}_\infty$ and $\mathbf{k}_{\mathrm{rt}}$ respectively satisfy one of the radius growth (Def. 4) and tail decay (Def. 7) conditions in the table below, then $\mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S}_{\mathrm{KT}}) = \mathcal{O}_d(\varepsilon_{\mathrm{MMD},n})$ in probability where $\varepsilon_{\mathrm{MMD},n}$ is the corresponding table entry.*

$\boldsymbol{\varepsilon_{\mathrm{MMD},n}}$	Compact $\mathcal{S}_\infty$ $R_{\mathcal{S}_n} \precsim_d 1$	SubGauss $\mathcal{S}_\infty$ $R_{\mathcal{S}_n} \precsim_d \sqrt{\log n}$	SubExp $\mathcal{S}_\infty$ $R_{\mathcal{S}_n} \precsim_d \log n$	HeavyTail(ρ) $\mathcal{S}_\infty$ $R_{\mathcal{S}_n} \precsim_d n^{1/\rho}$
Compact $\mathbf{k}_{\mathrm{rt}}$ $R^\dagger_{\mathbf{k}_{\mathrm{rt}},n} \precsim_d 1$	$\sqrt{\frac{\log n}{n}}$	$(\log n)^{\frac{d+2}{4}}\sqrt{\frac{\log\log n}{n}}$	$(\log n)^{\frac{d+1}{2}}\sqrt{\frac{\log\log n}{n}}$	$\frac{\log n}{\sqrt{n^{1-d/\rho}}}$
SubGauss $\mathbf{k}_{\mathrm{rt}}$ $R^\dagger_{\mathbf{k}_{\mathrm{rt}},n} \precsim_d \sqrt{\log n}$	$(\log n)^{\frac{d+2}{4}}\sqrt{\frac{\log\log n}{n}}$	$(\log n)^{\frac{d+2}{4}}\sqrt{\frac{\log\log n}{n}}$	$(\log n)^{\frac{d+1}{2}}\sqrt{\frac{\log\log n}{n}}$	$\frac{\log n}{\sqrt{n^{1-d/\rho}}}$
SubExp $\mathbf{k}_{\mathrm{rt}}$ $R^\dagger_{\mathbf{k}_{\mathrm{rt}},n} \precsim_d \log n$	$(\log n)^{\frac{d+1}{2}}\sqrt{\frac{\log\log n}{n}}$	$(\log n)^{\frac{d+1}{2}}\sqrt{\frac{\log\log n}{n}}$	$(\log n)^{\frac{d+1}{2}}\sqrt{\frac{\log\log n}{n}}$	$\frac{\log n}{\sqrt{n^{1-d/\rho}}}$
HeavyTail(ρ') $\mathbf{k}_{\mathrm{rt}}$ $R^\dagger_{\mathbf{k}_{\mathrm{rt}},n} \precsim_d n^{1/\rho'}$	$\frac{\log n}{\sqrt{n^{1-d/\rho'}}}$	$\frac{\log n}{\sqrt{n^{1-d/\rho'}}}$	$\frac{\log n}{\sqrt{n^{1-d/\rho'}}}$	$\frac{\log n}{\sqrt{n^{1-d/(\rho\wedge\rho')}}}$

Square-root kernel $\mathbf{k}_{\mathrm{rt}}$	$\\|\mathbf{k}_{\mathrm{rt}}\\|_\infty$	$(R_{\mathbf{k}_{\mathrm{rt}},n}, R'_{\mathbf{k}_{\mathrm{rt}},n})$ $\precsim$	$\mathfrak{M}_{\mathbf{k}_{\mathrm{rt}}}(n, \frac{1}{2}\log_2 n, d, \frac{\delta}{n}, \delta', R)$ $\precsim$
$\left(\frac{2}{\pi\sigma^2}\right)^{\frac{d}{4}}$**Gaussian**$\left(\frac{\sigma}{\sqrt{2}}\right)$	$\left(\frac{2}{\pi\sigma^2}\right)^{\frac{d}{4}}$	$(\sigma\sqrt{\log n}, \sigma\sqrt{d+\log n})$	$\sqrt{\log(\frac{n}{\delta})[\log(\frac{1}{\delta'})+d\log(\sqrt{\log n}+\frac{R}{\sigma})]}$
$A_{\nu,\gamma,d}$**Matérn**$(\frac{\nu}{2},\gamma)$	$\nu(\frac{\gamma^2}{2\pi(a-1)})^{\frac{d}{4}}$	$(\gamma^{-1}(\log n+a\log(1+a)),$ $\gamma^{-1}(a+\log n+E))$	$\sqrt{\log(\frac{n}{\delta})[\log(\frac{1}{\delta'})+d\log(\log n+B+\gamma R)]}$
$\widetilde{S}_{\beta,d}$**B-spline**(β)	c_β^d	$(\frac{1}{2}\sqrt{d}(\beta+1), \frac{1}{2}\sqrt{d}(\beta+1))$	$\sqrt{\log(\frac{n}{\delta})[\log(\frac{1}{\delta'})+d\log(d\beta+\sqrt{d}R)]}$

Table 2: Explicit bounds on Thm. 1 quantities for common kernels. Here, $A_{\nu,\gamma,d}$, and $\widetilde{S}_{\beta,d}$ are as in Tab. 1, $a \triangleq \frac{1}{2}(\nu-d)(>1), B \triangleq a\log(1+a), E \triangleq d\log(\frac{\sqrt{2e\pi}}{\gamma})+\log(\frac{(\nu-2)^{\nu-\frac{3}{2}}}{(2(a-1))^{2a-1}d^{\frac{d}{2}+1}})$, $c_1 = \frac{2}{\sqrt{3}}$, and $c_\beta < 1$ for $\beta > 1$ (see (120)). See App. O for our derivation.

Remark 5 (Probability parameters)**.** *The condition* $(\star)$ $\sum_{j=1}^{m} \frac{2^{j-1}}{m} \sum_{i=1}^{\lfloor n/2^j \rfloor} \delta_i = o(1)$ *is satisfied when* $\delta_1 = \cdots = \delta_{\lfloor \frac{n}{2} \rfloor} = o(\frac{1}{n})$. *Hence, both* $\log(1/\delta^\star) = \mathcal{O}(\log n)$ *and* $(\star)$ *are satisfied when, for example,* $\delta_1 = \cdots = \delta_{\lfloor \frac{n}{2} \rfloor} = \frac{1}{n \log\log n}$.

Cor. 1 shows that kernel thinning returns an $(n^{\frac{1}{2}}, \mathcal{O}_d(n^{-\frac{1}{2}}\sqrt{\log n}))$-MMD coreset in probability when $\mathcal{S}_\infty$ and $\mathbf{k}_{\mathrm{rt}}$ are compactly supported. For fixed d, this guarantee significantly improves upon the baseline $\Omega_p(n^{-\frac{1}{4}})$ rates of i.i.d. sampling and standard MCMC thinning and matches the minimax lower bounds of Sec. 8.1 up to a $\sqrt{\log n}$ term and constants depending on d. For example, when $\mathcal{S}_\infty$ is drawn i.i.d. from $\mathbb{P}$, kernel thinning is nearly minimax optimal amongst *all* distributional approximations (even weighted coresets and non-coreset approximations) that depend on $\mathbb{P}$ only through n i.i.d. input points (Tolstikhin et al., 2017, Thms. 1 and 6).

More generally, when $\mathcal{S}_\infty$ and $\mathbf{k}_{\mathrm{rt}}$ are SubGauss, SubExp, or HeavyTail(ρ) with $\rho > 2d$, Cor. 1 shows that the kernel thinning provides an MMD error of $\mathcal{O}_d(n^{-\frac{1}{2}}\sqrt{(\log n)^{d/2+1}\log\log n})$, $\mathcal{O}_d(n^{-\frac{1}{2}}\sqrt{(\log n)^{d+1}\log\log n})$), and $\mathcal{O}_d(n^{-\frac{1}{2}} n^{\frac{d}{2\rho}}\log n)$ in probability with output coresets of size $\sqrt{n}$. In each case, we find that kernel thinning significantly improves upon an $\Omega_p(n^{-\frac{1}{4}})$ baseline when n is sufficiently large relative to d and, by Rem. 3, is never significantly worse than the baseline when n is small. Our SubExp guarantees also resemble the classical quasi-Monte Carlo guarantees for the uniform distribution on $[0,1]^d$ (see Sec. 8.1) but allow for non-uniform and unbounded target distributions $\mathbb{P}$.

Thm. 1 also allows us to derive more precise, explicit error bounds for specific kernels. For example, for the popular Gaussian, Matérn, and B-spline kernels, Tab. 2 provides explicit bounds on each kernel-dependent quantity in Thm. 1: $\\|\mathbf{k}_{\mathrm{rt}}\\|_\infty$, the kernel radii $(R_{\mathbf{k}_{\mathrm{rt}},n}, R'_{\mathbf{k}_{\mathrm{rt}},\sqrt{n}})$, and the inflation factor $\mathfrak{M}_{\mathbf{k}_{\mathrm{rt}}}$.

4.2 MMD coresets from square-root L^∞ coresets

Thm. 1 builds on a second key result, proved in App. H, that bounds MMD error for $\mathbf{k}_\star$ in terms of L^∞ error for the *square-root* kernel $\mathbf{k}_{\mathrm{rt}}$.

Theorem 2 (MMD guarantee for square-root L^∞ approximations)**.** *Suppose* $\mathbf{k}_\star$ *satisfying Assump. 1 has a square-root kernel* $\mathbf{k}_{\mathrm{rt}}$ *satisfying Assump. 3. Then for any distributions* μ

and ν on $\mathbb{R}^d$ and scalars $r, a, b \geq 0$ with $a + b = 1$,

$$\mathrm{MMD}_{\mathbf{k}_\star}(\mu,\nu) \leq e_d r^{\frac{d}{2}} \cdot \|\mu\mathbf{k}_{\mathrm{rt}} - \nu\mathbf{k}_{\mathrm{rt}}\|_\infty + 2\tau_{\mathbf{k}_{\mathrm{rt}}}(ar) + 2\|\mathbf{k}_\star\|_\infty^{\frac{1}{2}} \cdot \max\{\tau_\mu(br), \tau_\nu(br)\}, \qquad (10)$$

where $e_d \triangleq (\pi^{d/2}/\Gamma(d/2+1))^{\frac{1}{2}}$ decreases super-exponentially in d and $\tau_\mu(r) \triangleq \mu(\mathcal{B}^c(0,r))$.

Importantly, Thm. 2 implies that *any* L^∞ coreset for the square-root kernel $\mathbf{k}_{\mathrm{rt}}$, even one not produced by kernel thinning, is also an MMD coreset for the target kernel $\mathbf{k}_\star$ with MMD error depending on the tail decay of $(\mathbf{k}_{\mathrm{rt}}, \mu, \nu)$. Cor. 2 summarizes the implications of Thm. 2 for common classes of tail decay. See App. I for the proof with explicit constants.

Corollary 2 (MMD error from square-root L^∞ error)**.** *Under the setting and assumptions of Thm. 2, define the L^∞ error $\varepsilon \triangleq \|\mu\mathbf{k}_{\mathrm{rt}} - \nu\mathbf{k}_{\mathrm{rt}}\|_\infty$ and the tail decay function $\widetilde{\tau}(r) \triangleq \tau_{\mathbf{k}_{\mathrm{rt}}}(r) + \|\mathbf{k}_\star\|_\infty^{1/2} \max\{\tau_\mu(r), \tau_\nu(r)\}$ for $r \geq 0$. Then the following implications hold for any nonnegative c and ρ.*

Tail Decay $\widetilde{\tau}(r) \precsim$	Compact $\mathbb{I}(r \leq c)$	SubGauss e^{-cr^2}	SubExp e^{-cr}	HeavyTail(ρ) $r^{-\rho}$
$\Rightarrow \mathrm{MMD}_{\mathbf{k}_\star}(\mu,\nu) \precsim_d$	ε	$\varepsilon(\log\frac{1}{\varepsilon})^{\frac{d}{4}}$	$\varepsilon(\log\frac{1}{\varepsilon})^{\frac{d}{2}}$	$\varepsilon^{\frac{2\rho}{d+2\rho}}$

Remark 6 (Tail decay from finite moments)**.** *By Markov's inequality (Durrett, 2019, Thm. 1.6.4), $\widetilde{\tau}$ has (i)* Compact *decay when $\mathbf{k}_{\mathrm{rt}}$ is* Compact *and μ and ν have compact support; (ii)* SubGauss *decay when $\mathbf{k}_{\mathrm{rt}}$ is* SubGauss *and $\mathbb{E}_{X\sim\mu}[e^{c\|X\|_2^2}], \mathbb{E}_{X\sim\nu}[e^{c\|X\|_2^2}] < \infty$; (iii)* SubExp *decay when $\mathbf{k}_{\mathrm{rt}}$ is* SubExp *and $\mathbb{E}_{X\sim\mu}[e^{c\|X\|_2}], \mathbb{E}_{X\sim\nu}[e^{c\|X\|_2}] < \infty$; and (iv)* HeavyTail$(\rho)$ *decay when $\mathbf{k}_{\mathrm{rt}}$ is* HeavyTail(ρ) *and $\mathbb{E}_{X\sim\mu}[\|X\|_2^\rho], \mathbb{E}_{X\sim\nu}[\|X\|_2^\rho] < \infty$.*

Cor. 2 highlights that MMD quality for $(\mathbf{k}_\star, \mu)$ is of the same order as L^∞ quality for $(\mathbf{k}_{\mathrm{rt}}, \mu)$ when $\mathbf{k}_{\mathrm{rt}}, \mu$, and the approximation ν have compact support. MMD quality then degrades naturally as the tail behavior worsens. In Secs. 5 and 6, we show that, with high probability, kt-split provides a high-quality L^∞ coreset for $(\mathbf{k}_{\mathrm{rt}}, \mathcal{S}_n)$ and hence, by Cor. 2, also provides a high-quality MMD coreset for $(\mathbf{k}_\star, \mathcal{S}_n)$.

5. Self-balancing Hilbert Walk

To exploit the L^∞-MMD connection revealed in Thm. 2, we now turn our attention to constructing high-quality thinned L^∞ coresets. Our strategy relies on a new Hilbert space generalization of the self-balancing walk of Alweiss et al. (2021). We dedicate this section to defining and analyzing this *self-balancing Hilbert walk*, and we detail its connection to kernel thinning in Sec. 6.

Alweiss et al. (2021, Thm. 1.2) introduced a randomized algorithm called the *self-balancing walk* that takes as input a streaming sequence of Euclidean vectors $x_i \in \mathbb{R}^d$ with $\|x_i\|_2 \leq 1$ and outputs a online sequence of random assignments $\eta_i \in \{-1, 1\}$ satisfying

$$\|\textstyle\sum_{i=1}^n \eta_i x_i\|_\infty \precsim \sqrt{\log(d/\delta)\log(n/\delta)} \quad \text{with probability at least } 1-\delta. \qquad (11)$$

Algorithm 3: Self-balancing Hilbert Walk

Input: sequence of functions $(f_i)_{i=1}^n$ in a Hilbert space $\mathcal{H}$, threshold sequence $(\mathfrak{a}_i)_{i=1}^n$

```
ψ_0 ← 0 ∈ H
for i = 1, 2, ..., n do
    α_i ← ⟨ψ_{i−1}, f_i⟩_H    // Compute Hilbert space inner product
    if |α_i| > a_i:
        ψ_i ← ψ_{i−1} − f_i · α_i/a_i
    else:
        η_i ← 1  with probability  ½(1 − α_i/a_i)  and  η_i ← −1  otherwise
        ψ_i ← ψ_{i−1} + η_i f_i
end
return ψ_n, combination of signed input functions
```

Since our ultimate aim is to combine kernel functions, we define a suitable Hilbert space generalization in Alg. 3.

Given a streaming sequence of functions f_i in an arbitrary Hilbert space $\mathcal{H}$ with a norm $\|\cdot\|_{\mathcal{H}}$, this *self-balancing Hilbert walk* (SBHW) outputs a streaming sequence of signed function combinations ψ_i satisfying the following desirable properties established in App. J.

Theorem 3 (Self-balancing Hilbert walk properties). *Consider the self-balancing Hilbert walk (Alg. 3) with each $f_i \in \mathcal{H}$ and $\mathfrak{a}_i > 0$ and define the sub-Gaussian constants*

$$\sigma_0^2 \triangleq 0 \quad \textit{and} \quad \sigma_i^2 \triangleq \sigma_{i-1}^2 + \|f_i\|_{\mathcal{H}}^2 \Big(1 + \tfrac{\sigma_{i-1}^2}{\mathfrak{a}_i^2}(\|f_i\|_{\mathcal{H}}^2 - 2\mathfrak{a}_i)\Big)_+ \quad \forall i \geq 1. \tag{12}$$

The following properties hold true.

(i) ***Functional sub-Gaussianity:*** *For each $i \in [n]$, ψ_i is σ_i* sub-Gaussian*:*

$$\mathbb{E}[\exp(\langle \psi_i, u\rangle_{\mathcal{H}})] \leq \exp\Big(\tfrac{\sigma_i^2 \|u\|_{\mathcal{H}}^2}{2}\Big) \quad \textit{for all} \quad u \in \mathcal{H}. \tag{13}$$

(ii) ***Signed sum representation:*** *If $\mathfrak{a}_i \geq \sigma_{i-1}\|f_i\|_{\mathcal{H}}\sqrt{2\log(2/\delta_i)}$ for $\delta_i \in (0, 1]$, then, with probability at least $1 - \sum_{i=1}^n \delta_i$,*

$$|\alpha_i| \leq \mathfrak{a}_i, \forall i \in [n], \quad \textit{and} \quad \psi_n = \textstyle\sum_{i=1}^n \eta_i f_i. \tag{14}$$

(iii) ***Exact halving via symmetrization:*** *If $\mathfrak{a}_i \geq \sigma_{i-1}\|f_i\|_{\mathcal{H}}\sqrt{2\log(2/\delta_i)}$ for $\delta_i \in (0, 1]$ and each $f_i = g_{2i-1} - g_{2i}$ for $g_1, \ldots, g_{2n} \in \mathcal{H}$, then, with prob. at least $1 - \sum_{i=1}^n \delta_i$,*

$$|\alpha_i| \leq \mathfrak{a}_i, \forall i \in [n], \quad \textit{and} \quad \tfrac{1}{2n}\psi_n = \tfrac{1}{2n}\textstyle\sum_{i=1}^{2n} g_i - \tfrac{1}{n}\sum_{i\in\mathcal{I}} g_i \quad \textit{for} \quad \mathcal{I} = \{2i + \tfrac{\eta_i - 1}{2} : i \in [n]\}.$$

(iv) ***Pointwise sub-Gaussianity in RKHS:*** *If $\mathcal{H}$ is the RKHS of a kernel $\mathbf{k} : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, then, for each $i \in [n]$ and $x \in \mathcal{X}$, $\psi_i(x)$ is $\sigma_i\sqrt{\mathbf{k}(x, x)}$ sub-Gaussian:*

$$\mathbb{E}[\exp(\psi_i(x))] \leq \exp\Big(\tfrac{\sigma_i^2 \mathbf{k}(x,x)}{2}\Big).$$

(v) ***Sub-Gaussian constant bound:*** *Fix any* $q \in [0, 1)$*, and suppose* $\frac{1}{2}\|f_i\|_{\mathcal{H}}^2 \leq \mathfrak{a}_i$ *for all* $i \in [n]$*. If* $\frac{\|f_i\|_{\mathcal{H}}^2}{1+q} \leq \mathfrak{a}_i \leq \frac{\|f_i\|_{\mathcal{H}}^2}{1-q}$ *whenever both* $\sigma_{i-1}^2 < \frac{\mathfrak{a}_i^2}{2\mathfrak{a}_i - \|f_i\|_{\mathcal{H}}^2}$ *and* $\|f_i\|_{\mathcal{H}} > 0$*, then*

$$\sigma_i^2 \leq \frac{\max_{j\in[i]} \|f_j\|_{\mathcal{H}}^2}{1-q^2} \quad \textit{for all} \quad i \in [n]. \tag{15}$$

(vi) ***Adaptive thresholding:*** *If* $\mathfrak{a}_i = \max(c_i \sigma_{i-1}\|f_i\|_{\mathcal{H}}, \|f_i\|_{\mathcal{H}}^2)$ *for* $c_i \geq 0$*, then*

$$\sigma_n^2 \leq \frac{\max_{i\in[n]} \|f_i\|_{\mathcal{H}}^2}{4}(c^\star + 1/c^\star)^2 \quad \textit{for} \quad c^\star \triangleq \max_{i\in[n]} c_i.$$

Remark 7. *The kernel* $\mathbf{k}$ *in Property (iv) can be arbitrary and need not be bounded.*

Property (i) ensures that the functions ψ_i produced by Alg. 3 are mean zero and unlikely to be large in any particular direction u. Property (ii) builds on this functional sub-Gaussianity to ensure that ψ_n is precisely a sum of the signed input functions $\pm f_i$ with high probability. The two properties together imply that, with high probability and an appropriate setting of $\mathfrak{a}_i$, Alg. 3 partitions the input functions f_i into two groups such that the function sums are nearly balanced across the two groups. Property (iii) uses the signed sum representation to construct a two-thinned coreset for any input function sequence $(g_i)_{i=1}^{2n}$. This is achieved by offering the consecutive function differences $g_{2i-1} - g_{2i}$ as the inputs f_i to Alg. 3. Property (iv) highlights that functional sub-Gaussianity also implies sub-Gaussianity of the function values $\psi_i(x)$ whenever the Hilbert space $\mathcal{H}$ is an RKHS. Finally, Properties (v) and (vi) provide explicit bounds on the sub-Gaussian constants σ_i when adaptive settings of the thresholds $\mathfrak{a}_i$ are employed. In Sec. 6, we will connect the SBHW to kernel halving and use Properties (iii), (iv), and (vi) together to show that kernel halving and hence also KT-SPLIT coresets have provably small L^∞ kernel error. Specifically, we will boost the pointwise sub-Gaussianity (Property (iv)) of the output function ψ_n into a high probability bound for $\|\psi_n\|_\infty = \sup_x |\psi_n(x)| = \sup_x |\langle\psi_n, \mathbf{k}(x, \cdot)\rangle|$ by constructing a finite cover for $(\psi_n(x))_{x\in\mathbb{R}^d}$ based on the decay and smoothness of the kernel $\mathbf{k}$.

Comparison with i.i.d. signs A simple alternative to Alg. 3 is to assign signs uniformly at random to each vector, that is, to output $\psi_n' = \sum_{i=1}^n \eta_i' f_i$ with independent Rademacher $\eta_i' \in \{\pm 1\}$. Since the minimal squared sub-Gaussian constant of a sum of independent weighted Rademachers is equal to its variance (Buldygin and Kozachenko, 1980, Lem. 5 & Ex. 1), the minimal squared sub-Gaussian constant of ψ_n' satisfies

$$\sigma_{\mathrm{IID}}^2 = \sup_{u\in\mathcal{H}} \mathrm{Var}(\langle\psi_n', u\rangle_{\mathcal{H}})/\|u\|_{\mathcal{H}}^2 = \sup_{u\in\mathcal{H}} \textstyle\sum_{i=1}^n \langle f_i, u\rangle_{\mathcal{H}}^2/\|u\|_{\mathcal{H}}^2.$$

In the best case, all f_i are orthogonal and bounded and $\sigma_{\mathrm{IID}}^2 = \max_{i\in[n]} \|f_i\|_{\mathcal{H}}^2$ does not grow with n; the reader can check that Alg. 3 also reduces to i.i.d. signing in this case. However, in the worst case, all f_i are equal, and $\sigma_{\mathrm{IID}}^2 = n \max_{i\in[n]} \|f_i\|_{\mathcal{H}}^2$. In contrast, if we choose $\mathfrak{a}_i$ as in Property (vi) with $c_i = \sqrt{2\log(2n/\delta)}$, then the SBHW output $\psi_n = \sum_{i=1}^n \eta_i f_i$ with probability $1 - \delta$ by Property (ii) and has squared sub-Gaussian constant $\sigma_{\mathrm{SBHW}}^2 = \mathcal{O}(\log(2n/\delta) \max_{i\in[n]} \|f_i\|_{\mathcal{H}}^2)$ in every case by Property (vi). This drop from $\Omega(n)$ to $\mathcal{O}(\log n)$ represents an exponential improvement in worst-case balance over employing i.i.d. signs.

We can attribute this gain to the carefully chosen updates of Alg. 3. Notice that, on round i, the function ψ_{i-1} is updated only in the f_i direction, so it suffices to examine the evolution of $\langle \psi_{i-1}, f_i \rangle_{\mathcal{H}}$. We show in App. J.1 that this evolution takes the form

$$\langle \psi_i, f_i \rangle_{\mathcal{H}} = (1 - \|f_i\|_{\mathcal{H}}^2/\mathfrak{a}_i)\langle \psi_{i-1}, f_i \rangle_{\mathcal{H}} + \varepsilon_i \|f_i\|_{\mathcal{H}}^2$$

where $\varepsilon_i \triangleq \mathbb{I}[|\alpha_i| \leq \mathfrak{a}_i](\eta_i + \alpha_i/\mathfrak{a}_i)$ is mean-zero and 1-sub-Gaussian given ψ_{i-1}. In other words, whenever $\mathfrak{a}_i \geq \|f_i\|_{\mathcal{H}}^2$ (as recommended in Property (vi)), Alg. 3 first shrinks the magnitude of ψ_{i-1} in the f_i direction before adding a sub-Gaussian variable in this direction. This targeted shrinkage is absent in the i.i.d. signing update,

$$\langle \psi_i', f_i \rangle_{\mathcal{H}} = \langle \psi_{i-1}', f_i \rangle_{\mathcal{H}} + \eta_i' \|f_i\|_{\mathcal{H}}^2,$$

which simply adds a sub-Gaussian variable in the f_i direction, and allows the SBHW to maintain a substantially smaller sub-Gaussian constant.

General recipe for exact halving The symmetrization construction introduced in Property (iii) can be used to convert any vector balancing algorithm (i.e., any algorithm which assigns ± 1 signs to a sequence of vectors) into an exact halving algorithm (i.e., one which assigns -1 to exactly half of the points) simply by running the algorithm on paired vector differences. We will use this property in the sequel to painlessly construct coresets of an exact target size.

Comparison with the self-balancing walk of Alweiss et al. (2021) In the Euclidean setting with $\mathcal{H} = \mathbb{R}^d$, constant thresholds $\mathfrak{a}_i = 30\log(n/\delta)$, and $\langle \psi_{i-1}, f_i \rangle_{\mathcal{H}}$ the usual Euclidean dot product, Alg. 3 recovers a slight variant of the Euclidean self-balancing walk of Alweiss et al. (2021, Proof of Thm. 1.2). The original algorithm differs only superficially by terminating with failure whenever $|\alpha_i| > \mathfrak{a}_i$. We allow the walk to continue with the update $\psi_i \leftarrow \psi_{i-1} - f_i \cdot \alpha_i/\mathfrak{a}_i$, as it streamlines our sub-Gaussianity analysis and avoids the reliance on distributional symmetry present in Sec. 2.1 of Alweiss et al. (2021). We show in App. R that Thm. 3 recovers the guarantee (11) of Alweiss et al. (2021, Thm. 1.2) with improved constants and a less conservative setting of $\mathfrak{a}_i$.

6. L^∞ Guarantees

In this section, we derive near-optimal L^∞ coreset guarantees for kernel halving (Alg. 2) and KT-SPLIT by relating the two algorithms to the self-balancing Hilbert walk (Alg. 3).

6.1 L^∞ guarantees for kernel halving

To make the connection between KH and SBHW more apparent, we have translated each line of Alg. 2 into the notation of Alg. 3. In this notation, we see that Alg. 2 forms signed combinations ψ_i of paired kernel differences $f_i = \mathbf{k}(x_{2i-1}, \cdot) - \mathbf{k}(x_{2i}, \cdot)$; that the inner product $\alpha_i = \langle \psi_{i-1}, f_i \rangle_{\mathbf{k}}$ has a simple explicit form in terms of kernel evaluations; and that, under the event $\mathcal{E} = \{|\alpha_i| \leq \mathfrak{a}_i \text{ for all } i = 1, \ldots, \lfloor n/2 \rfloor\}$, the function $\psi_{\lfloor n/2 \rfloor}$ of Alg. 2 exactly matches the output of SBHW. Indeed, the function `get_swap_params` serves to compute the sub-Gaussian constants σ_i exactly as defined in Thm. 3 and to adaptively select the thresholds $\mathfrak{a}_i$ exactly as recommended in Thm. 3(iii) and (vi): $\mathfrak{a}_i = \max(\|f_i\|_{\mathcal{H}} \sigma_{i-1} \sqrt{2\log(2/\delta_i)}, \|f_i\|_{\mathcal{H}}^2)$.

This choice of $\mathfrak{a}_i$ simultaneously ensures that the KH-SBHW equivalence event $\mathcal{E}$ occurs with high probability by Thm. 3(iii) and that the SBHW sub-Gaussian constant remains small by Thm. 3(vi). Hence, we can invoke the pointwise SBHW sub-Gaussianity revealed in Thm. 3(iv) to control the KH L^∞ coreset error with high probability.

Theorem 4 (L^∞ guarantees for kernel halving)**.** *Let $\mathcal{S}_{\mathrm{KH}}(\mathbf{k}, \mathcal{S}, \Delta)$ denote the output of kernel halving (Alg. 2) with kernel $\mathbf{k}$ satisfying Assumps. 2 and 3, input point sequence $\mathcal{S}$, and probability sequence Δ. Let $\mathbb{P}_n \triangleq \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_{x_i}$, and recall the definitions of $\mathfrak{M}_{\mathbf{k}}$ (8) and $R_{\mathcal{S}_n,\mathbf{k},n}$ (7). The following statements hold for any $m \in \mathbb{N}$ with $m \leq \log_2 n$ and $\delta' \in (0,1)$.*

(a) **Kernel halving yields a 2-thinned L^∞ coreset:** *The output $\mathcal{S}^{(1)} \triangleq \mathcal{S}_{\mathrm{KH}}(\mathbf{k}, \mathcal{S}_n, (\delta_i)_{i=1}^{\lfloor \frac{n}{2} \rfloor})$ has size $n_{\mathrm{out}} = \lfloor \frac{n}{2} \rfloor$ with $\mathbb{Q}_{\mathrm{KH}}^{(1)} \triangleq \frac{1}{n_{\mathrm{out}}}\sum_{x\in\mathcal{S}^{(1)}} \boldsymbol{\delta}_x$ satisfying*

$$\|\mathbb{P}_n\mathbf{k} - \mathbb{Q}_{\mathrm{KH}}^{(1)}\mathbf{k}\|_\infty \leq \tfrac{\|\mathbf{k}\|_\infty}{n_{\mathrm{out}}} \cdot \mathfrak{M}_{\mathbf{k}}(n, 1, d, \delta^\star, \delta', R_{\mathcal{S}_n,\mathbf{k},n}) \tag{16}$$

with probability at least $1-\delta'-\sum_{i=1}^{n_{\mathrm{out}}} \delta_i$ for $\delta^\star \triangleq \min_{i\leq n_{\mathrm{out}}} \delta_i$.

(b) **Repeated kernel halving yields a 2^m-thinned L^∞ coreset:** *For each $j > 1$, let $\mathcal{S}^{(j)} \triangleq \mathcal{S}_{\mathrm{KH}}(\mathbf{k}, \mathcal{S}^{(j-1)}, (\frac{2^{j-1}}{m}\delta_i)_{i=1}^{\lfloor n/2^j \rfloor})$ be the output of kernel halving recursively applied for j rounds. Then $\mathcal{S}^{(m)}$ has size $n_{\mathrm{out}} = \lfloor \frac{n}{2^m} \rfloor$ with $\mathbb{Q}_{\mathrm{KH}}^{(m)} \triangleq \frac{1}{n_{\mathrm{out}}}\sum_{x\in\mathcal{S}^{(m)}} \boldsymbol{\delta}_x$ satisfying*

$$\|\mathbb{P}_n\mathbf{k} - \mathbb{Q}_{\mathrm{KH}}^{(m)}\mathbf{k}\|_\infty \leq \tfrac{\|\mathbf{k}\|_\infty}{n_{\mathrm{out}}} \cdot \mathfrak{M}_{\mathbf{k}}(n, m, d, \delta^\star, \delta', R_{\mathcal{S}_n,\mathbf{k},n}) \tag{17}$$

with probability at least $1-\delta'-\sum_{j=1}^{m} \frac{2^{j-1}}{m} \sum_{i=1}^{2^{m-j}n_{\mathrm{out}}} \delta_i$ for $\delta^\star \triangleq \min_{i\leq 2^m n_{\mathrm{out}}} \delta_i$.

Thm. 4, proved in App. K, shows that L^∞ error for KH scales simply as the kernel thinning inflation factor $\mathfrak{M}_{\mathbf{k}}$ (8) divided by the size of the output. Our next corollary, an immediate consequence of Thm. 4(b) and the definition of $\mathfrak{M}_{\mathbf{k}}$ (8), translates these bounds into rates of decay depending on the radius growth of $\mathcal{S}_\infty$ and the tail decay of $\mathbf{k}$.

Corollary 3 (L^∞ rates for kernel halving)**.** *Under the notation and assumptions of Thm. 4, consider a sequence of repeated kernel halving runs (Alg. 2), indexed by $n \in \mathbb{N}$, with $m = \lfloor \frac{1}{2}\log_2 n \rfloor$ rounds, $\log(1/\delta^\star) = \mathcal{O}(\log n)$, and $\sum_{j=1}^m \frac{2^{j-1}}{m}\sum_{i=1}^{\lfloor n/2^j \rfloor} \delta_i = o(1)$ as $n \to \infty$. If $\mathcal{S}_\infty$ and $\mathbf{k}$ respectively satisfy one of the radius growth (Def. 4) and tail decay (Def. 7) conditions in the table below, then $\|\mathbb{P}_n\mathbf{k} - \mathbb{Q}_{\mathrm{KH}}^{(m)}\mathbf{k}\|_\infty = \mathcal{O}(\varepsilon_{\infty,n})$ in probability where $\varepsilon_{\infty,n}$ is the corresponding table entry and all hidden constants are independent of the dimension d.*

Type of $\mathcal{S}_\infty$ Type of $\mathbf{k}$	Compact with $R_{\mathcal{S}_n} = \mathcal{O}(\sqrt{d})$ Compact with $\frac{L_{\mathbf{k}}(R_{\mathbf{k},n}\wedge R_{\mathcal{S}_n})}{\|\mathbf{k}\|_\infty} = \mathcal{O}(1)$	Arbitrary SubPoly with $\log(\frac{L_{\mathbf{k}} R_{\mathbf{k},n}}{\|\mathbf{k}\|_\infty}) = \mathcal{O}(\log n)$
$\boldsymbol{\varepsilon_{\infty,n}}$	$\sqrt{\frac{d\log n}{n}}$	$\sqrt{\frac{d}{n}}\log n$

Remark 8 (Example settings for KH L^∞ rates)**.** *Rem. 5 specifies settings of $(\delta_i)_{i=1}^{\lfloor n/2 \rfloor}$ that meet the conditions of Cor. 3. For any radial kernel $\mathbf{k}(x,y) = \kappa(\|x-y\|_2/\sigma)$ with L-Lipschitz $\kappa : \mathbb{R} \to \mathbb{R}$, bandwidth $\sigma > 0$, and $\|\mathbf{k}\|_\infty = 1$, we have $\frac{L_{\mathbf{k}}}{\|\mathbf{k}\|_\infty} \leq \frac{L}{\sigma}$, $R_{\mathbf{k},n} \leq \sigma\kappa^\dagger(1/n)$,*

and therefore $\frac{L_{\mathbf{k}} R_{\mathbf{k},n}}{\|\mathbf{k}\|_\infty} \leq L\, \kappa^\dagger(1/n)$ *where* L *and* $\kappa^\dagger(u) \triangleq \sup\{r : \kappa(r) \geq u\}$ *are independent of* d *and* σ. *Hence Gaussian, Matérn (with* $\nu > \frac{d}{2}+1$*), and IMQ kernels satisfy the* SubPoly *requirements of Cor. 3 with any choice of bandwidth and satisfy the* Compact *requirements when restricted to a compact domain with* $\sigma = \sqrt{d}$. *See App. L for our proof.*

Near-optimal L^∞ coresets For any bounded, radial $\mathbf{k}$ satisfying mild decay and smoothness conditions, Phillips and Tai (2020, Thm. 3.1) proved that any procedure outputting a coreset of size $n^{\frac{1}{2}}$ must suffer $\Omega(\min(\sqrt{d}n^{-\frac{1}{2}}, n^{-\frac{1}{4}}))$ L^∞ error with constant probability for some $\mathbb{P}_n$. Hence, the repeated KH quality guarantees from Cor. 3 are within a $\log n$ factor of minimax optimality for this kernel family which includes Gaussian, Matérn, Wendland, and IMQ kernels.

Online vector balancing in an RKHS In the online vector balancing problem of Spencer (1977), one must assign signs η_i to Euclidean vectors f_i in an online fashion while keeping the norm of the signed sum $\|\sum_{i=1}^n \eta_i f_i\|_\infty$ as small as possible. As an immediate consequence of Thm. 4(a), we find that kernel halving solves an RKHS generalization of the online vector balancing problem.

Corollary 4 (Online vector balancing in an RKHS)**.** *Consider a sequence of kernel halving runs (Alg. 2), indexed by* $n \in \mathbb{N}$, *with* $\log(1/\delta^\star) = \mathcal{O}(\log n)$ *and* $\sum_{i=1}^{\lfloor n/2^j \rfloor} \delta_i = o(1)$ *as* $n \to \infty$. *If the kernel* $\mathbf{k}$ *is* SubPoly *(Def. 7) with* $\log(\frac{L_{\mathbf{k}} R_{\mathbf{k},n}}{\|\mathbf{k}\|_\infty}) = \mathcal{O}(\log n)$, *then* $\|\sum_{i=1}^{2n} \epsilon_i \mathbf{k}(x_i, \cdot)\|_\infty = \mathcal{O}(\sqrt{d}\log n)$ *in probability for the generated signs* $\epsilon_i \triangleq \eta_{\lceil i/2 \rceil}(-1)^i$.

Proof By Thm. 4(a) and the definition of $\mathfrak{M}_{\mathbf{k}}$ (8), $\|\sum_{i=1}^n \eta_i f_i\|_\infty = \mathcal{O}(\sqrt{d}\log n)$ in probability. Noting that $-\eta_i f_i = \eta_i(\mathbf{k}(x_{2i}, \cdot) - \mathbf{k}(x_{2i-1}, \cdot)) = \epsilon_{2i-1}\mathbf{k}(x_{2i-1}, \cdot) + \epsilon_{2i}\mathbf{k}(x_{2i}, \cdot)$, the claim follows. ■

Cor. 4 applies to *any* fixed input point sequence $\mathcal{S}_\infty$ and to a broad range of kernels including, by Rem. 8, Gaussian, Matérn, and IMQ kernels, as well as B-spline kernel since (122) and (123) imply that $\log(\frac{L_{\mathbf{k}} R_{\mathbf{k},n}}{\|\mathbf{k}\|_\infty}) = \mathcal{O}(\log d) = \mathcal{O}(\log n)$ for B-spline $\mathbf{k}$.

6.2 L^∞ and MMD guarantees for KT-SPLIT

We finally extend our L^∞ and MMD guarantees to the kt-split step of kernel thinning by observing that each candidate coreset generated by kt-split is the product of repeated kernel halving (Alg. 2) with $\mathbf{k}_{\mathrm{rt}}$ as the chosen kernel $\mathbf{k}$. Hence, Thm. 4(a) applies equally to the coreset $\mathcal{S}^{(1,1)}$ returned by kt-split with $m = 1$, and, when $m > 1$, Thm. 4(b) applies to the coreset $\mathcal{S}^{(m,1)}$ produced by kt-split. Combining these L^∞ bounds with our L^∞ to MMD conversion theorem (Thm. 2) yields the following corollary proved in App. M.

Corollary 5 (L^∞ and MMD guarantees for kt-split)**.** *Consider* kt-split *with* $\mathbf{k}_{\mathrm{rt}}$ *satisfying Assumps. 2 and 3,* $\delta^\star \triangleq \min_i \delta_i$, *and* $\mathbb{P}_n \triangleq \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}_{x_i}$. *The guarantees of Thm. 4 with* $\mathbf{k} = \mathbf{k}_{\mathrm{rt}}$ *hold if* $\mathbb{Q}_{\mathrm{KT}}^{(m,1)} \triangleq \frac{1}{n_{\mathrm{out}}}\sum_{x \in \mathcal{S}^{(m,1)}} \boldsymbol{\delta}_x$ *is substituted for* $\mathbb{Q}_{\mathrm{KH}}^{(m)}$, *and the guarantees of Thm. 1 hold if the output coreset* $\mathcal{S}^{(m,1)}$ *is substituted for* $\mathcal{S}_{\mathrm{KT}}$,

Cor. 5 ensures that, with high probability, at least one kt-split candidate is a high-quality L^∞ coreset for $(\mathbf{k}_{\mathrm{rt}}, \mathbb{P}_n)$ and hence also a high-quality MMD coreset for $(\mathbf{k}_\star, \mathbb{P}_n)$.

The proof of Thm. 1 in App. E establishes the same MMD guarantee for kernel thinning by noting that the subsequent KT-SWAP step directly minimizes the $\mathrm{MMD}_{\mathbf{k}_\star}$ to $\mathbb{P}_n$ and hence can only improve or maintain this MMD quality. Finally, exactly as in the proofs of Cors. 1 and 3 we can deduce both L^∞ and MMD rate bounds for KT-SPLIT coresets.

Corollary 6 (L^∞ and MMD rates for KT-SPLIT)**.** *Under the notation and assumptions of Cor. 5, consider a sequence of* KT-SPLIT *runs, indexed by* $n \in \mathbb{N}$, *with* $m = \lfloor \frac{1}{2}\log_2 n \rfloor$ *rounds,* $\log(1/\delta^\star) = \mathcal{O}(\log n)$, *and* $\sum_{j=1}^{m} \frac{2^{j-1}}{m} \sum_{i=1}^{\lfloor n/2^j \rfloor} \delta_i = o(1)$ *as* $n \to \infty$. *The guarantees of Cor. 3 with* $\mathbf{k} = \mathbf{k}_{\mathrm{rt}}$ *hold if* $\mathbb{Q}_{\mathrm{KT}}^{(m,1)}$ *is substituted for* $\mathbb{Q}_{\mathrm{KH}}^{(m)}$, *and the guarantees of Cor. 1 hold if* $\mathcal{S}^{(m,1)}$ *is substituted for* $\mathcal{S}_{\mathrm{KT}}$.

7. Vignettes

We complement our primary methodological and theoretical development with two vignettes illustrating the promise of kernel thinning for improving upon (a) i.i.d. sampling in dimensions $d = 2$ through 100 and (b) standard MCMC thinning when targeting challenging differential equation posteriors. See App. P for supplementary details and

`https://github.com/microsoft/goodpoints`

for a Python implementation of kernel thinning and code replicating each vignette.

7.1 Common settings

Throughout, we adopt a **Gaussian**(σ) target kernel $\mathbf{k}_\star(x, y) = \exp(-\frac{1}{2\sigma^2}\|x - y\|_2^2)$ and the corresponding square-root kernel $\mathbf{k}_{\mathrm{rt}}$ from Tab. 1. To output a coreset of size $n^{\frac{1}{2}}$ with n input points, we (a) take every $n^{\frac{1}{2}}$-th point for standard thinning and (b) run kernel thinning (KT) with $m = \frac{1}{2}\log_2 n$ using a standard thinning coreset as the base coreset in KT-SWAP. For each input sample size $n \in \{2^4, 2^6, \dots, 2^{14}, 2^{16}\}$ with $\delta_i = \frac{1}{2n}$, we report the mean coreset error $\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathcal{S}) \pm 1$ standard error across 10 independent replications of the experiment (the standard errors are too small to be visible in all experiments). We additionally regress the log mean MMD onto the log input size using ordinary least squares and display both the best linear fit and an empirical decay rate based on the slope of that fit, e.g., for a slope of -0.25, we report an empirical decay rate of $n^{-0.25}$ for the mean MMD.

7.2 Kernel thinning versus i.i.d. sampling

We first illustrate the benefits of kernel thinning over i.i.d. sampling from a target $\mathbb{P}$. We generate each input sequence $\mathcal{S}_n$ i.i.d. from $\mathbb{P}$, use squared kernel bandwidth $\sigma^2 = 2d$, and consider both Gaussian targets $\mathbb{P} = \mathcal{N}(0, \mathbf{I}_d)$ for $d \in \{2, 4, 10, 100\}$ and mixture of Gaussians (MoG) targets $\mathbb{P} = \frac{1}{M}\sum_{j=1}^{M} \mathcal{N}(\mu_j, \mathbf{I}_2)$ with $M \in \{4, 6, 8\}$ component locations $\mu_j \in \mathbb{R}^2$ defined in App. P.1.

Fig. 2 highlights the visible differences between the KT and i.i.d. coresets for the MoG targets. Even for small sample sizes, the KT coresets achieves better stratification across components with less clumping and fewer gaps within components, suggestive of a better approximation to $\mathbb{P}$. Indeed, when we examine MMD error as a function of coreset size in Fig. 3, we observe that kernel thinning provides a significant improvement across all

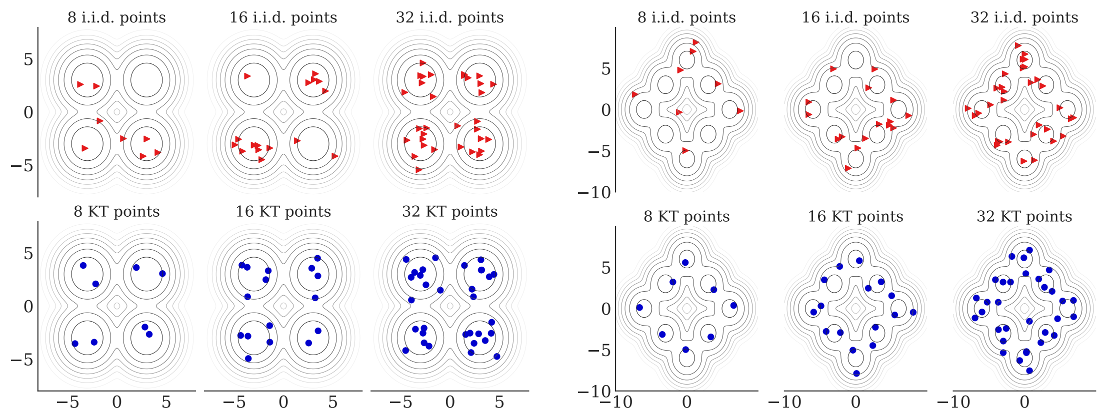


Figure 2: Kernel thinning (KT) and i.i.d. coresets for 4- and 8-component mixture of Gaussian targets with equidensity contours of the target underlaid. See Sec. 7.2 for more details.

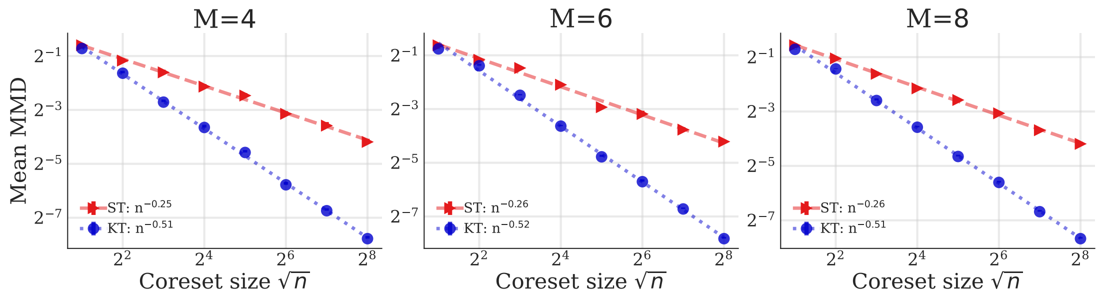
settings. For the Gaussian $d = 2$ target and each MoG target, the KT MMD error scales as $n^{-\frac{1}{2}}$, a quadratic improvement over the $\Omega_p(n^{-\frac{1}{4}})$ MMD error of i.i.d. sampling. Notably, we would not expect to see an exact empirical rate of $n^{-\frac{1}{2}}$ for larger d and small n due to the logarithmic factors in our MMD bounds. However, even for small sample sizes and high dimensions, we observe in Fig. 3(b) that KT significantly improves both the magnitude and the decay rate of MMD relative to i.i.d. sampling.

7.3 Kernel thinning versus standard MCMC thinning

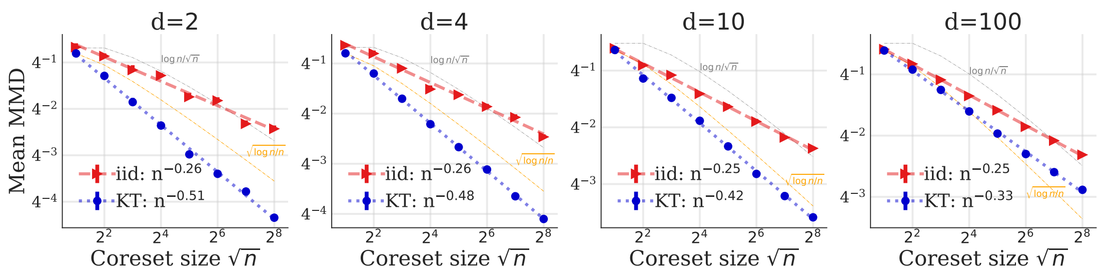
Next, we illustrate the benefits of kernel thinning over standard Markov chain Monte Carlo (MCMC) thinning on twelve posterior inference experiments conducted by Riabiz et al. (2021). We briefly describe each experiment here and refer the reader to Riabiz et al. (2021, Sec. 4) for more details.

Goodwin and Lotka-Volterra experiments From Riabiz et al. (2020), we obtain 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 target, Riabiz et al. (2020) provide 2×10^6 sample points from each of four MCMC algorithms: Gaussian random walk (RW), adaptive Gaussian random walk (adaRW, Haario et al., 1999), Metropolis-adjusted Langevin algorithm (MALA, Roberts and Tweedie, 1996), and pre-conditioned MALA (pMALA, Girolami and Calderhead, 2011).

Hinch experiments From Riabiz et al. (2020), we also obtain the output of two independent Gaussian random walk MCMC chains for 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 et al., 2004) and (2) a tempered version of the same posterior, as defined by Riabiz et al. (2021, App. S5.4). In computational cardiology, the calcium signalling model represents one component of a heart simulator, and one aims to propagate uncertainty in the signalling model through the whole heart simulation, an operation which



(a) Mixture of Gaussians $\mathbb{P}$



(b) Gaussian $\mathbb{P}$

Figure 3: Kernel thinning versus i.i.d. sampling. For (a) mixture of Gaussian targets with $M \in \{4, 6, 8\}$ components and (b) standard multivariate Gaussian targets in dimension $d \in \{2, 4, 10, 100\}$, kernel thinning (KT) reduces $\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathcal{S})$ significantly more quickly than the standard $n^{-\frac{1}{4}}$ rate for $n^{\frac{1}{2}}$ i.i.d. points, even in high dimensions. For reference, decay rates of $\sqrt{\log n/n}$ and $\log n/\sqrt{n}$ are plotted in each of the figures in panel (b).

requires thousands of CPU hours per sample point (Riabiz et al., 2021). In this setting, the costs of running kernel thinning are dwarfed by the time required to generate the input sample (two weeks) and more than offset by the cost savings in the downstream uncertainty propagation task.

Preprocessing and kernel settings We discard the initial points of each chain as burn-in using the maximum burn-in period reported in Riabiz et al. (2021, Tabs. S4 & S6, App. S5.4). and 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 form an input sequence $\mathcal{S}_n$ of length n for coreset construction, we downsample the remaining points using standard thinning. Since exact computation of $\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathcal{S})$ is intractable for these posterior targets, we report $\mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S})$—the error that is controlled directly in our theoretical results—for these experiments. We select the kernel bandwidth σ using the popular median heuristic (see, e.g., Garreau et al., 2017). Additional details can be found in App. P.3.

Results Fig. 4 compares the mean $\mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S})$ error of the generated kernel thinning and standard thinning coresets. In each of the twelve experiments, KT significantly improves both the rate of decay and the order of magnitude of mean MMD, in line with the guarantees

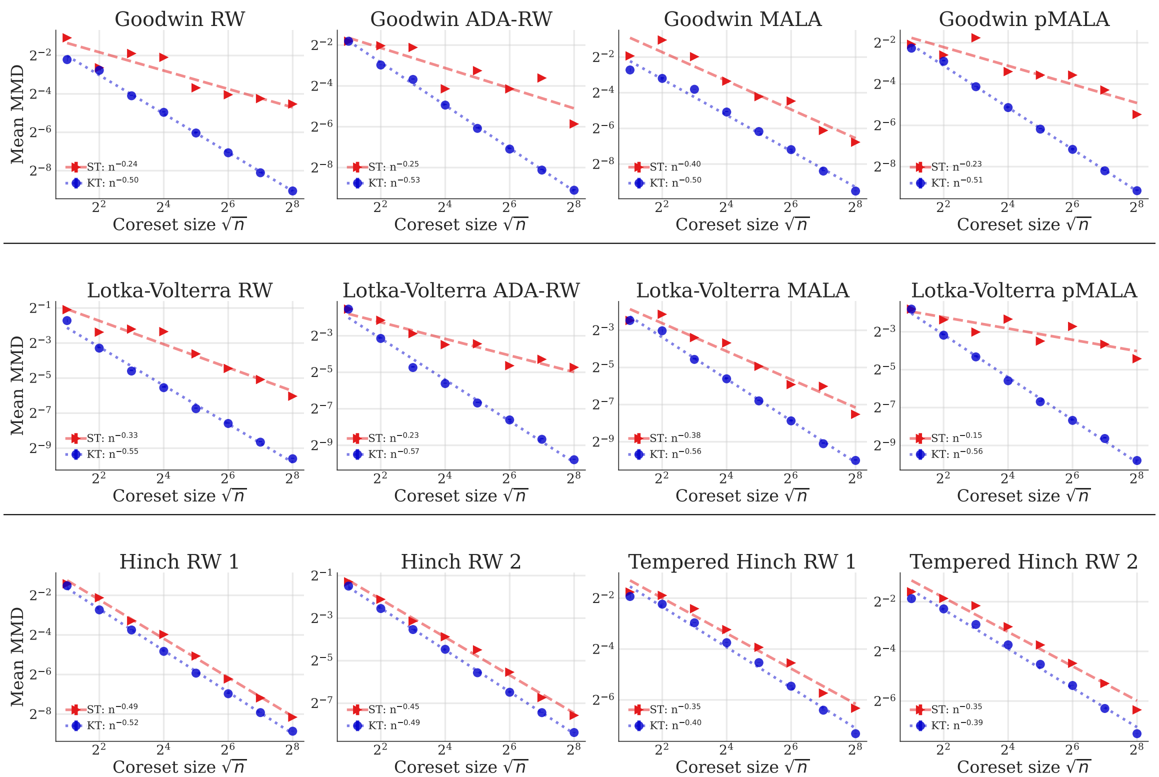


Figure 4: Kernel thinning versus standard MCMC thinning. Kernel thinning (KT) significantly improves both the rate of decay and the order of magnitude of mean $\mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S})$ in each posterior inference task, including eight tasks with 4-dimensional targets (Goodwin and Lotka-Volterra) and four tasks with 38-dimensional targets (Hinch). See Sec. 7.3 for more details.

of Thm. 1. Notably, in the $d = 38$-dimensional Hinch experiments, standard thinning already improves upon the $n^{-\frac{1}{4}}$ rate of i.i.d. subsampling but is outpaced by KT which consistently provides further improvements.

8. Discussion

We introduced kernel thinning (Alg. 1), a new, practical solution to the thinned MMD coreset problem that, given $\mathcal{O}(n^2)$ time and $\mathcal{O}(n \min(d, n))$ storage, improves upon the integration error of i.i.d. sampling and standard MCMC thinning. To achieve this we first showed that any L^∞ coreset for a square-root kernel $\mathbf{k}_{\mathrm{rt}}$ also provides an MMD coreset for its associated target kernel $\mathbf{k}_\star$ (Thm. 2). We next introduced and analyzed a self-balancing Hilbert walk for solving the online vector balancing problem in Hilbert spaces (Alg. 3 and Thm. 3). We then designed a symmetrized version of SBHW for RKHSes—kernel halving—that delivers 2-thinned coresets with small L^∞ error (Alg. 2 and Thm. 4). Our online algorithm, KT-SPLIT, recursively applies kernel halving to a square-root kernel to obtain near-optimal L^∞ coresets in $\mathcal{O}(n^2)$ time with $\mathcal{O}(n \min(d, n))$ space (Cors. 3 and 5). Kernel thinning then combines KT-SPLIT with a greedy refinement step (KT-SWAP) to yield coresets with better-than-i.i.d. MMD for a broad range of kernels and target distributions (Thm. 1 and Cor. 1). While our analysis is restricted to kernels that admit square-root dominating kernels, Dwivedi and Mackey (2022) recently generalized the KT algorithm and analysis to support arbitrary kernels. Separately, Shetty et al. (2022) have developed a distribution compression meta-algorithm, Compress++, which reduces the runtime of KT to near-linear $\mathcal{O}(n \log^3 n)$ time with MMD error that is worse by at most a factor of 4. Hence, KT-Compress++ can be practically deployed even for very large input sizes.

8.1 Related work on MMD coresets

While $(n^{\frac{1}{2}}, o_p(n^{-\frac{1}{4}}))$-MMD coresets have been developed for specific $(\mathbb{P}, \mathbf{k}_\star)$ pairings like the uniform distribution on the unit cube paired with a Sobolev kernel $\mathbf{k}_\star$, to the best of our knowledge, no prior $(n^{\frac{1}{2}}, o_p(n^{-\frac{1}{4}}))$-MMD coreset constructions were known for the range of $\mathbb{P}$ and $\mathbf{k}_\star$ studied in this work. For comparison, we review here both lower bounds and prior strategies for generating coresets with small MMD.

Lower bounds For any bounded and radial (i.e., $\mathbf{k}_\star(x, y) = \kappa(\|x-y\|_2^2)$) kernel satisfying mild decay and smoothness conditions, Phillips and Tai (2020, Thm. 3.1) showed that any procedure outputting coresets of size $n^{\frac{1}{2}}$ must suffer $\Omega(\min(\sqrt{d}n^{-\frac{1}{2}}, n^{-\frac{1}{4}}))$ $\mathrm{MMD}_{\mathbf{k}_\star}$ for some (discrete) target distribution $\mathbb{P}$. This lower bound applies, for example, to Matérn kernels and to infinitely smooth Gaussian kernels. For any continuous and shift-invariant (i.e., $\mathbf{k}_\star(x, y) = \kappa(x-y)$) kernel taking on at least two values, Tolstikhin et al. (2017, Thm. 1) showed that *any* estimator (including non-coreset estimators) based only on n i.i.d. draws from $\mathbb{P}$ must suffer $C_{\mathbf{k}_\star} n^{-\frac{1}{2}}$ $\mathrm{MMD}_{\mathbf{k}_\star}$ with probability at least $1/4$ for some discrete target $\mathbb{P}$ and a constant $C_{\mathbf{k}_\star}$ depending only $\mathbf{k}_\star$. If, in addition, $\mathbf{k}_\star$ is characteristic (i.e., $\mathrm{MMD}_{\mathbf{k}_\star}(\mu, \nu) \neq 0$ when $\mu \neq \nu$), then Tolstikhin et al. (2017, Thm. 6) establish the same lower bound for some continuous target $\mathbb{P}$ with infinitely differentiable density. These last two lower bounds hold, for example, for Gaussian, Matérn, and B-spline kernels and apply in particular to any thinning algorithm that compresses n i.i.d. sample points without additional knowledge of $\mathbb{P}$. For light-tailed $\mathbb{P}$ and $\mathbf{k}_{\mathrm{rt}}$, the kernel thinning guarantees of Thm. 1 match each of these lower bounds up to factors of $\sqrt{\log n}$ and constants depending on d.

Order $(n^{\frac{1}{2}}, n^{-\frac{1}{4}})$-MMD coresets for general target $\mathbb{P}$ By Prop. A.1 of Tolstikhin et al. (2017), an i.i.d. sample from $\mathbb{P}$ yields an order $(n^{\frac{1}{2}}, n^{-\frac{1}{4}})$-MMD coreset in probability. Chen et al. (2010) showed that *kernel herding* with a finite-dimensional kernel (like the linear $\mathbf{k}_\star(x, y) = \langle x, y\rangle$) finds an $(n^{\frac{1}{2}}, (C_{\mathbb{P},\mathbf{k}_\star,d} n)^{-\frac{1}{2}})$-MMD coreset for an inexplicit parameter $C_{\mathbb{P},\mathbf{k}_\star,d}$. However, Bach et al. (2012) showed that their analysis does not apply to any infinite-dimensional kernel (like the Gaussian, Matérn, and B-spline kernels studied in this work), as $C_{\mathbb{P},\mathbf{k}_\star,d}$ would necessarily equal 0. The best known rate for kernel herding with bounded infinite-dimensional kernels (Lacoste-Julien et al., 2015, Thm. G.1) guarantees an order $(n^{\frac{1}{2}}, n^{-\frac{1}{4}})$-MMD coreset, matching the i.i.d. guarantee. For bounded kernels, the same guarantee is available for *Stein Point MCMC* (Chen et al., 2019, Thm. 1) which greedily minimizes MMD[4] over random draws from $\mathbb{P}$ and for a variant of the *greedy sign selection*

4. To bound $\mathrm{MMD}_{\mathbf{k}_\star}$ using Chen et al. (2019, Thm. 1), choose $\mathbf{k}_{\star 0}(x, y) = \mathbf{k}_\star(x, y) - \mathbb{P}\mathbf{k}_\star(x) - \mathbb{P}\mathbf{k}_\star(y) + \mathbb{P}\mathbb{P}\mathbf{k}_\star$.

algorithm described in Karnin and Liberty (2019, Sec. 3.1).[5] Slightly inferior guarantees were established for Stein points (Chen et al., 2018, Thm. 1) and Stein thinning (Riabiz et al., 2021, Thm. 1), both of which accommodate unbounded kernels as well.

Finite-dimensional kernels Harvey and Samadi (2014) construct $(n^{\frac{1}{2}}, \sqrt{d}n^{-\frac{1}{2}}\log^{2.5} n)$-MMD coresets for finite-dimensional linear kernels on $\mathbb{R}^d$ but do not address infinite-dimensional kernels.

Uniform distribution on $[0, 1]^d$ The explicit low discrepancy *quasi-Monte Carlo (QMC)* construction of Chen and Skriganov (2002) provides a $(n^{\frac{1}{2}}, \mathcal{O}_d(n^{-\frac{1}{2}}\log^{\frac{d-1}{2}} n))$-MMD coreset for an L^2 discrepancy kernel when $\mathbb{P}$ is the uniform distribution on the unit cube $[0, 1]^d$. For the same target, the *online Haar strategy* of Dwivedi et al. (2019) yields an $(n^{\frac{1}{2}}, \mathcal{O}_d(n^{-\frac{1}{2}}\log^{2d} n))$-MMD coreset in probability. Dwivedi et al. (2019) also conjecture that a greedy variant of their Haar strategy would provide an improved $(n^{\frac{1}{2}}, \mathcal{O}_d(n^{-\frac{1}{2}}\log^{d} n))$-MMD coreset. These constructions satisfy our quality criteria but are tailored specifically to the uniform distribution on the unit cube.

Unknown coreset quality On compact manifolds, optimal coresets of size $n^{\frac{1}{2}}$ minimize the weighted Riesz energy (a form of relative MMD with a weighted Riesz kernel) at known rates (Borodachov et al., 2014); however, practical *minimum Riesz energy* (Borodachov et al., 2014) and *minimum energy design* (Joseph et al., 2015, 2019) constructions have not been analyzed. When $\mathbf{k}_\star$ is nonnegative and the kernel matrix $(\mathbf{k}_\star(x_i, x_j))_{i,j=1}^n$ satisfies a strong diagonal dominance condition, Kim et al. (2016, Cor. 3, Thm. 6) show that greedy optimization of $\mathrm{MMD}_{\mathbf{k}_\star}$ yields an *MMD-critic* coreset $\hat{\mathcal{S}}$ of size $n^{\frac{1}{2}}$ satisfying

$$\mathrm{MMD}^2_{\mathbf{k}_\star}(\mathbb{P}_n, \hat{\mathcal{S}}) \leq (1 - \tfrac{1}{e})\,\mathrm{MMD}^2_\star + \tfrac{1}{e}\mathbb{P}_n\mathbb{P}_n\mathbf{k}_\star \quad \text{for} \quad \mathrm{MMD}_\star = \min_{|\mathcal{S}|=\sqrt{n}} \mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}_n, \mathcal{S}).$$

In the usual case when $\mathbb{P}_n\mathbb{P}_n\mathbf{k}_\star = \Omega(1)$, this error bound does not decay to 0 with n. Paige et al. (2016) analyze the impact of approximating a kernel in *super-sampling with a reservoir* but do not analyze the quality of the constructed MMD coreset. For the conditionally positive definite energy distance kernel, Mak and Joseph (2018) establish that an optimal coreset of size $n^{\frac{1}{2}}$ has $o(n^{-\frac{1}{4}})$ MMD but do not provide a construction; in addition, Mak and Joseph (2018) propose two *support points convex-concave procedures* for constructing MMD coresets but do not establish their optimality and do not analyze their quality.

8.2 Related work on weighted MMD coresets

While coresets satisfy a number of valuable constraints that are critical for some downstream applications—exact approximation of constants, automatic preservation of convex integrand constraints, compatibility with unweighted downstream tasks, easy visualization, straightforward sampling, and increased numerical stability against errors in integral evaluations (Karvonen et al., 2019)—some applications also support *weighted coreset* approximations of $\mathbb{P}$ of the form $\sum_{i=1}^{\sqrt{n}} w_i \boldsymbol{\delta}_{x_i}$ for weights $w_i \in \mathbb{R}$ that need not be equal, need not be nonnegative, or need not sum to 1. Notably, weighted coresets that depend on $\mathbb{P}$ only through an i.i.d. sample of size n are subject to the same $\Omega(n^{-\frac{1}{2}})$ MMD lower bounds of Tolstikhin

5. The statement of Karnin and Liberty (2019, Thm. 24) bounds $\|\cdot\|_\infty$, but the proof bounds $\mathrm{MMD}_{\mathbf{k}_\star}$.

et al. (2017) described in Sec. 8.1. Any constructions that violate these bounds do so only by exploiting additional information about $\mathbb{P}$ (for example, exact knowledge of $\mathbb{P}\mathbf{k}_\star$) that is not generally available and not required for our kernel thinning guarantees. Moreover, while weighted coresets need not provide satisfactory solutions to the unweighted coreset problem studied in this work, kernel thinning coreset points can be converted into an optimally weighted coreset of no worse quality by explicitly minimizing $\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}_n, \sum_{i=1}^{\sqrt{n}} w_i \boldsymbol{\delta}_{x_i})$ or, if computable, $\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \sum_{i=1}^{\sqrt{n}} w_i \boldsymbol{\delta}_{x_i})$ over the weights w_i in $\mathcal{O}(n^{3/2})$ time.

With this context, we now review known weighted MMD coreset guarantees. We highlight that only one of the weighted $(n^{\frac{1}{2}}, o(n^{-\frac{1}{4}}))$-MMD guarantees covers the unbounded distributions addressed in this work and that the single unbounded guarantee relies on a restrictive uniformly bounded eigenfunction assumption that is typically not satisfied. In other words, our analysis establishes MMD improvements for practical $(\mathbf{k}_\star, \mathbb{P})$ pairings not covered by prior weighted analyses.

$\mathbb{P}$ with bounded support If the target $\mathbb{P}$ has bounded density and bounded, regular support and $\mathbf{k}_\star$ is a Gaussian or Matérn kernel, then *Bayesian quadrature* (O'Hagan, 1991) and *Bayes-Sard cubature* (Karvonen et al., 2018) with quasi-uniform unisolvent point sets yield weighted $(n^{\frac{1}{2}}, o(n^{-\frac{1}{4}}))$-MMD coresets by Wendland (2004, Thm. 11.22 and Cor. 11.33). If $\mathbb{P}$ has bounded support, and $\mathbf{k}_\star$ has more than d continuous derivatives, then the *P-greedy* algorithm (De Marchi et al., 2005) also yields weighted $(n^{\frac{1}{2}}, o(n^{-\frac{1}{4}}))$-MMD coresets by Santin and Haasdonk (2017, Thm. 4.1). For $(\mathbf{k}_\star, \mathbb{P})$ pairs with compact support and sufficiently rapid eigenvalue decay, approximate *continuous volume sampling kernel quadrature* (Belhadji et al., 2020) using the Gibbs sampler of Rezaei and Gharan (2019) yields weighted coresets with $o(n^{-\frac{1}{4}})$ root mean squared MMD.

Finite-dimensional kernels with compactly supported $\mathbb{P}$ For compactly supported $\mathbb{P}$, Briol et al. (2015, Thm. 1) and Bach et al. (2012, Prop. 1) show that *Frank-Wolfe Bayesian quadrature* and weighted variants of kernel herding respectively yield weighted $(n^{\frac{1}{2}}, o(n^{-\frac{1}{4}}))$-MMD coresets for continuous finite-dimensional kernels, but, by Bach et al. (2012, Prop. 2), these analyses do not extend to infinite-dimensional kernels, like the Gaussian, Matérn, and B-spline kernels studied in this work.

Eigenfunction restrictions For $(\mathbf{k}_\star, \mathbb{P})$ pairs with known Mercer eigenfunctions, Belhadji et al. (2019) bound the expected squared MMD of *determinantal point process (DPP) kernel quadrature* in terms of kernel eigenvalue decay and provide explicit rates for univariate Gaussian $\mathbb{P}$ and uniform $\mathbb{P}$ on $[0, 1]$. Their construction makes explicit use of the kernel eigenfunctions which are not available for most $(\mathbf{k}_\star, \mathbb{P})$ pairings. For $(\mathbf{k}_\star, \mathbb{P})$ pairs with $\mathbb{P}\mathbf{k}_\star = \mathbf{0}$, uniformly bounded eigenfunctions, and rapidly decaying eigenvalues, Liu and Lee (2017, App. B.2) prove that *black-box importance sampling* generates probability-weighted coresets with $o(n^{-\frac{1}{4}})$ root mean squared MMD but do not provide any examples verifying their assumptions. The uniformly bounded eigenfunction condition is considered particularly difficult to check (Steinwart and Scovel, 2012), does not hold for Gaussian kernels with Gaussian $\mathbb{P}$ (Minh, 2010, Thm. 1), and need not hold even for infinitely univariate smooth kernels on $[0, 1]$ (Zhou, 2002, Ex. 1).

Unknown coreset quality Huszár and Duvenaud (2012, Prop. 2) bound the MMD error of weighted *sequential Bayesian quadrature* coresets using weak submodularity, but this bound does not decay to zero with n. Khanna and Mahoney (2019, Thm. 2) prove that *weighted kernel herding* yields a weighted $(n^{\frac{1}{2}}, \exp(-n^{\frac{1}{2}}/\kappa_n))$-MMD coreset. However, the κ_n term in Khanna and Mahoney (2019, Thm.3, Assum. 2) is at least as large as the condition number of an $\sqrt{n} \times \sqrt{n}$ kernel matrix, which for typical kernels (including the Gaussian and Matérn kernels) is $\Omega(\sqrt{n})$ (Koltchinskii and Giné, 2000; El Karoui, 2010); the resulting MMD error bound therefore does not decay with n. The *ProtoGreedy* and *ProtoDash* algorithms of Gurumoorthy et al. (2019, Thm. IV.3, IV.5) yield nonnegative weighted coresets $\hat{\mathcal{S}}$ of size $n^{\frac{1}{2}}$ satisfying $\mathrm{MMD}^2_{\mathbf{k}_\star}(\mathbb{P}_n, \hat{\mathcal{S}}) \leq \mathrm{MMD}^2_\star + (\mathbb{P}_n\mathbb{P}_n\mathbf{k}_\star - \mathrm{MMD}^2_\star)e^{-\lambda_{\sqrt{n}}}$ where $\mathrm{MMD}_\star$ is the optimal MMD error to $\mathbb{P}_n$ for a nonnegatively weighted coreset of size $n^{\frac{1}{2}}$. However, careful inspection reveals that $\lambda_{\sqrt{n}} \leq 1$ for any kernel and any n. Hence, in the usual case in which $\mathbb{P}_n\mathbb{P}_n\mathbf{k}_\star = \Omega(1)$, this error bound does not decay to 0 with n. Campbell and Broderick (2019, Thm. 4.4) prove that *Hilbert coresets via Frank-Wolfe* with n input points yield weighted order $(n^{\frac{1}{2}}, \nu_n^{\sqrt{n}})$-MMD coresets for some $\nu_n < 1$ but do not analyze the dependence of ν_n on n.

Non-MMD guarantees For $\mathbb{P}$ with continuously differentiable Lebesgue density and $\mathbf{k}_\star$ a bounded Langevin Stein kernel with $\mathbb{P}\mathbf{k}_\star = \mathbf{0}$, Thm. 2 of Oates et al. (2017) does not bound MMD but does prove that a randomized *control functionals* weighted coreset satisfies $\sqrt{\mathbb{E}[(\mathbb{E}f - \sum_{i=1}^{\sqrt{n}} w_i f(x_i))^2]} \leq C_{\mathbb{P},\mathbf{k}_\star,d,f}/n^{\frac{7}{24}}$ for each f in the RKHS of $\mathbf{k}_\star$ and an unspecified $C_{\mathbb{P},\mathbf{k}_\star,d,f}$. This bound is asymptotically better than the $\Omega(n^{-\frac{1}{4}})$ guarantee for unweighted i.i.d. coresets but worse than the unweighted kernel thinning guarantees of Thm. 1. On compact domains, Thm. 1 of Oates et al. (2019) establishes improved rates for the same weighted coreset when both $\mathbb{P}$ and $\mathbf{k}_\star$ are sufficiently smooth. Bardenet and Hardy (2020) establish an $n^{-\frac{1}{4}-\frac{1}{4d}}$ asymptotic decay of $\mathbb{E}f - \sum_{i=1}^{\sqrt{n}} w_i f(x_i)$ for DPP kernel quadrature with $\mathbb{P}$ on $[-1, 1]^d$ and each f in the RKHS of a particular kernel.

8.3 Related work on L^∞ coresets

A number of alternative strategies are available for constructing coresets with L^∞ guarantees. For example, for any bounded $\mathbf{k}_{\mathrm{rt}}$, Cauchy-Schwarz and the reproducing property imply that

$$\|(\mathbb{P} - \mathbb{P}_n)\mathbf{k}_\star\|_\infty = \sup_{z\in\mathbb{R}^d}\left|\langle \mathbf{k}_\star(z,\cdot), \mathbb{P}\mathbf{k}_\star - \mathbb{P}_n\mathbf{k}_\star\rangle_{\mathbf{k}_\star}\right| \leq \mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathbb{P}_n) \cdot \|\mathbf{k}_\star\|_\infty^{\frac{1}{2}},$$

so that all of the order $(n^{\frac{1}{2}}, n^{-\frac{1}{4}})$-MMD coreset constructions discussed in Sec. 8.1 also yield order $(n^{\frac{1}{2}}, n^{-\frac{1}{4}})$-$L^\infty$ coresets. However, none of those constructions is known to provide a $(n^{\frac{1}{2}}, o(n^{-\frac{1}{4}}))$-$L^\infty$ coreset.

A series of breakthroughs due to Joshi et al. (2011); Phillips (2013); Phillips and Tai (2018, 2020); Tai (2020) has led to a sequence of increasingly compressed $(n^{\frac{1}{2}}, o(n^{-\frac{1}{4}}))$-$L^\infty$ coreset constructions, with the best known guarantees currently due to Phillips and Tai (2020) and Tai (2020). Given n input points, Phillips and Tai (2020) developed an offline, polynomial-time construction to find an $(n^{\frac{1}{2}}, \mathcal{O}_p(\sqrt{d}n^{-\frac{1}{2}}\sqrt{\log n}))$-$L^\infty$ coreset for Lipschitz kernels exhibiting suitable decay, while Tai (2020) developed an offline construction for

Gaussian kernels that runs in $\Omega(d^{5d})$ time and yields an $(n^{\frac{1}{2}}, \mathcal{O}_p(2^d n^{-\frac{1}{2}}\sqrt{\log(d\log n)}))$-$L^\infty$ coreset. More details on these constructions based on the Gram-Schmidt walk of Bansal et al. (2018) can be found in App. S. Notably, the Phillips and Tai (hereafter, PT) guarantee is tighter than that of Thm. 4 by a factor of $\sqrt{\log\log n}$ for sub-Gaussian kernels and input points and $\sqrt{\log n}$ for heavy-tailed kernels and input points. Similarly, the Tai guarantee provides an improvement when n is doubly-exponential in the dimension, that is, when $\sqrt{d\log n} = \Omega(2^d)$.

Moreover, by Thm. 2, we may apply the PT and Tai constructions to a square-root kernel $\mathbf{k}_{\mathrm{rt}}$ to obtain comparable MMD guarantees for the target kernel $\mathbf{k}_\star$ with high probability. However, kernel thinning has a number of practical advantages that lead us to recommend it. First with n input points, using standard matrix multiplication, the PT and Tai constructions have $\Omega(n^4)$ computational complexity and $\Omega(n^2)$ storage costs, a substantial increase over the $\mathcal{O}(n^2)$ running time and $\mathcal{O}(n\min(d,n))$ storage of kernel thinning. Second, KT-SPLIT is an online algorithm while the PT and Tai constructions require the entire set of input points to be available a priori. Finally, each halving round of KT-SPLIT splits the sample size exactly in half, allowing the user to run all m halving rounds simultaneously; the PT and Tai constructions require a rebalancing step after each round forcing the halving rounds to be conducted sequentially.

8.4 Future directions

Several other opportunities for future development recommend themselves. First, since our results cover any target $\mathbb{P}$ with at least $2d$ moments—even discrete and other non-smooth targets—a natural question is whether tighter error bounds with better sample complexities are available when $\mathbb{P}$ is also known to have a smooth Lebesgue density. Second, the MMD to L^∞ reduction in Thm. 2 applies also to weighted L^∞ coresets, and, in applications in which weighted point sets are supported, we would expect either quality or compression improvements from employing non-uniform weights (see, e.g. Turner et al., 2021).

Appendix

Table of contents

Appendix A. Appendix Notation

For each $p \geq 1$, we define L^p as the set of measurable $g : \mathbb{R}^d \to \mathbb{R}$ with $\|g\|_{L^p} \triangleq (\int |g(x)^p| dx)^{1/p} < \infty$ and C^p as the set of $g : \mathbb{R}^d \to \mathbb{R}$ for which all partial derivatives of order p exist and are continuous. For a kernel $\mathbf{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, we also write $\mathbf{k} \in L^{2,\infty}$ to indicate that $\mathbf{k}$ is measurable with finite

$$\|\mathbf{k}\|_{L^{2,\infty}} \triangleq \sup_{x\in\mathbb{R}^d} \big(\int \mathbf{k}^2(x, y)dy\big)^{\frac{1}{2}} = \sup_{x\in\mathbb{R}^d} \|\mathbf{k}(x, \cdot)\|_{L^2}. \tag{18}$$

Throughout, we follow the unitary angular frequency convention of Wendland (2004, Def. 5.15) and define the Fourier transform $\mathcal{F}(f)$ of an integrable complex function $f : \mathbb{R}^d \to \mathbb{C}$ via

$$\mathcal{F}(f)(\omega) \triangleq \tfrac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} f(x) e^{-i\langle x,\omega\rangle} dx \quad \text{for all} \quad \omega \in \mathbb{R}^d. \tag{19}$$

Appendix B. Proof of Prop. 1: MMD guarantee for MCMC

By Douc et al. (2018, Lem. 9.3.9, Cor. 9.2.16), a homogeneous ϕ-irreducible geometrically ergodic Markov chain with stationary distribution $\mathbb{P}$ is also aperiodic with a unique stationary distribution.[6] Since $(x_i)_{i=0}^{\infty}$ are the iterates of such a chain, there exist, by Gallegos-Herrada et al. (2023, Thm. 1xi), constants $\rho \in (0,1)$ and $\tau < \infty$ and a measurable $\mathbb{P}$-almost everywhere finite function $V : \mathbb{R}^d \to [1,\infty]$ satisfying $\mathbb{P}V < \infty$ and

$$\sup_{\text{measurable } h: \frac{|h(x)|}{V(x)} \leq 1, \forall x \in \mathbb{R}^d} |\mathbb{E}[h(x_i) \mid x_0 = x] - \mathbb{P}h| \leq \tau V(x)\rho^i, \text{ for all } x \in \mathbb{R}^d \text{ and } i \in \mathbb{N}. \quad (20)$$

Since V is finite $\mathbb{P}$-almost everywhere, we will choose $c(x) = \infty \Leftrightarrow V(x) = \infty$ to ensure that our claim is (vacuously) true whenever $V(x_0) = \infty$.

Hereafter, suppose $V(x_0) < \infty$. Since the Markov chain is irreducible and aperiodic with a unique stationary distribution $\mathbb{P}$, Assump. H1 of Havet et al. (2020) is satisfied. Hence, by an application of Havet et al. (2020, Prop. 2.1) with $V' = V\rho^g$ and $\zeta = 2/\rho^g$ for sufficiently large $g \in \mathbb{N}$, there exists a set $C(x_0) \subseteq \mathbb{R}^d$ that contains x_0 and satisfies Assumps. H2 and H3 of Havet et al. (2020).

Now fix any $y_1, \ldots, y_n, z_1, \ldots, z_n \in \mathbb{R}^d$. We invoke the definition of MMD (1), the triangle inequality, the reproducing property of an RKHS (Steinwart and Christmann, 2008, Def. 4.18), and Cauchy-Schwarz in turn to deduce a bounded differences property for MMD:

$$\begin{aligned}
&\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \tfrac{1}{n}\textstyle\sum_{i=1}^n \boldsymbol{\delta}_{y_i}) - \mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \tfrac{1}{n}\textstyle\sum_{i=1}^n \boldsymbol{\delta}_{z_i}) \\
&= \sup_{\|f\|_{\mathbf{k}_\star} \leq 1} |\mathbb{P}f - \tfrac{1}{n}\textstyle\sum_{i=1}^n f(y_i)| - \sup_{\|f\|_{\mathbf{k}_\star} \leq 1} |\mathbb{P}f - \tfrac{1}{n}\textstyle\sum_{i=1}^n f(z_i)| \\
&\leq \sup_{\|f\|_{\mathbf{k}_\star} \leq 1} |\tfrac{1}{n}\textstyle\sum_{i=1}^n f(y_i) - f(z_i)| = \sup_{\|f\|_{\mathbf{k}_\star} \leq 1} \tfrac{1}{n}\textstyle\sum_{i=1}^n |\langle \mathbf{k}_\star(y_i,\cdot) - \mathbf{k}_\star(z_i,\cdot), f\rangle_{\mathbf{k}_\star}| \\
&\leq \sup_{\|f\|_{\mathbf{k}_\star} \leq 1} \tfrac{1}{n}\textstyle\sum_{i=1}^n \|\mathbf{k}_\star(y_i,\cdot) - \mathbf{k}_\star(z_i,\cdot)\|_{\mathbf{k}_\star} \|f\|_{\mathbf{k}_\star} \\
&= \sup_{\|f\|_{\mathbf{k}_\star} \leq 1} \tfrac{1}{n}\textstyle\sum_{i=1}^n \sqrt{\mathbf{k}_\star(y_i,y_i) + \mathbf{k}_\star(z_i,z_i) - 2\mathbf{k}_\star(y_i,z_i)}\|f\|_{\mathbf{k}_\star} \leq \tfrac{2}{n}\|\mathbf{k}_\star\|_\infty^{\frac{1}{2}} \textstyle\sum_{i=1}^n \mathbb{I}[y_i \neq z_i].
\end{aligned}$$

Since x_0 belongs to a set $C(x_0)$ satisfying Assumps. H2 and H3 of Havet et al. (2020), McDiarmid's inequality for geometrically ergodic Markov chains (Havet et al., 2020, Thm. 3.1) implies that, with probability at least $1-\delta$ conditional on x_0,

$$\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathbb{P}_n) \leq \mathbb{E}[\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathbb{P}_n) \mid x_0] + \sqrt{c_1(x_0)\|\mathbf{k}_\star\|_\infty \log(1/\delta)/n}$$

where $c_1(x_0)$ is a finite value depending only on the transition probabilities of the chain and the set $C(x_0)$.

Now, define the $\mathbb{P}$ centered kernel $\mathbf{k}_{\star\mathbb{P}}(x,y) = \mathbf{k}_\star(x,y) - \mathbb{P}\mathbf{k}_\star(x) - \mathbb{P}\mathbf{k}_\star(y) + \mathbb{P}\mathbb{P}\mathbf{k}_\star$. To bound the expectation, we will use a slight modification of Lem. 3 of Riabiz et al. (2021). The original lemma used the assumption of V-uniform ergodicity (Meyn and Tweedie, 2012, Defn. (16.0.1)) and the assumption $V(x) \geq \sqrt{\mathbf{k}_{\star\mathbb{P}}(x,x)}$ solely to argue that, for some $R > 0$,

$$|\mathbb{E}[f(x_i) \mid x_0 = x] - \mathbb{P}f| \leq RV(x)\rho^i \quad \text{for all} \quad x \in \mathbb{R}^d \quad \text{and} \quad f \in \mathcal{H}_{\mathbf{k}_{\star\mathbb{P}}} \text{ with } \|f\|_{\mathbf{k}_{\star\mathbb{P}}} = 1.$$

6. In Havet et al. (2020); Douc et al. (2018, Def. 9.2.1) the term *irreducible* is synonymous with ϕ-irreducible as defined by Gallegos-Herrada et al. (2023, Sec. 2).

In our case, since $\mathbf{k}_{\star\mathbb{P}}$ is bounded and any $f \in \mathcal{H}_{\mathbf{k}_{\star\mathbb{P}}}$ with $\|f\|_{\mathbf{k}_{\star\mathbb{P}}} = 1$ satisfies

$$|f(x)| = |\langle \mathbf{k}_{\star\mathbb{P}}(x, \cdot), f\rangle_{\mathbf{k}_{\star\mathbb{P}}}| \leq \|\mathbf{k}_{\star\mathbb{P}}(x, \cdot)\|_{\mathbf{k}_{\star\mathbb{P}}} \|f\|_{\mathbf{k}_{\star\mathbb{P}}} = \sqrt{\mathbf{k}_{\star\mathbb{P}}(x, x)} \leq \sqrt{\|\mathbf{k}_{\star\mathbb{P}}\|_\infty} \text{ for all } x \in \mathbb{R}^d$$

by the reproducing property and Cauchy-Schwarz, the geometric ergodicity property (27) implies the analogous bound

$$|\mathbb{E}[f(x_i) \mid x_0 = x] - \mathbb{P}f| \leq \tau\sqrt{\|\mathbf{k}_{\star\mathbb{P}}\|_\infty} V(x)\rho^i \text{ for all } x \in \mathbb{R}^d \text{ and } f \in \mathcal{H}_{\mathbf{k}_{\star\mathbb{P}}} \text{ with } \|f\|_{\mathbf{k}_{\star\mathbb{P}}} = 1.$$

Hence, the conclusions of Riabiz et al. (2021, Lem. 3) with $R = \tau\sqrt{\|\mathbf{k}_{\star\mathbb{P}}\|_\infty}$ hold under our assumptions. Jensen's inequality and the conclusion of Lem. 3 of Riabiz et al. (2021) now yield the sure bound

$$\begin{aligned}
&\mathbb{E}[\operatorname{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathbb{P}_n) \mid x_0]^2 \leq \mathbb{E}[\operatorname{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathbb{P}_n)^2 \mid x_0] \\
&\quad = \mathbb{E}[\tfrac{1}{n^2}\textstyle\sum_{i=1}^n \mathbf{k}_{\star\mathbb{P}}(x_i, x_i) + \tfrac{1}{n^2}\sum_{i=1}^n \sum_{j\neq i} \mathbf{k}_{\star\mathbb{P}}(x_i, x_j) \mid x_0] \\
&\quad \leq \tfrac{1}{n}\|\mathbf{k}_{\star\mathbb{P}}\|_\infty (1 + \tfrac{2\tau\rho}{1-\rho}\tfrac{1}{n}\textstyle\sum_{i=1}^{n-1} \mathbb{E}[V(x_i) \mid x_0]) \leq \tfrac{4}{n}\|\mathbf{k}_\star\|_\infty (1 + \tfrac{2\tau\rho}{1-\rho}\tfrac{1}{n}\sum_{i=1}^{n-1} \mathbb{E}[V(x_i) \mid x_0]).
\end{aligned}$$

Now, define $c_2(x_0) \triangleq \sup_{n\in\mathbb{N}} \frac{1}{n}\sum_{i=1}^{n-1} \mathbb{E}[V(x_i) \mid x_0]$. Since $V(x_0) < \infty$, the geometric ergodicity property (27) and the fact that $\mathbb{P}V < \infty$ imply

$$c_2(x_0) \leq \mathbb{P}V + V(x_0)\sup_{n\in\mathbb{N}} \tfrac{1}{n}\textstyle\sum_{i=1}^{n-1}\rho^i \leq \mathbb{P}V + V(x_0)\rho < \infty.$$

Taking $c(x_0) = \sqrt{2}\max(c_1(x_0), 4c_2(x_0)(1 + \frac{2\tau\rho}{1-\rho}))$ completes the proof.

Appendix C. Proof of Prop. 2: Almost sure radius growth

We prove this result for the identically distributed case in App. C.1 and for the Markov chain case in App. C.2.

C.1 Radius growth for identically distributed sequences

Suppose x_0 and $\mathcal{S}_\infty$ are drawn identically from $\mathbb{P}$. Claim (a) is true by definition. To establish the remaining claims, we use the following more general result proved in App. C.1.1.

Lemma 5 (Growth rate for identically distributed sequence)**.** *Consider a sequence of identically distributed random variables $(Y_i)_{i=1}^\infty$ on $\mathbb{R}$ and a measurable function $\psi : \mathbb{R} \to \mathbb{R}$ with an increasing inverse function ψ^{-1}. If $\psi(Y_1) \geq 0$ almost surely, then*

$$\Pr(\max_{i\leq n} Y_i > \psi^{-1}(n) \textit{ for some } n \in \mathbb{N}) \leq \mathbb{E}(\psi(Y_1)).$$

Consequently, if $\mathbb{E}(\psi(Y_1)) < \infty$, then, for any $\delta \in (0, 1]$,

$$\Pr(\max_{i\leq n} Y_i \leq \psi^{-1}(\tfrac{n\mathbb{E}[\psi(Y_1)]}{\delta}) \textit{ for all } n \in \mathbb{N}) \geq 1 - \delta.$$

Fix any $\delta \in (0, 1]$. Thm. 5 with $Y_i = \|x_i\|_2$ implies that, with probability $1 - \delta$,

$$R_{\mathcal{S}_n} \leq \psi^{-1}(\tfrac{n\mathbb{E}[\psi(\|x_1\|_2)]}{\delta}) \text{ for all } n \in \mathbb{N}$$

for $\psi^{-1}(r) = \frac{\sqrt{\log r}}{\sqrt{c}}$ in case (b), $\psi^{-1}(r) = \frac{\log r}{c}$ in case (c), and $\psi^{-1}(r) = r^{1/\rho}$ in case (d). As a result we have, with probability $1 - \delta$, $R_{\mathcal{S}_n} = \mathcal{O}_d(\sqrt{\log n})$ in case (b), $R_{\mathcal{S}_n} = \mathcal{O}_d(\log n)$ in case (c), and $R_{\mathcal{S}_n} = \mathcal{O}_d(n^{1/\rho})$ in case (d). Since δ is arbitrary, these orders hold with probability 1 as claimed.

C.1.1 Proof of Thm. 5: Growth rate for identically distributed sequence

We make use of three lemmas. The first rewrites maximum exceedance events in terms of individual variable exceedance events when thresholds are nondecreasing.

Lemma 6 (Exceedance equivalence)**.** *For any real-valued* $(a_i)_{i=1}^{\infty}$ *and nondecreasing* $(b_i)_{i=1}^{\infty}$*,*

$$\max_{i\leq n} a_i > b_n \text{ for some } n \in \mathbb{N} \quad\iff\quad a_i > b_i \text{ for some } i \in \mathbb{N}. \tag{21}$$

Proof The $\Leftarrow$ part follows immediately. To prove the $\Rightarrow$ part, suppose $\max_{i\leq n^\star} a_i > b_{n^\star}$ for some $n^\star$. Then there exists an $i \leq n^\star$ with $a_i > b_{n^\star} \geq b_i$ since $(b_i)_{i=1}^{\infty}$ is nondecreasing. ■

The second bounds the probability of growth rate violation for *any* sequence of random variables in terms of a sum of exceedance probabilities.

Lemma 7 (Growth rate for arbitrary sequence)**.** *For any sequence of random variables* $(Y_i)_{i=1}^{\infty}$ *on* $\mathbb{R}$ *and a nondecreasing real-valued sequence* $(b_i)_{i=1}^{\infty}$*, we have*

$$\Pr(\max_{i\leq n} Y_i > b_n \text{ for some } n \in \mathbb{N}) = \Pr(Y_i > b_i \text{ for some } i \in \mathbb{N}) \leq \textstyle\sum_{i=1}^{\infty} \Pr(Y_i > b_i).$$

Proof The result follows from immediately from Thm. 6 and the union bound. ■

The third lemma bounds a sum of exceedance probabilities whenever the random variables are identically distributed and nonnegative.

Lemma 8 (Bounding exceedances with expectations)**.** *If the random variables* $(Z_i)_{i=1}^{\infty}$ *are identically distributed and almost surely nonnegative, then*

$$\textstyle\sum_{i=1}^{\infty} \Pr(Z_i > i) \leq \mathbb{E}(Z_1). \tag{22}$$

Proof Since Z_1 is almost surely nonnegative, we have $Z_1 = \int_0^{Z_1} dt = \int_0^{\infty} \mathbb{I}(Z_1 > t)dt$ almost surely. Tonelli's theorem (Mukherjea, 1972, Thm. 1) therefore implies that

$$\begin{aligned}\mathbb{E}(Z_1) &= \mathbb{E}(\textstyle\int_0^{\infty} \mathbb{I}(Z_1 > t)dt) = \int_0^{\infty} \Pr(Z_1 > t)dt \\ &\geq \textstyle\int_0^{\infty} \Pr(Z_1 > \lceil t \rceil)dt = \sum_{i=1}^{\infty} \Pr(Z_1 > i) = \sum_{i=1}^{\infty} \Pr(Z_i > i)\end{aligned}$$

where the final inequality uses the identically distributed assumption. ■

Since $\psi(Y_1) \geq 0$ almost surely and ψ^{-1} is increasing, we invoke Thm. 8 with $Z_i = \psi(Y_i)$, the invertibility of ψ, and Thm. 7 with $b_i = \psi^{-1}(i)$ in turn to conclude

$$\begin{aligned}\mathbb{E}(\psi(Y_1)) &\overset{(26)}{\geq} \textstyle\sum_{i=1}^{\infty} \Pr(\psi(Y_i) > i) = \sum_{i=1}^{\infty} \Pr(Y_i > \psi^{-1}(i)) \\ &\overset{(21)}{\geq} \Pr(\max_{i\leq n} Y_i > \psi^{-1}(n) \text{ for some } n \in \mathbb{N}).\end{aligned}$$

C.2 Radius growth for MCMC

Now suppose x_0 and $\mathcal{S}_\infty$ are the iterates of a homogeneous ϕ-irreducible geometrically ergodic Markov chain with initial state x_0, subsequent iterates $\mathcal{S}_\infty$, and stationary distribution $\mathbb{P}$. Our claims will follow from the following more detailed result proved in App. C.2.1.

Lemma 9 (Growth rate for MCMC). *Consider a homogeneous ϕ-irreducible geometrically ergodic Markov chain with initial state x_0, subsequent iterates $(x_i)_{i=1}^\infty$, and stationary distribution $\mathbb{P}$. There exist constants $\rho \in (0,1)$ and $\tau < \infty$ and a measurable $\mathbb{P}$-almost everywhere finite function $V : \mathbb{R}^d \to [1,\infty]$ such that, for any index $j \in \mathbb{N}$, measurable function $g : \mathbb{R}^d \to \mathbb{R}$, measurable nonnegative function ψ on $\mathbb{R}$ with increasing inverse function ψ^{-1}, and $X \sim \mathbb{P}$,*

$$\begin{aligned}&\Pr(\max_{i\leq n} g(x_i) > \psi^{-1}(n) \text{ for some } n \in \mathbb{N} \mid x_0) \leq \\ &\mathbb{E}(\psi(g(X))) + \tfrac{\tau\rho^{j+1}}{1-\rho} V(x_0) + \textstyle\sum_{i=1}^j \Pr(g(x_i) > \psi^{-1}(i) \mid x_0).\end{aligned} \tag{23}$$

Now suppose $V(x_0) < \infty$, and fix any measurable nonnegative ψ on $\mathbb{R}$ with increasing ψ^{-1} and any measurable $g : \mathbb{R}^d \to \mathbb{R}$. If $\mathbb{E}(\psi(g(X))) < \infty$, then, for any $\delta \in (0,1]$, there exists a constant $c_{\delta,\psi\circ g}(x_0) \in (0,\infty)$ such that

$$\Pr(\max_{i\leq n} g(x_i) \leq \psi^{-1}(c_{\delta,\psi\circ g}(x_0)n) \text{ for all } n \in \mathbb{N} \mid x_0) \geq 1-\delta. \tag{24}$$

Moreover, if $\mathbb{P}$ is compactly supported, then for any $\delta \in (0,1]$, there exists a constant $c_\delta(x_0) \in (0,\infty)$ such that

$$\Pr(\sup_{i\in\mathbb{N}} \|x_i\|_2 \leq c_\delta(x_0) \mid x_0) \geq 1-\delta. \tag{25}$$

Instantiate the function V from Thm. 9, and suppose that $V(x_0) < \infty$, an event that holds for $\mathbb{P}$-almost every x_0. Claims (b) to (d) then follow by invoking the time-uniform tail bound (24) with $g = \|\cdot\|_2$ and proceeding as in App. C.1. Finally, claim (a) follows from the bound (25), which establishes $\|x_i\|_2 = \mathcal{O}_d(1)$ with probability 1 conditional on x_0.

C.2.1 Proof of Thm. 9: Growth rate for MCMC

The proof closely parallels that of Thm. 5 except that we substitute the following estimate for Thm. 8.

Lemma 10 (Bounding MCMC exceedances with expectations). *Consider a homogeneous ϕ-irreducible geometrically ergodic Markov chain with initial state x_0, subsequent iterates $(x_i)_{i=1}^\infty$, and stationary distribution $\mathbb{P}$. There exist constants $\rho \in (0,1)$ and $\tau < \infty$ and a measurable $\mathbb{P}$-almost everywhere finite function $V : \mathbb{R}^d \to [1,\infty]$ such that, for any index $j \in \mathbb{N}$, measurable nonnegative function f on $\mathbb{R}^d$, and $X \sim \mathbb{P}$,*

$$\textstyle\sum_{i=j}^\infty \mathbb{P}(f(x_i) > i \mid x_0) \leq \mathbb{E}(f(X)) + \tfrac{\tau\rho^j}{1-\rho} V(x_0). \tag{26}$$

Proof By Douc et al. (2018, Lem. 9.3.9), a homogeneous ϕ-irreducible geometrically ergodic Markov chain with stationary distribution $\mathbb{P}$ is also aperiodic. By Gallegos-Herrada et al.

(2023, Thm. 1xi), there exist constants $\rho \in (0,1)$ and $\tau < \infty$ and a measurable $\mathbb{P}$-almost everywhere finite function $V : \mathbb{R}^d \to [1,\infty]$ satisfying

$$\sup_{h:|h(x)|\leq V(x),\forall x\in\mathbb{R}^d} |\mathbb{E}[h(x_i) \mid x_0 = x] - \mathbb{P}h| \leq \tau V(x)\rho^i \quad \text{for all} \quad x \in \mathbb{R}^d. \tag{27}$$

Applying this result to the functions $h_i(x) \triangleq \mathbb{I}[f(x) > i]$, we find that

$$\textstyle\sum_{i=j}^{\infty} \mathbb{P}(f(x_i) > i \mid x_0) \leq \sum_{i=j}^{\infty} \mathbb{P}(f(X) > i) + \sum_{i=j}^{\infty} \tau V(x_0)\rho^i \leq \mathbb{E}[f(X)] + \frac{\tau\rho^j}{1-\rho} V(x_0)$$

where the final inequality uses Thm. 8. ■

Fix any V, ρ, and τ satisfying the conclusions of Thm. 10, any measurable $g : \mathbb{R}^d \to \mathbb{R}$, and any measurable nonnegative ψ on $\mathbb{R}$ with increasing ψ^{-1}. The first claim (23) follows by applying Thm. 7 with $b_i = \psi^{-1}(i)$, the assumed invertiblity and strict monotonicity of ψ^{-1}, and Thm. 10 with $f = \psi \circ g$ in turn to find that

$$\begin{aligned}
&\mathbb{P}(\max_{i\leq n} g(x_i) > \psi^{-1}(n) \text{ for some } n \mid x_0) \leq \textstyle\sum_{i=1}^{\infty} \mathbb{P}(g(x_i) > \psi^{-1}(i) \mid x_0) \\
&= \textstyle\sum_{i=1}^{\infty} \mathbb{P}(\psi(g(x_i)) > i \mid x_0) \leq \sum_{i=1}^{j} \mathbb{P}(\psi(g(x_i)) > i \mid x_0) + \mathbb{E}(\psi(g(X))) + \frac{\tau\rho^{j+1}}{1-\rho} V(x_0).
\end{aligned}$$

Now suppose $V(x_0) < \infty$, fix any $\delta \in (0,1]$, and let j be the smallest positive index with $\frac{\tau\rho^{j+1}}{1-\rho} V(x_0) < \frac{\delta}{3}$. Since each $\psi(g(x_i))$ is a tight random variable given x_0, we can additionally choose a constant $c_{\delta,\psi\circ g}(x_0) \in (0,\infty)$ satisfying $\sum_{i=1}^{j} \mathbb{P}(\psi(g(x_i))/c_{\delta,\psi\circ g}(x_0) > i \mid x_0) < \frac{\delta}{3}$ and $\mathbb{E}(\psi(g(X)))/c_{\delta,\psi\circ g}(x_0) < \frac{\delta}{3}$. The claim (24) now follows by applying the initial result (23) to the function $\psi/c_{\delta,\psi\circ g}(x_0)$.

To establish the final claim (25), suppose that $\mathbb{P}$ is compactly supported. Since $\mathbb{P}$ has compact support and each $\|x_i\|_2$ is a tight random variable, there exists a constant $c_\delta \in (0,\infty)$ satisfying $\Pr_{X\sim\mathbb{P}}(\|X\|_2 > c_\delta) = 0$ and $\sum_{i=1}^{j} \mathbb{P}(\psi(\|x_i\|_2) > c_\delta \mid x_0) < \frac{\delta}{2}$. The union bound and the geometric ergodicity property (27) applied to the function $h(x) = \mathbb{I}(\|x\|_2 > c_\delta)$ with $\mathbb{P}h = 0$ now imply

$$\begin{aligned}
\Pr(\sup_{i\in\mathbb{N}} \|x_i\|_2 > c_\delta \mid x_0) &\leq \textstyle\sum_{i=1}^{\infty} \Pr(\|x_i\|_2 > c_\delta \mid x_0) \\
&\leq \textstyle\sum_{i=1}^{j} \Pr(\|x_i\|_2 > c_\delta \mid x_0) + \tau V(x_0) \sum_{i=j+1} \rho^i \\
&= \textstyle\sum_{i=1}^{j} \Pr(\|x_i\|_2 > c_\delta \mid x_0) + \frac{\tau\rho^{j+1}}{1-\rho} V(x_0) < \frac{\delta}{2} + \frac{\delta}{3} < \delta.
\end{aligned}$$

Appendix D. Proof of Prop. 3: Shift-invariant square-root kernels

Bochner's theorem (Bochner, 1933; Wendland, 2004, Thm. 6.6) implies that $\mathbf{k}_{\mathrm{rt}}$ is a kernel since κ_{rt} is the Fourier transform of a finite Borel measure with Lebesgue density $\sqrt{\widehat{\kappa}}$. Moreover, as $\mathbf{k}_{\mathrm{rt}}(x,\cdot) = \frac{1}{(2\pi)^{d/4}} \mathcal{F}(e^{-i\langle\cdot,x\rangle} \sqrt{\widehat{\kappa}})$ and $e^{-i\langle\cdot,x\rangle}\sqrt{\widehat{\kappa}}$ is integrable and square integrable, the Plancherel-Parseval identity (Wendland, 2004, Proof of Thm. 5.23) implies that

$$\begin{aligned}
\textstyle\int_{\mathbb{R}^d} \mathbf{k}_{\mathrm{rt}}(x,z)\mathbf{k}_{\mathrm{rt}}(y,z)dz &= \textstyle\int_{\mathbb{R}^d} \frac{1}{(2\pi)^{d/4}} e^{-i\langle\omega,x\rangle} \sqrt{\widehat{\kappa}(\omega)} \; \frac{1}{(2\pi)^{d/4}} e^{i\langle\omega,y\rangle} \sqrt{\widehat{\kappa}(\omega)} d\omega \\
&= \textstyle\frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\langle\omega,x-y\rangle} \widehat{\kappa}(\omega) d\omega = \mathbf{k}_\star(x,y)
\end{aligned}$$

confirming that $\mathbf{k}_{\mathrm{rt}}$ is a square-root kernel of $\mathbf{k}_\star$.

Appendix E. Proof of Thm. 1: MMD guarantee for kernel thinning

By design, KT-SWAP ensures

$$\mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S}_{\mathrm{KT}}) \leq \mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S}^{(m,1)}),$$

where $\mathcal{S}^{(m,1)}$ denotes the first coreset returned by KT-SPLIT. Next, applying Cor. 5, in particular, the bound (85) yields the desired claim.

Appendix F. Proof of Rem. 4: Finite-time and anytime guarantees

We prove the three claims one by one.

Finite time guarantee For the case with known n, the claim follows simply by noting that

$$\textstyle\sum_{j=1}^m \frac{2^{j-1}}{m} \sum_{i=1}^{2^{m-j}\lfloor n/2^m \rfloor} \frac{\delta}{n} = \sum_{j=1}^m \frac{2^m}{m} \lfloor \frac{n}{2^m} \rfloor \frac{\delta}{2n} \leq \sum_{j=1}^m \frac{2^m}{m} \frac{n}{2^m} \frac{\delta}{2n} = \frac{\delta}{2}.$$

Any time guarantee When the input size n is not known in advance but is chosen independently of the randomness in kernel thinning, we first note that

$$\textstyle\sum_{i=1}^\infty \frac{1}{(i+1)\log^2(i+1)} \stackrel{(i)}{\leq} 2, \quad \text{and} \quad \sum_{j=1}^m 2^j = 2^{m+1} - 2 \leq 2^{m+1}. \tag{28}$$

where step (i) can be verified using mathematical programming software. Therefore, for any $n \in \mathbb{N}$, with $\delta_i = \frac{m\delta}{2^{m+2}(i+1)\log^2(i+1)}$, we have

$$\begin{aligned}\textstyle\sum_{j=1}^m \frac{2^{j-1}}{m} \sum_{i=1}^{2^{m-j}\lfloor n/2^m \rfloor} \delta_i \leq \sum_{j=1}^m \frac{2^{j-1}}{m} \sum_{i=1}^\infty \delta_i &= \textstyle\sum_{j=1}^m \frac{2^{j-1}}{m} \sum_{i=1}^\infty \frac{m\delta}{2^{m+2}(i+1)\log^2(i+1)} \\ &= \textstyle\frac{\delta}{2^{m+3}} \Big(\sum_{j=1}^m 2^j\Big)\Big(\sum_{i=1}^\infty \frac{1}{(i+1)\log^2(i+1)}\Big) \\ &\stackrel{(28)}{\leq} \textstyle\frac{\delta}{2^{m+3}} \cdot 2^{m+1} \cdot 2 \leq \frac{\delta}{2}.\end{aligned}$$

Upper bound on $\delta^\star$ The probability lower bound in Thm. 1 is $1-\delta'-\sum_{j=1}^m \frac{2^{j-1}}{m} \sum_{i=1}^{2^{m-j}\lfloor n/2^m \rfloor} \delta_i$, which is non-negative only if

$$\textstyle 1 \geq \sum_{j=1}^m \frac{2^{j-1}}{m} \sum_{i=1}^{2^{m-j}\lfloor n/2^m \rfloor} \delta_i \geq \frac{2^m}{2} \lfloor \frac{n}{2^m} \rfloor \delta^\star,$$

which holds only if $\delta^\star \leq \frac{2}{2^m \lfloor n/2^m \rfloor} \leq \frac{6m}{2^m}$ since $m \in [1, \log_2 n]$. The claim follows.

Appendix G. Proof of Cor. 1: MMD rates for kernel thinning

Repeating arguments similar to those deriving (125) in App. N, we find that for the advertised choices of m and for any fixed δ such that $\delta' = \frac{\delta}{2}$ and $\log(1/\delta^\star) = \mathcal{O}(\log(n/\delta))$, the

RHS of the bound (9) on MMD from Thm. 1 can be simplified as follows:

$$
\begin{aligned}
&\mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S}_{\mathrm{KT}}) \\
&\leq c\|\mathbf{k}_{\mathrm{rt}}\|_\infty \Big(c' \frac{\max(R'_{\mathbf{k}_{\mathrm{rt}},n}, R_{\mathcal{S}_n})^2}{d}\Big)^{\frac{d}{4}} d^{\frac{1}{4}} \sqrt{\tfrac{\log(n/\delta)}{n}\Big[\log(\tfrac{8}{\delta}) + \log\Big(2 + \tfrac{L_{\mathbf{k}_{\mathrm{rt}}}(R_{\mathbf{k}_{\mathrm{rt}},n} + R_{\mathcal{S}_n})}{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}\Big)\Big]}, \\
&= c_{\delta,d}\|\mathbf{k}_{\mathrm{rt}}\|_\infty \Big(\max(R'_{\mathbf{k}_{\mathrm{rt}},n}, R_{\mathcal{S}_n})\Big)^{\frac{d}{2}} \sqrt{\tfrac{\log n}{n}} \sqrt{\log(1 + \max(R_{\mathbf{k}_{\mathrm{rt}},n}, R_{\mathcal{S}_n})) + \log(1 + \tfrac{L_{\mathbf{k}_{\mathrm{rt}}}}{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}}),
\end{aligned} \tag{29}
$$

for some universal constants c, c' where to simplify the expressions, we have used the fact that $R_{\mathcal{S}_n,\mathbf{k}_{\mathrm{rt}},n} \leq R_{\mathcal{S}_n}$ (7). Noting that (29) holds with probability at least $1-\delta$, Cor. 1 now follows from plugging the assumed growth rate bounds into the estimate (29), and treating $\|\mathbf{k}_{\mathrm{rt}}\|_\infty$ and $\frac{L_{\mathbf{k}_{\mathrm{rt}}}}{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}$ as some constant while n grows.

Appendix H. Proof of Thm. 2: MMD guarantee for square-root L^∞ approximations

Our proof will use the following two lemmas proved in Apps. H.1 and H.2 respectively.

Lemma 11 (Square-root representation of MMD)**.** *For $\mathbf{k}_\star$ satisfying Assump. 1 with square-root kernel $\mathbf{k}_{\mathrm{rt}} \in L^{2,\infty}$ we have, for any distributions μ and ν on $\mathbb{R}^d$,*

$$\mathrm{MMD}_{\mathbf{k}_\star}(\mu, \nu) = \sup_{g \in L^2 : \|g\|_{L^2} \leq 1} \big|\int g(y)(\mu\mathbf{k}_{\mathrm{rt}}(y) - \nu\mathbf{k}_{\mathrm{rt}}(y))dy\big|.$$

Lemma 12 (L^∞ bound on L^2 kernel error)**.** *Consider any kernel $\mathbf{k} \in L^{2,\infty}$ satisfying Assump. 1, distributions μ, ν on $\mathbb{R}^d$, and function $g \in L^2$ with $\|g\|_{L^2} \leq 1$. For any $r, a, b \geq 0$ with $a + b = 1$,*

$$\big|\int g(y)(\mu\mathbf{k}(y) - \nu\mathbf{k}(y))dy\big| \leq \|\mu\mathbf{k} - \nu\mathbf{k}\|_\infty \mathrm{Vol}^{\frac{1}{2}}(r) + 2\tau_{\mathbf{k}}(ar) + 2\|\mathbf{k}\|_{L^{2,\infty}} \max\{\tau_\mu(br), \tau_\nu(br)\},$$

where $\mathrm{Vol}(r) \triangleq \pi^d/\Gamma(d/2+1)r^d$ *denotes the volume of the Euclidean ball* $\mathcal{B}(0; r)$.

We first note that, by the square-root kernel definition (Def. 5), $\|\mathbf{k}_{\mathrm{rt}}(x, \cdot)\|_{L^2} = \sqrt{\mathbf{k}_\star(x,x)}$ for each $x \in \mathbb{R}^d$. Since $\mathbf{k}_\star$ is bounded, we therefore have $\mathbf{k}_{\mathrm{rt}} \in L^{2,\infty}$ with $\|\mathbf{k}_{\mathrm{rt}}\|_{L^{2,\infty}} = \sqrt{\|\mathbf{k}_\star\|_\infty}$. The result now follows by invoking Thms. 11 and 12 with $\mathbf{k} = \mathbf{k}_{\mathrm{rt}}$.

H.1 Proof of Thm. 11: Square-root representation of MMD

Let $\mathcal{H}_{\mathbf{k}_\star}$ represent the RKHS of $\mathbf{k}_\star$. By Saitoh (1999, Thms. 1 and 2) and the definition (3) of $\mathbf{k}_{\mathrm{rt}}$, for any $f \in \mathcal{H}_{\mathbf{k}_\star}$, there exists a function $g \in L^2$ such that

$$\|f\|_{\mathbf{k}_\star} = \|g\|_{L^2} \quad \text{and} \quad f(x) = \int g(y)\mathbf{k}_{\mathrm{rt}}(x,y)dy \quad \text{for all} \quad x \in \mathbb{R}^d, \tag{30}$$

and, for any $g \in L^2$, there exists an $f \in \mathcal{H}_{\mathbf{k}_\star}$ such that (30) holds. Note that the integral in (30) is well defined for each x since $g \in L^2$ and $\mathbf{k}_{\mathrm{rt}} \in L^{2,\infty}$. Hence, we have

$$
\begin{aligned}
\mathrm{MMD}_{\mathbf{k}_\star}(\mu, \nu) &= \sup_{f \in \mathcal{H}_{\mathbf{k}_\star} : \|f\|_{\mathbf{k}_\star} \leq 1} |\mu f - \nu f| \\
&\overset{(i)}{=} \sup_{g \in L^2 : \|g\|_{L^2} \leq 1} \big|\int\int g(y)\mathbf{k}_{\mathrm{rt}}(y,x)dyd\mu(x) - \int\int g(y)\mathbf{k}_{\mathrm{rt}}(y,x)dyd\nu(x)\big| \\
&\overset{(ii)}{=} \sup_{g \in L^2 : \|g\|_{L^2} \leq 1} \big|\int g(y)\mu\mathbf{k}_{\mathrm{rt}}(y)dy - \int g(y)\nu\mathbf{k}_{\mathrm{rt}}(y)dy\big|.
\end{aligned}
$$

where step (i) follows from (30), and we can swap the order of integration to obtain step (ii) using Fubini's theorem along with the following fact justified by Hölder's inequality:

$$\int\int|g(y)\mathbf{k}_{\mathrm{rt}}(y,x)|dyd\tilde{\mu}(x) \le \|g\|_{L^2}\|\mathbf{k}_{\mathrm{rt}}\|_{L^{2,\infty}}\int d\tilde{\mu}(x) < \infty \text{ for any distribution. } \tilde{\mu} \tag{31}$$

H.2 Proof of Thm. 12: L^∞ bound on L^2 kernel error

Fix any $r \ge 0$, introduce the shorthand $\mathcal{B}(r) = \mathcal{B}(0;r)$, and define the restrictions

$$g_r(x) = g(x)\mathbb{I}_{\mathcal{B}(r)}(x), \quad \mathbf{k}_r(x,z) \triangleq \mathbf{k}(x,z)\cdot\mathbb{I}_{\mathcal{B}(r)}(z), \quad \text{and} \quad \mathbf{k}_r^{(c)} \triangleq \mathbf{k}-\mathbf{k}_r,$$

so that $\mu\mathbf{k} = \mu\mathbf{k}_r + \mu\mathbf{k}_r^{(c)}$. We first note that, by Cauchy-Schwarz, $g_r \in L^1 \cap L^2$ with

$$\|g_r\|_{L^1} \le \|g_r\|_{L^2}\cdot\sqrt{\mathrm{Vol}(r)} \le \|g\|_{L^2}\cdot\sqrt{\mathrm{Vol}(r)} \le \sqrt{\mathrm{Vol}(r)} \tag{32}$$

and that, exactly as in (31), $\int\int|g(y)\mathbf{k}(y,x)|dyd\tilde{\mu}(x) < \infty$ for any distribution $\tilde{\mu}$ so that each of the integrals to follow is well defined. We now apply the triangle inequality and Hölder's inequality to obtain

$$\begin{aligned}
\left|\int g(y)(\mu\mathbf{k}(y)-\nu\mathbf{k}(y))dy\right| &= \left|\int g(y)(\mu\mathbf{k}_r(y)-\nu\mathbf{k}_r(y))dy + \int g(y)(\mu\mathbf{k}_r^{(c)}(y)-\nu\mathbf{k}_r^{(c)}(y))dy\right| \\
&\le \left|\int g_r(y)(\mu\mathbf{k}_r(y)-\nu\mathbf{k}_r(y))dy\right| + \left|\int_{\|y\|_2\ge r} g(y)(\mu\mathbf{k}_r^{(c)}(y)-\nu\mathbf{k}_r^{(c)}(y))dy\right| \\
&\le \|g_r\|_{L^1}\cdot\|\mu\mathbf{k}-\nu\mathbf{k}\|_\infty + \left|\int_{\|y\|_2\ge r} g(y)(\mu\mathbf{k}(y)-\nu\mathbf{k}(y))dy\right|. \qquad (33)
\end{aligned}$$

Next, we bound the second term in (33). For any $x, y \in \mathbb{R}^d$ with $\|y\|_2 \ge r$ and scalars $a, b \in [0,1]$ such that $a+b=1$, either $\|x-y\|_2 \ge ar$ or $\|x\|_2 \ge br$. Hence,

$$\begin{aligned}
\left|\int_{\|y\|_2\ge r} g(y)(\mu\mathbf{k}(y)-\nu\mathbf{k}(y))dy\right| &= \left|\int_{\|y\|_2\ge r} g(y)\int_{x\in\mathbb{R}^d}\mathbf{k}(x,y)(d\mu(x)-d\nu(x))dy\right| \\
&\le \left|\int_{\|y\|_2\ge r}\int_{\|x-y\|_2\ge ar} g(y)\mathbf{k}(x,y)(d\mu(x)-d\nu(x))dy\right| \\
&\quad + \left|\int_{\|y\|_2\ge r}\int_{\|x\|_2\ge br} g(y)\mathbf{k}(x,y)(d\mu(x)-d\nu(x))dy\right| \\
&=: T_1 + T_2.
\end{aligned}$$

Note that both $T_1, T_2 < \infty$ since $g \in L^2$, $\mathbf{k}_{\mathrm{rt}} \in L^{2,\infty}$ and μ, ν are probability measures. We now bound the terms T_1 and T_2 separately in (34a) and (34b) below. These bounds, together with the estimates (32) and (33), yield our claim.

Bounding T_1 Substituting $x - y = z$, we have

$$\begin{aligned}
T_1 &\le \int_{\|x-z\|_2\ge r}\int_{\|z\|_2\ge ar}|g(x-z)\mathbf{k}(x,x-z)||d\mu(x)-d\nu(x)dz| \\
&\le \int\int_{\|z\|_2\ge ar}|g(x-z)\mathbf{k}(x,x-z)|dz|d\mu(x)-d\nu(x)| \\
&\overset{(i)}{\le} \int\|g(x-\cdot)\|_{L^2}\sup_{x'}\left(\int_{\|z\|_2\ge ar}\mathbf{k}^2(x',x'-z)dz\right)^{1/2}|d\mu(x)-d\nu(x)| \\
&\overset{(ii)}{=} \int\|g\|_{L^2}\tau_{\mathbf{k}}(ar)|d\mu(x)-d\nu(x)| \overset{(iii)}{\le} 2\|g\|_{L^2}\tau_{\mathbf{k}}(ar), \qquad (34a)
\end{aligned}$$

where step (i) follows from Cauchy-Schwarz, step (ii) from the definition (5) of $\tau_{\mathbf{k}}$, and step (iii) from the fact $\int|d\mu(x)-d\nu(x)| \le 2$.

Bounding T_2 We have

$$
\begin{aligned}
T_2 &\leq \int_{\|x\|_2 \geq br} \Big(\int_{\|y\|_2 \geq r} |g(y)\mathbf{k}(x,y)| dy \Big) |d\mu(x) - d\nu(x)| \\
&\leq \int_{\|x\|_2 \geq br} \|g\|_{L^2} \sup_{x'} \|\mathbf{k}(x',\cdot)\|_{L^2} |d\mu(x) - d\nu(x)| \\
&\overset{(iv)}{\leq} 2\|g\|_{L^2} \cdot \|\mathbf{k}\|_{L^{2,\infty}} \max\{\tau_\mu(br), \tau_\nu(br)\}.
\end{aligned} \tag{34b}
$$

where step (iv) follows from the definitions (5) and (18) of (τ_μ, τ_ν) and $\|\mathbf{k}\|_{L^{2,\infty}}$.

Appendix I. Proof of Cor. 2: MMD error from square-root L^∞ error

Let $e'_d \triangleq 2^{\frac{d}{2}} e_d$ for e_d defined in Thm. 2. Notice that the bound (10) for the choice of $a = b = \frac{1}{2}$ can be rewritten as

$$
\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathbb{Q}) \leq \inf_r (e'_d r^{\frac{d}{2}} \varepsilon + 2\widetilde{\tau}(r)). \tag{35}
$$

Then the claims of Cor. 2 follow by optimizing the RHS of (35) over the choice of r depending on the tail decay of $\widetilde{\tau}$. Throughout the proofs c, c', ρ denote the (exactly same) constants underlying the assumed tail decay of $\widetilde{\tau}$.

Proof for COMPACT part Choosing $r \downarrow c'$, we obtain that

$$
\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathbb{Q}) \leq \inf_r (e'_d r^{\frac{d}{2}} \varepsilon + 2c\mathbb{I}(r \leq c')) \leq \varepsilon \cdot e'_d (c')^{\frac{d}{2}}.
$$

Proof for SUBGAUSS part Choosing $r = \sqrt{\frac{1}{c'} \log(1 \vee \frac{8cc'}{de'_d \varepsilon})}$, we obtain that

$$
\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathbb{Q}) \leq \inf_r \left(e'_d r^{\frac{d}{2}} \varepsilon + 2ce^{-c'r^2} \right) \leq \varepsilon \cdot e'_d \big[\big(\frac{\log(1 \vee \frac{8cc'}{de'_d \varepsilon})}{c'} \big)^{\frac{d}{4}} + \frac{d}{4c'} \big].
$$

Proof for SUBEXP part Choosing $r = \frac{1}{c'} \log(1 \vee \frac{4cc'}{de'_d \varepsilon})$, we obtain that

$$
\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathbb{Q}) \leq \inf_r \left(e'_d r^{\frac{d}{2}} \varepsilon + 2ce^{-c'r} \right) \leq \varepsilon \cdot e'_d \big[\big(\frac{\log(1 \vee \frac{4cc'}{de'_d \varepsilon})}{c'} \big)^{\frac{d}{2}} + \frac{d}{2c'} \big].
$$

Proof for HEAVYTAIL(ρ) part Choosing $r = (\frac{4c\rho}{de'_d \varepsilon})^{\frac{2}{d+2\rho}}$, we obtain that

$$
\mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}, \mathbb{Q}) \leq \inf_r \left(e'_d r^{\frac{d}{2}} \varepsilon + 2cr^{-\rho} \right) \leq (\varepsilon \cdot e'_d)^{\frac{2\rho}{d+2\rho}} \cdot (\frac{4c\rho}{d})^{\frac{d}{d+2\rho}} (1 + \frac{2\rho}{d}).
$$

Appendix J. Proof of Thm. 3: Self-balancing Hilbert walk properties

We prove each property from Thm. 3 one by one.

J.1 Property (i): Functional sub-Gaussianity

We prove the functional sub-Gaussianity claim (13) by induction on the iteration $i \in \{0, \ldots, n\}$. Our proof uses the following lemma proved in App. J.7, which supplies a convenient decomposition for the self-balancing Hilbert walk iterates.

Lemma 13 (Alternate representation of ψ_i)**.** *For each $i \in [n]$, the iterate ψ_i of the self-balancing Hilbert walk (Alg. 3) satisfies*

$$\langle \psi_i, u\rangle_{\mathcal{H}} = \left\langle \psi_{i-1}, u - f_i \tfrac{\langle f_i, u\rangle_{\mathcal{H}}}{\mathfrak{a}_i} \right\rangle_{\mathcal{H}} + \varepsilon_i \langle f_i, u\rangle_{\mathcal{H}} \quad \textit{for all} \quad u \in \mathcal{H} \tag{36}$$

for the random variable $\varepsilon_i \triangleq \mathbb{I}[|\alpha_i| \leq \mathfrak{a}_i](\eta_i + \alpha_i/\mathfrak{a}_i)$ which satisfies

$$\mathbb{E}[\varepsilon_i|\psi_{i-1}] = 0, \quad \varepsilon_i \in [-2, 2], \quad \textit{and} \quad \mathbb{E}[e^{t\varepsilon_i}|\psi_{i-1}] \leq e^{t^2/2} \quad \textit{for all} \quad t \in \mathbb{R}. \tag{37}$$

Now we proceed with our induction argument.

Base case For $i = 1$, noting that $\psi_0 = 0$, we have

$$\mathbb{E}[\exp(\langle \psi_1, u\rangle_{\mathcal{H}})] \overset{(36)}{=} \mathbb{E}[\exp(\varepsilon_1 \langle f_1, u\rangle_{\mathcal{H}})|] \overset{(37)}{\leq} \exp\Big(\tfrac{1}{2}\langle f_1, u\rangle_{\mathcal{H}}^2\Big) \leq \exp\big(\tfrac{1}{2}\|f_1\|_{\mathcal{H}}^2 \|u\|_{\mathcal{H}}^2\big),$$

where the last step follows from Cauchy-Schwarz, and thus ψ_1 is sub-Gaussian with parameter $\sigma_1 = \|f_1\|_{\mathcal{H}}$ as desired.

Inductive step Fix any $i \in [n]$ with $i \geq 2$ and assume that the functional sub-Gaussianity claim (13) holds for ψ_{i-1} with σ_{i-1}. We have

$$\begin{aligned}
\mathbb{E}[\exp(\langle \psi_i, u\rangle_{\mathcal{H}})] &= \mathbb{E}[\mathbb{E}[\exp(\langle \psi_i, u\rangle_{\mathcal{H}})|\psi_{i-1}]] \\
&\overset{(36)}{=} \mathbb{E}\Big[\exp\Big(\Big\langle \psi_{i-1}, u - f_i \tfrac{\langle f_i, u\rangle_{\mathcal{H}}}{\mathfrak{a}_i}\Big\rangle_{\mathcal{H}}\Big) \cdot \mathbb{E}[\exp(\varepsilon_i \langle f_i, u\rangle_{\mathcal{H}})|\psi_{i-1}]\Big] \\
&\overset{(37)}{\leq} \mathbb{E}\Big[\exp\Big(\Big\langle \psi_{i-1}, u - f_i \tfrac{\langle f_i, u\rangle_{\mathcal{H}}}{\mathfrak{a}_i}\Big\rangle_{\mathcal{H}}\Big) \cdot \exp\Big(\tfrac{1}{2}\langle f_i, u\rangle_{\mathcal{H}}^2\Big)\Big] \\
&= \exp\Big(\tfrac{1}{2}\langle f_i, u\rangle_{\mathcal{H}}^2\Big) \cdot \mathbb{E}\Big[\exp\Big(\Big\langle \psi_{i-1}, u - f_i \tfrac{\langle f_i, u\rangle_{\mathcal{H}}}{\mathfrak{a}_i}\Big\rangle_{\mathcal{H}}\Big)\Big] \\
&\overset{(i)}{\leq} \exp\Big(\tfrac{1}{2}\langle f_i, u\rangle_{\mathcal{H}}^2 + \tfrac{\sigma_{i-1}^2}{2}\Big\| u - f_i \tfrac{\langle f_i, u\rangle_{\mathcal{H}}}{\mathfrak{a}_i}\Big\|_{\mathcal{H}}^2\Big),
\end{aligned} \tag{38}$$

where step (i) follows from the induction hypothesis. Simplifying the exponent in the display (38) using Cauchy-Schwarz and the definition (12) of σ_i, we have

$$\begin{aligned}
\tfrac{1}{2}\langle f_i, u\rangle_{\mathcal{H}}^2 + \tfrac{\sigma_{i-1}^2}{2}\Big\| u - f_i \tfrac{\langle f_i, u\rangle_{\mathcal{H}}}{\mathfrak{a}_i}\Big\|_{\mathcal{H}}^2 &= \tfrac{1}{2}\langle f_i, u\rangle_{\mathcal{H}}^2 + \tfrac{\sigma_{i-1}^2}{2}\Big(\|u\|_{\mathcal{H}}^2 + \langle f_i, u\rangle_{\mathcal{H}}^2\Big(\tfrac{\|f_i\|_{\mathcal{H}}^2}{\mathfrak{a}_i^2} - \tfrac{2}{\mathfrak{a}_i}\Big)\Big) \\
&= \tfrac{\sigma_{i-1}^2}{2}\|u\|_{\mathcal{H}}^2 + \langle f_i, u\rangle_{\mathcal{H}}^2 \cdot \Big(\tfrac{1}{2} + \tfrac{\sigma_{i-1}^2\|f_i\|_{\mathcal{H}}^2}{2\mathfrak{a}_i^2} - \tfrac{\sigma_{i-1}^2}{\mathfrak{a}_i}\Big) \\
&\leq \tfrac{\sigma_{i-1}^2}{2}\|u\|_{\mathcal{H}}^2 + \langle f_i, u\rangle_{\mathcal{H}}^2 \cdot \Big(\tfrac{1}{2} + \tfrac{\sigma_{i-1}^2\|f_i\|_{\mathcal{H}}^2}{2\mathfrak{a}_i^2} - \tfrac{\sigma_{i-1}^2}{\mathfrak{a}_i}\Big)_+ \\
&\leq \tfrac{\|u\|_{\mathcal{H}}^2}{2}\Big(\sigma_{i-1}^2 + \|f_i\|_{\mathcal{H}}^2\Big(1 + \tfrac{\sigma_{i-1}^2}{\mathfrak{a}_i^2}(\|f_i\|_{\mathcal{H}}^2 - 2\mathfrak{a}_i)\Big)_+\Big) \\
&= \tfrac{\sigma_i^2}{2}\|u\|_{\mathcal{H}}^2.
\end{aligned}$$

J.2 Property (ii): Signed sum representation

Since Alg. 3 adds $\pm f_i$ to ψ_{i-1} whenever $|\alpha_i| = |\langle \psi_{i-1}, f_i \rangle_{\mathcal{H}}| \leq \mathfrak{a}_i$, by the union bound, it suffices to lower bound the probability of this event by $1 - \delta_i$ for each i. The following lemma establishes this bound using the functional sub-Gaussianity (13) of each ψ_{i-1}.

Lemma 14 (Self-balancing Hilbert walk success probability)**.** *The self-balancing Hilbert walk (Alg. 3) with threshold $\mathfrak{a}_i \geq \sigma_{i-1}\|f_i\|_{\mathcal{H}}\sqrt{2\log(2/\delta_i)}$ for $\delta_i \in (0, 1]$ satisfies*

$$\Pr(\mathcal{E}_i) \geq 1 - \delta_i \quad \textit{for} \quad \mathcal{E}_i = \{|\langle \psi_{i-1}, f_i \rangle_{\mathcal{H}}| \leq \mathfrak{a}_i\}.$$

Proof Instantiate the notation of Thm. 3. The sub-Gaussian Hoeffding inequality (Wainwright, 2019, Prop. 2.5), the functional sub-Gaussianity of ψ_{i-1} (13), and the choice of $\mathfrak{a}_i$ imply that

$$\Pr(\mathcal{E}_i^c) = \Pr(|\langle \psi_{i-1}, f_i \rangle_{\mathcal{H}}| > \mathfrak{a}_i) \leq 2\exp(-\mathfrak{a}_i^2/(2\sigma_{i-1}^2\|f_i\|_{\mathcal{H}}^2)) \leq 2\exp(-\log(2/\delta_i)) = \delta_i.$$

■

J.3 Property (iii): Exact halving via symmetrization

Whenever the signed sum representation (14) holds, we have

$$\psi_n = \textstyle\sum_{i=1}^n \eta_i f_i = \sum_{i=1}^n (\eta_i g_{2i-1} - \eta_i g_{2i}) = \sum_{i=1}^{2n} g_i - 2\sum_{i\in\mathcal{I}} g_i$$

where the last step follows from the definition of $\mathcal{I}$.

J.4 Property (iv): Pointwise sub-Gaussianity in RKHS

The reproducing property of the kernel $\mathbf{k}_\star$ and the established functional sub-Gaussianity (13) yield

$$\mathbb{E}[\exp(\psi_i(x))] = \mathbb{E}[\exp(\langle \psi_i, \mathbf{k}_\star(x, \cdot)\rangle_{\mathcal{H}})] \leq \exp\Big(\frac{\sigma_i^2\|\mathbf{k}_\star(x,\cdot)\|_{\mathcal{H}}^2}{2}\Big) = \exp\Big(\frac{\sigma_i^2\mathbf{k}_\star(x,x)}{2}\Big), \quad \forall x \in \mathcal{X}.$$

J.5 Property (v): Sub-Gaussian constant bound

We establish the bound (15) for all $i \in \{0, \ldots, n\}$ by induction on the iteration i.

Base case The claim (15) holds for the base case, $i = 0$, since $\sigma_0 = 0$.

Inductive step Fix any $i \in [n]$ and assume that the claim (15) holds for σ_{i-1}.

If either $\|f_i\|_{\mathcal{H}} = 0$ or $\sigma_{i-1}^2 \geq \frac{\mathfrak{a}_i^2}{2\mathfrak{a}_i - \|f_i\|_{\mathcal{H}}^2}$, then $\sigma_i^2 = \sigma_{i-1}^2$ by the definition (12) of σ_i and the assumption that $\frac{\|f_i\|_{\mathcal{H}}^2}{2} \leq \mathfrak{a}_i$, completing the inductive step.

If, alternatively, $\sigma_{i-1}^2 < \frac{\mathfrak{a}_i^2}{2\mathfrak{a}_i - \|f_i\|_{\mathcal{H}}^2}$ and $\|f_i\|_{\mathcal{H}} > 0$, then our assumptions $\frac{\|f_i\|_{\mathcal{H}}^2}{1+q} \leq \mathfrak{a}_i \leq \frac{\|f_i\|_{\mathcal{H}}^2}{1-q}$ imply that $(\frac{\|f_i\|_{\mathcal{H}}^2}{\mathfrak{a}_i} - 1)^2 \leq q^2$. Hence, by the definition (12) of σ_i and the inductive hypothesis,

$$\begin{aligned}\sigma_i^2 &= \sigma_{i-1}^2 + \|f_i\|_{\mathcal{H}}^2\Big(1 + \frac{\sigma_{i-1}^2}{\mathfrak{a}_i^2}(\|f_i\|_{\mathcal{H}}^2 - 2\mathfrak{a}_i)\Big) \\ &= \|f_i\|_{\mathcal{H}}^2 + \sigma_{i-1}^2(\frac{\|f_i\|_{\mathcal{H}}^2}{\mathfrak{a}_i} - 1)^2 \\ &\leq (1-q^2)\frac{\|f_i\|_{\mathcal{H}}^2}{1-q^2} + q^2\frac{\max_{j\in[i-1]}\|f_j\|_{\mathcal{H}}^2}{1-q^2} \leq \frac{\max_{j\in[i]}\|f_j\|_{\mathcal{H}}^2}{1-q^2},\end{aligned}$$

completing the inductive step.

J.6 Property (vi): Adaptive thresholding

Define $c_1^\star = \max(c^\star, 1)$ and let $q = \frac{(c_1^\star)^2-1}{(c_1^\star)^2+1} \in [0,1)$ so that

$$\frac{1}{1-q} = \frac{1+(c_1^\star)^2}{2} \quad \text{and} \quad \frac{1}{1-q^2} = \frac{(c_1^\star+1/c_1^\star)^2}{4} \leq \frac{(c^\star+1/c^\star)^2}{4},$$

since $1 = \operatorname{argmin}_{c\geq 0} c + 1/c$. By assumption, $\mathfrak{a}_i \geq \|f_i\|_{\mathcal{H}}^2 \geq \frac{\|f_i\|_{\mathcal{H}}^2}{1+q}$ for all $i \in [n]$.

Now suppose that $\sigma_{i-1}^2 < \frac{\mathfrak{a}_i^2}{2\mathfrak{a}_i - \|f_i\|_{\mathcal{H}}^2}$ and $\|f_i\|_{\mathcal{H}} > 0$. If $\mathfrak{a}_i \leq \|f_i\|_{\mathcal{H}}^2$, then $\mathfrak{a}_i \leq \frac{\|f_i\|_{\mathcal{H}}^2}{1-q}$. If, alternatively, $\mathfrak{a}_i \leq c_i\sigma_{i-1}\|f_i\|_{\mathcal{H}}$, then

$$\mathfrak{a}_i < \tfrac{1}{2}\|f_i\|_{\mathcal{H}}^2(1+c_i^2) \leq \frac{\|f_i\|_{\mathcal{H}}^2}{1-q}.$$

The conclusion now follows from the sub-Gaussian constant bound (15).

J.7 Proof of Thm. 13: Alternate representation of ψ_i

Alg. 3 and our definition $\varepsilon_i \triangleq \mathbb{I}[|\alpha_i| \leq \mathfrak{a}_i](\eta_i + \alpha_i/\mathfrak{a}_i)$ give

$$\psi_i = \psi_{i-1} - f_i\,\alpha_i/\mathfrak{a}_i + \mathbb{I}[|\alpha_i| \leq \mathfrak{a}_i](f_i\,\alpha_i/\mathfrak{a}_i + \eta_i f_i) = \psi_{i-1} - f_i\,\frac{\langle f_i, \psi_{i-1}\rangle_{\mathcal{H}}}{\mathfrak{a}_i} + \varepsilon_i f_i.$$

Taking an inner product with $u \in \mathcal{H}$ now yields the equality (36). By construction, $\varepsilon_i \in [c_{min}, c_{max}] \subseteq [-2, 2]$ for

$$c_{min} = \max(-2, \min(0, -1 + \alpha_i/\mathfrak{a}_i)) \quad \text{and} \quad c_{max} = \min(2, \max(0, 1 + \alpha_i/\mathfrak{a}_i))$$

by construction. Moreover,

$$\begin{aligned}&\mathbb{E}[\varepsilon_i \mid \psi_{i-1}, |\alpha_i| > \mathfrak{a}_i] = 0 \\ &\mathbb{E}[\varepsilon_i \mid \psi_{i-1}, |\alpha_i| \leq \mathfrak{a}_i] = \Big(1 + \frac{\alpha_i}{\mathfrak{a}_i}\Big)\cdot\frac{1}{2}\Big(1 - \frac{\alpha_i}{\mathfrak{a}_i}\Big) + \Big(-1 + \frac{\alpha_i}{\mathfrak{a}_i}\Big)\cdot\frac{1}{2}\Big(1 + \frac{\alpha_i}{\mathfrak{a}_i}\Big) = 0,\end{aligned}$$

so that $\mathbb{E}[\varepsilon_i \mid \psi_{i-1}] = 0$ as claimed. The conditional sub-Gaussianity claim

$$\mathbb{E}[e^{t\varepsilon_i} \mid \psi_{i-1}] \leq e^{t^2/2} \quad \text{for all} \quad t \in \mathbb{R},$$

now follows from Hoeffding's lemma (Hoeffding, 1963, (4.16)) since ε_i is bounded with $c_{max} - c_{min} \leq 2$ and mean-zero conditional on ψ_{i-1}.

Appendix K. Proof of Thm. 4: L^∞ guarantees for kernel halving

We start by showing that after m rounds of kernel halving the output size is $n_{\mathrm{out}} = \lfloor \frac{n}{2^m} \rfloor$ (also see Footnote 3). We prove this by induction. The base case of $m = 1$ can be directly verified. Let $n_j \triangleq |\mathcal{S}^{(j)}|$ denote the output size after j rounds of kernel halving and assume that the hypothesis is true for round j, i.e., $n_j = \lfloor \frac{n}{2^j} \rfloor$. Then for round $j+1$, we have $n_{j+1} = \lfloor n_j/2 \rfloor = \lfloor \frac{\lfloor n/2^j \rfloor}{2} \rfloor \stackrel{(i)}{=} \lfloor \frac{n}{2^{j+1}} \rfloor$, where step (i) follows from the fact (Graham et al., 1994, Eqn. (3.11)) that $\lfloor \frac{\lfloor \ell \rfloor}{2} \rfloor = \lfloor \frac{\ell}{2} \rfloor$ for any integer ℓ. Our desired claim follows.

Next, define $n_{\mathrm{in}} = 2^m n_{\mathrm{out}} = 2^m \lfloor \frac{n}{2^m} \rfloor$. In Apps. K.1 and K.2 we will show that the respective claims (16) and (17) hold whenever $n = n_{\mathrm{in}}$, that is, whenever 2^m divides n evenly. Now suppose that $n \neq n_{\mathrm{in}}$. Since the output of m rounds of kernel halving depends only on the first n_{in} points, the evenly divisible case of App. K.1 implies that, for part (a),

$$\|\mathbb{P}_{n_{\mathrm{in}}}\mathbf{k} - \mathbb{Q}_{\mathrm{KH}}^{(1)}\mathbf{k}\|_\infty \leq \frac{\|\mathbf{k}\|_\infty}{n_{\mathrm{out}}} \cdot \mathfrak{M}_{\mathbf{k}}(n_{\mathrm{in}},1,d,\delta^\star,\delta',R_{\mathcal{S}_{n_{\mathrm{in}}},\mathbf{k},n_{\mathrm{in}}}) \text{ for } n_{\mathrm{in}} = 2\lfloor \tfrac{n}{2} \rfloor = 2n_{\mathrm{out}}, \tag{39}$$

with probability at least $1-\delta'-\sum_{i=1}^{n_{\mathrm{in}}/2}\delta_i$, and App. K.1 implies that, for part (b),

$$\|\mathbb{P}_{n_{\mathrm{in}}}\mathbf{k}-\mathbb{Q}_{\mathrm{KH}}^{(m)}\mathbf{k}\|_\infty \leq \frac{\|\mathbf{k}\|_\infty}{n_{\mathrm{out}}} \cdot \mathfrak{M}_{\mathbf{k}}(n_{\mathrm{in}},m,d,\delta^\star,\delta',R_{\mathcal{S}_{n_{\mathrm{in}}},\mathbf{k},n_{\mathrm{in}}}) \text{ for } n_{\mathrm{in}}=2^m\lfloor \tfrac{n}{2^m} \rfloor=2^m n_{\mathrm{out}}, \tag{40}$$

with probability at least $1 - \delta' - \sum_{j=1}^{m} \frac{2^{j-1}}{m} \sum_{i=1}^{n_{\mathrm{in}}/2^j} \delta_i$. Next, to recover the bounds (16) and (17), we use the following deterministic inequalities:

$$\|\mathbb{P}_n\mathbf{k} - \mathbb{P}_{n_{\mathrm{in}}}\mathbf{k}\|_\infty \leq \frac{2(n-n_{\mathrm{in}})\|\mathbf{k}\|_\infty}{n} \leq \frac{2(2^m-1)\|\mathbf{k}\|_\infty}{n} \leq \frac{2\|\mathbf{k}\|_\infty}{n_{\mathrm{out}}}, \tag{41}$$

$$R_{\mathcal{S}_{n_{\mathrm{in}}},\mathbf{k},n_{\mathrm{in}}} \leq R_{\mathcal{S}_n,\mathbf{k},n}, \quad \text{and} \tag{42}$$

$$\begin{aligned}\mathfrak{M}_{\mathbf{k}}(n_{\mathrm{in}},m,d,\delta,\delta',R_{\mathcal{S}_{n_{\mathrm{in}}},\mathbf{k},n_{\mathrm{in}}})+2 &\stackrel{(i)}{\leq} \mathfrak{M}_{\mathbf{k}}(n_{\mathrm{in}},m,d,\delta,\delta',R_{\mathcal{S}_n,\mathbf{k},n})+2 \\ &\stackrel{(ii)}{\leq} \mathfrak{M}_{\mathbf{k}}(n,m,d,\delta,\delta',R_{\mathcal{S}_n,\mathbf{k},n}),\end{aligned} \tag{43}$$

where (42) follows directly from the definition (7), and step (i) follows from (42) and the fact that $\mathfrak{M}_{\mathbf{k}}(n,m,d,\delta,\delta',R)$ (8) is non-decreasing in R, and step (ii) follows since $\mathfrak{M}_{\mathbf{k}}(n,m,d,\delta,\delta',R)$ is non-decreasing in n and $n_{\mathrm{in}} \neq n$. For part (a) we conclude, by (39),

$$\begin{aligned}\|\mathbb{P}_n\mathbf{k}-\mathbb{Q}_{\mathrm{KH}}^{(1)}\mathbf{k}\|_\infty &\leq \|\mathbb{P}_n\mathbf{k} - \mathbb{P}_{n_{\mathrm{in}}}\mathbf{k}\|_\infty + \|\mathbb{P}_{n_{\mathrm{in}}}\mathbf{k}-\mathbb{Q}_{\mathrm{KH}}^{(1)}\mathbf{k}\|_\infty \\ &\stackrel{(41)}{\leq} \frac{2\|\mathbf{k}\|_\infty}{n_{\mathrm{out}}} + \frac{\|\mathbf{k}\|_\infty}{n_{\mathrm{out}}} \cdot \mathfrak{M}_{\mathbf{k}}(n_{\mathrm{in}}, 1, d, \delta^\star, \delta', R_{\mathcal{S}_{n_{\mathrm{in}}},\mathbf{k},n_{\mathrm{in}}}) \\ &\stackrel{(43)}{\leq} \frac{\|\mathbf{k}\|_\infty}{n_{\mathrm{out}}} \cdot \mathfrak{M}_{\mathbf{k}}(n, 1, d, \delta, \delta', R_{\mathcal{S}_n,\mathbf{k},n}).\end{aligned}$$

For part (b), we conclude, by (40),

$$\begin{aligned}\|\mathbb{P}_n\mathbf{k}-\mathbb{Q}_{\mathrm{KH}}^{(m)}\mathbf{k}\|_\infty &\leq \|\mathbb{P}_n\mathbf{k} - \mathbb{P}_{n_{\mathrm{in}}}\mathbf{k}\|_\infty + \|\mathbb{P}_{n_{\mathrm{in}}}\mathbf{k}-\mathbb{Q}_{\mathrm{KH}}^{(m)}\mathbf{k}\|_\infty \\ &\stackrel{(41)}{\leq} \frac{2\|\mathbf{k}\|_\infty}{n_{\mathrm{out}}} + \frac{\|\mathbf{k}\|_\infty}{n_{\mathrm{out}}} \cdot \mathfrak{M}_{\mathbf{k}}(n_{\mathrm{in}}, 1, d, \delta^\star, \delta', R_{\mathcal{S}_{n_{\mathrm{in}}},\mathbf{k},n_{\mathrm{in}}}) \\ &\stackrel{(43)}{\leq} \frac{\|\mathbf{k}\|_\infty}{n_{\mathrm{out}}} \cdot \mathfrak{M}_{\mathbf{k}}(n, m, d, \delta, \delta', R_{\mathcal{S}_n,\mathbf{k},n}).\end{aligned}$$

K.1 Proof of part (a): Kernel halving yields a 2-thinned L^∞ coreset

As noted earlier, we prove this part assuming n is even. Consider a self-balancing Hilbert walk (Alg. 3) with inputs $(f_i)_{i=1}^{n/2}$ and $(\mathfrak{a}_i)_{i=1}^{n/2}$, where $f_i = \mathbf{k}(x_{2i-1}, \cdot) - \mathbf{k}(x_{2i}, \cdot)$ and $\mathfrak{a}_i = \max(\|f_i\|_{\mathbf{k}_\star}\sigma_{i-1}\sqrt{2\log(2/\delta_i)}, \|f_i\|_{\mathbf{k}_\star}^2)$, and σ_i was defined in (12). Property (iii) of Thm. 3 implies that for a self-balancing walk with the choices summarized above, the event $\mathcal{E}_{\text{half}} = \{|\alpha_i| \leq \mathfrak{a}_i \text{ for all } i \in [n/2]\}$ satisfies

$$\Pr(\mathcal{E}_{\text{half}}) \geq 1 - \textstyle\sum_{i=1}^{n/2} \delta_i. \tag{44}$$

Consider kernel halving coupled with the above instantiation of self-balancing Hilbert walk. Due to the equivalence with kernel halving on the event $\mathcal{E}_{\text{half}}$, we conclude that the output $\psi_{n/2}$ of self-balancing Hilbert walk matches with that of the kernel halving, and satisfies

$$\tfrac{1}{n}\psi_{n/2} = \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \mathbf{k}(x_i, \cdot) - \tfrac{1}{n/2}\textstyle\sum_{x\in\mathcal{S}^{(1)}} \mathbf{k}(x, \cdot) = \mathbb{P}_n\mathbf{k} - \mathbb{Q}_{\text{KH}}^{(1)}\mathbf{k}, \tag{45}$$

on the event $\mathcal{E}_{\text{half}}$. Furthermore, on the event $\mathcal{E}_{\text{half}}$, we can also write that the kernel halving coreset satisfies $\mathcal{S}^{(1)} = (x_i)_{i\in\mathcal{I}}$ for $\mathcal{I} = \{2i - \frac{\eta_i - 1}{2} : i \in [n/2]\}$ and η_i defined in Alg. 3. Finally, applying property (vi) of Thm. 3 for $\sigma_{n/2}$ with $c_i = \sqrt{2\log(2/\delta_i)}$, we obtain that

$$\sigma_{n/2}^2 \leq \frac{\max_{i\leq n/2}\|f_i\|_{\mathbf{k}_\star}^2}{4} \cdot 2\log(\tfrac{2}{\delta^\star})(1 + \tfrac{1}{2\log(2/\delta^\star)})^2 \stackrel{(i)}{\leq} 4\|\mathbf{k}\|_\infty \log(\tfrac{2}{\delta^\star}) \tag{46}$$

where in step (i), we use the fact that $f_i = \mathbf{k}(x_{2i-1}, \cdot) - \mathbf{k}(x_{2i}, \cdot)$ satisfies

$$\|f_i\|_{\mathbf{k}_\star}^2 = \mathbf{k}(x_{2i-1}, x_{2i-1}) + \mathbf{k}(x_{2i}, x_{2i}) - 2\mathbf{k}(x_{2i-1}, x_{2i}) \leq 4\|\mathbf{k}\|_\infty.$$

Next, we split the proof in two parts: **Case (I)** When $R_{\mathcal{S}_n,\mathbf{k},n} < R_{\mathcal{S}_n}$, and **Case (II)** when $R_{\mathcal{S}_n,\mathbf{k},n} = R_{\mathcal{S}_n}$, where $R_{\mathcal{S}_n,\mathbf{k},n}$ was defined in (7). We prove the results for these two cases in Apps. K.1.1 and K.1.2 respectively. In the sequel, we make use of the following tail quantity of the kernel:

$$\overline{\tau}_{\mathbf{k}}(R') \triangleq \sup\{|\mathbf{k}(x, y)| : \|x - y\|_2 \geq R'\}. \tag{47}$$

K.1.1 Proof for case (I): When $R_{\mathcal{S}_n,\mathbf{k},n} < R_{\mathcal{S}_n}$

By definition (7), for this case,

$$R_{\mathcal{S}_n,\mathbf{k},n} = n^{1+\frac{1}{d}} R_{\mathbf{k},n} + n^{\frac{1}{d}} \tfrac{\|\mathbf{k}\|_\infty}{L_{\mathbf{k}}}. \tag{48}$$

On the event $\mathcal{E}_{\text{half}}$, the following lemma provides a high probability bound on $\|\psi_{n/2}\|_\infty$ in terms of the kernel parameters, the sub-Gaussianity parameter $\sigma_{n/2}$, and the size of the *cover* (Wainwright, 2019, Def. 5.1) of a neighborhood of the input points $(x_i)_{i=1}^n$.

Lemma 15 (A direct covering bound on $\|\psi_{n/2}\|_\infty$)**.** *Fix $R \geq r > 0$ and $\delta' > 0$, and suppose $\mathcal{C}^n(r, R)$ is a set of minimum cardinality satisfying*

$$\textstyle\bigcup_{i=1}^n \mathcal{B}(x_i, R) \subseteq \bigcup_{z\in\mathcal{C}^n(r,R)} \mathcal{B}(z, r). \tag{49}$$

If $\mathbf{k}$ satisfies Assumps. 1 and 2, then, conditional on the event $\mathcal{E}_{\text{half}}$ (45), the event

$$\mathcal{E}_\infty \triangleq \big\{\|\psi_{n/2}\|_\infty \leq \max\big(n\overline{\tau}_{\mathbf{k}}(R), \quad nL_{\mathbf{k}}r + \sigma_{n/2}\sqrt{2\|\mathbf{k}\|_\infty \log(2|\mathcal{C}^n(r, R)|/\delta')}\big)\big\}, \tag{50}$$

occurs with probability at least $1-\delta'$, i.e., $\Pr(\mathcal{E}_\infty|\mathcal{E}_{\text{half}}) \geq 1-\delta'$, where $\overline{\tau}_{\mathbf{k}}$ was defined in (47).

Thm. 15 succeeds in controlling $\psi_{n/2}(x)$ for all $x \in \mathbb{R}^d$ since either x lies far from every input point x_i so that each $\mathbf{k}(x_i, x)$ in the expansion (45) is small or x lies near some x_i, in which case $\psi_{n/2}(x)$ is well approximated by $\psi_{n/2}(z)$ for $z \in \mathcal{C}^n(r, R)$. The proof inspired by the covering argument of Phillips and Tai (2020, Lem. 2.1) and using the pointwise sub-Gaussianity property of Thm. 3 over the finite cover $\mathcal{C}^n$ can be found in App. K.3.

Now we put together the pieces to prove Thm. 4.

First, (Wainwright, 2019, Lem. 5.7) implies that $|\mathcal{C}^1(r, R)| \leq (1 + 2R/r)^d$ (i.e., any ball of radius R in $\mathbb{R}^d$ can be covered by $(1 + 2R/r)^d$ balls of radius r). Thus for an arbitrary R, we can conclude that

$$|\mathcal{C}^n(r, R)| \leq n(1 + 2R/r)^d = (n^{1/d} + 2n^{1/d}R/r)^d. \tag{51}$$

Second we fix R and r such that $n\overline{\tau}_{\mathbf{k}}(R) = \|\mathbf{k}\|_\infty$ and $nL_{\mathbf{k}}r = \|\mathbf{k}\|_\infty$, so that $\frac{R}{r} \leftarrow nR_{\mathbf{k},n}\frac{L_{\mathbf{k}}}{\|\mathbf{k}\|_\infty}$ (c.f. (6) and (47)). Substituting these choices of radii in the bound (50) of Thm. 15, we find that conditional to $\mathcal{E}_{\mathrm{half}} \cap \mathcal{E}_\infty$, we have

$$\begin{aligned}
\|\psi_{n/2}\|_\infty &\leq \max\big(n\overline{\tau}_{\mathbf{k}}(R), \quad nL_{\mathbf{k}}r + \sigma_{n/2}\sqrt{2\|\mathbf{k}\|_\infty \log(2|\mathcal{C}^n(r, R)|/\delta')}\big) \\
&\overset{(46)}{\leq} \|\mathbf{k}\|_\infty + 2\sqrt{2}\|\mathbf{k}\|_\infty\sqrt{\log(\tfrac{2}{\delta^\star})\Big[\log(\tfrac{2}{\delta'}) + d\log\Big(n^{\frac{1}{d}} + \tfrac{2L_{\mathbf{k}}}{\|\mathbf{k}\|_\infty}\cdot n^{1+\frac{1}{d}}R_{\mathbf{k},n}\Big)\Big]} \quad (52)\\
&\overset{(48)}{\leq} \|\mathbf{k}\|_\infty + 2\sqrt{2}\|\mathbf{k}\|_\infty\sqrt{\log(\tfrac{2}{\delta^\star})\Big[\log(\tfrac{2}{\delta'}) + d\log\Big(\tfrac{2L_{\mathbf{k}}}{\|\mathbf{k}\|_\infty}(R_{\mathbf{k},n} + R_{\mathcal{S}_n,\mathbf{k},n})\Big)\Big]} \\
&\overset{(8)}{\leq} \|\mathbf{k}\|_\infty \cdot 2\mathfrak{M}_{\mathbf{k}}(n, 1, d, \delta^\star, \delta', R_{\mathcal{S}_n,\mathbf{k},n}). \qquad (53)
\end{aligned}$$

Putting (45) and (53) together, we conclude

$$\begin{aligned}
\Pr\big(\|\mathbb{P}_n\mathbf{k} - \mathbb{Q}_{\mathrm{KH}}^{(1)}\mathbf{k}\|_\infty > \frac{1}{n/2}\mathfrak{M}_{\mathbf{k}}(n, 1, d, \delta^\star, \delta', R_{\mathcal{S}_n,\mathbf{k},n})\big) &\leq \Pr((\mathcal{E}_{\mathrm{half}} \cap \mathcal{E}_\infty)^c) \\
&= \Pr(\mathcal{E}_{\mathrm{half}}^c \cup \mathcal{E}_\infty^c) \\
&= \Pr(\mathcal{E}_{\mathrm{half}}^c) + \Pr(\mathcal{E}_{\mathrm{half}} \cap \mathcal{E}_\infty^c) \\
&= \Pr(\mathcal{E}_{\mathrm{half}}^c) + \Pr(\mathcal{E}_\infty^c|\mathcal{E}_{\mathrm{half}}) \\
&\leq \delta' + \textstyle\sum_{i=1}^{n/2}\delta_i,
\end{aligned}$$

where the last step follows from (44) and Thm. 15. The claim now follows.

K.1.2 Proof for case (II): When $R_{\mathcal{S}_n,\mathbf{k},n} = R_{\mathcal{S}_n}$

In this case, we split the proof for bounding $\left\|\psi_{n/2}\right\|_\infty$ into two lemmas. First, we relate the $\left\|\psi_{n/2}\right\|_\infty$ in terms of the tail behavior of $\mathbf{k}$ and the supremum of differences for $\psi_{n/2}$ between any pair points on a Euclidean ball (see App. K.4 for the proof):

Lemma 16 (A basic bound on $\|\psi_{n/2}\|_\infty$)**.** *If $\mathbf{k}$ satisfies Assump. 1, then, conditional on the event $\mathcal{E}_{\mathrm{half}}$ (45), for any fixed $R = R' + R_{\mathcal{S}_n}$ with $R' > 0$ and any fixed $\delta' \in (0, 1)$,*

$$\widetilde{\mathcal{E}}_\infty = \Big\{\|\psi_{n/2}\|_\infty \leq \max\big(n\overline{\tau}_{\mathbf{k}}(R'), \sigma_{n/2}\|\mathbf{k}\|_\infty^{\frac{1}{2}}\sqrt{2\log(\tfrac{4}{\delta'})} + \sup_{x,x' \in \mathcal{B}(0,R)}|\psi_{n/2}(x) - \psi_{n/2}(x')|\big)\Big\} \tag{54}$$

occurs with probability at least $1 - \frac{\delta'}{2}$, i.e., $\Pr(\widetilde{\mathcal{E}}_\infty|\mathcal{E}_{\mathrm{half}}) \geq 1 - \frac{\delta'}{2}$, where $\overline{\tau}_{\mathbf{k}}$ was defined in (47).

Next, to control the supremum term on the RHS of the display (54), we establish a high probability bound in the next lemma. Its proof in App. K.5 proceeds by showing that $\psi_{n/2}$ is an Orlicz process with a suitable metric and then applying standard concentration arguments for such processes.

Lemma 17 (A high probability bound on supremum of $\psi_{n/2}$ differences)**.** *If* $\mathbf{k}$ *satisfies Assumps. 1 and 2, then, for any fixed* $R > 0, \delta' \in (0,1)$*, the event*

$$\mathcal{E}_{\mathrm{sup}} \triangleq \left\{ \sup_{x,x'\in\mathcal{B}(0,R)} |\psi_{n/2}(x) - \psi_{n/2}(x')| \leq 8D_R\left(\sqrt{\log(\tfrac{4}{\delta'})} + 6\sqrt{d\log\left(2 + \tfrac{L_{\mathbf{k}}R}{\|\mathbf{k}\|_\infty}\right)}\right)\right\} \tag{55}$$

occurs with probability at least $1 - \frac{\delta'}{2}$*, where* $D_R \triangleq \sqrt{\frac{32}{3}}\sigma_{n/2}\|\mathbf{k}\|_\infty^{\frac{1}{2}} \min\left(1, \sqrt{\frac{1}{2}L_{\mathbf{k}}R}\right)$.

We now turn to the rest of the proof for Thm. 4. We apply both Thms. 16 and 17 with $R = R_{\mathcal{S}_n} + R_{\mathbf{k},n}$ (6) and (7). For this R, we have $R' = R_{\mathbf{k},n}$ in Thm. 16 and hence $n\overline{\tau}_{\mathbf{k}}(R') \leq \|\mathbf{k}\|_\infty$. Now, condition on the event $\mathcal{E}_{\mathrm{half}} \cap \widetilde{\mathcal{E}}_\infty \cap \mathcal{E}_{\mathrm{sup}}$. Then, we have

$$\begin{aligned}
&n\|\mathbb{P}_n\mathbf{k} - \mathbb{Q}_{\mathrm{KH}}^{(1)}\mathbf{k}\|_\infty \\
&\overset{(45)}{=} \left\|\psi_{n/2}\right\|_\infty \\
&\overset{(54)}{\leq} \max(\|\mathbf{k}\|_\infty,\ \sigma_{n/2}\|\mathbf{k}\|_\infty^{\frac{1}{2}}\sqrt{2\log(\tfrac{4}{\delta'})} + \sup\nolimits_{x,x'\in\mathcal{B}(0,R)} |\psi_{n/2}(x) - \psi_{n/2}(x')|) \\
&\overset{(55)}{\leq} \max\Big(\|\mathbf{k}\|_\infty,\ \sigma_{n/2}\|\mathbf{k}\|_\infty^{\frac{1}{2}}\sqrt{2\log(\tfrac{4}{\delta'})} + \\
&\qquad\qquad 32\sqrt{\tfrac{2}{3}}\sigma_{n/2}\|\mathbf{k}\|_\infty^{\frac{1}{2}}\Big(\sqrt{\log(\tfrac{4}{\delta'})} + 6\sqrt{d\log(2 + \tfrac{L_{\mathbf{k}}(R_{\mathcal{S}_n}+R_{\mathbf{k},n})}{\|\mathbf{k}\|_\infty})}\Big)\Big) \\
&\overset{(i)}{\leq} \max(\|\mathbf{k}\|_\infty,\ 32\sigma_{n/2}\|\mathbf{k}\|_\infty^{\frac{1}{2}}\Big(\sqrt{\log(\tfrac{4}{\delta'})} + 5\sqrt{d\log\Big(2 + \tfrac{L_{\mathbf{k}}(R_{\mathcal{S}_n}+R_{\mathbf{k},n})}{\|\mathbf{k}\|_\infty}\Big)}\Big)) \qquad (56) \\
&\overset{(46)}{\leq} \|\mathbf{k}\|_\infty \cdot 64\sqrt{\log\left(\tfrac{2}{\delta^\star}\right)}\left[\sqrt{\log(\tfrac{4}{\delta'})} + 5\sqrt{d\log\Big(2 + \tfrac{L_{\mathbf{k}}(R_{\mathcal{S}_n}+R_{\mathbf{k},n})}{\|\mathbf{k}\|_\infty}\Big)}\right] \\
&\overset{(ii)}{\leq} \|\mathbf{k}\|_\infty 2\,\mathfrak{M}_{\mathbf{k}}(n, 1, d, \delta^\star, \delta', R_{\mathcal{S}_n,\mathbf{k},n}), \qquad (57)
\end{aligned}$$

where step (i) follows from the fact that $D_R \leq \sqrt{\frac{32}{3}}\sigma_{n/2}\|\mathbf{k}\|_\infty^{\frac{1}{2}}$, and in step (ii) we have used the working assumption for this case, i.e., $R_{\mathcal{S}_n,\mathbf{k},n} = R_{\mathcal{S}_n}$. As a result,

$$\begin{aligned}
\Pr(\|\mathbb{P}_n\mathbf{k} - \mathbb{Q}_{\mathrm{KH}}^{(1)}\mathbf{k}\|_\infty > \tfrac{2}{n}\mathfrak{M}_{\mathbf{k}}(n,1,d,\delta^\star,\delta',R_{\mathcal{S}_n,\mathbf{k},n}) &\leq \Pr((\mathcal{E}_{\mathrm{half}} \cap \widetilde{\mathcal{E}}_\infty \cap \mathcal{E}_{\mathrm{sup}})^c) \\
&= \Pr(\mathcal{E}_{\mathrm{half}}^c \cup \widetilde{\mathcal{E}}_\infty^c \cup \mathcal{E}_{\mathrm{sup}}^c) \\
&= \Pr(\mathcal{E}_{\mathrm{half}}^c \cup \mathcal{E}_{\mathrm{sup}}^c) + \Pr((\mathcal{E}_{\mathrm{half}}^c \cup \mathcal{E}_{\mathrm{sup}}^c)^c \cap \widetilde{\mathcal{E}}_\infty^c) \\
&= \Pr(\mathcal{E}_{\mathrm{half}}^c \cup \mathcal{E}_{\mathrm{sup}}^c) + \Pr(\mathcal{E}_{\mathrm{half}} \cap \mathcal{E}_{\mathrm{sup}} \cap \widetilde{\mathcal{E}}_\infty^c) \\
&\leq \Pr(\mathcal{E}_{\mathrm{half}}^c \cup \mathcal{E}_{\mathrm{sup}}^c) + \Pr(\mathcal{E}_{\mathrm{half}} \cap \widetilde{\mathcal{E}}_\infty^c) \\
&\leq \Pr(\mathcal{E}_{\mathrm{half}}^c) + \Pr(\mathcal{E}_{\mathrm{sup}}^c) + \Pr(\widetilde{\mathcal{E}}_\infty^c|\mathcal{E}_{\mathrm{half}}) \\
&\leq \textstyle\sum_{i=1}^{n/2}\delta_i + \frac{\delta'}{2} + \frac{\delta'}{2},
\end{aligned}$$

where the last step follows from (44) and Thms. 16 and 17. The desired claim follows.

K.2 Proof of part (b): Repeated kernel halving yields a 2^m-thinned L^∞ coreset

As noted earlier, we prove this part assuming $n/2^m \in \mathbb{N}$. The proof in this section follows by applying the arguments from the previous section, separately for each round and then invoking the sub-Gaussianity of a weighted sum of the output functions from each round.

Note that coreset $\mathcal{S}^{(j-1)}$ is independent of the randomness for the j-th round of kernel havling and thus can be treated as fixed for that round. When running kernel halving with the input $\mathcal{S}^{(j-1)}$, let $f_{i,j}, \mathfrak{a}_{i,j}, \psi_{i,j}, \alpha_{i,j}, \eta_{i,j}$ denote the analog of the quantities $f_i, \mathfrak{a}_i, \psi_i, \alpha_i, \eta_i$ defined in Alg. 2. Like in part (a), we would couple the j-th round of kernel halving with an instantiation of self-balancing Hilbert walk with inputs $(f_{i,j}, \mathfrak{a}_{i,j})_{i=1}^{n/2^j}$ and final output $\psi_{n/2^j,j}$, where $\mathfrak{a}_{i,j} = \max(\|f_{i,j}\|_{\mathbf{k}}\sigma_{i-1,j}\sqrt{2\log(\frac{4m}{2^j\delta_i})}, \|f_{i,j}\|_{\mathbf{k}}^2)$ and we define $\sigma_{i,j}$ in a recursive manner as in (12) with $f_{i,j}, \mathfrak{a}_{i,j}, \sigma_{i,j}$ taking the role of $f_i, \mathfrak{a}_i, \sigma_i$ respectively.

With this set-up, first we apply property (i) of Thm. 3 which implies that given $\mathcal{S}^{(j-1)}$, the function $\psi_{n/2^j,j}$ is $\sigma_{n/2^j,j}$ sub-Gaussian, where

$$\sigma_{n/2^j,j}^2 \leq 4\|\mathbf{k}\|_\infty \log(\tfrac{4m}{2^j\delta_j^\star}) \quad \text{with} \quad \delta_j^\star = \min(\delta_i)_{i=1}^{n/2^j}, \tag{58}$$

using an argument similar to (46) with property (vi) of Thm. 3. Next, we note that for j-th round, the event $\mathcal{E}_{\mathrm{half}}^{(j)} = |\alpha_{i,j}| \leq \mathfrak{a}_{i,j}$ for all $i \in [n/2^j]$, satisfies

$$\mathbb{P}(\mathcal{E}_{\mathrm{half}}^{(j)}) \geq 1 - \textstyle\sum_{i=1}^{n/2^j} \delta_i \frac{2^{j-1}}{m}; \tag{59}$$

and this event serves as the analog the equivalence event $\mathcal{E}_{\mathrm{half}}$ from part (a) for the j-th round of kernel halving. Thus on the event $\mathcal{E}_{\mathrm{half}}^{(j)}$, we can write that the kernel halving coreset $\mathcal{S}^{(j)} = (x_i)_{i\in\mathcal{I}_j}$ for $\mathcal{I}_j = \{2i - \frac{\eta_{i,j}-1}{2} : i \in [n/2^j]\}$ and $\eta_{i,j}$ as defined while running Alg. 3, so that

$$\tfrac{2^{j-1}}{n}\psi_{n/2^j,j} = \tfrac{1}{n/2^{j-1}} \textstyle\sum_{x\in\mathcal{S}^{(j-1)}} \mathbf{k}(x,\cdot) - \tfrac{1}{n/2^j} \textstyle\sum_{x\in\mathcal{S}^{(j)}} \mathbf{k}(x,\cdot). \tag{60}$$

Now conditional on the event $\cap_{j=1}^m \mathcal{E}_{\mathrm{half}}^{(j)}$, we conclude that the output of all m kernel halving rounds are equivalent to the output of m different self-balancing Hilbert walks (each with exact two-thinning in each round). Putting the pieces together, we conclude that on the event $\cap_{j=1}^m \mathcal{E}_{\mathrm{half}}^{(j)}$, we have

$$\begin{aligned}
\mathbb{P}_n\mathbf{k} - \mathbb{Q}_{\mathrm{KH}}^{(m)}\mathbf{k} &= \tfrac{1}{n} \textstyle\sum_{x\in\mathcal{S}_n} \mathbf{k}(x,\cdot) - \tfrac{1}{n/2^m} \textstyle\sum_{x\in\mathcal{S}^{(m)}} \mathbf{k}(x,\cdot) \\
&\overset{(i)}{=} \textstyle\sum_{j=1}^m \tfrac{1}{n/2^{j-1}} \textstyle\sum_{x\in\mathcal{S}^{(j-1)}} \mathbf{k}(x,\cdot) - \tfrac{1}{n/2^j} \textstyle\sum_{x\in\mathcal{S}^{(j)}} \mathbf{k}(x,\cdot) \\
&= \textstyle\sum_{j=1}^m \tfrac{2^{j-1}}{n} \psi_{n/2^j,j}.
\end{aligned} \tag{61}$$

where we abuse notation $\mathcal{S}^{(0)} \triangleq \mathcal{S}_n$ in step (i) for simplicity of expressions.

Next, we use the following basic fact:

Lemma 18 (Sub-Gaussian additivity). *For a sequence of random variables $(Z_j)_{j=1}^m$ such that Z_j is a σ_j sub-Gaussian variable conditional on $(Z_1, \ldots, Z_{j-1})$, the random variable $\mathcal{Z} = \sum_{j=1}^m \theta_j Z_j$ is $(\sum_{j=1}^m \theta_j^2\sigma_j^2)^{1/2}$ sub-Gaussian.*

Proof We will prove the result for $\mathcal{Z}_s = \sum_{j=1}^s \theta_j Z_j$ by induction on $s \leq m$. The result holds for the base case of $s = 0$ as $\mathcal{Z}_s = 0$ is 0 sub-Gaussian. For the inductive case, suppose the result holds for s. Then we may apply the tower property, our conditional sub-Gaussianity assumption, and our inductive hypothesis in turn to conclude

$$\mathbb{E}[e^{\sum_{j=1}^{s+1}\theta_j Z_j}] = \mathbb{E}[e^{\sum_{j=1}^{s+1}\theta_j Z_j}\mathbb{E}[e^{\theta_{s+1}Z_{s+1}} \mid Z_{1:s}]] \leq \mathbb{E}[e^{\sum_{j=1}^{s}\theta_j Z_j}]e^{\frac{\theta_{s+1}^2\sigma_{s+1}^2}{2}} = e^{\frac{\sum_{j=1}^{s+1}\theta_j^2\sigma_j^2}{2}}.$$

Hence, $\mathcal{Z}_{s+1}$ is $\sqrt{\sum_{j=1}^{s+1}\theta_j^2\sigma_j^2}$ sub-Gaussian, and the proof is complete. ■

Applying Thm. 18 to the output sequence $(\psi_{n/2^j,j})_{j=1}^m$ for the m rounds of self-balancing Hilbert walks, we conclude that, the random variable

$$\mathcal{W}_m \triangleq \sum_{j=1}^m \frac{2^{j-1}}{n}\psi_{n/2^j,j} \tag{62}$$

is sub-Gaussian with parameter

$$\sigma_{\mathcal{W}_m} \triangleq \frac{2}{\sqrt{3}}\frac{2^m}{n}\sqrt{\|\mathbf{k}\|_\infty \log(\tfrac{6m}{2^m\delta^\star})} \overset{(i)}{\geq} \sqrt{\|\mathbf{k}\|_\infty \sum_{j=1}^m \tfrac{4^j}{n^2}\log(\tfrac{4m}{2^j\delta^\star})} \overset{(ii)}{\geq} \sqrt{\sum_{j=1}^m (\tfrac{2^{j-1}}{n})^2\sigma_{n/2^j,j}^2}, \tag{63}$$

conditional to the input $\mathcal{S}_n$, where step (ii) follows since

$$\|\mathbf{k}\|_\infty \sum_{j=1}^m 4^j \log(\tfrac{4m}{2^j\delta^\star}) \overset{(iii)}{\geq} 4\|\mathbf{k}\|_\infty \sum_{j=1}^m 4^{j-1}\log(\tfrac{4m}{2^j\delta^\star}) \overset{(58)}{\geq} \sum_{j=1}^m 4^{j-1}\sigma_{n/2^j,j}^2,$$

and step (iii) follows from the fact that $\delta^\star = \min(\delta_j^\star)_{j=1}^m$ (58). Now to prove step (i), we note that

$$\mathfrak{S}_1 \triangleq \sum_{j=1}^m 4^j = \tfrac{4}{3}(4^m - 1) \leq \tfrac{4}{3}4^m, \quad \text{and} \quad \mathfrak{S}_2 \triangleq \sum_{j=1}^m j4^j = \tfrac{4m}{3}\cdot 4^m - \tfrac{\mathfrak{S}_1}{3},$$

which in turn implies that

$$\begin{aligned}\sum_{j=1}^m 4^j \log(\tfrac{4m}{2^j\delta^\star}) = \mathfrak{S}_1 \log(\tfrac{4m}{\delta^\star}) - \log 2 \cdot \mathfrak{S}_2 &= \mathfrak{S}_1\Big(\log(\tfrac{4m}{\delta^\star}) + \tfrac{\log 2}{3}\Big) - \tfrac{4m\log 2}{3}\cdot 4^m \\ &\leq \mathfrak{S}_1 \cdot \log(\tfrac{6m}{\delta^\star}) - \tfrac{4m\log 2}{3}\cdot 4^m \\ &\leq \tfrac{4}{3}4^m \cdot \log(\tfrac{6m}{\delta^\star}) - \tfrac{4}{3}4^m \cdot \log 2^m = \tfrac{4}{3}4^m \cdot \log(\tfrac{6m}{2^m\delta^\star}),\end{aligned}$$

thereby establishing step (i).

Next, analogous to the proof of part (a), we split the proof in two parts: **Case (I)** When $R_{\mathcal{S}_n,\mathbf{k},n} < R_{\mathcal{S}_n}$, in which case, we proceed with a direct covering argument to bound $\|\mathcal{W}_m\|_\infty$ using a lemma analogous to Thm. 15; and **Case (II)** when $R_{\mathcal{S}_n,\mathbf{k},n} = R_{\mathcal{S}_n}$, in which case, we proceed with a metric-entropy based argument to bound $\|\mathcal{W}_m\|_\infty$ using two lemmas that are analogous Thms. 16 and 17.

K.2.1 Proof for case (I): When $R_{\mathcal{S}_n,\mathbf{k},n} < R_{\mathcal{S}_n}$

Recall that for this case, $R_{\mathcal{S}_n,\mathbf{k},n}$ is given by (48). The next lemma, a straightforward extension of Thm. 15, provides a high probability control on $\|\mathcal{W}_m\|_\infty$. Its proof is omitted for brevity.

Lemma 19 (A direct covering bound on $\|\mathcal{W}_m\|_\infty$). *Fix $R \geq r > 0$ and $\delta' > 0$, and recall the definition (49) of $\mathcal{C}^n(r, R)$. If $\mathbf{k}$ satisfies Assumps. 1 and 2, then, conditional on the event $\cap_{j=1}^m \mathcal{E}_{\mathrm{half}}^{(j)}$ (60), the event*

$$\mathcal{E}_\infty^{(m)} \triangleq \big\{\|\mathcal{W}_m\|_\infty \leq \max\big(2^m \overline{\tau}_{\mathbf{k}}(R), \quad 2^m L_{\mathbf{k}} r + \sigma_{\mathcal{W}_m} \sqrt{2\|\mathbf{k}\|_\infty \log(2|\mathcal{C}^n(r, R)|/\delta')}\big)\big\}, \quad (64)$$

occurs with probability at least $1-\delta'$, i.e., $\mathrm{Pr}(\mathcal{E}_\infty^{(m)} | \cap_{j=1}^m \mathcal{E}_{\mathrm{half}}^{(j)}) \geq 1-\delta'$.

Next, we repeat arguments similar to those used earlier around the display (52) and (53). Fix R and r such that $\overline{\tau}_{\mathbf{k}}(R) = \|\mathbf{k}\|_\infty/n$ and $L_{\mathbf{k}} r = \|\mathbf{k}\|_\infty/n$, so that $\frac{R}{r} \leftarrow n R_{\mathbf{k},n} \frac{L_{\mathbf{k}}}{\|\mathbf{k}\|_\infty}$ (c.f. (6) and (47)). Substituting these choices of radii in the bound (64) of Thm. 19, we find that conditional to $\cap_{j=1}^m \mathcal{E}_{\mathrm{half}}^{(j)} \cap \mathcal{E}_\infty^{(m)}$, we have

$$\begin{aligned}
\|\mathbb{P}_n \mathbf{k} - \mathbb{Q}_{\mathrm{KH}}^{(m)} \mathbf{k}\|_\infty &\overset{(61)}{=} \|\textstyle\sum_{j=1}^m \frac{2^{j-1}}{n} \psi_{n/2^j, j}\|_\infty \\
&\overset{(62)}{=} \|\mathcal{W}_m\|_\infty \\
&\overset{(64)}{\leq} \max\big(2^m \overline{\tau}_{\mathbf{k}}(R), \quad 2^m L_{\mathbf{k}} r + \sigma_{\mathcal{W}_m} \sqrt{2\|\mathbf{k}\|_\infty \log(2|\mathcal{C}^n(r, R)|/\delta')}\big) \\
&\overset{(i)}{\leq} \tfrac{2^m}{n} \|\mathbf{k}\|_\infty (1 + 2\sqrt{\tfrac{2}{3}} \sqrt{\log(\tfrac{6m}{2^m \delta^\star}) \Big[\log(\tfrac{2}{\delta'}) + d \log\Big(n^{\frac{1}{d}} + \tfrac{2L_{\mathbf{k}}}{\|\mathbf{k}\|_\infty} \cdot n^{1+\frac{1}{d}} R_{\mathbf{k},n}\Big)\Big]} \\
&\overset{(48)}{\leq} \tfrac{2^m}{n} \|\mathbf{k}\|_\infty (1 + 2\sqrt{\tfrac{2}{3}} \sqrt{\log(\tfrac{6m}{2^m \delta^\star}) \Big[\log(\tfrac{2}{\delta'}) + d \log\Big(\tfrac{2L_{\mathbf{k}}}{\|\mathbf{k}\|_\infty} (R_{\mathbf{k},n} + R_{\mathcal{S}_n, \mathbf{k}, n})\Big)\Big]}) \\
&\overset{(8)}{\leq} \|\mathbf{k}\|_\infty \cdot \tfrac{2^m}{n} \mathfrak{M}_{\mathbf{k}}(n, m, d, \delta^\star, \delta', R_{\mathcal{S}_n, \mathbf{k}, n}),
\end{aligned}$$

where in step (i), we have used the bounds (51) and (63). Putting the pieces together, we conclude

$$\begin{aligned}
&\mathrm{Pr}\big(\|\mathbb{P}_n \mathbf{k} - \mathbb{Q}_{\mathrm{KH}}^{(m)} \mathbf{k}\|_\infty > \|\mathbf{k}\|_\infty \cdot \frac{2^m}{n} \mathfrak{M}_{\mathbf{k}}(n, m, d, \delta^\star, \delta', R_{\mathcal{S}_n, \mathbf{k}, n})\big) \\
&\leq \mathrm{Pr}((\cap_{j=1}^m \mathcal{E}_{\mathrm{half}}^{(j)} \cap \mathcal{E}_\infty^{(m)})^c) \\
&= \mathrm{Pr}(\cup_{j=1}^m (\mathcal{E}_{\mathrm{half}}^{(j)})^c \cup (\mathcal{E}_\infty^{(m)})^c) \\
&= \mathrm{Pr}(\cup_{j=1}^m (\mathcal{E}_{\mathrm{half}}^{(j)})^c) + \mathrm{Pr}(\cap_{j=1}^m \mathcal{E}_{\mathrm{half}}^{(j)} \cap (\mathcal{E}_\infty^{(m)})^c) \\
&\leq \sum_{j=1}^m \mathrm{Pr}((\mathcal{E}_{\mathrm{half}}^{(j)})^c) + \mathrm{Pr}((\mathcal{E}_\infty^{(m)})^c | \cap_{j=1}^m (\mathcal{E}_{\mathrm{half}}^{(j)})) \\
&\leq \textstyle\sum_{j=1}^m \sum_{i=1}^{n/2^j} \delta_i \cdot \frac{2^{j-1}}{m} + \delta',
\end{aligned}$$

where the last step follows from (59) and Thm. 19. The claim follows.

K.2.2 Proof for case (II): When $R_{\mathcal{S}_n, \mathbf{k}, n} = R_{\mathcal{S}_n}$

In this case, we makes use of two lemmas. Their proofs (omitted for brevity) can be derived essentially by replacing $\psi_{n/2}$ and $\sigma_{n/2}$ with $\mathcal{W}_m$ (62) and $\sigma_{\mathcal{W}_m}$ (63) respectively, and repeating the proof arguments from Thms. 16 and 17.

Lemma 20 (A basic bound on $\|\mathcal{W}_m\|_\infty$)**.** *If* $\mathbf{k}$ *satisfies Assump. 1, then, conditional on the event* $\cap_{j=1}^m \mathcal{E}_{\text{half}}^{(j)}$ (60)*, for any fixed* $R = R' + R_{\mathcal{S}_n}$ *with* $R' > 0$ *and any fixed* $\delta' \in (0,1)$*,*

$$\widetilde{\mathcal{E}}_\infty^{(m)} = \left\{ \|\mathcal{W}_m\|_\infty \leq \max\left(2^m \overline{\tau}_{\mathbf{k}}(R'),\ \sigma_{\mathcal{W}_m} \|\mathbf{k}\|_\infty^{\frac{1}{2}} \sqrt{2\log(\frac{4}{\delta'})} + \sup_{x,x' \in \mathcal{B}(0,R)} |\mathcal{W}_m(x) - \mathcal{W}_m(x')| \right) \right\},$$

occurs with probability at least $1 - \frac{\delta'}{2}$*, i.e.,* $\Pr(\widetilde{\mathcal{E}}_\infty^{(m)} | \cap_{j=1}^m \mathcal{E}_{\text{half}}^{(j)}) \geq 1 - \frac{\delta'}{2}$.

Lemma 21 (A high probability bound on supremum of $\mathcal{W}_m$ differences)**.** *If* $\mathbf{k}$ *satisfies Assumps. 1 and 2, then, for any fixed* $R > 0, \delta' \in (0,1)$*, given* $\mathcal{S}_n$*, the event*

$$\mathcal{E}_{\sup}^{(m)} \triangleq \left\{ \sup_{x,x' \in \mathcal{B}(0,R)} |\mathcal{W}_m(x) - \mathcal{W}_m(x')| \leq 8 D_R \left(\sqrt{\log(\tfrac{4}{\delta'})} + 6\sqrt{d \log\big(2 + \tfrac{L_{\mathbf{k}} R}{\|\mathbf{k}\|_\infty}\big)} \right) \right\}$$

occurs with probability at least $1 - \delta'/2$*, where* $D_R \triangleq \sqrt{\frac{32}{3}} \sigma_{\mathcal{W}_m} \|\mathbf{k}\|_\infty^{\frac{1}{2}} \min\big(1, \sqrt{\frac{1}{2} L_{\mathbf{k}} R}\big)$.

Mimicking the arguments like those in display (56) and (57), with $R = R_{\mathcal{S}_n} + R_{\mathbf{k},n}$, and $R' = R_{\mathbf{k},n}$, we find that conditional on the event $\widetilde{\mathcal{E}}_\infty^{(m)} \cap \mathcal{E}_{\sup}^{(m)} \cap_{j=1}^m \mathcal{E}_{\text{half}}^{(j)}$,

$$\begin{aligned}
\|\mathbb{P}_n \mathbf{k} - \mathbb{Q}_{\text{KH}}^{(m)} \mathbf{k}\|_\infty &\overset{(61)}{=} \| \textstyle\sum_{j=1}^m \frac{2^{j-1}}{n} \psi_{n/2^j, j} \|_\infty \\
&\overset{(62)}{=} \|\mathcal{W}_m\|_\infty \\
&\leq \max\big(\tfrac{2^m}{n} \|\mathbf{k}\|_\infty,\ 32 \sigma_{\mathcal{W}_m} \|\mathbf{k}\|_\infty^{\frac{1}{2}} \Big(\sqrt{\log(\tfrac{4}{\delta'})} + 5\sqrt{d \log\Big(2 + \tfrac{L_{\mathbf{k}}(R_{\mathcal{S}_n} + R_{\mathbf{k},n})}{\|\mathbf{k}\|_\infty}\Big)} \Big)\big) \\
&\overset{(63)}{\leq} \|\mathbf{k}\|_\infty \cdot \tfrac{2^m}{n} \cdot 37 \sqrt{\log\big(\tfrac{6m}{2^m \delta^\star}\big)} \Big[\sqrt{\log(\tfrac{4}{\delta'})} + 5\sqrt{d \log\Big(2 + \tfrac{L_{\mathbf{k}}(R_{\mathcal{S}_n} + R_{\mathbf{k},n})}{\|\mathbf{k}\|_\infty}\Big)} \Big] \\
&\overset{(8)}{=} \|\mathbf{k}\|_\infty \cdot \tfrac{2^m}{n} \mathfrak{M}_{\mathbf{k}}(n, m, d, \delta^\star, \delta', R_{\mathcal{S}_n}),
\end{aligned}$$

which does not happed with probability at most

$$\begin{aligned}
\mathbb{P}((\cap_{j=1}^m \mathcal{E}_{\text{half}}^{(m)} \cap \widetilde{\mathcal{E}}_\infty^{(m)} \cap \mathcal{E}_{\sup}^{(m)})^c) &\leq \textstyle\sum_{j=1}^m \Pr((\mathcal{E}_{\text{half}}^{(j)})^c) + \Pr((\mathcal{E}_{\sup}^{(m)})^c) + \Pr((\widetilde{\mathcal{E}}_\infty^{(m)})^c | \cap_{j=1}^m (\mathcal{E}_{\text{half}}^{(j)})) \\
&\leq \textstyle\sum_{j=1}^m \sum_{i=1}^{\lfloor n/2^j \rfloor} \delta_i \frac{2^{j-1}}{m} + \frac{\delta'}{2} + \frac{\delta'}{2},
\end{aligned}$$

where the last step follows from (59) and Thms. 20 and 21 as claimed. The proof is now complete.

K.3 Proof of Thm. 15: A direct covering bound on $\|\psi_{n/2}\|_\infty$

We claim that conditional on the event $\mathcal{E}_{\text{half}}$ (45), we deterministically have

$$\|\psi_{n/2}\|_\infty \leq \max\big\{ n \overline{\tau}_{\mathbf{k}}(R),\ \ n L_{\mathbf{k}} r + \max_{z \in \mathcal{C}^n(r,R)} |\psi_{n/2}(z)| \big\}, \tag{65}$$

and the event

$$\big\{ \max_{z \in \mathcal{C}^n(r,R)} \big|\psi_{n/2}(z)\big| \leq \sigma_{n/2} \sqrt{2\|\mathbf{k}\|_\infty \log(2|\mathcal{C}^n(r,R)|/\delta')} \big\} \tag{66}$$

occurs with probability at least $1 - \delta'$. Putting these two claims together yields the lemma. We now prove these two claims separately.

K.3.1 Proof of (65)

Note that on the event $\mathcal{E}_{\text{half}}$ (45) we have

$$\psi_{n/2} = n(\mathbb{P}_n\mathbf{k} - \widetilde{\mathbb{Q}}_{n/2}\mathbf{k}) = \sum_{i=1}^n \eta_i \mathbf{k}(x_i, \cdot) \tag{67}$$

for some $\eta_i \in \{-1, 1\}$. Now fix any $x \in \mathbb{R}^d$, and introduce the shorthand $\mathcal{C}^n = \mathcal{C}^n(r, R)$. The result follows by considering two cases.

(Case 1) $\boldsymbol{x \notin \bigcup_{i=1}^n \mathcal{B}(x_i, R)}$ In this case, we have $\|x - x_i\|_2 \geq R$ for all $i \in [n]$ and therefore, representation (67) yields that

$$\left|\psi_{n/2}(x)\right| = |\sum_{i=1}^n \eta_i \mathbf{k}(x_i, x)| \leq \sum_{i=1}^n |\eta_i||\mathbf{k}(x_i, x)| = \sum_{i=1}^n |\mathbf{k}(x_i, x)| \leq n\overline{\tau}_{\mathbf{k}}(R),$$

by Cauchy-Schwarz's inequality and the definition (47) of $\overline{\tau}_{\mathbf{k}}$.

(Case 2) $\boldsymbol{x \in \bigcup_{i=1}^n \mathcal{B}(x_i, R)}$ By the definition (49) of our cover $\mathcal{C}^n$, there exists $z \in \mathcal{C}^n$ such that $\|x - z\|_2 \leq r$. Therefore, on the event $\mathcal{E}_{\text{half}}$, using representation (67), we find that

$$\begin{aligned}
\left|\psi_{n/2}(x)\right| \leq |\sum_{i=1}^n \eta_i \mathbf{k}(x_i, x)| &= |\sum_{i=1}^n \eta_i (\mathbf{k}(x_i, x) - \mathbf{k}(x_i, z) + \mathbf{k}(x_i, z))| \\
&\leq |\sum_{i=1}^n \eta_i (\mathbf{k}(x_i, x) - \mathbf{k}(x_i, z))| + |\sum_{i=1}^n \eta_i \mathbf{k}(x_i, z)| \\
&\leq \sum_{i=1}^n |\mathbf{k}(x_i, x) - \mathbf{k}(x_i, z)| + \left|\psi_{n/2}(z)\right| \\
&\leq nL_{\mathbf{k}}(r) + \sup_{z' \in \mathcal{C}^n} \left|\psi_{n/2}(z')\right|
\end{aligned}$$

by Cauchy-Schwarz's inequality.

K.3.2 Proof of (66)

Introduce the shorthand $\mathcal{C}^n = \mathcal{C}^n(r, R)$. Then applying the union bound, the pointwise sub-Gaussian property (iv) of Thm. 3, and the sub-Gaussian Hoeffding inequality (Wainwright, 2019, Prop. 2.5), we find that

$$\begin{aligned}
\Pr(\max_{z \in \mathcal{C}^n} |\psi_{n/2}(z)| > t\,|) &\leq \sum_{z \in \mathcal{C}^n} 2\exp(-t^2/(2\sigma_{n/2}^2 \mathbf{k}(z, z))) \\
&\leq 2|\mathcal{C}^n| \exp(-t^2/(2\sigma_{n/2}^2 \|\mathbf{k}\|_\infty)) = \delta' \\
\text{for} \quad t &\triangleq \sigma_{n/2}\sqrt{2\|\mathbf{k}\|_\infty \log(2|\mathcal{C}^n|/\delta')},
\end{aligned}$$

as claimed.

K.4 Proof of Thm. 16: A basic bound on $\|\psi_{n/2}\|_\infty$

For any $R > 0$, we deterministically have

$$\left\|\psi_{n/2}\right\|_\infty \leq \max(\sup_{x \in \mathcal{B}(0,R)} \left|\psi_{n/2}(x)\right|, \sup_{x \in \mathcal{B}^c(0,R)} \left|\psi_{n/2}(x)\right|) \tag{68}$$

Since $R = R' + R_{\mathcal{S}_n}$, for any $x \in \mathcal{B}^c(0, R)$, we have $\|x - x_i\|_2 \geq R'$ for all $i \in [n]$. Thus conditional on the event $\mathcal{E}_{\text{half}}$, applying property (ii) from Thm. 3, we find that

$$\left|\psi_{n/2}(x)\right| = |\sum_{i=1}^n \eta_i \mathbf{k}(x_i, x)| \leq \sum_{i=1}^n |\eta_i||\mathbf{k}(x_i, x)| = \sum_{i=1}^n |\mathbf{k}(x_i, x)| \leq n\overline{\tau}_{\mathbf{k}}(R'),$$

by Cauchy-Schwarz's inequality and the definition (47) of $\overline{\tau}_{\mathbf{k}}$.

For the first term on the RHS of (68), we have

$$\begin{aligned}\sup_{x\in\mathcal{B}(0,R)}\left|\psi_{n/2}(x)\right| &\leq \left|\psi_{n/2}(0)\right| + \sup_{x\in\mathcal{B}(0,R)}\left|\psi_{n/2}(x)-\psi_{n/2}(0)\right| \\ &\leq \left|\psi_{n/2}(0)\right| + \sup_{x,x'\in\mathcal{B}(0,R)}\left|\psi_{n/2}(x)-\psi_{n/2}(x')\right|.\end{aligned}$$

Now, the sub-Gaussianity of $\psi_{n/2}$ (property (iv) of Thm. 3) with $x=0$, and the sub-Gaussian Hoeffding inequality (Wainwright, 2019, Prop. 2.5) imply that

$$\left|\psi_{n/2}(0)\right| \leq \sigma_{n/2}\sqrt{2\mathbf{k}(0,0)\log(4/\delta)} \leq \sigma_{n/2}\sqrt{2\|\mathbf{k}\|_\infty \log(4/\delta')},$$

with probability at least $1-\delta'/2$. Putting the pieces together completes the proof.

K.5 Proof of Thm. 17: A high probability bound on supremum of $\psi_{n/2}$ differences

The proof proceeds by using concentration arguments for Orlicz processes (Wainwright, 2019, Def. 5.5). Given a set $\mathbb{T}\subseteq\mathbb{R}^d$, a random process $\{Z_x, x\in\mathbb{T}\}$ is called an Orlicz Ψ_2-process with respect to the metric ρ if

$$\|Z_x - Z_{x'}\|_{\Psi_2} \leq \rho(x,x') \quad \text{for all} \quad x,x'\in\mathbb{T},$$

where for any random variable Z, its Orlicz Ψ_2-norm is defined as

$$\|Z\|_{\Psi_2} = \inf\left\{\lambda>0 : \mathbb{E}[\exp(Z^2/\lambda^2)] \leq 2\right\}.$$

Our next result (see App. K.6 for the proof) establishes that $\psi_{n/2}$ is an Orlicz process with respect to a suitable metric. (For clarity, we use $\mathcal{B}_2$ to denote the Euclidean ball.)

Lemma 22 ($\psi_{n/2}$ is an Orlicz Ψ_2-process)**.** *If $\mathbf{k}$ satisfies Assumps. 1 and 2, then, given any fixed $R>0$, the random process $\left\{\psi_{n/2}(x), x\in\mathcal{B}_2(0;R)\right\}$ is an Orlicz Ψ_2-process with respect to the metric ρ defined in (70), i.e.,*

$$\left\|\psi_{n/2}(x)-\psi_{n/2}(x')\right\|_{\Psi_2} \leq \rho(x,x') \quad \textit{for all} \quad x,x'\in\mathcal{B}_2(0;R), \tag{69}$$

where the metric ρ is defined as

$$\rho(x,x') = \sqrt{\tfrac{8}{3}}\cdot\sigma_{n/2}\cdot\min\left(\sqrt{2L_{\mathbf{k}}\|x-x'\|_2}, 2\sqrt{\|\mathbf{k}\|_\infty}\right). \tag{70}$$

Given Thm. 22, we can invoke high probability bounds for Orlicz processes. For the Orlicz process $\left\{\psi_{n/2}(x), x\in\mathbb{T}\right\}$, given any fixed $\delta'\in(0,1]$, Wainwright (2019, Thm 5.36) implies that

$$\sup_{x,x'\in\mathbb{T}}\left|\psi_{n/2}(x)-\psi_{n/2}(x')\right| \leq 8\Big(\mathcal{J}_{\mathbb{T},\rho}(D) + D\sqrt{\log(4/\delta')}\Big), \tag{71}$$

with probability at least $1-\delta'/2$, where $D\triangleq\sup_{x,x'\in\mathbb{T}}\rho(x,x')$ denotes the ρ-diameter of $\mathbb{T}$, and the quantity $\mathcal{J}_{\mathbb{T},\rho}(D)$ is defined as

$$\mathcal{J}_{\mathbb{T},\rho}(D) \triangleq \int_0^D \sqrt{\log(1+\mathcal{N}_{\mathbb{T},\rho}(u))}du. \tag{72}$$

Here $\mathcal{N}_{\mathbb{T},\rho}(u)$ denotes the u-covering number of $\mathbb{T}$ with respect to metric ρ, namely the cardinality of the smallest cover $\mathcal{C}_{\mathbb{T},\rho}(u) \subseteq \mathbb{R}^d$ such that

$$\mathbb{T} \subseteq \cup_{z\in\mathcal{C}_{\mathbb{T},\rho}(u)} \mathcal{B}_\rho(z;u) \quad \text{where} \quad \mathcal{B}_\rho(z;u) \triangleq \{x \in \mathbb{R}^d : \rho(z,x) \le u\}. \tag{73}$$

To avoid confusion, let $\mathcal{B}_2$ denote the Euclidean ball (metric induced by $\|\cdot\|_2$). Then in our setting, we have $\mathbb{T} = \mathcal{B}_2(0;R)$ and hence

$$D_R \triangleq \sup_{x,x'\in\mathcal{B}_2(0;R)} \rho(x,x') = \min\Big(\sqrt{\tfrac{\beta_R}{2}}, \sigma_{n/2}\sqrt{\tfrac{32\|\mathbf{k}\|_\infty}{3}}\Big), \text{ with } \beta_R = \tfrac{32}{3}\sigma_{n/2}^2 L_{\mathbf{k}} R. \tag{74}$$

Some algebra establishes that this definition of D_R is identical to that specified in Thm. 17. Next, we derive bounds for $\mathcal{N}_{\mathbb{T},\rho}$ and $\mathcal{J}_{\mathbb{T},\rho}$ (see App. K.7 for the proof).

Lemma 23 (Bounds on $\mathcal{N}_{\mathbb{T},\rho}$ and $\mathcal{J}_{\mathbb{T},\rho}$)**.** *If $\mathbf{k}$ satisfies Assumps. 1 and 2, then, for $\mathbb{T} = \mathcal{B}_2(0;R)$ with $R > 0$ and ρ defined in* (70)*, we have*

$$\mathcal{N}_{\mathbb{T},\rho}(u) \le \left(1 + \beta_R/u^2\right)^d, \quad \textit{and} \tag{75}$$

$$\mathcal{J}_{\mathbb{T},\rho}(D_R) \le \sqrt{d} D_R\Big[3 + \sqrt{2\log(\beta_R/D_R^2)}\Big], \tag{76}$$

where D_R and β_R are defined in (74)*.*

Doing some algebra, we find that

$$\tfrac{\beta_R}{D_R^2} = \max\Big(2, \tfrac{L_{\mathbf{k}} R}{\|\mathbf{k}\|_\infty}\Big) \le 2 + \tfrac{L_{\mathbf{k}} R}{\|\mathbf{k}\|_\infty}. \tag{77}$$

Moreover, note that $3 + \sqrt{2\log(2+a)} \le 6\sqrt{\log(2+a)}$ for all $a \ge 0$, so that the bound (76) can be simplified to

$$\mathcal{J}_{\mathbb{T},\rho}(D_R) \le 6\sqrt{d} D_R \sqrt{\log(2 + \tfrac{L_{\mathbf{k}} R}{\|\mathbf{k}\|_\infty})}. \tag{78}$$

Substituting the bound (78) in (71) yields that the event $\mathcal{E}_{\sup}$ defined in (55) holds with probability at least $1 - \delta'/2$, as claimed.

K.6 Proof of Thm. 22: $\psi_{n/2}$ is an Orlicz Ψ_2-process

The proof of this lemma follows from the sub-Gaussianity of $\psi_{n/2}$ established in Thm. 3. Introduce the shorthand $Y \triangleq \psi_{n/2}(x) - \psi_{n/2}(x')$. Applying property (i) of Thm. 3 with $u = \mathbf{k}(x,\cdot) - \mathbf{k}(x',\cdot)$ along with the reproducing property of the kernel $\mathbf{k}$, we find that for any $t \in \mathbb{R}$,

$$\mathbb{E}[\exp(tY)] = \mathbb{E}[\exp(t\langle \psi_{n/2}, \mathbf{k}(x,\cdot) - \mathbf{k}(x',\cdot)\rangle_{\mathbf{k}_\star})] \le \exp(\tfrac{1}{2}t^2\sigma_{n/2}^2 \|\mathbf{k}(x,\cdot) - \mathbf{k}(x',\cdot)\|_{\mathbf{k}_\star}^2).$$

That is, the random variable Y is sub-Gaussian with parameter $\sigma_Y \triangleq \sigma_{n/2}\|\mathbf{k}(x,\cdot) - \mathbf{k}(x',\cdot)\|_{\mathbf{k}_\star}$. Next, Wainwright (2019, Thm 2.6 (iv)) yields that $\mathbb{E}[\exp(\frac{3Y^2}{8\sigma_Y^2})] \le 2$, which in turn implies that $\|Y\|_{\Psi_2} \le \sqrt{\frac{8}{3}}\sigma_Y$. Moreover, we have

$$\|\mathbf{k}(x,\cdot) - \mathbf{k}(x',\cdot)\|_{\mathbf{k}_\star}^2 = \mathbf{k}(x,x) - \mathbf{k}(x,x') + \mathbf{k}(x',x') - \mathbf{k}(x',x) \le \min(2L_{\mathbf{k}}\|x - x'\|_2, 4\|\mathbf{k}\|_\infty),$$

using the Lipschitz continuity of $\mathbf{k}$ and the definition of $\|\mathbf{k}\|_\infty$ (4). Putting the pieces together along with the definition (70) of ρ, we find that $\|Y\|_{\Psi_2} \le \rho(x,x')$ thereby yielding the claim (69).

K.7 Proof of Thm. 23: Bounds on $\mathcal{N}_{\mathbb{T},\rho}$ and $\mathcal{J}_{\mathbb{T},\rho}$

We use the relation of ρ to the Euclidean norm $\|\cdot\|_2$ to establish the bound (75). The definitions (70) and (73) imply that

$$\mathcal{B}_2(z; \tfrac{u^2}{\alpha}) \subseteq \mathcal{B}_\rho(z; u) \quad \text{for} \quad \alpha \triangleq \tfrac{16}{3}\sigma_{n/2}^2 L_{\mathbf{k}},$$

since $\rho(x, x') \leq \sqrt{\alpha\|x - x'\|_2}$. Consequently, any u^2/α-cover $\mathcal{C}$ of $\mathbb{T}$ in the Euclidean ($\|\cdot\|_2$) metric automatically yields a u-cover of $\mathbb{T}$ in ρ metric as we note that

$$\mathcal{B}_2(0; R) \subseteq \cup_{z\in\mathcal{C}} \mathcal{B}_2(z; \tfrac{u^2}{\alpha}) \subseteq \cup_{z\in\mathcal{C}} \mathcal{B}_\rho(z; u).$$

Consequently, the smallest cover $\mathcal{C}_{\mathbb{T},\rho}(u)$ would not be larger than than the smallest cover $\mathcal{C}_{\mathbb{T},\|\cdot\|_2}(u^2/\alpha)$, or equivalently:

$$\mathcal{N}_{\mathbb{T},\rho}(u) \leq \mathcal{N}_{\mathbb{T},\|\cdot\|_2}(\tfrac{u^2}{\alpha})) \stackrel{(i)}{\leq} \left(1 + 2R\alpha/u^2\right)^d,$$

where inequality (i) follows from Wainwright (2019, Lem 5.7) using the fact that $\mathbb{T} = \mathcal{B}_2(0; R)$. Noting that $\beta_R = 2R\alpha$ yields the bound (75).

We now use the bound (75) on the covering number to establish the bound (76).

Proof of bound on $\mathcal{J}_{\mathbb{T},\rho}$ Applying the definition (72) and the bound (75), we find that

$$\begin{aligned}\mathcal{J}_{\mathbb{T},\rho}(D_R) &\leq \textstyle\int_0^{D_R} \sqrt{\log(1 + (1 + \beta_R/u^2)^d)}du \\ &\stackrel{(i)}{=} \sqrt{\tfrac{\beta_R}{2}} \textstyle\int_{\beta_R/D_R^2}^\infty s^{-3/2}\sqrt{\log(1 + (1+s)^d)}ds, \end{aligned} \tag{79}$$

where step (i) follows from a change of variable $s \leftarrow \beta_R/u^2$. Note that $\log(1 + a) \leq \log a + 1/a$, and $\sqrt{a + b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$. Applying these inequalities, we find that

$$\log(1 + (1 + s)^d) \leq d\log(2 + s) \leq d(\log s + \tfrac{1}{s} + \tfrac{1}{1+s}) \leq d(\log s + \tfrac{2}{s}),$$

for $s \in [\beta_R/D_R^2, \infty) \stackrel{(77)}{\subseteq} [2, \infty)$. Consequently,

$$\textstyle\int_{\beta_R/D_R^2}^\infty s^{-3/2}\sqrt{\log(1 + (1 + s)^d)}ds \leq \sqrt{d} \int_{\beta_R/D_R^2}^\infty \Big(\frac{\sqrt{\log s}}{s^{3/2}} + \frac{\sqrt{2}}{s^2}\Big)ds. \tag{80}$$

Next, we note that

$$\textstyle\int_{\beta_R/D_R^2}^\infty \frac{\sqrt{2}}{s^2}ds = \sqrt{2}\frac{D_R^2}{\beta_R}, \quad \text{and} \tag{81}$$

$$\begin{aligned}\textstyle\int_{\beta_R/D_R^2}^\infty \frac{\sqrt{\log s}}{s^{3/2}}ds &= -2\sqrt{\tfrac{\log s}{s}} - \sqrt{2}\Gamma(\tfrac{1}{2}, \tfrac{1}{2}\log s)\Big|_{s=\beta_R/D_R^2}^{s=\infty} \\ &\stackrel{(i)}{\leq} \tfrac{2D_R}{\sqrt{\beta_R}}\bigg(\sqrt{\log(\beta_R/D_R^2)} + \frac{1}{\sqrt{\log(\beta_R/D_R^2)}}\bigg),\end{aligned} \tag{82}$$

where step (i) follows from the following bound on the incomplete Gamma function:

$$\Gamma(\tfrac{1}{2}, a) = \textstyle\int_a^\infty \frac{1}{\sqrt{t}}e^{-t}dt \leq \frac{1}{\sqrt{a}}\int_a^\infty e^{-t}dt = a^{-\frac{1}{2}}e^{-a} \quad \text{for any} \quad a > 0.$$

Putting the bounds (79) to (82) together, we find that

$$\mathcal{J}_{\mathbb{T},\rho}(D_R) \leq \sqrt{d}D_R\Big(\tfrac{D_R}{\sqrt{\beta_R}} + \sqrt{2}\Big(\sqrt{\log(\beta_R/D_R^2)} + 1/\sqrt{\log(\beta_R/D_R^2)}\Big)\Big).$$

Note that by definition (74) $\frac{D_R}{\sqrt{\beta_R}} \leq \frac{1}{\sqrt{2}}$. Using this observation and the fact that $\frac{1}{\sqrt{2}} + \sqrt{\frac{2}{\log 2}} \leq 3$ yields the claimed bound (76).

Appendix L. Proof of Rem. 8: Example settings for KH L^∞ rates

Note that when the kernel is restricted to a compact domain, we have $R_{\mathbf{k},n} \leq \mathcal{O}(\sigma)$ so that in this case, $\frac{L_{\mathbf{k}}(R_{\mathbf{k},n} \wedge R_{\mathcal{S}_n})}{\|\mathbf{k}\|_\infty} \leq \mathcal{O}(L)$. Next, we establish the bounds on L and $\kappa^\dagger(1/n)$ for each of the kernels, which immediately imply the respective claims both when supported on $\mathbb{R}^d$ and on a compact domain with $\sigma = \sqrt{d}$.

- For a Gaussian kernel, we have $\kappa(r) = e^{-r^2/2}$, which when put together with (98) and (99) implies that $L = \sqrt{2/e}$, and $\kappa^\dagger(1/n) = \sqrt{2\log n}$.

- For a Matérn kernel, we have $\kappa(r) = c_b r^b K_b(r)$ for $b = \nu - \frac{d}{2}$, for suitable constant c_b, and function K_b; see Tab. 1. Moreover, (107) and (108) imply that $L = \frac{C_1}{\sqrt{C_2+b}} \leq \frac{C_1}{\sqrt{C_2+1}}$ for universal constants C_1, C_2, and $\kappa^\dagger(1/n) \leq \mathcal{O}(\max(b\log b, \log n)) = \mathcal{O}(\log n)$.

- For a IMQ kernels, we have $\kappa(r) = \frac{1}{(\gamma^2+r^2)^\nu}$, so that

  $$L \leq \sup_r \kappa'(r) = \sup_r \tfrac{2\nu r}{(\gamma^2+r^2)^{\nu+1}} = \tfrac{2\nu\gamma}{\gamma^{2(\nu+1)}} \tfrac{1}{\sqrt{1+2\nu}} (\tfrac{2\nu+1}{2\nu+2})^{\nu+1},$$

  and $\kappa^\dagger(1/n) = \sqrt{n^{1/\nu} - \gamma^2} \leq n^{1/2\nu}$.

Appendix M. Proof of Cor. 5: L^∞ and MMD guarantees for KT-SPLIT

First applying Thm. 4(b) with $\mathbf{k} = \mathbf{k}_{\mathrm{rt}}$ yields that

$$\|\mathbb{P}_n\mathbf{k}_{\mathrm{rt}} - \mathbb{Q}_{\mathrm{KT}}^{(m,1)}\mathbf{k}_{\mathrm{rt}}\|_\infty \leq \tfrac{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{n_{\mathrm{out}}} \cdot \mathfrak{M}_{\mathbf{k}_{\mathrm{rt}}}(n, m, d, \delta^\star, \delta', R_{\mathcal{S}_n,\mathbf{k}_{\mathrm{rt}},n}) \tag{83}$$

with probability at least $1 - \delta' - \sum_{j=1}^m \frac{2^{j-1}}{m} \sum_{i=1}^{2^{m-j}\lfloor n/2^m \rfloor} \delta_i$. Call this event $\mathcal{E}$.

Let $r' = R'_{\mathcal{S}_n,\mathbf{k}_{\mathrm{rt}},n_{\mathrm{out}}} + \varepsilon$ denote a shorthand notation, where $\varepsilon > 0$ is an arbitrary scalar. Applying Thm. 2 with $r = 2r'$ and $a = b = \frac{1}{2}$, we find that on event $\mathcal{E}$,

$$\begin{aligned}
&\mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S}^{(m,1)}) \\
&= \mathrm{MMD}_{\mathbf{k}_\star}(\mathbb{P}_n, \mathbb{Q}_{\mathrm{KT}}^{(m,1)}) \\
&\leq \|\mathbb{P}_n\mathbf{k}_{\mathrm{rt}} - \mathbb{Q}_{\mathrm{KT}}^{(m,1)}\mathbf{k}_{\mathrm{rt}}\|_\infty e_d r^{d/2} + 2\tau_{\mathbf{k}_{\mathrm{rt}}}(r') + 2\|\mathbf{k}_\star\|_\infty^{\frac{1}{2}} \cdot \max(\tau_{\mathbb{P}_n}(r'), \tau_{\mathbb{Q}_{\mathrm{KT}}^{(m,1)}}(r')) \\
&\overset{(i)}{=} \|\mathbb{P}_n\mathbf{k}_{\mathrm{rt}} - \mathbb{Q}_{\mathrm{KT}}^{(m,1)}\mathbf{k}_{\mathrm{rt}}\|_\infty \cdot \sqrt{\tfrac{(4\pi)^{d/2}}{\Gamma(\frac{d}{2}+1)}}(r')^{\frac{d}{2}} + \tfrac{2\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{n_{\mathrm{out}}}, \\
&\overset{(83)}{\leq} \tfrac{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{n_{\mathrm{out}}} \cdot \mathfrak{M}_{\mathbf{k}_{\mathrm{rt}}}(n, m, d, \delta^\star, \delta', R_{\mathcal{S}_n,\mathbf{k}_{\mathrm{rt}},n}) \cdot \sqrt{\tfrac{(4\pi)^{d/2}}{\Gamma(\frac{d}{2}+1)}}(r')^{\frac{d}{2}} + \tfrac{2\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{n_{\mathrm{out}}} \\
&= \tfrac{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{n_{\mathrm{out}}}\left[2 + \sqrt{\tfrac{(4\pi)^{d/2}}{\Gamma(\frac{d}{2}+1)}}(r')^{\frac{d}{2}}\mathfrak{M}_{\mathbf{k}_{\mathrm{rt}}}(n, m, d, \delta^\star, \delta', R_{\mathcal{S}_n,\mathbf{k}_{\mathrm{rt}},n})\right] \qquad (84)
\end{aligned}$$

where step (i) uses the following arguments: (a) $e_d r^{d/2} = \sqrt{\frac{(4\pi)^{d/2}}{\Gamma(\frac{d}{2}+1)}}(r')^{\frac{d}{2}}$; (b) $\mathbb{Q}_{\mathrm{KT}}^{(m,1)}$ is supported on a subset of points from $\mathcal{S}_n$ and hence

$$\max\{\tau_{\mathbb{P}_n}(r'), \tau_{\mathbb{Q}_{\mathrm{KT}}^{(m,1)}}(r')\} = \tau_{\mathbb{P}_n}(r') = 0 \quad \text{for any} \quad r' > R_{\mathcal{S}_n};$$

and (c) $\tau_{\mathbf{k}_{\mathrm{rt}}}(r') \leq \frac{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{n_{\mathrm{out}}}$ for any $r' > R'_{\mathbf{k}_{\mathrm{rt}},n_{\mathrm{out}}}$. Noting that (84) applies simultaneously for all $\varepsilon > 0$ under event $\mathcal{E}$, taking the limit $\varepsilon \to 0$ yields that

$$\mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S}^{(m,1)}) \leq \tfrac{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{n_{\mathrm{out}}}\left[2 + \sqrt{\tfrac{(4\pi)^{d/2}}{\Gamma(\frac{d}{2}+1)}}(R'_{\mathcal{S}_n,\mathbf{k}_{\mathrm{rt}},n_{\mathrm{out}}})^{\frac{d}{2}}\mathfrak{M}_{\mathbf{k}_{\mathrm{rt}}}(n, m, d, \delta^\star, \delta', R_{\mathcal{S}_n,\mathbf{k}_{\mathrm{rt}},n})\right] \qquad (85)$$

with probability at least $1-\delta'-\sum_{j=1}^{m}\frac{2^{j-1}}{m}\sum_{i=1}^{2^{m-j}\lfloor n/2^m\rfloor}\delta_i$, as claimed.

Appendix N. Derivation of Tab. 1: Square-root kernels $\mathbf{k}_{\mathrm{rt}}$ for common target kernels $\mathbf{k}_\star$

In this appendix, we derive the results stated in Tab. 1.

N.1 General proof strategy

Let $\mathcal{F}$ denote the Fourier operator (19). In Tab. 3, we state the continuous κ such that $\mathbf{k}_\star(x, y) = \kappa(x - y)$ in the first column, its Fourier transform (19) $\widehat{\kappa}$ in the second column, the square-root Fourier transform in the third column, and the square-root kernel in the fourth column, given by $\mathbf{k}_{\mathrm{rt}}(x, y) = \frac{1}{(2\pi)^{d/4}}\kappa_{\mathrm{rt}}(x - y)$ with $\kappa_{\mathrm{rt}} = \mathcal{F}(\sqrt{\widehat{\kappa}})$. Prop. 3 along with expressions in Tab. 3 directly establishes the validity of the square-root kernels for the Gaussian and (scaled) B-spline kernels. For completeness, we also illustrate the remaining calculus for the B-spline kernels in App. N.3. We do a similar calculation in App. N.2 for the Matérn kernel for better exposition of the involved expressions.

Expression for $\kappa(z)$	Expression for Fourier transform of κ: $\widehat{\kappa}(\omega)$	Square-root Fourier transform: $\sqrt{\widehat{\kappa}(\omega)}$	Expression for $\kappa_{\mathrm{rt}}(z) \triangleq (\frac{1}{2\pi})^{\frac{d}{4}} \mathcal{F}(\sqrt{\widehat{\kappa}})(z)$
$\exp\left(-\frac{\Vert z\Vert_2^2}{2\sigma^2}\right)$	$\sigma^d \exp\left(-\frac{\sigma^2\Vert\omega\Vert_2^2}{2}\right)$	$\sigma^{\frac{d}{2}} \exp\left(-\frac{\sigma^2\Vert\omega\Vert_2^2}{4}\right)$	$\left(\frac{2}{\pi\sigma^2}\right)^{\frac{d}{4}} \exp\left(-\frac{\Vert z\Vert_2^2}{\sigma^2}\right)$
$\prod_{j=1}^d \circledast^{2\beta+2}\mathbb{I}_{[-\frac{1}{2},\frac{1}{2}]}(z_j)$	$\left(\frac{4^{\beta+1}}{\sqrt{2\pi}}\right)^d \prod_{j=1}^d \frac{\sin^{2\beta+2}(\omega_j/2)}{\omega_j^{2\beta+2}}$	$\left(\frac{4^{\beta+1}}{\sqrt{2\pi}}\right)^{\frac{d}{2}} \prod_{j=1}^d \frac{\sin^{\beta+1}(\omega_j/2)}{\omega_j^{\beta+1}}$	$\prod_{j=1}^d \circledast^{\beta+1}\mathbb{I}_{[-\frac{1}{2},\frac{1}{2}]}(z_j)$

Table 3: Fourier transforms of kernels $\mathbf{k}_\star(x,y) = \kappa(x-y)$ and square-root kernels $\mathbf{k}_{\mathrm{rt}}(x,y) = \kappa_{\mathrm{rt}}(x-y)$ from Tab. 1. Here $\mathcal{F}$ denotes the Fourier operator (19), and $\circledast^\ell$ denotes the convolution operator applied $\ell-1$ times, with the convention $\circledast^1 f \triangleq f$ and $\circledast^2 f = f \circledast f$ for $(f \circledast g)(x) = \int f(y)g(x-y)dy$. Each Fourier transform is derived from Sriperumbudur et al. (2010, Tab. 2).

N.2 Deriving $\mathbf{k}_{\mathrm{rt}}$ for the Matérn kernel

For $\mathbf{k}_\star = \textbf{Matérn}(\nu,\gamma)$ from Tab. 1, we have

$$\mathbf{k}_\star(x,y) = \phi_{d,\nu,\gamma} \cdot \Phi_{\nu,\gamma}(x-y) \quad \text{where} \quad \Phi_{\nu,\gamma}(z) \triangleq c_\nu \left(\frac{\Vert z\Vert_2}{\gamma}\right)^{\nu-\frac{d}{2}} K_{\nu-\frac{d}{2}}(\gamma\Vert z\Vert_2), \tag{86}$$
$$\phi_{d,\nu,\gamma} \triangleq \frac{c_{\nu-d/2}}{c_\nu}\gamma^{2\nu-d} \quad \text{and} \quad c_\nu \triangleq \frac{2^{1-\nu}}{\Gamma(\nu)},$$

and K_a denotes the modified Bessel function of third kind of order a (Wendland, 2004, Def. 5.10). That is, for Matérn kernel with $\mathbf{k}_\star(x,y) = \kappa(x-y)$, the function κ is given by $\kappa = \phi_{d,\nu,\gamma}\Phi_{\nu,\gamma}$. Now applying (Wendland, 2004, Thm. is 8.15), we find that

$$\mathcal{F}(\Phi_{\nu,\gamma}) = \frac{1}{(\gamma^2+\Vert\omega\Vert_2^2)^\nu} \quad \Longrightarrow \quad \mathcal{F}(\sqrt{\mathcal{F}(\Phi_{\nu,\gamma})}) = \frac{1}{(2\pi)^{d/4}}\Phi_{\nu/2,\gamma},$$

where in the last step we also use the facts that $\Phi_{\nu,\gamma}$ is an even function and $\sqrt{\mathcal{F}(\Phi_{\nu,\gamma})} \in L^1$ for all $\nu > d/2$. Thus, by Prop. 3, a valid square-root kernel $\mathbf{k}_{\mathrm{rt}}(x,y) = \kappa_{\mathrm{rt}}(x-y)$ is defined via

$$\kappa_{\mathrm{rt}} = \sqrt{\phi_{d,\nu,\gamma}}\frac{1}{(2\pi)^{d/4}}\Phi_{\nu/2,\gamma}, \quad \Longrightarrow \quad \mathbf{k}_{\mathrm{rt}} = A_{\nu,\gamma,d} \cdot \textbf{Matérn}(\nu/2,\gamma), \tag{87}$$
$$\text{where} \quad A_{\nu,\gamma,d} \triangleq \left(\frac{1}{4\pi}\gamma^2\right)^{d/4}\sqrt{\frac{\Gamma(\nu)}{\Gamma(\nu-d/2)}} \cdot \frac{\Gamma((\nu-d)/2)}{\Gamma(\nu/2)}. \tag{88}$$

N.3 Deriving $\mathbf{k}_{\mathrm{rt}}$ for the B-Spline kernel

For positive integers β, d, define the constants

$$\mathfrak{B}_\beta \triangleq \frac{1}{(\beta-1)!}\sum_{j=0}^{\lfloor\beta/2\rfloor}(-1)^j\binom{\beta}{j}(\tfrac{\beta}{2}-j)^{\beta-1}, \qquad S_{\beta,d} \triangleq \mathfrak{B}_\beta^{-d}, \quad \text{and} \quad \widetilde{S}_{\beta,d} \triangleq \frac{\sqrt{S_{2\beta+2,d}}}{S_{\beta+1,d}}. \tag{89}$$

Define the function $\chi_\beta : \mathbb{R}^d \to \mathbb{R}$ as follows:

$$\chi_\beta(z) \triangleq S_{\beta,d}\prod_{i=1}^d f_\beta(z_i). \tag{90}$$

Then for kernel $\mathbf{k}_\star = \textbf{B-spline}(2\beta+1)$, we have

$$\mathbf{k}_\star(x,y) = \chi_{2\beta+2}(x-y). \tag{91}$$

The second row of Tab. 3 implies that

$$(\tfrac{1}{2\pi})^{\frac{d}{4}}\mathcal{F}\big(\sqrt{\mathcal{F}(\chi_{2\beta+2})}\big) = \frac{\sqrt{S_{2\beta+2,d}}}{S_{\beta+1,d}}\cdot \chi_{\beta+1} \overset{(89)}{=} \widetilde{S}_{\beta,d}\cdot\chi_{\beta+1},$$

where we also use the fact that χ_β is an even function. Putting the pieces together with Prop. 3, we conclude that

$$\mathbf{k}_{\mathrm{rt}} = \widetilde{S}_{\beta,d}\cdot \textbf{B-spline}(\beta), \tag{92}$$

(with $\kappa_{\mathrm{rt}} = \widetilde{S}_{\beta,d}\chi_{\beta+1}$) is a valid square-root kernel for $\mathbf{k}_\star = \textbf{B-spline}(2\beta+1)$.

Appendix O. Derivation of Tab. 2: Explicit bounds on Thm. 1 quantities for common kernels

We start by collecting some common tools in App. O.1 that we later use for our derivations for the Gaussian, Matérn and B-spline kernels in Apps. O.2 to O.4 respectively. Finally, in App. O.5, we put the pieces together to derive explicit MMD rates, as a function of d, n, δ and kernel parameters for the three kernels, and summarize the rates in Tab. 4. (This table is complementary to the generic results stated in Cor. 1.)

O.1 Common tools for our derivations

We collect some simplified expressions, and techniques that come handy in our derivations to follow.

Simplified bounds on $\mathfrak{M}_{\mathbf{k}_{\mathrm{rt}}}$ From (8), we have

$$\begin{aligned}
&\mathfrak{M}_{\mathbf{k}_{\mathrm{rt}}}(n, \tfrac{1}{2}\log_2 n, d, \tfrac{\delta}{n}, \delta', R) \precsim \sqrt{\log\tfrac{n}{\delta}\cdot\Big[\log(\tfrac{1}{\delta'}) + d\log\Big(\tfrac{L_{\mathbf{k}_{\mathrm{rt}}}}{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}\cdot(R_{\mathbf{k}_{\mathrm{rt}},n}+R)\Big)\Big]}\\
\Longrightarrow\quad &\mathfrak{M}_{\mathbf{k}_{\mathrm{rt}}}(n, \tfrac{1}{2}\log_2 n, d, \tfrac{\delta}{n}, \delta, R) \precsim_\delta \sqrt{d\log n\cdot\log\Big(\tfrac{L_{\mathbf{k}_{\mathrm{rt}}}}{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}\cdot(R_{\mathbf{k}_{\mathrm{rt}},n}+R)\Big)}.
\end{aligned} \tag{93}$$

Thus, given (93), to get a bound on $\mathfrak{M}_{\mathbf{k}_{\mathrm{rt}}}$, we need to derive bounds on $(L_{\mathbf{k}_{\mathrm{rt}}}, \|\mathbf{k}_{\mathrm{rt}}\|_\infty, R_{\mathbf{k}_{\mathrm{rt}},n})$ for various kernels to obtain the desired guarantees on MMD and L^∞ coresets.

Bounds on Gamma function Our proofs make use of the bounds from Batir (2017, Thm 2.2) on the Gamma function:

$$\Gamma(b+1) \geq (b/e)^b\sqrt{2\pi b} \text{ for any } b\geq 1, \text{ and } \Gamma(b+1)\leq (b/e)^b\sqrt{e^2 b} \text{ for any } b \geq 1.1. \tag{94}$$

General tools for bounding the Lipschitz constant To bound the Lipschitz constant $L_{\mathbf{k}_{\mathrm{rt}}}$, the following two observations come in handy. For a radial kernel $\mathbf{k}_{\mathrm{rt}}(x,y) = \widetilde{\kappa}_{\mathrm{rt}}(\|x-y\|_2)$ with $\widetilde{\kappa}_{\mathrm{rt}}:\mathbb{R}_+\to\mathbb{R}$, we note

$$\begin{aligned}
\sup_{\substack{x,y,z:\\ \|y-z\|_2\leq r}} |\mathbf{k}_{\mathrm{rt}}(x,y)-\mathbf{k}_{\mathrm{rt}}(x,z)| &\leq \sup\nolimits_{a>0,b\leq r}|\widetilde{\kappa}_{\mathrm{rt}}(a)-\widetilde{\kappa}_{\mathrm{rt}}(a+b)| \leq \|\widetilde{\kappa}'_{\mathrm{rt}}\|_\infty r\\
\Longrightarrow\quad L_{\mathbf{k}_{\mathrm{rt}}} &\leq \|\widetilde{\kappa}'_{\mathrm{rt}}\|_\infty.
\end{aligned} \tag{95}$$

For a translation-invariant kernel $\mathbf{k}_{\mathrm{rt}}(x, y) = \kappa_{\mathrm{rt}}(x-y)$ with $\kappa_{\mathrm{rt}} : \mathbb{R}^d \to \mathbb{R}$, we use the bound

$$\begin{aligned}
\sup_{\substack{x,y,z: \\ \|y-z\|_2 \leq r}} |\mathbf{k}_{\mathrm{rt}}(x, y) - \mathbf{k}_{\mathrm{rt}}(x, z)| &= \sup_{z', z'', \|z-z'\|_2 \leq r} |\kappa_{\mathrm{rt}}(z') - \kappa_{\mathrm{rt}}(z'')| \\
&\leq \sup_{z' \in \mathbb{R}^d} \|\nabla \kappa_{\mathrm{rt}}(z')\|_2 r, \\
\implies \quad L_{\mathbf{k}_{\mathrm{rt}}} &\leq \sup_{z' \in \mathbb{R}^d} \|\nabla \kappa_{\mathrm{rt}}(z')\|_2 \stackrel{(i)}{\leq} \sqrt{d} \sup_{z' \in \mathbb{R}^d} \|\nabla \kappa_{\mathrm{rt}}(z')\|_\infty, \quad (96)
\end{aligned}$$

where step (i) follows from Cauchy-Schwarz's inequality (and is handy when coordinate wise control on $\nabla \kappa_{\mathrm{rt}}$ is easier to derive). We later apply the inequality (95) for the Gaussian and Matérn kernels, and (96) for the B-spline kernel.

O.2 Proofs for Gaussian kernel

For the kernel $\mathbf{k}_\star = \mathbf{Gauss}(\sigma)$ and its square-root kernel $\mathbf{k}_{\mathrm{rt}} = \left(\frac{2}{\pi\sigma^2}\right)^{\frac{d}{4}} \mathbf{Gauss}(\frac{\sigma}{\sqrt{2}})$, we claim the following bounds on various quantities

$$\begin{aligned}
\|\mathbf{k}_{\mathrm{rt}}\|_\infty &= \left(\tfrac{2}{\pi\sigma^2}\right)^{\frac{d}{4}} \quad \text{and} \quad \|\mathbf{k}_\star\|_\infty = 1, & (97) \\
L_{\mathbf{k}_{\mathrm{rt}}} &\leq \left(\tfrac{2}{\pi\sigma^2}\right)^{\frac{d}{4}} \tfrac{\sqrt{2/e}}{\sigma} \quad \implies \quad L_{\mathbf{k}_{\mathrm{rt}}} / \|\mathbf{k}_{\mathrm{rt}}\|_\infty \stackrel{(97)}{=} \tfrac{1}{\sigma}\sqrt{2/e} & (98) \\
R_{\mathbf{k}_{\mathrm{rt}}, n} &= \sigma\sqrt{\log n}, & (99) \\
R'_{\mathbf{k}_{\mathrm{rt}}, \sqrt{n}} &= \mathcal{O}(\sigma\sqrt{d + \log n}). & (100)
\end{aligned}$$

Substituting these expressions in (93), we find that

$$\mathfrak{M}^{\mathbf{Gauss}}_{\mathbf{k}_{\mathrm{rt}}}(n, \tfrac{1}{2}\log_2 n, d, \tfrac{\delta}{2n}, \delta', R) \precsim \sqrt{\log(\tfrac{n}{\delta})\left[\log\tfrac{1}{\delta'} + d\log\left(\sqrt{\log n} + \tfrac{R}{\sigma}\right)\right]},$$

as claimed in Tab. 2.

Proof of claims (97) to (100) The claim (97) follows directly from the definition of $\mathbf{k}_{\mathrm{rt}}$. The bound (98) on $L_{\mathbf{k}_{\mathrm{rt}}}$ follows from the fact $\|\frac{d}{dr} e^{-r^2/\sigma^2}\|_\infty = \frac{\sqrt{2/e}}{\sigma}$ and invoking the relation (95). Next, recalling the definition (6) and noting that

$$\sup_{\substack{x,y: \\ \|x-y\|_2 \geq r}} |\mathbf{k}_{\mathrm{rt}}(x, y)| \leq \left(\tfrac{2}{\pi\sigma^2}\right)^{\frac{d}{4}} e^{-r^2/\sigma^2}$$

implies the bound (99) on $R_{\mathbf{k}_{\mathrm{rt}}, n}$.

Next, we have

$$\begin{aligned}
\tau^2_{\mathbf{k}_{\mathrm{rt}}}(R) &= \int_{\|z\|_2 \geq R} \left(\tfrac{2}{\pi\sigma^2}\right)^{\frac{d}{2}} \exp(-2\|z\|_2^2/\sigma^2) dz \\
&= \Pr_{X \sim \mathcal{N}(0, \sigma^2/4 \cdot \mathbf{I}_d)}(\|X\|_2 \geq R) \stackrel{(i)}{\leq} e^{-R^2/\sigma^2}, \quad (101)
\end{aligned}$$

for $R \geq \sigma\sqrt{2d}$, where step (i) follows from the standard tail bound for a chi-squared random variable Y with k degree of freedom (Laurent and Massart, 2000, Lem. 1): $\Pr(Y - k > 2\sqrt{kt} + 2t) \leq e^{-t}$, wherein we substitute $k = d$, $Y = \frac{4}{\sigma^2}\|X\|_2^2$, $t = d\alpha$ with $\alpha \geq 2$, and $R^2 = \sigma^2 t$. Using the tail bound (101), the bound (97) on $\|\mathbf{k}_{\mathrm{rt}}\|_\infty$ and the definition (6) of $R'_{\mathbf{k}_{\mathrm{rt}}, \sqrt{n}}$, we find that

$$R'_{\mathbf{k}_{\mathrm{rt}}, \sqrt{n}} = \mathcal{O}(\sigma \max(\sqrt{(\log n - \tfrac{d}{2}\log(\sigma\sqrt{\tfrac{\pi}{2}}))_+}, \sqrt{d})) = \mathcal{O}(\sigma\sqrt{d + \log n}).$$

yielding the claim (100).

O.3 Proofs for Matérn kernel

Recall the notations from (86) to (88) for the Matérn kernel related quantities. Let K_b denote the modified Bessel function of the third kind with order b (Wendland, 2004, Def. 5.10). Let $a \triangleq \frac{\nu-d}{2}$. Then, the kernel $\mathbf{k}_\star = \mathbf{Mat\acute{e}rn}(\nu,\gamma)$ and its square-root kernel $\mathbf{k}_{\mathrm{rt}} = A_{\nu,\gamma,d}\cdot\mathbf{Mat\acute{e}rn}(a,\gamma)$ satisfy

$$\mathbf{k}_\star(x,y) = \widetilde{\kappa}_{\nu-d/2}(\gamma\|x-y\|_2), \quad \text{and} \tag{102}$$

$$\mathbf{k}_{\mathrm{rt}}(x,y) = A_{\nu,\gamma,d}\widetilde{\kappa}_a(\gamma\|x-y\|_2), \tag{103}$$

$$\text{where } \widetilde{\kappa}_b(r) \triangleq c_b r^b K_b(r), \text{ and } c_b = \tfrac{2^{1-b}}{\Gamma(b)} \text{ for } b>0. \tag{104}$$

We claim the following bounds on various quantities assuming $a \geq 2.2$, and $d \geq 2$:[7]

$$\|\mathbf{k}_{\mathrm{rt}}\|_\infty = A_{\nu,\gamma,d} \quad \text{and} \quad \|\mathbf{k}_\star\|_\infty = 1, \tag{105}$$

$$A_{\nu,\gamma,d} \leq 5\nu\big(\tfrac{\gamma^2}{2\pi(a-1)}\big)^{\frac{d}{4}}, \tag{106}$$

$$L_{\mathbf{k}_{\mathrm{rt}}} \leq A_{\nu,\gamma,d}\tfrac{C_1\gamma}{\sqrt{a+C_2}} \quad \Longrightarrow \quad \tfrac{L_{\mathbf{k}_{\mathrm{rt}}}}{\|\mathbf{k}_{\mathrm{rt}}\|_\infty} \stackrel{(105)}{\leq} \tfrac{C_1\gamma}{\sqrt{a+C_2}}, \tag{107}$$

$$R_{\mathbf{k}_{\mathrm{rt}},n} = \tfrac{1}{\gamma}\mathcal{O}(\max(\log n - a\log(1+a), C_2 a\log(1+a))), \tag{108}$$

$$R'_{\mathbf{k}_{\mathrm{rt}},\sqrt{n}} = \tfrac{1}{\gamma}\cdot\mathcal{O}(a+\log n+d\log(\tfrac{\sqrt{2e\pi}}{\gamma})+\log(\tfrac{(\nu-2)^{\nu-\frac{3}{2}}}{(2(a-1))^{2a-1}d^{\frac{d}{2}+1}})) \tag{109}$$

We prove these claims in App. O.3.1 through App. O.3.5. Putting these bounds together with (93) yields that

$$\begin{aligned}\mathfrak{M}^{\mathbf{Mat\acute{e}rn}}_{\mathbf{k}_{\mathrm{rt}}}(n, \tfrac{1}{2}\log_2 n, d, \tfrac{\delta}{2n}, \delta', R) &\precsim \begin{cases} \sqrt{\log(\frac{n}{\delta})\Big[\log\frac{1}{\delta} + d\log\Big(\frac{1}{\sqrt{1+a}}\cdot(\log n+\gamma R)\Big)\Big]} & \text{if } a = o(\log n) \\ \sqrt{\log(\frac{n}{\delta})\big[\log\frac{1}{\delta'} + d\log(\sqrt{a}\log(1+a)+\gamma R)\big]} & \text{if } a = \Omega(\log n)\end{cases} \\ &\precsim \sqrt{\log(\tfrac{n}{\delta})\big[\log\tfrac{1}{\delta'} + d\log(\log n + B + \gamma R)\big]}\end{aligned}$$

with $B = a\log(1+a)$, as claimed in Tab. 2.

O.3.1 Set-up for proofs of Matérn kernel quantities

Before proceeding to the proofs of the claims (105) to (109), we collect some handy inequalities. Applying Wendland (2004, Lem. 5.13, 5.14), we have

$$\widetilde{\kappa}_a(\gamma r) \leq \min\Big(1, \sqrt{2\pi}c_a(\gamma r)^{a-\frac{1}{2}}e^{-\gamma r+\frac{a^2}{2\gamma r}}\Big) \quad \text{for} \quad r>0. \tag{110}$$

For a large enough r, we also establish the following bound:

$$\widetilde{\kappa}_a(\gamma r) \leq \min\big(1, 4c_a(\gamma r)^{a-1}e^{-\gamma r/2}\big) \quad \text{for} \quad \gamma r \geq 2(a-1), a\geq 1. \tag{111}$$

7. When $a \in (1, 2.2)$, and $d \in [1, 2)$, analogous bounds with slightly different constants follow from employing the Gamma function upper bound of Batir (2017, Thm 2.3) in place of our upper bound (94). For brevity, we omit these derivations.

Proof of (111) Noting the definition (104) of κ_a, it suffices to show the following bound:

$$|K_a(r)| \leq \tfrac{4}{r} e^{-r/2} \quad \text{for} \quad r/2 \geq a-1,$$

where K_a is the modified Bessel function of the third kind. Using Wendland (2004, Def. 5.10), we have

$$K_a(r) = \tfrac{1}{2}\int_0^\infty e^{-r\cosh t}\cosh(at)dt \stackrel{(i)}{\leq} \int_0^\infty e^{-\frac{r}{2}e^t} e^{at} dt \stackrel{(ii)}{=} \int_{r/2}^\infty e^{-s}(\tfrac{2s}{r})^a \cdot \tfrac{1}{s} ds \stackrel{(iii)}{\leq} \tfrac{4}{r} e^{-r/2}$$

where step (i) uses the following inequalities: $\cosh t \triangleq \frac{1}{2}(e^t + e^{-t}) \geq \frac{1}{2}e^t$ and $\cosh(at) \leq e^{at}$ for $a > 0, t > 0$, step (ii) follows from a change of variable $s \leftarrow \frac{r}{2}e^t$, and finally step (iii) uses the following bound on the incomplete Gamma function obtained by substituting $B = 2$ and $A = a$ in Borwein and Chan (2009, Eq. 1.5):

$$\int_r^\infty t^{a-1} e^{-t} dt \leq 2r^{a-1}e^{-r} \quad \text{for} \quad r \geq a-1.$$

The proof is now complete.

O.3.2 Proof of the bound (105) on $\|\mathbf{k}_{\mathrm{rt}}\|_\infty$ and $\|\mathbf{k}_\star\|_\infty$

We claim that

$$\|\widetilde{\kappa}_b\|_\infty = 1 \quad \text{for all} \quad b > 0. \tag{112}$$

where was defined in (104). Putting the equality (112) with (102), (103), and (110) immediately implies the bounds (105) on the $\|\mathbf{k}_{\mathrm{rt}}\|_\infty$ and $\|\mathbf{k}_\star\|_\infty$. To prove (112), we follow the steps from the proof of Wendland (2004, Lem. 5.14). Using Wendland (2004, Def. 5.10), for $b > 0$ we have

$$K_b(r) = \tfrac{1}{2}\int_{-\infty}^\infty e^{-r\cosh t} e^{bt} dt = \tfrac{1}{2}\int_{-\infty}^\infty e^{-\frac{r}{2}(e^t+e^{-t})} e^{bt} dt = k^{-b}\tfrac{1}{2}\int_0^\infty e^{-r/2(u/k+k/u)} u^{b-1} du$$

where the last step follows by substituting $u = ke^t$. Setting $k = r/2$, we find that

$$r^b K_b(r) = 2^{b-1}\int_0^\infty e^{-u} e^{-r^2/(4u)} u^{b-1} du \leq 2^{b-1}\Gamma(b),$$

where we achieve equality in the last step when we take the limit $r \to 0$. Noting that $\widetilde{\kappa}_b(r) = \frac{2^{1-b}}{\Gamma(b)} r^b K_b(r)$ (104) yields the claim (112).

O.3.3 Proof of the bound (106) on $A_{\nu,\gamma,d}$

Using the definition (88) we have $A_{\nu,\gamma,d} = \left(\frac{1}{4\pi}\gamma^2\right)^{\frac{d}{4}} \cdot A'_{\nu,\gamma,d}$, where

$$\begin{aligned}
A'_{\nu,\gamma,d} &= \sqrt{\frac{\Gamma(\nu)}{\Gamma(\nu-d/2)}} \cdot \frac{\Gamma((\nu-d)/2)}{\Gamma(\nu/2)} \stackrel{(94)}{\leq} \sqrt{\frac{e}{\sqrt{2\pi}} \frac{(\frac{\nu-1}{e})^{\nu-1}\sqrt{\nu-1}}{(\frac{\nu-\frac{d}{2}-1}{e})^{\nu-\frac{d}{2}-1}\sqrt{\nu-\frac{d}{2}-1}}} \cdot \frac{e}{\sqrt{2\pi}} \frac{(\frac{\nu-d-2}{2e})^{\frac{\nu-d-2}{2}}\sqrt{\frac{\nu-d-2}{2}}}{(\frac{\nu-2}{2e})^{\frac{\nu-2}{2}}\sqrt{\frac{\nu-2}{2}}} \\
&\leq \left(\frac{e^2}{2\pi}\right)^{\frac{3}{4}} (2\sqrt{e})^{\frac{d}{2}} \cdot \left(\frac{\nu-1}{\nu-2}\right)^{\frac{\nu-2}{2}} \cdot \sqrt{\frac{\nu-1}{\nu-2}} \cdot (\nu-1)^{\frac{1}{4}} \cdot \left(\frac{\nu-d-2}{\nu-\frac{d}{2}-1}\right)^{\frac{\nu-\frac{d}{2}-2}{2}} \cdot \frac{\sqrt{(\nu-d-2)/\sqrt{(\nu-\frac{d}{2}-1)}}}{(\nu-d-2)^{\frac{1}{2}(\frac{d}{2}-1)}} \\
&\leq \left(\frac{e^2}{2\pi}\right)^{\frac{3}{4}} (2\sqrt{e})^{\frac{d}{2}} \cdot \sqrt{e} \cdot \sqrt{2} \cdot (\nu-1)^{\frac{1}{4}} \cdot (\sqrt{e})^{-\frac{d}{2}+1} \cdot \frac{(\nu-d-2)/(\nu-\frac{d}{2}-1)^{\frac{1}{4}}}{(\nu-d-2)^{\frac{d}{4}}} \\
&\leq \frac{e^2\sqrt{2e}}{(2\pi)^{\frac{3}{4}}} (\nu-1)^{\frac{1}{4}} \cdot (\nu - \tfrac{d}{2} - 1)^{\frac{3}{4}} \left(\tfrac{\nu-d-2}{4}\right)^{-\frac{d}{4}}.
\end{aligned}$$

As a result, we have

$$A_{\nu,\gamma,d} = \left(\tfrac{1}{4\pi}\gamma^2\right)^{d/4} \cdot A'_{\nu,\gamma,d} \leq 5\nu(\tfrac{\gamma^2}{\pi(\nu-d-2)})^{\frac{d}{4}} = 5\nu(\tfrac{\gamma^2}{2\pi(a-1)})^{\frac{d}{4}},$$

as claimed.

O.3.4 Proof of the bound (107) on $L_{\mathbf{k}_{\mathrm{rt}}}$

To derive a bound on the Lipschitz constant $L_{\mathbf{k}_{\mathrm{rt}}}$, we bound the derivative of $\widetilde{\kappa}_a$. Using $(r^a K_a(r))' = -r^a K_{a-1}(r)$ (Wendland, 2004, Eq. 5.2), we find that

$$(\widetilde{\kappa}_a(\gamma r))' = c_a \tfrac{d}{dr}((\gamma r)^a K_a(\gamma r)) = -\tfrac{c_a}{c_{a-1}}\gamma^2 r \widetilde{\kappa}_{a-1}(\gamma r).$$

Using (110), we have

$$\tfrac{c_a}{c_{a-1}}\gamma^2 r\widetilde{\kappa}_{a-1}(\gamma r) \leq \tfrac{c_a\gamma}{c_{a-1}}\min(\gamma r, \sqrt{2\pi}c_{a-1}(\gamma r)^{a-\frac{1}{2}}e^{-\gamma r+\frac{(a-1)^2}{2\gamma r}}) \quad \text{for} \quad r > 0.$$

And (111) implies that

$$\tfrac{c_a}{c_{a-1}}\gamma^2 r\widetilde{\kappa}_{a-1}(\gamma r) \leq \tfrac{c_a\gamma}{c_{a-1}}\min\bigl(\gamma r, 4c_{a-1}(\gamma r)^{a-1}e^{-\gamma r/2}\bigr) \quad \text{for} \quad r \geq 2(a-2), a \geq 2.$$

Next, we make use of the following observations: The function $t^b e^{-t/2}$ achieves its maximum at $t = 2b$. Then for any function $f : \mathbb{R}_+ \to \mathbb{R}_+$ such that $f(t) \leq t$ for all $t \geq 0$ and $f(t) \leq \min(t, Ct^b e^{-t/2})$ for $t > t^\dagger$ with $t^\dagger < 2b$, we have

$$\sup_{t\geq 0} f(t) \leq \min(2b, C(2b)^b e^{-b}).$$

As a result, we conclude that

$$\sup_{r\geq 0}\tfrac{c_a}{c_{a-1}}\gamma^2 r\widetilde{\kappa}_{a-1}(\gamma r) \leq \tfrac{c_a\gamma}{c_{a-1}}\min\bigl(2(a-1), 4c_{a-1}(2(a-1))^{a-1}e^{-a+1}\bigr). \tag{113}$$

Substituting the value of c_{a-1} and the bound (94), we can bound the second argument inside the min on the RHS of (113) (for $a \geq 3$) as follows:

$$\begin{aligned} 4c_{a-1}(2(a-1))^{a-1}e^{-a+1} &\leq 4\cdot 2^{2-a}\tfrac{1}{\sqrt{2\pi(a-2)}}(\tfrac{e}{a-2})^{a-2}2^{a-1}(a-1)^{a-1}e^{1-a} \\ &\leq \tfrac{8}{e\sqrt{2\pi}}(1+\tfrac{1}{a-2})^{a-2}\Bigl(\sqrt{a-2}+\tfrac{1}{\sqrt{a-2}}\Bigr) \leq \tfrac{8}{\sqrt{\pi}}\sqrt{a-2} \end{aligned} \tag{114}$$

for $a \geq 3$. When $a \in [2, 3)$, one can directly show that $4c_{a-1}(2(a-1))^{a-1}e^{-a+1} \leq 8/e$. Putting the pieces together, and noting that $\frac{c_a}{c_{a-1}} = \frac{1}{\max(2(a-1),1)}$, we find that

$$\sup_{r\geq 0}\tfrac{c_a}{c_{a-1}}\gamma^2 r\widetilde{\kappa}_{a-1}(\gamma r) \leq \tfrac{\gamma}{\max(2(a-1),1)}\min\Bigl(2(a-1), \tfrac{8}{e}+\tfrac{8}{\sqrt{\pi}}\sqrt{a-2}\Bigr) \leq \tfrac{C_1\gamma}{\sqrt{a+C_2}}$$

for any $a \geq 2$. And hence

$$L_{\mathbf{k}_{\mathrm{rt}}} \leq A_{\nu,\gamma,d}\sup_{r\geq 0}|\kappa_a'(\gamma r)| \leq A_{\nu,\gamma,d}\tfrac{C_1\gamma}{\sqrt{a+C_2}} \quad \Longrightarrow \quad \tfrac{L_{\mathbf{k}_{\mathrm{rt}}}}{\|\mathbf{k}_{\mathrm{rt}}\|_\infty} \overset{(105)}{\leq} \tfrac{C_1\gamma}{\sqrt{a+C_2}},$$

as claimed.

O.3.5 Proof of the bound (108) on $R_{\mathbf{k}_{\mathrm{rt}},n}$

Using arguments similar to those used to obtain (114), we find that

$$\max_{\gamma r \geq 2(a-1)} \widetilde{\kappa}_a(\gamma r) \leq 4c_a(2(a-1))^{a-1}e^{-(a-1)} \leq \sqrt{\tfrac{4}{a-1}} = \sqrt{\tfrac{8}{\nu-d-2}}.$$

Next, note that

$$(\gamma r)^{a-\frac{1}{2}}e^{-\gamma r+\frac{a^2}{2\gamma r}} \precsim e^{-\gamma r/2} \quad \text{for} \quad \gamma r \succsim a\log(1+a), \tag{115}$$
$$c_a = \tfrac{2^{1-a}}{\Gamma(a)} \stackrel{(94)}{\leq} (\tfrac{2e}{a-1})^{a-1}\tfrac{1}{\sqrt{2\pi(a-1)}},$$

where (115) follows from standard algebra. Thus, we have

$$\begin{aligned}
&c_a\widetilde{\kappa}_a(\gamma r) \precsim c_a\exp(-\gamma r/2) \quad \text{for} \quad \gamma r \succsim a\log(1+a)\\
\stackrel{(94)}{\Longrightarrow} \quad &R_{\mathbf{k}_{\mathrm{rt}},n} \precsim \tfrac{1}{\gamma}\max(\log n - a\log(1+a), C_1 a\log(1+a)),
\end{aligned}$$

as claimed.

O.3.6 Proof of the bound (109) on $R'_{\mathbf{k}_{\mathrm{rt}},\sqrt{n}}$

Let $V_d = \pi^{d/2}/\Gamma(d/2+1)$ denote the volume of unit Euclidean ball in $\mathbb{R}^d$. Using (103), we have

$$\begin{aligned}
\tfrac{1}{A^2_{\nu,\gamma,d}}\tau^2_{\mathbf{k}_{\mathrm{rt}}}(R) = \textstyle\int_{\|z\|_2\geq R}\widetilde{\kappa}^2_a(\gamma\|z\|_2)dz &= dV_d\textstyle\int_{r>R} r^{d-1}\widetilde{\kappa}^2_a(\gamma r)dr\\
&\stackrel{(110)}{\leq} 2\pi c_a^2 dV_d\textstyle\int_{r>R} r^{d-1}(\gamma r)^{2a-1}\exp(-2\gamma r+\tfrac{a^2}{\gamma r})dr\\
&= 2\pi c_a^2 dV_d\gamma^{1-d}\textstyle\int_{r>R}(\gamma r)^{\nu-2}\exp(-2\gamma r+\tfrac{a^2}{\gamma r})dr \quad (116)
\end{aligned}$$

where we have also used the fact that $2a+d=\nu$. Next, we note that

$$\exp(-2\gamma r+\tfrac{a^2}{\gamma r}) \leq \exp(-\tfrac{3}{2}\gamma r) \quad \text{for} \quad \gamma r \geq \sqrt{2}a \tag{117}$$

For any integer $b>1$, noticing that $f(t)=\frac{1}{\Gamma(b)}t^{b-1}e^{-t}$ is the density function of an Erlang distribution with shape parameter b and rate parameter 1, and using the expression for its complementary cumulative distribution function, we find that

$$\textstyle\int_r^\infty t^{b-1}e^{-t}dt = \Gamma(b)\sum_{i=0}^{b-1}\frac{r^i}{i!}e^{-r} \tag{118}$$

Thus, for $R > \frac{\sqrt{2}}{\gamma}a$, we have

$$\begin{aligned}
\textstyle\int_{r>R}(\gamma r)^{\nu-2}\exp(-2\gamma r+\tfrac{a^2}{\gamma r})dr &\stackrel{(117)}{\leq} \textstyle\int_{r>R}(\gamma r)^{\nu-2}\exp(-\tfrac{3}{2}\gamma r)dr\\
&= \tfrac{1}{\gamma}\big(\tfrac{2}{3}\big)^{\nu-1}\textstyle\int_{3\gamma R/2}^\infty t^{\nu-2}e^{-t}dt\\
&\stackrel{(118)}{=} \tfrac{1}{\gamma}\big(\tfrac{2}{3}\big)^{\nu-1}\Gamma(\nu-1)e^{-\frac{3}{2}\gamma R}\textstyle\sum_{i=0}^{\nu-1}\frac{(3\gamma R/2)^i}{i!}\\
&\leq \tfrac{1}{\gamma}\Gamma(\nu-1)e^{-\frac{3}{2}\gamma R}e^{\gamma R} \leq \tfrac{1}{\gamma}\Gamma(\nu-1)e^{-\frac{1}{2}\gamma R}. \quad (119)
\end{aligned}$$

Putting the bounds (116) and (119) together for $R \geq \frac{\nu-d}{\sqrt{2}\gamma}$, we find that

$$\tau^2_{\mathbf{k}_{\mathrm{rt}}}(R) \leq A^2_{\nu,\gamma,d} \cdot \gamma^{-d} \cdot 2\pi \tfrac{2^{2-2a}}{\Gamma^2(a)} \cdot \tfrac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \cdot \Gamma(\nu-1)\exp(-\tfrac{1}{2}\gamma R).$$

Using (94), we have

$$\begin{aligned} 2\pi \tfrac{2^{2-2a}}{\Gamma^2(a)} \cdot \tfrac{\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2}+1)} \cdot \Gamma(\nu-1) &\leq 2\pi \cdot \tfrac{2^{2-2a}e^{2(a-1)}}{2\pi(a-1)(a-1)^{2(a-1)}} \cdot \tfrac{\pi^{d/2}(2e)^{d/2}}{\sqrt{\pi d}\cdot d^{d/2}} \cdot e\sqrt{\nu-2}(\tfrac{\nu-2}{e})^{\nu-2} \\ &= \tfrac{2}{\sqrt{\pi}} e^{2a-2+d/2+1-\nu}(2\pi)^{d/2}(2a-2)^{-(2a-1)}d^{-d/2-1}(\nu-2)^{\nu-3/2} \\ &= \tfrac{2}{e\sqrt{\pi}}(2e\pi)^{d/2}(\nu-d-2)^{-(\nu-d-1)}d^{-d/2-1}(\nu-2)^{\nu-3/2} \\ &= \tfrac{2}{e\sqrt{\pi}}(2e\pi)^{d/2}(2(a-1))^{-(2a-1)}d^{-d/2-1}(\nu-2)^{\nu-3/2}. \end{aligned}$$

Putting the pieces together, and solving for $\tau_{\mathbf{k}_{\mathrm{rt}}}(R) \leq \|\mathbf{k}_{\mathrm{rt}}\|_\infty/\sqrt{n} \overset{(105)}{=} A_{\nu,\gamma,d}/\sqrt{n}$, we find that $R'_{\mathbf{k}_{\mathrm{rt}},\sqrt{n}}$ (6) can be bounded as

$$\begin{aligned} R'_{\mathbf{k}_{\mathrm{rt}},\sqrt{n}} &\leq \tfrac{2}{\gamma} \cdot \max(\tfrac{a}{\sqrt{2}}, \log n + \log(\tfrac{2}{e\sqrt{\pi}}) + d\log(\tfrac{\sqrt{2e\pi}}{\gamma}) + \log(\tfrac{(\nu-2)^{\nu-\frac{3}{2}}}{(2(a-1))^{2a-1}d^{\frac{d}{2}+1}})) \\ &\precsim \tfrac{1}{\gamma} \cdot (a + \log n + d\log(\tfrac{\sqrt{2e\pi}}{\gamma}) + \log(\tfrac{(\nu-2)^{\nu-\frac{3}{2}}}{(2(a-1))^{2a-1}d^{\frac{d}{2}+1}}), \end{aligned}$$

as claimed.

O.4 Proofs for B-spline kernel

Recall the notation from (89) to (92). Then, the kernel $\mathbf{k}_\star = \textbf{B-spline}(2\beta+1)$ and its square-root kernel $\mathbf{k}_{\mathrm{rt}} = \widetilde{S}_{\beta,d} \cdot \textbf{B-spline}(\beta)$ satisfy

$$\|\mathbf{k}_{\mathrm{rt}}\|_\infty = \widetilde{S}_{\beta,d} \overset{(i)}{=} c_\beta^d \quad \text{where} \quad c_\beta \triangleq \frac{\mathfrak{B}_{\beta+1}}{\sqrt{\mathfrak{B}_{2\beta+2}}} \begin{cases} \overset{(ii)}{=} \frac{2}{\sqrt{3}} \text{ when } \beta = 1 \\ \overset{(iii)}{\leq} 1 \text{ when } \beta > 1 \end{cases}, \tag{120}$$

$$\|\mathbf{k}_\star\|_\infty = 1, \tag{121}$$

$$\tfrac{L_{\mathbf{k}_{\mathrm{rt}}}}{\|\mathbf{k}_{\mathrm{rt}}\|_\infty} \leq \tfrac{4}{3}\sqrt{d}, \tag{122}$$

$$R_{\mathbf{k}_{\mathrm{rt}},n} \leq \sqrt{d}(\beta+1)/2, \quad \text{and} \quad R'_{\mathbf{k}_{\mathrm{rt}},\sqrt{n}} \leq \sqrt{d}(\beta+1)/2. \tag{123}$$

While claims (i) and (ii) follow directly from the definitions in the display (89), claim (iii) can be verified numerically, e.g., using SciPy (Virtanen et al., 2020). From numerical verification, we also find that the constant c_β in (120) is decreasing with β. See App. O.4.1 for the proofs of the remaining claims. Finally, substituting various quantities from (120), (122), and (123) in (93), we find that

$$\mathfrak{M}^{\textbf{B-spline}}_{\mathbf{k}_{\mathrm{rt}}}(n, \tfrac{1}{2}\log_2 n, d, \tfrac{\delta}{2n}, \delta', R) \precsim \sqrt{\log(\tfrac{n}{\delta})\Big(\log(\tfrac{1}{\delta}) + d\log(d\beta + \sqrt{d}R)\Big)}.$$

O.4.1 Proofs of the bounds on B-spline kernel quantities

We start with some basic set-up. Consider the (unnormalized) univariate B-splines

$$f_\beta : \mathbb{R} \to [0,1] \quad \text{with} \quad f_\beta(a) = \circledast^\beta \mathbb{I}_{[-\frac{1}{2},\frac{1}{2}]}(a) \overset{(i)}{=} \tfrac{1}{(\beta-1)!} \textstyle\sum_{j=0}^{\beta} (-1)^j \binom{\beta}{j} (a + \frac{\beta}{2} - j)_+^{\beta-1}$$

where step (i) follows from Schumaker (2007, Eqn 4.46, 4.47, 4.59, p.135, 138). Noting that f_β is an even function with a unique global maxima at 0 (see Schumaker (2007, Ch 4.) for more details), we find that

$$\|f_\beta\|_\infty = f_\beta(0) = \tfrac{1}{(\beta-1)!} \textstyle\sum_{j=0}^{\lfloor \beta/2 \rfloor} (-1)^j \binom{\beta}{j} (\frac{\beta}{2} - j)^{\beta-1} \overset{(89)}{=} \mathfrak{B}_\beta. \tag{124}$$

Bounds on $\|\mathbf{k}_\star\|_\infty$ and $\|\mathbf{k}_{\mathrm{rt}}\|_\infty$ Recalling (90) and (91), we find that

$$\|\mathbf{k}_\star\|_\infty = \|\chi_{2\beta+2}\|_\infty = S_{2\beta+2,d} \|f_{2\beta+2}\|_\infty^d = S_{2\beta+2,d} \mathfrak{B}_{2\beta+2}^d = 1,$$

thereby establishing (121).

Bounds on $L_{\mathbf{k}_{\mathrm{rt}}}$ We have

$$f'_{\beta+1}(a) = \textstyle\int f_\beta(b) \frac{d}{da} \mathbb{I}_{[-\frac{1}{2},\frac{1}{2}]}(a-b) dy = f_\beta(a - \frac{1}{2}) - f_\beta(a + \frac{1}{2})$$

and hence $\|f'_{\beta+1}\|_\infty = \sup_a |f_\beta(a - \frac{1}{2}) - f_\beta(a + \frac{1}{2})| \leq \|f_\beta\|_\infty$ since f_β is non-negative. Putting the pieces together, we have

$$\begin{aligned}
\tfrac{L_{\mathbf{k}_{\mathrm{rt}}}}{\|\mathbf{k}_{\mathrm{rt}}\|_\infty} = \tfrac{1}{\widetilde{S}_{\beta,d}} L_{\mathbf{k}_{\mathrm{rt}}} \leq \sup_z \|\nabla \chi_{\beta+1}(z)\|_2 &\overset{(90)}{\leq} \sqrt{d} \cdot S_{\beta+1,d} \|f'_{\beta+1}\|_\infty \cdot \|f_{\beta+1}\|_\infty^{d-1} \\
&\overset{(89,124)}{\leq} \sqrt{d} \cdot \mathfrak{B}_{\beta+1}^{-d} \cdot \mathfrak{B}_\beta \cdot \mathfrak{B}_{\beta+1}^{d-1} \\
&= \sqrt{d} \tfrac{\mathfrak{B}_\beta}{\mathfrak{B}_{\beta+1}} \overset{(i)}{\leq} \sqrt{d} \tfrac{4}{3},
\end{aligned}$$

where step (i) can be verified numerically.

Bounds on $R_{\mathbf{k}_{\mathrm{rt}},n}$, and $R'_{\mathbf{k}_{\mathrm{rt}},\sqrt{n}}$ Using the property of convolution, we find that $f_{\beta+1}(a) = 0$ if $|a| \geq \frac{1}{2}(\beta+1)$. Hence $\kappa_{\mathrm{rt}}(z) = 0$ for $\|z\|_\infty > (\beta+1)/2$ and applying the definitions (6), we find that

$$R_{\mathbf{k}_{\mathrm{rt}},n} \leq \sqrt{d}(\beta+1)/2 \quad \text{and} \quad R'_{\mathbf{k}_{\mathrm{rt}},\sqrt{n}} \leq \sqrt{d}(\beta+1)/2$$

as claimed in (123).

O.5 Explicit MMD rates for common kernels

Putting the quantities from Tab. 2 together with Thm. 1 (with $\delta' = \delta$, and δ treated as a constant) yields the MMD rates summarized in Tab. 4.

For completeness, we illustrate a key simplification that can readily yield the results stated in Tab. 4. Define $\zeta_{\mathbf{k}_{\mathrm{rt}},\mathcal{S}_n}$ as follows:

$$\zeta_{\mathbf{k}_{\mathrm{rt}},\mathcal{S}_n} \triangleq \tfrac{1}{d} \max(R'_{\mathbf{k}_{\mathrm{rt}},\sqrt{n}}, R_{\mathcal{S}_n})^2,$$

Kernel $\mathbf{k}_\star$	$\mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S}_{\mathrm{KT}}) \precsim$
Gaussian(σ)	$C_1^d \cdot \sqrt{\frac{\log n}{n} \cdot [1+\frac{1}{d}(\log n + (\frac{R_{\mathcal{S}_n}}{\sigma})^2)]^{\frac{d}{2}} \cdot \log(\sqrt{\log n} + \frac{R_{\mathcal{S}_n}}{\sigma})}$
Matérn(ν, γ)	$C_2^d \cdot \sqrt{\frac{\nu^2 \log n}{n}\Big[\frac{1}{d(a-1)}(a+\log^2 n + d^2\log^2(\frac{\sqrt{2e\pi}}{\gamma}) + G_{\nu,d}^2 + \gamma^2 R_{\mathcal{S}_n}^2)\Big]^{\frac{d}{2}} \cdot \log(\log n + a + \gamma R_{\mathcal{S}_n})}$
B-spline$(2\beta+1)$	$C_3^d \cdot \sqrt{\frac{\log n}{n} \cdot [\beta^2 + \frac{R_{\mathcal{S}_n}^2}{d}]^{\frac{d}{2}} \cdot \log(\beta + \frac{R_{\mathcal{S}_n}}{\sqrt{d}})}$

Table 4: Explicit kernel thinning MMD guarantees for common kernels. Here, $a \triangleq \frac{1}{2}(\nu - d)$, $G_{\nu,d} \triangleq \log(\frac{(\nu-2)^{\nu-\frac{3}{2}}}{(2(a-1))^{2a-1} d^{\frac{d}{2}+1}})$ and each C_i denotes a universal constant. See App. O.5 for more details on deriving these bounds from Thm. 1.

where $R_{\mathcal{S}_n}$, and $R'_{\mathbf{k}_{\mathrm{rt}},\sqrt{n}}$ were defined in (6) and (7) respectively. Then applying Thm. 1, and substituting the simplified bound (93) in (9), we find that

$$\begin{aligned}\mathrm{MMD}_{\mathbf{k}_\star}(\mathcal{S}_n, \mathcal{S}_{\mathrm{KT}}) &\leq 2\frac{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{\sqrt{n}} + \frac{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{\sqrt{n}}(B\zeta_{\mathbf{k}_{\mathrm{rt}},\mathcal{S}_n})^{\frac{d}{4}} \cdot d^{-\frac{1}{4}}\sqrt{\log\tfrac{n}{\delta}\cdot\Big[\log(\tfrac{1}{\delta}) + d\log\Big(\tfrac{L_{\mathbf{k}_{\mathrm{rt}}}(R_{\mathbf{k}_{\mathrm{rt}},n}+R)}{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}\Big)\Big]} \\ &\overset{(i)}{\precsim_\delta} 2\frac{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}{\sqrt{n}} + \|\mathbf{k}_{\mathrm{rt}}\|_\infty(B\zeta_{\mathbf{k}_{\mathrm{rt}},\mathcal{S}_n})^{\frac{d}{4}} \cdot d^{\frac{1}{4}}\sqrt{\frac{\log n}{n}\log\Big(\frac{L_{\mathbf{k}_{\mathrm{rt}}}(R_{\mathbf{k}_{\mathrm{rt}},n}+R_{\mathcal{S}_n})}{\|\mathbf{k}_{\mathrm{rt}}\|_\infty}\Big)}, \quad (125)\end{aligned}$$

with probability at least $1-2\delta$, where $B \triangleq 8e\pi$ is a universal constant, and in step (i) we use the following bound: For any $r = \sqrt{\alpha d}$, we have

$$c_d r^{\frac{d}{2}} = \frac{(4\pi)^{\frac{d}{4}}}{(\Gamma(\frac{d}{2}+1))^{\frac{1}{2}}}(\alpha d)^{\frac{d}{4}} \overset{(ii)}{\leq} \frac{(8e\pi\alpha)^{\frac{d}{4}}}{(\pi d)^{\frac{1}{4}}} \leq (B\alpha)^{\frac{d}{4}} d^{-\frac{1}{4}} \quad \text{where} \quad B \triangleq 8e\pi,$$

and step (ii) follows (for $d \geq 2$) from the Gamma function bounds (94). Now the results in Tab. 4 follow by simply substituting the various quantities from Tab. 2 in (125).

Appendix P. Supplementary Details for Vignettes of Sec. 7

This section provides supplementary details for the vignettes of Sec. 7.

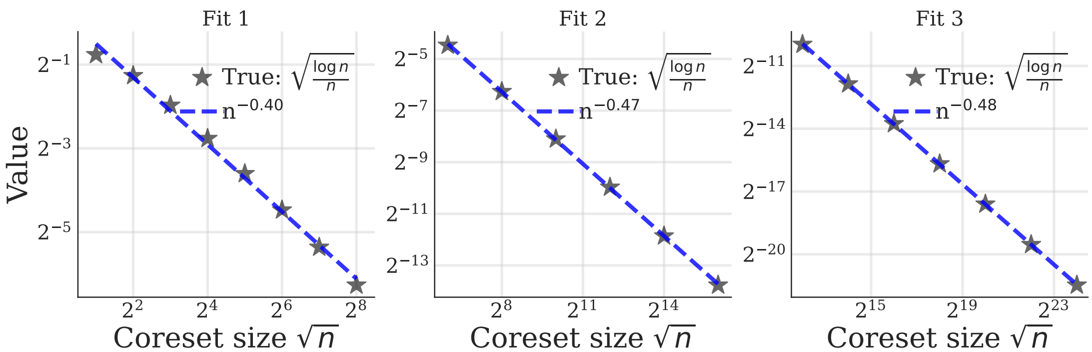
P.1 Mixture of Gaussians target

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

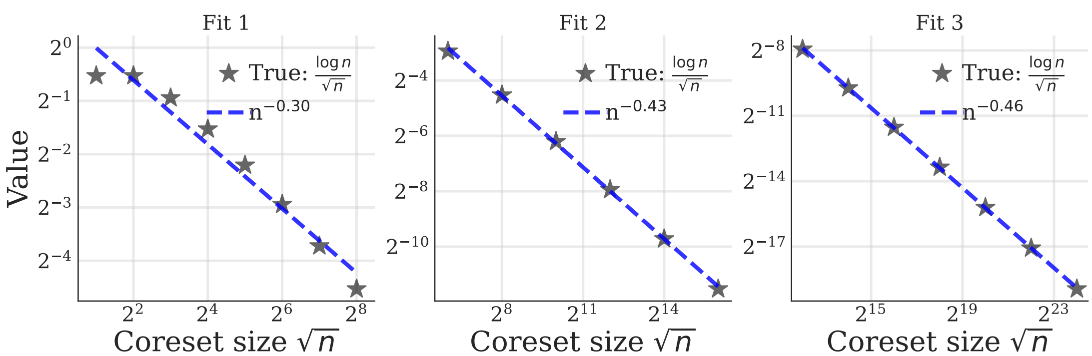
$$\begin{aligned}&\mu_1 = [-3,3]^\top, \quad \mu_2 = [-3,3]^\top, \quad \mu_3 = [-3,-3]^\top, \quad \mu_4 = [3,-3]^\top, \\ &\mu_5 = [0,6]^\top, \qquad \mu_6 = [-6,0]^\top, \quad \mu_7 = [6,0]^\top, \qquad \mu_8 = [0,-6]^\top.\end{aligned}$$

P.2 Empirical decay rates

In Fig. 5, we provide results for regressing the true error rate of $\sqrt{\frac{\log n}{n}}$ and $\frac{\log n}{\sqrt{n}}$ to the input size n (on the log-log scale as in all of mean MMD plots) for various ranges of coreset sizes. Notably, the observed empirical rates are significantly slower than $n^{-\frac{1}{2}}$ for the smallest coreset range (the same coreset range used in the panels in Fig. 3) but approach $n^{-\frac{1}{2}}$ as the coreset size increases. Hence, the observed empirical rates of Fig. 3 are consistent with $\frac{\log^c n}{\sqrt{n}}$ rates of decay.



(a) True error scaling of $\sqrt{\frac{\log n}{n}}$



(b) True error scaling of $\frac{\log n}{\sqrt{n}}$

Figure 5: Empirical decay rates when true error $= \frac{\log^c n}{\sqrt{n}}$**.** We display the ordinary least squares fits for regressing the log of true error onto the log of input size n when the true error equals either (a) $\sqrt{\frac{\log n}{n}}$ or (b) $\frac{\log n}{\sqrt{n}}$. As the input range increases, the fitted decay rate tends towards $n^{-\frac{1}{2}}$ despite appearing significantly slower for smaller input ranges.

P.3 MCMC vignette details

For complete details on the targets and sampling algorithms we refer the reader to Riabiz et al. (2021, Sec. 4). When applying standard thinning to any Markov chain output, we adopt the convention of keeping the final sample point. For all experiments, we used only the odd indices of the post burn-in sample points when thinning to form $\mathcal{S}_n$.

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. For the Hinch experiments, we discard the first 10^6 points as burn-in following Riabiz et al. (2021, App. S5.4). For all n, the parameter σ for the Gaussian kernel is set to the median distance between all pairs of 4^7 points obtained by standard thinning the post-burn-in odd indices. The resulting values for the Goodwin chains were 0.02 for RW; 0.0201 for adaRW; 0.0171 for MALA; and 0.0205 for pMALA. The respective numbers for Lotka-Volterra task were 0.0274 for RW; 0.0283 for adaRW; 0.023 for MALA; and 0.0288 for pMALA. Finally, the numbers for the Hinch task were 8.0676 for RW; 8.3189 for adaRW; 8.621 for MALA; and 8.6649 for pMALA.

Appendix Q. Kernel Thinning with Square-root Dominating Kernels

As alluded to in Sec. 2, it is not necessary to identify an exact square-root kernel $\mathbf{k}_{\mathrm{rt}}$ to run kernel thinning. Rather, it suffices to identify any *square-root dominating kernel* $\widetilde{\mathbf{k}}_{\mathrm{rt}}$ defined as follows.

Definition 8 (Square-root dominating kernel). *We say a reproducing kernel* $\widetilde{\mathbf{k}}_{\mathrm{rt}} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *is a* square-root dominating kernel *for a reproducing kernel* $\mathbf{k}_\star : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ *if* $\widetilde{\mathbf{k}}_{\mathrm{rt}}(x, \cdot)$ *is square-integrable for all* $x \in \mathbb{R}^d$ *and either of the following equivalent conditions hold.*

(a) The RKHS of $\mathbf{k}_\star$ *belongs to the RKHS of* $\widetilde{\mathbf{k}}(x, y) \triangleq \int_{\mathbb{R}^d} \widetilde{\mathbf{k}}_{\mathrm{rt}}(x, z)\widetilde{\mathbf{k}}_{\mathrm{rt}}(y, z)dz$.

(b) The function $\widetilde{\mathbf{k}} - c\mathbf{k}_\star$ *is positive definite for some* $c > 0$.

Remark 9 (Controlling $\mathrm{MMD}_{\mathbf{k}_\star}$). *A square-root dominating kernel* $\widetilde{\mathbf{k}}_{\mathrm{rt}}$ *is a suitable surrogate for* $\mathbf{k}_{\mathrm{rt}}$ *as, by Zhang and Zhao (2013, Lem. 2.2, Prop. 2.3),* $\mathrm{MMD}_{\mathbf{k}_\star}(\mu, \nu) \le \sqrt{1/c} \cdot \mathrm{MMD}_{\widetilde{\mathbf{k}}}(\mu, \nu)$ *for c the constant appearing in Def. 8(b) and all distributions* μ *and* ν.

Def. 8 and Rem. 9 enable us to use convenient dominating surrogates whenever an exact square-root kernel is inconvenient to derive or deploy. For example, our next result, proved in App. Q.1, shows that a standard Matérn kernel is a square-root dominating kernel for every sufficiently-smooth shift-invariant and absolutely integrable $\mathbf{k}_\star$.

Proposition 4 (Dominating smooth kernels). *If* $\mathbf{k}_\star(x, y) = \kappa(x - y)$ *and* $\kappa \in L^1 \cap C^{2\nu}$ *for* $\nu > d$*, then, for any* $\gamma > 0$*, the* ***Matérn***$(\frac{\nu}{2}, \gamma)$ *kernel of Tab. 1 is a square-root dominating kernel for* $\mathbf{k}_\star$.

Checking the square-root dominating condition is also particularly simple for any pair of continuous shift-invariant kernels as the next result, proved in App. Q.2, shows.

Proposition 5 (Dominating shift-invariant kernels). *Suppose that the real-valued kernels* $\mathbf{k}_\star(x, y) = \kappa(x - y)$ *and* $\widetilde{\mathbf{k}}_{\mathrm{rt}}(x, y) = \widetilde{\kappa}_{\mathrm{rt}}(x - y)$ *have respective spectral densities (Def. 6)* $\widehat{\kappa}$ *and* $\widehat{\kappa}_{\mathrm{rt}}$*. If* $\widehat{\kappa}_{\mathrm{rt}} \in L^2$*, then* $\widetilde{\mathbf{k}}_{\mathrm{rt}}$ *is a square-root dominating kernel for* $\mathbf{k}_\star$ *if and only if*

$$\operatorname*{ess\,sup}_{\omega \in \mathbb{R}^d :\, \widehat{\kappa}(\omega) > 0} \frac{\widehat{\kappa}(\omega)}{\widehat{\kappa}_{\mathrm{rt}}(\omega)^2} < \infty. \tag{126}$$

In Tab. 5, we use Prop. 5 to derive convenient tailored square-root dominating kernels $\widetilde{\mathbf{k}}_{\mathrm{rt}}$ for standard inverse multiquadric kernels, hyberbolic secant (sech) kernels, and Wendland's compactly supported kernels. In each case, we can identify a square-root dominating kernel from the same family.

Name of kernel $\mathbf{k}_\star(x,y)=\kappa(x-y)$	Expression for $\kappa(z)$	Name of square-root dominating kernel $\widetilde{\mathbf{k}}_{\mathrm{rt}}(x,y)=\widetilde{\kappa}_{\mathrm{rt}}(x-y)$
InverseMultiquadric(ν,γ) $\nu>0,\gamma>0$	$(\gamma^2+\|z\|_2^2)^{-\nu}$	**InverseMultiquadric**(ν',γ') $(\nu',\gamma')\in\mathcal{C}_{\nu,\gamma,d}$; see (131)
Sech(a) $a>0$	$\prod_{j=1}^d \mathrm{sech}\big(\sqrt{\frac{\pi}{2}}az_j\big)$	**Sech**$(2a)$
Wendland(s) $s\in\mathbb{N},s\geq\frac{1}{2}(d+1)$	$\phi_{d,s}(\|z\|_2)$; see (127)	**Wendland**(s') $s'\in\mathbb{N}_0,s'\leq\frac{1}{4}(2s-1-d)$

Table 5: Square-root dominating kernels $\widetilde{\mathbf{k}}_{\mathrm{rt}}$ for common kernels $\mathbf{k}_\star$ (see Def. 8). Here $\mathbb{N}_0$ denotes the non-negative integers. See App. Q.3 for our derivation.

Expressions for Wendland kernels The **Wendland**(s) kernel is a compactly-supported radial kernel $\phi_{d,s}(\|x-y\|_2)$ on $\mathbb{R}^d$ where $\phi_{d,s}:\mathbb{R}_+\to\mathbb{R}$ is a truncated minimal-degree polynomial with $2s$ continuous derivatives. We collect here the expressions for $\phi_{d,s}$ for $s=0,1,2$ and refer the readers to Wendland (2004, Ch. 9, Thm. 9.13, Tab. 9.1) for more general s. Let $(r)_+=\max(0,r)$ and $\ell\triangleq\lfloor d/2\rfloor+3$, then we have

$$\begin{aligned}\phi_{d,0}(r)&=(1-r)_+^{\lfloor d/2\rfloor+1},\quad \phi_{d,1}(r)=(1-r)_+^{\lfloor d/2\rfloor+2}[(\lfloor d/2\rfloor+2)r+1] \\ \phi_{d,2}(r)&=(1-r)_+^{\lfloor d/2\rfloor+3}[(\ell^2+4\ell+3)r^2+(3\ell+6)r+3].\end{aligned} \tag{127}$$

Q.1 Proof of Prop. 4: Dominating smooth kernels

Since $\kappa\in L^1$, $\mathcal{F}(\kappa)$ is bounded by the Babenko-Beckner inequality (Beckner, 1975) and nonnegative by Bochner's theorem (Bochner, 1933; Wendland, 2004, Thm. 6.6). Moreover, since $\kappa\in C^{2\nu}$, Sun (1993, Thm. 4.1) implies that $\int\|\omega\|_2^{2\nu}\mathcal{F}(\kappa)(\omega)d\omega<\infty$. By Wendland (2004, Theorem 8.15), the **Matérn**(ν,γ) kernel $\widetilde{\mathbf{k}}(x,y)\propto\Phi_{\nu,\gamma}(x-y)$ for $\Phi_{\nu,\gamma}$ continuous with $\mathcal{F}(\Phi_{\nu,\gamma})(\omega)=(\gamma^2+\|\omega\|_2^2)^{-\nu}$. Since we have established that

$$\int\frac{\mathcal{F}(\kappa)(\omega)^2}{\mathcal{F}(\Phi_{\nu,\gamma})(\omega)}d\omega=\int(\gamma^2+\|\omega\|_2^2)^{\nu}\mathcal{F}(\kappa)(\omega)^2d\omega\leq\|\mathcal{F}(\kappa)\|_\infty\int(\gamma^2+\|\omega\|_2^2)^{\nu}\mathcal{F}(\kappa)(\omega)d\omega<\infty,$$

Wendland (2004, Thm. 10.12) implies that κ belongs to $\mathcal{H}_{\widetilde{\mathbf{k}}}$ and hence that $\mathcal{H}_{\mathbf{k}_\star}\subseteq\mathcal{H}_{\widetilde{\mathbf{k}}}$. Finally, by App. N.1, **Matérn**$(\frac{\nu}{2},\gamma)$ is a valid square-root dominating kernel for $\widetilde{\mathbf{k}}$ and therefore for $\mathbf{k}_\star$.

Q.2 Proof of Prop. 5: Dominating shift-invariant kernels

Def. 6 implies that

$$\mathbf{k}_\star(x,y)=\frac{1}{(2\pi)^{d/2}}\int e^{-i\langle\omega,x-y\rangle}\widehat{\kappa}(\omega)d\omega\quad\text{and}\quad\widetilde{\mathbf{k}}_{\mathrm{rt}}(x,y)=\frac{1}{(2\pi)^{d/2}}\int e^{-i\langle\omega,x-y\rangle}\widehat{\kappa}_{\mathrm{rt}}(\omega)d\omega,$$

Moreover, since $\mathbf{k}_{\mathrm{rt}}(x,\cdot)=\mathcal{F}(e^{-i\langle\cdot,x\rangle}\widehat{\kappa}_{\mathrm{rt}})$ for $e^{-i\langle\cdot,x\rangle}\widehat{\kappa}_{\mathrm{rt}}\in L^1\cap L^2$, the Plancherel-Parseval identity (Wendland, 2004, Proof of Thm. 5.23) implies that

$$\begin{aligned}\widetilde{\mathbf{k}}(x,y)\triangleq\int_{\mathbb{R}^d}\widetilde{\mathbf{k}}_{\mathrm{rt}}(x,z)\widetilde{\mathbf{k}}_{\mathrm{rt}}(y,z)dz&=\int_{\mathbb{R}^d}e^{-i\langle\omega,x\rangle}\widehat{\kappa}_{\mathrm{rt}}(\omega)\ e^{i\langle\omega,y\rangle}\widehat{\kappa}_{\mathrm{rt}}(\omega)d\omega\\&=\int_{\mathbb{R}^d}e^{-i\langle\omega,x-y\rangle}\widehat{\kappa}_{\mathrm{rt}}(\omega)^2d\omega,\end{aligned}$$

and Bochner's theorem (Bochner, 1933; Wendland, 2004, Thm. 6.6) implies that $\widetilde{\mathbf{k}}$ is a kernel. Finally, Prop. 3.1 of Zhang and Zhao (2013) now implies that the RKHS of $\mathbf{k}_\star$ belongs to the RKHS of $\widetilde{\mathbf{k}}$ if and only if the ratio condition (126) holds.

Q.3 Derivation of Tab. 5: Square-root dominating kernels $\widetilde{\mathbf{k}}_{\mathrm{rt}}$ for common kernels $\mathbf{k}_\star$

Thanks to Prop. 5, to establish the validity of the square root dominating kernels stated in Tab. 5, it suffices to verify that

$$\widehat{\kappa}_{\mathrm{rt}} \in L^2 \quad \text{and} \quad \widehat{\kappa} \precsim_{d,\kappa,\widetilde{\kappa}_{\mathrm{rt}}} \widehat{\kappa}_{\mathrm{rt}}^2, \tag{128}$$

where the functions κ and $\widehat{\kappa}_{\mathrm{rt}}$ are the spectral densities of κ and $\widetilde{\kappa}_{\mathrm{rt}}$.

Inverse multiquadric Consider the positive definite function $\Phi_{\nu,\gamma}$ (86) underlying the Matérn kernel, which is continuous on $\mathbb{R}^d \backslash \{\mathbf{0}\}$. When $\kappa(z) = (\gamma^2 + \|z\|_2^2)^{-\nu}$ and $\widetilde{\kappa}_{\mathrm{rt}}(z) = ({\gamma'}^2 + \|z\|_2^2)^{-\nu'}$, Wendland (2004, Theorem 8.15) implies that

$$\widehat{\kappa}(\omega) = \Phi_{\nu,\gamma}(\omega) \quad \text{and} \quad \Phi^2_{\nu',\gamma'}(\omega) = \widehat{\kappa}_{\mathrm{rt}}(\omega)^2.$$

Let $a(\omega) \asymp_{d,\nu,\gamma} b(\omega)$ denote asymptotic equivalence up to a constant depending on d, ν, γ. Then, by (DLMF, Eqs. 10.25.3 & 10.30.2), we have

$$\Phi_{\nu,\gamma}(\omega) \asymp_{d,\nu,\gamma} \|\omega\|_2^{\nu-\frac{d}{2}-\frac{1}{2}} e^{-\gamma\|\omega\|_2} \quad \text{as} \quad \|\omega\|_2 \to \infty, \tag{129}$$

$$\Phi_{\nu,\gamma}(\omega) \asymp_{d,\nu,\gamma} \|\omega\|_2^{-(d-2\nu)_+} \quad \text{as} \quad \|\omega\|_2 \to 0. \tag{130}$$

Hence, $\widehat{\kappa}_{\mathrm{rt}} \in L^2$ whenever $\nu' > \frac{d}{4}$.

Moreover, applying (129), we find that for $\|\omega\|_2 \to \infty$,

$$\begin{aligned}\frac{\Phi_{\nu,\gamma}(\omega)}{\Phi^2_{\nu',\gamma'}(\omega)} &\asymp_{d,\nu,\gamma,\nu',\gamma'} \|\omega\|_2^{\nu-\frac{d}{2}-\frac{1}{2}} e^{-\gamma\|\omega\|_2} \cdot \|\omega\|_2^{-2\nu'+d+1} e^{2\gamma'\|\omega\|_2} \\ &= \|\omega\|_2^{\nu+\frac{d}{2}+\frac{1}{2}-2\nu'} e^{(2\gamma'-\gamma)\|\omega\|_2}.\end{aligned}$$

If $2\gamma' - \gamma < 0$, this expression is bounded for any value of ν'. If $2\gamma' - \gamma = 0$, this expression is bounded when $\nu' \geq \frac{\nu}{2} + \frac{d}{4} + \frac{1}{4}$.

Applying (130), we find that for $\|\omega\|_2 \to 0$,

$$\frac{\Phi_{\nu,\gamma}(\omega)}{\Phi^2_{\nu',\gamma'}(\omega)} \asymp_{d,\nu,\gamma,\nu',\gamma'} \|\omega\|_2^{-(d-2\nu)_+} \cdot \|\omega\|_2^{2(d-2\nu')_+} = \|\omega\|_2^{2(d-2\nu')_+-(d-2\nu)_+},$$

If $\nu \geq \frac{d}{2}$, this expression is finite for any value of ν'. If $\nu < \frac{d}{2}$, this expression is finite when $\nu' \leq \frac{\nu}{2} + \frac{d}{4}$. Hence, our condition (128) is verified whenever (ν', γ') belongs to the set

$$\begin{aligned}\mathcal{C}_{\nu,\gamma,d} \triangleq \big\{(\nu', \gamma') : &(1)\ \nu' > \tfrac{d}{4} \quad \text{and} \quad \gamma' \leq \tfrac{\gamma}{2}, \quad \text{and} \\ &(2)\ \nu' \leq \tfrac{d}{4} + \tfrac{\nu}{2} \quad \text{if} \quad \nu < \tfrac{d}{2}, \quad \text{and} \\ &(3)\ \nu' \geq \tfrac{d}{4} + \tfrac{\nu}{2} + \tfrac{1}{4} \quad \text{if} \quad \gamma' = \tfrac{\gamma}{2}\big\}.\end{aligned} \tag{131}$$

Sech Define $\kappa_a(z) \triangleq \prod_{j=1}^d \mathrm{sech}\big(\sqrt{\frac{\pi}{2}} a z_j\big)$, and suppose $\kappa = \kappa_a$ and $\widetilde{\kappa}_{\mathrm{rt}} = \kappa_{2a}$. Huggins and Mackey (2018, Ex. 3.2) yields that

$$\widehat{\kappa}(\omega) = \mathcal{F}(\kappa_a)(\omega) = \frac{1}{a^d} \prod_{j=1}^d \mathrm{sech}\big(\sqrt{\tfrac{\pi}{2}} \tfrac{\omega_j}{a}\big) = a^{-d} \cdot \kappa_{1/a}(\omega), \quad \text{and} \tag{132}$$

$$\widehat{\kappa}_{\mathrm{rt}}(\omega) = \mathcal{F}(\widetilde{\kappa}_{\mathrm{rt}})(\omega) = (2a)^{-d} \kappa_{1/(2a)}(\omega). \tag{133}$$

Since $\mathrm{sech}^2(b/2) = \frac{4}{e^b + e^{-b} + 2} > \frac{4}{e^b + e^{-b}} = 2\,\mathrm{sech}(b)$, we have

$$\kappa_{1/a}(\omega) \leq 2^{-d} \cdot \kappa_{1/(2a)}(\omega)^2. \tag{134}$$

Putting the pieces together, we further have

$$\widehat{\kappa}(\omega) \stackrel{(132)}{=} a^{-d} \cdot \kappa_{1/a}(\omega) \stackrel{(134)}{\leq} (2a)^{-d} \cdot \kappa_{1/(2a)}(\omega)^2 \stackrel{(133)}{=} (2a)^d \cdot \widehat{\kappa}_{\mathrm{rt}}(\omega)^2.$$

Since $\kappa_{1/(2a)} \in L^2$, we have verified the condition (128).

Wendland When $\kappa(z) = \phi_{d,s}(\|z\|_2)$ and $\widetilde{\kappa}_{\mathrm{rt}}(z) = \phi_{d,s'}(\|z\|_2)$, Wendland (2004, Thm. 10.35) implies that

$$\widehat{\kappa}(\omega) \precsim_{d,s,s'} \frac{1}{(1+\|\omega\|_2^2)^{d+2s+1}} \leq \frac{1}{(1+\|\omega\|_2^2)^{2d+4s'+2}} \precsim_{d,s,s'} \widehat{\kappa}_{\mathrm{rt}}(\omega)^2 \precsim_{d,s,s'} \frac{1}{(1+\|\omega\|_2^2)^{2d+4s'+2}},$$

with $s' \in \mathbb{N}_0, s' \leq \frac{(2s-1-d)}{4}$, thereby establishing the condition (128).

Appendix R. Online Vector Balancing in Euclidean Space

Using Thm. 3, we recover the online vector balancing result of Alweiss et al. (2021, Thm. 1.2) with improved constants and a less conservative setting of the thresholds $\mathfrak{a}_i$. Note that we capture the usual obliviousness assumption by treating the sequence $(f_i)_{i=1}^n$ as a fixed, deterministic input to Alg. 3.

Corollary 7 (Online vector balancing in Euclidean space)**.** *If $\mathcal{H} = \mathbb{R}^d$ equipped with the Euclidean dot product, each $\|f_i\|_2 \leq 1$, and each $\mathfrak{a}_i = \frac{1}{2} + \log(4n/\delta)$, then, with probability at least $1 - \delta$, the self-balancing Hilbert walk (Alg. 3) returns ψ_n satisfying*

$$\|\psi_n\|_\infty = \|\textstyle\sum_{i=1}^n \eta_i f_i\|_\infty \leq \sqrt{2\log(4d/\delta)\log(4n/\delta)}.$$

Proof Instantiate the notation of Thm. 3. With our choice of $\mathfrak{a}_i$, the signed sum representation property (ii) implies that $\psi_n = \sum_{i=1}^n \eta_i f_i$ with probability at least $1-\delta/2$. Moreover, the union bound, the functional sub-Gaussianity property (i), and the sub-Gaussian Hoeffding inequality (Wainwright, 2019, Prop. 2.5) now imply

$$\Pr(\|\psi_n\|_\infty > t \mid \mathcal{F}_n) \leq \textstyle\sum_{j=1}^d \mathbb{P}(|\langle \psi_n, e_j \rangle| > t \mid \mathcal{F}_n) \leq 2d \exp(-t^2/(2\sigma_n^2)) = \delta/2$$

for $t \triangleq \sigma_n \sqrt{2\log(4d/\delta)}$. We claim that

$$\sigma_n^2 \leq \alpha_n \triangleq \max_{j \leq n} \max\big(r_j, \|f_j\|_{\mathcal{H}}^2\big) \quad \text{with} \quad r_j \triangleq \frac{\mathfrak{a}_j^2}{2\mathfrak{a}_j - \|f_j\|_{\mathcal{H}}^2}. \tag{135}$$

Assuming this claim as given at the moment, we finish the proof. Since $\|f_j\|_{\mathcal{H}} = \|f_j\|_2 \leq 1$ for each j and $\frac{\frac{1}{2}+\log(4/\delta)}{\log(4/\delta)}$ is increasing in $\delta \in (0, 1]$,

$$\begin{aligned}\sigma_n^2 &\leq 1 \vee \max_{j\leq n} \frac{\mathfrak{a}_j^2}{2\mathfrak{a}_j - \|f_j\|_{\mathcal{H}}^2} \leq 1 \vee \max_{j\leq n} \frac{\mathfrak{a}_j^2}{2\mathfrak{a}_j - 1} = \log(4n/\delta) \frac{(\frac{1}{2}+\log(4n/\delta))^2}{2(\log(4n/\delta))^2} \\ &\leq \log(4n/\delta) \frac{(\frac{1}{2}+\log(4))^2}{2(\log(4))^2} \leq \log(4n/\delta).\end{aligned}$$

The advertised result now follows from the union bound.

We will now establish our earlier claim (135) using induction. For the base case, we have $\sigma_1^2 = \|f_1\|_{\mathcal{H}}^2 \leq \alpha_1$ from the definition of σ_1^2 (12).

Now, for the inductive case, assume that $\sigma_{i-1}^2 \leq \alpha_{i-1}$ for some $i > 1$, and define the function $g(x) = x + \|f_i\|_{\mathcal{H}}^2 (1 - x/r_i)_+$ for $x \geq 0$. Using the definition of σ_i^2 (12), we find

$$\sigma_i^2 = \sigma_{i-1}^2 + \|f_i\|_{\mathcal{H}}^2 \Big(1 + \frac{\sigma_{i-1}^2}{\mathfrak{a}_i^2}(\|f_i\|_{\mathcal{H}}^2 - 2\mathfrak{a}_i)\Big)_+ = \sigma_{i-1}^2 + \|f_i\|_{\mathcal{H}}^2 (1 - \sigma_{i-1}^2/r_i)_+ = g(\sigma_{i-1}^2).$$

Since $\alpha_{i-1} \leq \alpha_i$, our inductive assumption implies $\sigma_{i-1}^2 \leq \alpha_i$ and hence

$$\sigma_i^2 = g(\sigma_{i-1}^2) \leq \max_{x\in[0,\alpha_i]} g(x) = \max(\max_{x\in[0,r_i]} g(x), \max_{x\in(r_i,\alpha_i]} g(x)).$$

Next, we observe that g is continuous and that the derivative of g is $1 - \|f_i\|_{\mathcal{H}}^2/r_i$ for $x < r_i$ and 1 for $x > r_i$. Consequently, whenever $\|f_i\|_{\mathcal{H}}^2 < r_i$, the function g is increasing so that $\max_{x\in[0,\alpha_i]} g(x) \leq g(\alpha_i)$. On the other hand, when $\|f_i\|_{\mathcal{H}}^2 \geq r_i$, the function g is non-increasing in $[0, r_i]$ and increasing on (r_i, ∞) so that $\max_{x\in[0,r_i]} g(x) \leq g(0)$ and $\max_{x\in(r_i,\alpha_i]} \leq g(\alpha_i)$. Moreover, $g(\alpha_i) = \alpha_i$ since $\alpha_i \geq r_i$. Therefore, we always have

$$\sigma_i^2 \leq \max_{x\in[0,\alpha_i]} g(x) \leq \max(g(0), g(\alpha_i)) = \max(\|f_i\|_{\mathcal{H}}^2, \alpha_i) = \alpha_i.$$

Since $i > 1$ was arbitrary, the desired claim (135) follows by induction. ■

Appendix S. L^∞ Coresets of Phillips and Tai (2020) and Tai (2020)

Here we provide more details on the L^∞ coreset construction of Phillips and Tai (2020) and Tai (2020) discussed in Sec. 8.3. Given input points $(x_i)_{i=1}^n$, the Phillips-Tai (PT) construction forms the matrix $K = (\mathbf{k}_\star(x_i, x_j))_{i,j=1}^n$ of pairwise kernel evaluations, finds a matrix square-root $V \in \mathbb{R}^{n\times n}$ satisfying $K = VV^\top$, and augments V with a row of ones (to encourage near-halving in the next step). Then, the PT construction runs the Gram-Schmidt (GS) walk of Bansal et al. (2018) to identify approximately half of the columns of V as coreset members and *rebalances* the coreset until exactly half of the input points belong to the coreset. The GS walk and rebalancing steps are recursively repeated $\Omega(\log(n))$ times to obtain an $(n^{\frac{1}{2}}, \mathcal{O}_p(\sqrt{d}n^{-\frac{1}{2}}\sqrt{\log n}))$-$L^\infty$ coreset. The low-dimensional Gaussian kernel construction of Tai (2020) first partitions the input points into balls of radius $2\sqrt{\log n}$ and then applies the PT construction separately to each ball. The result is an $(n^{\frac{1}{2}}, \mathcal{O}_p(2^d n^{-\frac{1}{2}}\sqrt{\log(d\log n)}))$-$L^\infty$ coreset with an additional superexponential $\Omega(d^{5d})$ running time dependence.

Acknowledgments

RD acknowledges support by National Science Foundation under Grant No. DMS2023528 for the Foundations of Data Science Institute (FODSI). The authors thank François-Xavier Briol, Lucas Janson, Lingxiao Li, Chris Oates, Art Owen, and the anonymous reviewers for their valuable feedback on this work and Jeffrey Rosenthal for helpful discussions surrounding geometric ergodicity. Part of this work was done when RD was interning at Microsoft Research New England.

References

Ryan Alweiss, Yang P Liu, and Mehtaab Sawhney. Discrepancy minimization via a self-balancing walk. In *Proceedings of the 53rd Annual ACM SIGACT Symposium on Theory of Computing*, pages 14–20, 2021.

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.

Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. In *Proceedings of the 29th International Coference on International Conference on Machine Learning*, ICML'12, page 1355–1362, Madison, WI, USA, 2012. Omnipress. ISBN 9781450312851.

Nikhil Bansal, Daniel Dadush, Shashwat Garg, and Shachar Lovett. The Gram-Schmidt walk: A cure for the Banaszczyk blues. In *Proceedings of the 50th Annual ACM SIGACT Symposium on Theory of Computing*, pages 587–597, 2018.

Rémi Bardenet and Adrien Hardy. Monte Carlo with determinantal point processes. *The Annals of Applied Probability*, 30(1):368–417, 2020.

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`.

William Beckner. Inequalities in Fourier analysis. *Annals of Mathematics*, pages 159–182, 1975.

Ayoub Belhadji, Rémi Bardenet, and Pierre Chainais. Kernel quadrature with DPPs. In *Advances in Neural Information Processing Systems*, volume 32, pages 12927–12937. Curran Associates, Inc., 2019.

Ayoub Belhadji, Rémi Bardenet, and Pierre Chainais. Kernel interpolation with continuous volume sampling. In *International Conference on Machine Learning*, pages 725–735. PMLR, 2020.

Salomon Bochner. Monotone funktionen, stieltjessche integrale und harmonische analyse. *Mathematische Annalen*, 108(1):378–410, 1933.

Sergiy V Borodachov, Douglas P Hardin, and Edward B Saff. Low complexity methods for discretizing manifolds via Riesz energy minimization. *Foundations of Computational Mathematics*, 14(6):1173–1208, 2014.

Jonathan M Borwein and O-Yeat Chan. Uniform bounds for the complementary incomplete gamma function. *Mathematical Inequalities and Applications*, 12:115–121, 2009.

François-Xavier Briol, Chris J. Oates, Mark Girolami, and Michael A. Osborne. Frank-Wolfe Bayesian Quadrature: Probabilistic integration with theoretical guarantees. In *Advances in Neural Information Processing Systems*, pages 1162–1170, 2015.

Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov chain Monte Carlo*. CRC press, 2011.

V. V. Buldygin and Yu. V. Kozachenko. Sub-gaussian random variables. *Ukrainian Mathematical Journal*, 32(6):483–489, 1980. doi: 10.1007/BF01087176. URL `https://doi.org/10.1007/BF01087176`.

Trevor Campbell and Tamara Broderick. Automated scalable Bayesian inference via Hilbert coresets. *The Journal of Machine Learning Research*, 20(1):551–588, 2019.

W WL Chen and MM Skriganov. Explicit constructions in the classical mean squares problem in irregularities of point distribution. *Journal fur die Reine und Angewandte Mathematik*, (545):67–95, 2002.

W. Y. Chen, L. Mackey, J. Gorham, F-X. Briol, and C. J. Oates. Stein points. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

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*, pages 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, page 109–116, Arlington, Virginia, USA, 2010. AUAI Press. ISBN 9780974903965.

Stefano De Marchi, Robert Schaback, and Holger Wendland. Near-optimal data-independent point locations for radial basis function interpolation. *Advances in Computational Mathematics*, 23(3):317–330, 2005.

DLMF. *NIST Digital Library of Mathematical Functions*. http://dlmf.nist.gov/, Release 1.1.1 of 2021-03-15. F. W. J. Olver, A. B. Olde Daalhuis, D. W. Lozier, B. I. Schneider, R. F. Boisvert, C. W. Clark, B. R. Miller, B. V. Saunders, H. S. Cohl, and M. A. McClain, eds.

Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains*. Springer, 2018.

Rick Durrett. *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition, 2019. doi: 10.1017/9781108591034.

Raaz Dwivedi and Lester Mackey. Generalized kernel thinning. In *International Conference on Learning Representations*, 2022.

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.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

M. A. Gallegos-Herrada, D. Ledvinka, and J. S. Rosenthal. Equivalences of geometric ergodicity of markov chains, 2023.

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

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.

Ronald L. Graham, Donald Ervin Knuth, and Oren Patashnik. *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Reading, MA, Second edition, 1994. URL `https://www.csie.ntu.edu.tw/~r97002/temp/Concrete%20Mathematics%202e.pdf`.

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.

Karthik S Gurumoorthy, Amit Dhurandhar, Guillermo Cecchi, and Charu Aggarwal. Efficient data representation by selecting prototypes with importance weights. In *2019 IEEE International Conference on Data Mining (ICDM)*, pages 260–269. IEEE, 2019.

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*, pages 1165–1182, 2014.

Antoine Havet, Matthieu Lerasle, Eric Moulines, and Elodie Vernet. A quantitative McDiarmid's inequality for geometrically ergodic Markov chains. *Electronic Communications in Probability*, 25(none):1 – 11, 2020. doi: 10.1214/20-ECP286. URL `https://doi.org/10.1214/20-ECP286`.

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.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

Jonathan Huggins and Lester Mackey. Random feature Stein discrepancies. In *Advances in Neural Information Processing Systems*, pages 1899–1909, 2018.

Ferenc Huszár and David Kristjanson Duvenaud. Optimally-weighted herding is bayesian quadrature. *ArXiv*, abs/1408.2049, 2012.

V Roshan Joseph, Tirthankar Dasgupta, Rui Tuo, and CF Jeff Wu. Sequential exploration of complex surfaces using minimum energy designs. *Technometrics*, 57(1):64–74, 2015.

V Roshan Joseph, Dianpeng Wang, Li Gu, Shiji Lyu, and Rui Tuo. Deterministic sampling of expensive posteriors using minimum energy designs. *Technometrics*, 2019.

Sarang Joshi, Raj Varma Kommaraji, Jeff M Phillips, and Suresh Venkatasubramanian. Comparing distributions and shapes using the kernel distance. In *Proceedings of the twenty-seventh annual symposium on Computational geometry*, pages 47–56, 2011.

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

Toni Karvonen, Chris J Oates, and Simo Sarkka. A Bayes-Sard cubature method. In *Advances in Neural Information Processing Systems*, pages 5882–5893, 2018.

Toni Karvonen, Motonobu Kanagawa, and Simo Särkkä. On the positivity and magnitudes of Bayesian quadrature weights. *Statistics and Computing*, 29(6):1317–1333, 2019.

Rajiv Khanna and Michael W Mahoney. On linear convergence of weighted kernel herding. *arXiv preprint arXiv:1907.08410*, 2019.

Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. Examples are not enough, learn to criticize! Criticism for interpretability. *Advances in neural information processing systems*, 29, 2016.

Vladimir Koltchinskii and Evarist Giné. Random matrix approximation of spectra of integral operators. *Bernoulli*, pages 113–167, 2000.

Simon Lacoste-Julien, Fredrik Lindsten, and Francis Bach. Sequential kernel herding: Frank-Wolfe optimization for particle filtering. In *Artificial Intelligence and Statistics*, pages 544–552. PMLR, 2015.

Beatrice Laurent and Pascal Massart. Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics*, pages 1302–1338, 2000.

Qiang Liu and Jason Lee. Black-box importance sampling. In *Artificial Intelligence and Statistics*, pages 952–961. PMLR, 2017.

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.

Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability.* Springer Science & Business Media, 2012.

Ha Quang Minh. Some properties of gaussian reproducing kernel hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32(2):307–338, 2010.

Arunava Mukherjea. A remark on tonelli's theorem on integration in product spaces. *Pacific Journal of Mathematics*, 42(1):177–185, 1972.

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.

Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 79(3):695–718, 2017.

Chris J Oates, Jon Cockayne, François-Xavier Briol, and Mark Girolami. Convergence rates for a class of estimators based on Stein's method. *Bernoulli*, 25(2):1141–1159, 2019.

Anthony O'Hagan. Bayes–Hermite quadrature. *Journal of statistical planning and inference*, 29(3):245–260, 1991.

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*, pages 567–576, 2016.

Jeff M Phillips. ε-samples for kernels. In *Proceedings of the twenty-fourth annual ACM-SIAM symposium on Discrete algorithms*, pages 1622–1632. SIAM, 2013.

Jeff M Phillips and Wai Ming Tai. Improved coresets for kernel density estimates. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 2718–2727. SIAM, 2018.

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

Alireza Rezaei and Shayan Oveis Gharan. A polynomial time mcmc method for sampling from continuous determinantal point processes. In *International Conference on Machine Learning*, pages 5438–5447. PMLR, 2019.

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, 2020. URL `https://doi.org/10.7910/DVN/MDKNWM`. Accessed on Mar 23, 2021.

Marina Riabiz, Wilson Chen, Jon Cockayne, Pawel Swietach, Steven A Niederer, Lester Mackey, and Chris Oates. Optimal thinning of MCMC output. *To appear: Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2021.

Gareth O Roberts and Jeffrey S Rosenthal. General state space Markov chains and MCMC algorithms. *Probability surveys*, 1:20–71, 2004.

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

Saburou Saitoh. Applications of the general theory of reproducing kernels. In *Reproducing Kernels and Their Applications*, pages 165–188. Springer, 1999.

Gabriele Santin and Bernard Haasdonk. Convergence rate of the data-independent p-greedy algorithm in kernel-based approximation. *Dolomites Research Notes on Approximation*, 10(Special_Issue), 2017.

Larry Schumaker. *Spline functions: basic theory.* Cambridge University Press, 2007.

Abhishek Shetty, Raaz Dwivedi, and Lester Mackey. Distribution compression in near-linear time. In *International Conference on Learning Representations*, 2022.

Joel Spencer. Balancing games. *Journal of Combinatorial Theory, Series B*, 23(1):68–74, 1977.

Bharath K Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert RG Lanckriet. Hilbert space embeddings and metrics on probability measures. *Journal of Machine Learning Research*, 11(Apr):1517–1561, 2010.

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

Ingo Steinwart and Clint Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constructive Approximation*, 35(3):363–417, 2012.

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.

Xingping Sun. Conditionally positive definite functions and their application to multivariate interpolations. *Journal of approximation theory*, 74(2):159–180, 1993.

Wai Ming Tai. New nearly-optimal coreset for kernel density estimation. *arXiv preprint arXiv:2007.08031*, 2020.

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.

Paxton Turner, Jingbo Liu, and Philippe Rigollet. A statistical perspective on coreset density estimation. In *International Conference on Artificial Intelligence and Statistics*, pages 2512–2520. PMLR, 2021.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, and SciPy 1. 0 others. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 2020. doi: https://doi.org/10.1038/s41592-019-0686-2.

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. The covering number in learning theory. *Journal of Complexity*, 18(3): 739–767, 2002.