Debiased Distribution Compression

Lingxiao Li [1] **Raaz Dwivedi** [2] **Lester Mackey** [3]

Abstract

Modern compression methods can summarize a target distribution $\mathbb{P}$ more succinctly than i.i.d. sampling but require access to a low-bias input sequence like a Markov chain converging quickly to $\mathbb{P}$. We introduce a new suite of compression methods suitable for compression with biased input sequences. Given n points targeting the wrong distribution and quadratic time, Stein Kernel Thinning (SKT) returns $\sqrt{n}$ equal-weighted points with $\widetilde{O}(n^{-1/2})$ maximum mean discrepancy (MMD) to $\mathbb{P}$. For larger-scale compression tasks, Low-rank SKT achieves the same feat in sub-quadratic time using an adaptive low-rank debiasing procedure that may be of independent interest. For downstream tasks that support simplex or constant-preserving weights, Stein Recombination and Stein Cholesky achieve even greater parsimony, matching the guarantees of SKT with as few as $\text{poly-log}(n)$ weighted points. Underlying these advances are new guarantees for the quality of simplex-weighted coresets, the spectral decay of kernel matrices, and the covering numbers of Stein kernel Hilbert spaces. In our experiments, our techniques provide succinct and accurate posterior summaries while overcoming biases due to burn-in, approximate Markov chain Monte Carlo, and tempering.

1. Introduction

Distribution compression is the problem of summarizing a target probability distribution $\mathbb{P}$ with a small set of representative points. Such compact summaries are particularly valuable for tasks that incur substantial downstream computation costs per summary point, like organ and tissue modeling in which each simulation consumes thousands of CPU hours (Niederer et al., 2011).

Remarkably, modern compression methods can summarize a distribution more succinctly than i.i.d. sampling. For example, kernel thinning (KT) (Dwivedi and Mackey, 2021; 2022), Compress++ (Shetty et al., 2022), recombination (Hayakawa et al., 2023), and randomly pivoted Cholesky (Epperly and Moreno, 2024) all provide $\widetilde{O}(1/m)$ approximation error using m points, a significant improvement over the $\Omega(1/\sqrt{m})$ approximation provided by i.i.d. sampling from $\mathbb{P}$. However, each of these constructions relies on access to an accurate input sequence, like an i.i.d. sample from $\mathbb{P}$ or a Markov chain converging quickly to $\mathbb{P}$.

Much more commonly, one only has access to n *biased* sample points approximating a wrong distribution $\mathbb{Q}$. Such biases are a common occurrence in Markov chain Monte Carlo (MCMC)-based inference due to tempering (where one targets a less peaked and more dispersed distribution to achieve faster convergence, Gramacy et al., 2010), burn-in (where the initial state of a Markov chain biases the distribution of chain iterates, Cowles and Carlin, 1996), or approximate MCMC (where one runs a cheaper approximate Markov chain to avoid the prohibitive costs of an exact MCMC algorithm, e.g., Ahn et al., 2012). The Stein thinning (ST) method of Riabiz et al. (2022) was developed to provide accurate compression even when the input sample sequence provides a poor approximation to the target. ST operates by greedily thinning the input sample to minimize the maximum mean discrepancy (MMD, Gretton et al., 2012) to $\mathbb{P}$. However, ST is only known to provide an $O(1/\sqrt{m})$ approximation to $\mathbb{P}$; this guarantee is no better than that of i.i.d. sampling and a far cry from the $\widetilde{O}(1/m)$ error achieved with unbiased coreset constructions.

In this work, we address this deficit by developing new, efficient coreset constructions that provably yield better-than-i.i.d. error even when the input sample is biased. For $\mathbb{P}$ on $\mathbb{R}^d$, our primary contributions are fourfold and summarized in Tab. 1. First, for the task of equal-weighted compression, we introduce *Stein Kernel Thinning* (SKT, Alg. 1), a strategy that combines the greedy bias correction properties of ST with the unbiased compression of KT to produce $\sqrt{n}$ summary points with error $\widetilde{O}(n^{-1/2})$ in $O(n^2)$ time. In contrast, ST would require $\Omega(n)$ points to guarantee this error. Second, for larger-scale compression problems, we propose *Low-rank SKT* (Alg. 3), a strategy that combines the scalable summarization of Compress++ with

[1]MIT CSAIL [2]Cornell Tech [3]Microsoft Research New England. Correspondence to: Lingxiao Li <lingxiao@mit.edu>, Raaz Dwivedi <dwivedi@cornell.edu>, Lester Mackey <lmackey@microsoft.com>.

Preliminary work. Under review.

Table 1: **Methods for debiased distribution compression.** For each method, we report the smallest coreset size m and running time, up to logarithmic factors, sufficient to guarantee $\widetilde{O}(n^{-1/2})$ $\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}$ to $\mathbb{P}$ given a LogGrowth kernel $\boldsymbol{k}_{\mathbb{P}}$ and n slow-growing input points $\mathcal{S}_n = (x_i)_{i=1}^n$ from a fast-mixing Markov chain targeting $\mathbb{Q}$ with tails no lighter than $\mathbb{P}$ (see Thm. 1 and Def. 3). For generic slow-growing $\mathcal{S}_n$, identical guarantees hold for excess $\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}$ (2) relative to the best simplex reweighting of $\mathcal{S}_n$.

Method		Compression Type	Coreset Size m	Runtime	Source
Stein Thinning (Riabiz et al., 2022)		equal-weighted	n	$d_{\boldsymbol{k}_{\mathbb{P}}} n^2$	App. D.1
Stein Kernel Thinning	Greedy (Alg. 1)	equal-weighted	$\sqrt{n}$	$d_{\boldsymbol{k}_{\mathbb{P}}} n^2$	Thm. 3
	Low-rank (Alg. 3)			$d_{\boldsymbol{k}_{\mathbb{P}}} n^{1.5}$	Thm. 5
Stein Recombination	Greedy (Alg. 5)	simplex-weighted	poly-log(n)	$d_{\boldsymbol{k}_{\mathbb{P}}} n^2$	Thm. 6
	Low-rank (Alg. 5)			$d_{\boldsymbol{k}_{\mathbb{P}}} n + n^{1.5}$	
Stein Cholesky	Greedy (Alg. 7)	constant-preserving	poly-log(n)	$d_{\boldsymbol{k}_{\mathbb{P}}} n^2$	Thm. 7
	Low-rank (Alg. 7)			$d_{\boldsymbol{k}_{\mathbb{P}}} n + n^{1.5}$	

a new low-rank debiasing procedure (Alg. 2) to match the SKT guarantees in sub-quadratic $o(n^2)$ time.

Third, for the task of simplex-weighted compression, in which summary points are accompanied by weights in the simplex, we propose greedy and low-rank *Stein Recombination* (Alg. 5) constructions that match the guarantees of SKT with as few as poly-log(n) points. Finally, for the task of constant-preserving compression, in which summary points are accompanied by real-valued weights summing to 1, we introduce greedy and low-rank *Stein Cholesky* (Alg. 7) constructions that again match the guarantees of SKT using as few as poly-log(n) points.

Underlying these advances are new guarantees for the quality of simplex-weighted coresets (Thms. 1 and 2), the spectral decay of kernel matrices (Cor. B.1), and the covering numbers of Stein kernel Hilbert spaces (Prop. 1) that may be of independent interest. In Sec. 5, we employ our new procedures to produce compact summaries of complex target distributions given input points biased by burn-in, approximate MCMC, or tempering.

Notation We assume Borel-measurable sets and functions and define $[n] \triangleq \{1,\dots,n\}$, $\Delta_{n-1} \triangleq \{w \in \mathbb{R}^n : w \geq 0, \mathbf{1}^\top w = 1\}$, $\|x\|_0 \triangleq |\{i : x_i \neq 0\}|$, and $\|x\|_p^p \triangleq \sum_i |x_i|^p$ for $x \in \mathbb{R}^d$ and $p \geq 1$. For $x \in \mathbb{R}^d$, δ_x denotes the delta measure at x. We let $\mathcal{H}_{\boldsymbol{k}}$ denote the reproducing kernel Hilbert space (RKHS) of a kernel $\boldsymbol{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ (Aronszajn, 1950) and $\|f\|_{\boldsymbol{k}}$ denote the RKHS norm of $f \in \mathcal{H}_{\boldsymbol{k}}$. For a measure μ and separately μ-integrable $\boldsymbol{k}$ and f, we write $\mu f \triangleq \int f(x) \mathrm{d}\mu(x)$ and $\mu \boldsymbol{k}(x) \triangleq \int \boldsymbol{k}(x,y)\mathrm{d}\mu(y)$. The divergence of a differentiable matrix-valued function A is $(\nabla_x \cdot A(x))_j = \sum_i \partial_{x_i} A_{ij}(x)$. For random variables $(X_n)_{n\in\mathbb{N}}$, we say $X_n = O(f(n,\delta))$ holds with probability $\geq 1-\delta$ if $\Pr(X_n \leq C f(n,\delta)) \geq 1-\delta$ for a constant C independent of (n,δ) and all n sufficiently large. When using this notation, we view all algorithm parameters except δ as functions of n. For $A \in \mathbb{R}^{n\times n}$ and $v \in \mathbb{R}^n$, $\mathrm{diag}(A)$ and $\mathrm{diag}(v)$ are $n \times n$ diagonal matrices with A_{ii} and v_i respectively as the i-th diagonal entry.

2. Debiased Distribution Compression

Throughout, we aim to summarize a fixed target distribution $\mathbb{P}$ on $\mathbb{R}^d$ using a sequence $\mathcal{S}_n \triangleq (x_i)_{i=1}^n$ of potentially biased candidate points in $\mathbb{R}^d$.[1] Correcting for unknown biases in $\mathcal{S}_n$ requires some auxiliary knowledge of $\mathbb{P}$. For us, this knowledge comes in the form of a kernel function $\boldsymbol{k}_{\mathbb{P}}$ with known expectation under $\mathbb{P}$. Without loss of generality, we can take this kernel mean to be identically zero.[2]

Assumption 1 (Mean-zero kernel)**.** *For some* $\mathfrak{p} \geq 1/2$*,* $\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}_{\mathbb{P}}(x,x)^{\mathfrak{p}}] < \infty$ *and* $\mathbb{P}\boldsymbol{k}_{\mathbb{P}} \equiv 0$.

Given a target compression size m, our goal is to output an weight vector $w \in \mathbb{R}^n$ with $\|w\|_0 \leq m$, $\mathbf{1}_n^\top w = 1$, and $o(m^{-1/2})$ (better-than-i.i.d.) maximum mean discrepancy (MMD) to $\mathbb{P}$:

$$\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\textstyle\sum_{i=1}^n w_i \delta_{x_i}, \mathbb{P}) \triangleq \sqrt{\textstyle\sum_{i,j=1}^n w_i w_j \boldsymbol{k}_{\mathbb{P}}(x_i,x_j)}.$$

We consider three standard compression tasks with $\|w\|_0 \leq m$. In *equal-weighted compression* one selects m possibly repeated points from $\mathcal{S}_n$ and assigns each a weight of $\frac{1}{m}$; because of repeats, the induced weight vector over $\mathcal{S}_n$ satisfies $w \in \Delta_{n-1} \cap (\frac{\mathbb{N}_0}{m})^n$. In *simplex-weighted compression* we allow any $w \in \Delta_{n-1}$, and in *constant-preserving compression* we simply enforce $\mathbf{1}_n^\top w = 1$.

[1] Our coreset constructions will in fact apply to any sample space, but our analysis will focus on $\mathbb{R}^d$.

[2] For $\mathbb{P}\boldsymbol{k}_{\mathbb{P}} \not\equiv 0$, the kernel $\boldsymbol{k}_{\mathbb{P}}'(x,y) = \boldsymbol{k}_{\mathbb{P}}(x,y) - \mathbb{P}\boldsymbol{k}_{\mathbb{P}}(x) - \mathbb{P}\boldsymbol{k}_{\mathbb{P}}(y) + \mathbb{P}\mathbb{P}\boldsymbol{k}_{\mathbb{P}}$ satisfies $\mathbb{P}\boldsymbol{k}_{\mathbb{P}}' \equiv 0$ and $\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}'} = \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}$.

When making big O statements, we will treat $\mathcal{S}_n$ as the prefix of an infinite sequence $\mathcal{S}_\infty \triangleq (x_i)_{i\in\mathbb{N}}$. We also write $\boldsymbol{k}_{\mathbb{P}}(\mathcal{S}_n[\mathrm{J}], \mathcal{S}_n[\mathrm{J}]) \triangleq [\boldsymbol{k}_{\mathbb{P}}(x_i, x_j)]_{i,j\in\mathrm{J}}$ for the principal kernel submatrix with indices $\mathrm{J} \subseteq [n]$.

2.1. Kernel assumptions

Many practical *Stein kernel* constructions are available for generating mean-zero kernels for a target $\mathbb{P}$ (Chwialkowski et al., 2016; Liu et al., 2016; Gorham and Mackey, 2017; Gorham et al., 2019; Barp et al., 2019; Yang et al., 2018; Afzali and Muthukumarana, 2023). We will use the most prominent of these Stein kernels as a running example:

Definition 1 (Stein kernel)**.** *Given a differentiable base kernel $\boldsymbol{k}$ and a symmetric positive semidefinite matrix M, the* Stein kernel *$\boldsymbol{k}_p : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ for $\mathbb{P}$ with positive differentiable Lebesgue density p is defined as*

$$\boldsymbol{k}_p(x, y) \triangleq \tfrac{1}{p(x)p(y)} \nabla_x \cdot \nabla_y \cdot (p(x) M \boldsymbol{k}(x, y) p(y)).$$

While our algorithms apply to any mean zero kernel, our guarantees adapt to the underlying smoothness of the kernels. Our next definition and assumption make this precise.

Definition 2 (Covering number)**.** *For a kernel $\boldsymbol{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $\mathcal{B}_{\boldsymbol{k}} \triangleq \{f \in \mathcal{H}_{\boldsymbol{k}} : \|f\|_{\boldsymbol{k}} \le 1\}$, a set $A \subset \mathbb{R}^d$, and $\varepsilon > 0$, the covering number $\mathcal{N}_{\boldsymbol{k}}(A, \varepsilon)$ is the minimum cardinality of all sets $\mathcal{C} \subset \mathcal{B}_{\boldsymbol{k}}$ satisfying*

$$\mathcal{B}_{\boldsymbol{k}} \subset \textstyle\bigcup_{h\in\mathcal{C}} \{g \in \mathcal{B}_{\boldsymbol{k}} : \sup_{x\in A} |h(x) - g(x)| \le \varepsilon\}.$$

Assumption $(\boldsymbol{\alpha}, \boldsymbol{\beta})$-kernel. *For some $\mathfrak{C}_d > 0$, all $r > 0$ and $\varepsilon \in (0, 1)$, and $\mathcal{B}_2(r) \triangleq \{x \in \mathbb{R}^d : \|x\|_2 \le r\}$, a kernel $\boldsymbol{k}$ is either* POLYGROWTH(α, β)*, i.e.,*

$$\log \mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(r), \varepsilon) \le \mathfrak{C}_d (1/\varepsilon)^\alpha (r + 1)^\beta,$$

with $\alpha < 2$ or LOGGROWTH(α, β)*, i.e.,*

$$\log \mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(r), \varepsilon) \le \mathfrak{C}_d \log(e/\varepsilon)^\alpha (r + 1)^\beta.$$

In Cor. B.1 we show that the eigenvalues of kernel matrices with POLYGROWTH and LOGGROWTH kernels have polynomial and exponential decay respectively. Dwivedi and Mackey (2022, Prop. 2) showed that all sufficiently differentiable kernels satisfy the POLYGROWTH condition and that bounded radially analytic kernels are LOGGROWTH. Our next result, proved in App. B.2, shows that a Stein kernel $\boldsymbol{k}_p$ can inherit the growth properties of its base kernel even if $\boldsymbol{k}_p$ is itself unbounded and non-smooth.

Proposition 1 (Stein kernel growth rates)**.** *A Stein kernel $\boldsymbol{k}_p$ with $\sup_{\|x\|_2 \le r} \|\nabla \log p(x)\|_2 = O(r^{d_\ell})$ for $d_\ell \ge 0$ is*

(a) LOGGROWTH$(d+1, 2d+\delta)$ *for any $\delta > 0$ if the base kernel $\boldsymbol{k}$ is radially analytic (Def. B.3) and*

(b) POLYGROWTH$(\frac{d}{s-1}, (1+\frac{d_\ell}{s})d)$ *if the base kernel $\boldsymbol{k}$ is s-times continuously differentiable (Def. B.2) for $s > 1$.*

Notably, the popular Gaussian (Ex. B.1) and inverse multiquadric (Ex. B.2) base kernels satisfy the LOGGROWTH preconditions, while Matérn, B-spline, sinc, sech, and Wendland's compactly supported kernels satisfy the POLYGROWTH precondition (Dwivedi and Mackey, 2022, Prop. 3). To our knowledge, Prop. 1 provides the first covering number bounds and eigenvalue decay rates for the (typically unbounded) Stein kernels $\boldsymbol{k}_p$.

2.2. Input point desiderata

Our primary desideratum for the input points is that they can be debiased into an accurate estimate of $\mathbb{P}$. Indeed, our high-level strategy for debiased compression is to first use $\boldsymbol{k}_{\mathbb{P}}$ to debias the input points into a more accurate approximation of $\mathbb{P}$ and then compress that approximation into a more succinct representation. Fortunately, even when the input $\mathcal{S}_n$ targets a distribution $\mathbb{Q} \ne \mathbb{P}$, effective debiasing is often achievable via simplex reweighting, i.e., by solving the convex optimization problem

$$w_{\mathrm{OPT}} \in \arg\min_{w\in\Delta_{n-1}} \textstyle\sum_{i,j=1}^n w_i w_j \boldsymbol{k}_{\mathbb{P}}(x_i, x_j) \qquad (1)$$
$$\text{with}\ \ \mathrm{MMD}_{\mathrm{OPT}} \triangleq \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\textstyle\sum_{i=1}^n w_{\mathrm{OPT}i} \delta_{x_i}, \mathbb{P}).$$

For example, Hodgkinson et al. (2020, Thm. 1b) showed that simplex reweighting can correct for biases due to off-target i.i.d. or MCMC sampling. Our next result (proved in App. C.2) significantly relaxes their conditions.

Theorem 1 (Debiasing via simplex reweighting)**.** *Consider a kernel $\boldsymbol{k}_{\mathbb{P}}$ satisfying Assum. 1 with $\mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$ separable, and suppose $(x_i)_{i=1}^\infty$ are the iterates of a homogeneous ϕ-irreducible geometrically ergodic Markov chain (Gallegos-Herrada et al., 2023, Thm. 1) with stationary distribution $\mathbb{Q}$ and initial distribution absolutely continuous with respect to $\mathbb{P}$. If $\mathbb{E}_{x\sim\mathbb{P}}[\frac{d\mathbb{P}}{d\mathbb{Q}}(x)^{2\mathfrak{q}-1} \boldsymbol{k}_{\mathbb{P}}(x, x)^{\mathfrak{q}}] < \infty$ for some $\mathfrak{q} > 1$ then $\mathrm{MMD}_{\mathrm{OPT}} = O(n^{-1/2})$ in probability.*

Remark 1. $\mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$ is separable whenever $\boldsymbol{k}_{\mathbb{P}}$ is continuous (Steinwart and Christmann, 2008, Lem. 4.33).

Since n points sampled i.i.d. from $\mathbb{P}$ have $\Theta(n^{-1/2})$ root mean squared MMD (see Prop. C.1), Thm. 1 shows that a debiased off-target sample can be as accurate as a direct sample from $\mathbb{P}$. Moreover, Thm. 1 applies to many practical examples. The simplest example of a geometrically ergodic chain is i.i.d. sampling from $\mathbb{Q}$, but geometric ergodicity has also been established for a variety of popular Markov chains including random walk Metropolis (Roberts and Tweedie, 1996, Thm. 3.2), independent Metropolis-Hastings (Atchadé and Perron, 2007, Thm. 2.2), the unadjusted Langevin algorithm (Durmus and Moulines, 2017, Prop. 8), the Metropolis-adjusted

Langevin algorithm (Durmus and Moulines, 2022, Thm. 1), Hamiltonian Monte Carlo (Durmus et al., 2020, Thm. 10 and Thm. 11), stochastic gradient Langevin dynamics (Li et al., 2023, Thm. 2.1), and the Gibbs sampler (Johnson, 2009). Moreover, for $\mathbb{Q}$ absolutely continuous with respect to $\mathbb{P}$, the importance weight $\frac{d\mathbb{P}}{d\mathbb{Q}}$ is typically bounded or slowly growing when the tails of $\mathbb{Q}$ are not much lighter than those of $\mathbb{P}$.

Remarkably, under more stringent conditions, Thm. 2 (proved in App. C.3) shows that simplex reweighting can decrease MMD to $\mathbb{P}$ at an even-faster-than-i.i.d. rate.

Theorem 2 (Better-than-i.i.d. debiasing via simplex reweighting)**.** *Consider a kernel $\boldsymbol{k}_{\mathbb{P}}$ satisfying Assum. 1 with $\mathfrak{p} = 2$ and points $(x_i)_{i=1}^{\infty}$ drawn i.i.d. from a distribution $\mathbb{Q}$ with $\frac{d\mathbb{P}}{d\mathbb{Q}}$ bounded. If $\mathbb{E}[\boldsymbol{k}_{\mathbb{P}}(x_1, x_1)^{\mathfrak{q}}] < \infty$ for some $\mathfrak{q} > 3$, then $\mathbb{E}[\mathrm{MMD}_{\mathrm{OPT}}^2] = o(n^{-1})$.*

The work of Liu and Lee (2017, Thm. 3.3) also established $o(n^{-1/2})$ MMD error for simplex reweighting but only under a uniformly bounded eigenfunctions assumption that is often violated (Minh, 2010, Thm. 1, Zhou, 2002, Ex. 1) and difficult to verify (Steinwart and Scovel, 2012).

Our remaining results make no particular assumption about the input points but rather upper bound the excess MMD

$$\Delta\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(w) \triangleq \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\textstyle\sum_{i\in[n]} w_i \delta_{x_i}, \mathbb{P}) - \mathrm{MMD}_{\mathrm{OPT}} \tag{2}$$

of a candidate weighting w in terms of the input point radius $R_n \triangleq \max_{i\in[n]} \|x_i\|_2 \vee 1$ and kernel radius $\|\boldsymbol{k}_{\mathbb{P}}\|_n \triangleq \max_{i\in[n]} \boldsymbol{k}_{\mathbb{P}}(x_i, x_i)$. While these results apply to *any* input points, we will consider the following running example of *slow-growing* input points throughout the paper.

Definition 3 (Slow-growing input points)**.** *We say $\mathcal{S}_n$ is γ-slow-growing if $R_n = O((\log n)^{\gamma})$ for some $\gamma \geq 0$ and $\|\boldsymbol{k}_{\mathbb{P}}\|_n = \widetilde{O}(1)$.*

Notably, $\mathcal{S}_n$ is 1-slow-growing with probability 1 when $\boldsymbol{k}_{\mathbb{P}}(x, x)$ is polynomially bounded by $\|x\|_2$ and the input points are drawn from a homogeneous ϕ-irreducible geometrically ergodic Markov chain with a sub-exponential target $\mathbb{Q}$, i.e., $\mathbb{E}[e^{c\|x\|_2}] < \infty$ for some $c > 0$ (Dwivedi and Mackey, 2021, Prop. 2). For a Stein kernel $\boldsymbol{k}_p$ (Def. 1), by Prop. B.3, $\boldsymbol{k}_p(x, x)$ is polynomially bounded by $\|x\|_2$ if $\boldsymbol{k}(x, x)$, $\|\nabla_x \nabla_y \boldsymbol{k}(x, x)\|_2$, and $\|\nabla \log p(x)\|_2$ are all polynomially bounded by $\|x\|_2$. Moreover, $\|\nabla \log p(x)\|_2$ is automatically polynomially bounded by $\|x\|_2$ when $\nabla \log p$ is Lipschitz or, more generally, pseudo-Lipschitz (Erdogdu et al., 2018, Eq. (2.5)).

2.3. Debiased compression via Stein Kernel Thinning

Off-the-shelf solvers based on mirror descent and Frank Wolfe can solve the convex debiasing program (1) in $O(n^3)$ time by generating weights with $O(n^{-1/2}\|\boldsymbol{k}_{\mathbb{P}}\|_n)$ excess MMD (Liu and Lee, 2017). We instead employ a more efficient, greedy debiasing strategy based on Stein thinning (ST). After n rounds, ST outputs an equal-weighted coreset of size n with $O(n^{-1/2}\|\boldsymbol{k}_{\mathbb{P}}\|_n)$ excess MMD (Riabiz et al., 2022, Thm. 1). Moreover, while the original implementation of Riabiz et al. (2022) has cubic runtime, our implementation (Alg. D.1) based on sufficient statistics improves the runtime to $O(n^2 d_{\boldsymbol{k}_{\mathbb{P}}})$ where $d_{\boldsymbol{k}_{\mathbb{P}}}$ denotes the runtime of a single kernel evaluation.[3]

The equal-weighted output of ST serves as the perfect input for the kernel thinning (KT) algorithm which compresses an equal-weighted sample of size n into a coreset of any target size $m \leq n$ in $O(n^2 d_{\boldsymbol{k}_{\mathbb{P}}})$ time. We adapt the target KT algorithm slightly to target MMD error to $\mathbb{P}$ and to include a baseline ST coreset of size m in the KT-SWAP step (see Alg. D.3). Combining the two routines we obtain Stein Kernel Thinning (SKT), our first solution for equal-weighted debiased distribution compression:

Algorithm 1 Stein Kernel Thinning (SKT)

Input: mean-zero kernel $\boldsymbol{k}_{\mathbb{P}}$, points $\mathcal{S}_n$, output size m, KT failure probability δ
$n' \leftarrow m\, 2^{\lceil \log_2 \frac{n}{m} \rceil}$
$w \leftarrow \mathrm{SteinThinning}(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, n')$
$w_{\mathrm{SKT}} \leftarrow \mathrm{KernelThinning}(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, n', w, m, \delta)$
Return: $w_{\mathrm{SKT}} \in \Delta_{n-1} \cap (\frac{\mathbb{N}_0}{m})^n$ $\quad\triangleright$ *hence* $\|w_{\mathrm{SKT}}\|_0 \leq m$

Our next result, proved in App. D.3, shows that SKT yields better-than-i.i.d. excess MMD whenever the radii (R_n and $\|\boldsymbol{k}_{\mathbb{P}}\|_n$) and kernel covering number exhibit slow growth.

Theorem 3 (MMD guarantee for SKT)**.** *Given a kernel $\boldsymbol{k}_{\mathbb{P}}$ satisfying Assums. 1 and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$-kernel, Stein Kernel Thinning (Alg. 1) outputs w_{SKT} in $O(n^2 d_{\boldsymbol{k}_{\mathbb{P}}})$ time satisfying*

$$\Delta\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(w_{\mathrm{SKT}}) = O\Big(\frac{\sqrt{\|\boldsymbol{k}_{\mathbb{P}}\|_n \ell_\delta \cdot \log n \cdot R_n^{\beta} G_m^{\alpha}}}{\min(m, \sqrt{n})}\Big)$$

with probability at least $1 - \delta$, where $\ell_\delta \triangleq \log^2(\frac{e}{\delta})$ and

$$G_m \triangleq \begin{cases} \log(em) & \textsc{LogGrowth}(\alpha, \beta), \\ m & \textsc{PolyGrowth}(\alpha, \beta). \end{cases}$$

Example 1. Under the assumptions of Thm. 3 with γ-slow-growing input points (Def. 3), LOGGROWTH $\boldsymbol{k}_{\mathbb{P}}$, and a coreset size $m \leq \sqrt{n}$, SKT delivers $\widetilde{O}(m^{-1})$ excess MMD with high probability, a significant improvement over the $\Omega(m^{-1/2})$ error rate of i.i.d. sampling.

Remark 2. When $m < \sqrt{n}$, we can uniformly subsample or, in the case of MCMC inputs, *standard thin* (i.e., keep only every $\frac{n}{m^2}$-th point of) the input sequence down to size

[3] Often, $d_{\boldsymbol{k}_{\mathbb{P}}} = \Theta(d)$ as in the case of Stein kernels (App. I.1).

m^2 before running SKT to reduce runtime while incurring only $O(m^{-1})$ excess error. The same holds for the LSKT algorithm introduced in Sec. 3.

3. Accelerated Debiased Compression

To enable larger-scale debiased compression, we next introduce a sub-quadratic-time version of SKT built via a new low-rank debiasing scheme and the near-linear-time compression algorithm of Shetty et al. (2022).

3.1. Fast bias correction via low-rank approximation

At a high level, our approach to accelerated debiasing involves four components. First, we form a rank-r approximation $FF^\top$ of the kernel matrix $K = \boldsymbol{k}_{\mathbb{P}}(\mathcal{S}_n, \mathcal{S}_n)$ in $O(nrd_{\boldsymbol{k}_{\mathbb{P}}} + nr^2)$ time using a weighted extension (WeightedRPCholesky, Alg. F.1) of the randomly pivoted Cholesky algorithm of Chen et al. (2022, Alg. 2.1). Second, we correct the diagonal to form $K' = FF^\top + \mathrm{diag}(K - FF^\top)$. Third, we solve the reweighting problem (1) with K' substituted for K using T iterations of accelerated entropic mirror descent (AMD, Wang et al., 2023, Alg. 14 with $\phi(w) = \sum_i w_i \log w_i$). The acceleration ensures $O(1/T^2)$ suboptimality after T iterations, and each iteration takes only $O(nr)$ time thanks to the low-rank plus diagonal approximation. Finally, we repeat this three-step procedure Q times, each time using the weights outputted by the prior round to update the low-rank approximation $\widehat{K}$. On these subsequent adaptive rounds, WeightedRPCholesky approximates the leading subspace of a *weighted* kernel matrix $\mathrm{diag}(\sqrt{\tilde{w}}) K \,\mathrm{diag}(\sqrt{\tilde{w}})$ before undoing the row and column reweighting. Since each round's weights are closer to optimal, this adaptive updating has the effect of upweighting more relevant subspaces for subsequent debiasing. For added sparsity, we prune the weights outputted by the prior round using stratified residual resampling (Resample, Alg. E.3, Douc and Cappé, 2005). Our complete Low-rank Debiasing (LD) scheme, summarized in Alg. 2, enjoys $o(n^2)$ runtime whenever $r = o(n^{1/2})$, $T = O(n^{1/2})$, and $Q = O(1)$.

Moreover, our next result, proved in App. F.1, shows that LD provides i.i.d.-level precision whenever $T \geq \sqrt{n}$, $Q = O(1)$, and r grows appropriately with the input radius and kernel covering number.

Assumption $(\boldsymbol{\alpha},\boldsymbol{\beta})$-params. *The kernel $\boldsymbol{k}_{\mathbb{P}}$ satisfies Assums. 1 and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$-kernel, the output size and rank $m, r \geq (\frac{\mathfrak{C}_d R_n^\beta + 1}{\sqrt{\log 2}} + 2\sqrt{\log 2})^2$, the AMD step count $T \geq \sqrt{n}$, and the adaptive round count $Q = O(1)$.*[4]

[4]To unify the presentation of our results, Assum. $(\boldsymbol{\alpha},\boldsymbol{\beta})$-params constrains all common algorithm input parameters with the understanding that the conditions are enforced only when the input is relevant to a given algorithm.

Algorithm 2 Low-rank Debiasing (LD)

Input: mean-zero kernel $\boldsymbol{k}_{\mathbb{P}}$, points $\mathcal{S}_n = (x_i)_{i=1}^n$, rank r, AMD steps T, adaptive rounds Q
$w^{(0)} \leftarrow (\frac{1}{n}, \dots, \frac{1}{n}) \in \mathbb{R}^n$
for $q = 1$ **to** Q **do**
  $\tilde{w} \leftarrow$ Resample$(w^{(q-1)}, n)$
  $\mathtt{I}, F \leftarrow$ WeightedRPCholesky$(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, \tilde{w}, r)$
  $K' \leftarrow FF^\top + \mathrm{diag}(\boldsymbol{k}_{\mathbb{P}}(\mathcal{S}_n, \mathcal{S}_n)) - \mathrm{diag}(FF^\top)$
  $w^{(q)} \leftarrow$ AMD$(K', T, \tilde{w}, \mathtt{AGG} = \mathbb{1}_{q>1})$
  if $(w^{(q)})^\top K' w^{(q)} > \tilde{w}^\top K' \tilde{w}$ **then** $w^{(q)} \leftarrow \tilde{w}$
end for
Return: $w_{\mathrm{LD}} \leftarrow w^{(Q)} \in \Delta_{n-1}$

Theorem 4 (Debiasing guarantee for LD)**.** *Under Assum. $(\boldsymbol{\alpha},\boldsymbol{\beta})$-params, Low-rank Debiasing (Alg. 2) takes $O((d_{\boldsymbol{k}_{\mathbb{P}}} + r + T)nr)$ time to output w_{LD} satisfying*

$$\Delta\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(w_{\mathrm{LD}}) = O\Bigg(\sqrt{\tfrac{\|\boldsymbol{k}_{\mathbb{P}}\|_n \max(\log n, 1/\delta)}{n}} + \sqrt{\tfrac{nH_{n,r}}{\delta}}\Bigg)$$

with probability at least $1 - \delta$, for any $\delta \in (0,1)$ and $H_{n,r}$ defined in (46) that satisfies

$$H_{n,r} = \begin{cases} O\big(\sqrt{r}\,(\frac{R_n^{2\beta}}{r})^{\frac{1}{\alpha}}\big) & \textsc{PolyGrowth}(\alpha,\beta), \\ O\big(\sqrt{r}\exp(-(\frac{0.83\sqrt{r} - 2.39}{\mathfrak{C}_d R_n^\beta})^{\frac{1}{\alpha}})\big) & \textsc{LogGrowth}(\alpha,\beta). \end{cases}$$

Example 2. Under the assumptions of Thm. 4 with γ-slow-growing input points (Def. 3), $\textsc{LogGrowth}$ $\boldsymbol{k}_{\mathbb{P}}$, $T = \Theta(\sqrt{n})$, and $r = (\log n)^{2(\alpha + \beta\gamma) + \epsilon}$ for any $\epsilon > 0$, LD delivers $\widetilde{O}(n^{-1/2})$ excess MMD with high probability in $\widetilde{O}(n^{1.5})$ time.

3.2. Fast debiased compression via Low-rank Stein KT

To achieve debiased compression in sub-quadratic time, we next propose Low-rank SKT (Alg. 3). LSKT debiases the input using LD, converts the LD output into an equal-weighted coreset using Resample, and finally combines KT with the divide-and-conquer Compress++ framework (Shetty et al., 2022) to compress n equal-weighted points into $\sqrt{n}$ in near-linear time.

Algorithm 3 Low-rank Stein Kernel Thinning (LSKT)

Input: mean-zero kernel $\boldsymbol{k}_{\mathbb{P}}$, points $\mathcal{S}_n = (x_i)_{i=1}^n$, rank r, AGM steps T, adaptive rounds Q, oversampling parameter $\mathfrak{g}$, failure prob. δ
$w \leftarrow$ Low-rankDebiasing$(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, r, T, Q)$
$n' \leftarrow 4^{\lceil \log_4 n \rceil}$, $m \leftarrow \sqrt{n'}$ ▷ *output size* $\sqrt{n} \leq m < 2\sqrt{n}$
$w \leftarrow$ Resample(w, n')
$w_{\mathrm{LSKT}} \leftarrow$ KT-Compress++$(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, n', w, \mathfrak{g}, \frac{\delta}{3})$
Return: $w_{\mathrm{LSKT}} \in \Delta_{n-1} \cap (\frac{\mathbb{N}_0}{m})^n$ ▷ *hence* $\|w_{\mathrm{LSKT}}\|_0 \leq m$

Our next result (proved in App. F) shows that LSKT can provide better-than-i.i.d. excess MMD in $o(n^2)$ time.

Theorem 5 (MMD guarantee for LSKT). *Under Assum. $(\boldsymbol{\alpha},\boldsymbol{\beta})$-params, Low-rank SKT (Alg. 3) with $\mathfrak{g} \in [\log_2 \log(n+1) + 3.1, \log_4(\sqrt{n}/\log n)]$ and $\delta \in (0,1)$ outputs w_{LSKT} in $O((d_{\boldsymbol{k}_{\mathbb{P}}}+r+T)nr+d_{\boldsymbol{k}_{\mathbb{P}}}n^{1.5})$ time satisfying, with probability at least $1-\delta$,*

$$\Delta\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(w_{\mathrm{LSKT}}) = O\left(\sqrt{\tfrac{\|\boldsymbol{k}_{\mathbb{P}}\|_n \max(1/\delta, \ell_\delta(\log n) n^{\gamma\beta} G^{\alpha}_{\sqrt{n}})}{n}} + \sqrt{\tfrac{nH_{n,r}}{\delta}}\right),$$

for G_m, $H_{n,r}$ as in Thms. 3 and 5.

Example 3. Under the assumptions of Thm. 5 with γ-slow-growing input points (Def. 3), LogGrowth $\boldsymbol{k}_{\mathbb{P}}$, $T = \Theta(\sqrt{n})$, and $r = (\log n)^{2(\alpha+\beta\gamma)+\epsilon}$ for any $\epsilon > 0$, LSKT delivers $\widetilde{O}(n^{-1/2})$ excess MMD with high probability in $\widetilde{O}(n^{1.5})$ time with a coreset of size $m \in [\sqrt{n}, 2\sqrt{n})$.

4. Weighted Debiased Compression

The prior sections developed debiased equal-weighted coresets with better-than-i.i.d. compression guarantees. In this section, we match those guarantees with significantly smaller weighted coresets.

4.1. Simplex-weighted coresets via Stein Recombination

Algorithm 4 Recombination Thinning (RT)

Input: mean-zero kernel $\boldsymbol{k}_{\mathbb{P}}$, points $\mathcal{S}_n = (x_i)_{i=1}^n$, weights $w \in \Delta_{n-1}$, output size m
$\tilde{w} \leftarrow$ Resample(w, n)
$\mathtt{I}, F \leftarrow$ WeightedRPCholesky$(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, \tilde{w}, m-1)$
$w' \leftarrow$ Recombination$([F, \mathbf{1}_n]^\top, \tilde{w})$ ▷ $[F, \mathbf{1}_n]^\top \in \mathbb{R}^{m\times n}$
▷ $F^\top \tilde{w} = F^\top w'$, $w' \in \Delta_{n-1}$, *and* $\|w'\|_0 \le m$
$w'' \leftarrow$ KT-Swap-LS$(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, w', \mathrm{SPLX})$; $\mathtt{J} \leftarrow \{i\colon w''_i > 0\}$
$w''[\mathtt{J}] \leftarrow \operatorname{argmin}_{w' \in \Delta_{|\mathtt{J}|-1}} w'^\top \boldsymbol{k}_{\mathbb{P}}(\mathcal{S}_n[\mathtt{J}], \mathcal{S}_n[\mathtt{J}])w'$ ▷ *use any $O(|\mathtt{J}|^3)$ quadratic programming solver*
Return: $w_{\mathrm{RT}} \leftarrow w'' \in \Delta_{n-1}$ with $\|w_{\mathrm{RT}}\|_0 \le m$

Inspired by the coreset constructions of Hayakawa et al. (2022; 2023), we first introduce a simplex-weighted compression algorithm, RecombinationThinning (RT, Alg. 4), suitable for summarizing a debiased input sequence. To produce a coreset given input weights $w \in \Delta_{n-1}$, RT first prunes small weights using Resample and then uses WeightedRPCholesky to identify $m-1$ test vectors that capture most of the variability in the weighted kernel matrix. Next, Recombination (Alg. G.1) (Tchernychova, 2016, Alg. 1) identifies a sparse simplex vector w' with $\|w'\|_0 \le m$ that exactly matches the inner product of its input with each of the test vectors. Then, we run KT-Swap-LS (Alg. G.2), a new, line-search version of KT-SWAP (Dwivedi and Mackey, 2021, Alg. 1b) that greedily improves MMD to $\mathbb{P}$ while maintaining both the sparsity and simplex constraint of its input. Finally, we optimize the weights of the remaining support points using any cubic-time quadratic programming solver.

In Prop. G.1 we show that RT runs in time $O((d_{\boldsymbol{k}_{\mathbb{P}}} + m)nm + m^3 \log n)$ and nearly preserves the MMD of its input whenever m grows appropriately with the kernel covering number. Combining RT with SteinThinning or Low-rankDebiasing in Alg. 5, we obtain Stein Recombination (SR) and Low-rank SR (LSR), our approaches to debiased simplex-weighted compression. Remarkably, SR and LSR can match the MMD error rates established for SKT and LSKT using substantially fewer coreset points, as our next result (proved in App. G.2) shows.

Algorithm 5 (Low-rank) Stein Recombination (SR / LSR)

Input: mean-zero kernel $\boldsymbol{k}_{\mathbb{P}}$, points $\mathcal{S}_n$, output size m, rank r, AGM steps T, adaptive rounds Q

$$w \leftarrow \begin{cases} \text{Low-rankDebiasing}(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, r, T, Q) & \text{if low-rank} \\ \text{SteinThinning}(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n) & \text{otherwise} \end{cases}$$

$w_{\mathrm{SR}} \leftarrow$ RecombinationThinning$(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, w, m)$
Return: $w_{\mathrm{SR}} \in \Delta_{n-1}$ with $\|w_{\mathrm{SR}}\|_0 \le m$

Theorem 6 (MMD guarantee for SR/ LSR). *Under Assum. $(\boldsymbol{\alpha},\boldsymbol{\beta})$-params, Stein Recombination (Alg. 5) takes $O(d_{\boldsymbol{k}_{\mathbb{P}}}n^2+(d_{\boldsymbol{k}_{\mathbb{P}}}+m)nm+m^3\log n)$ to output w_{SR}, and Low-rank SR takes $O((d_{\boldsymbol{k}_{\mathbb{P}}}+r+T)nr+(d_{\boldsymbol{k}_{\mathbb{P}}}+m)nm+m^3\log n)$ time to output w_{LSR}. Moreover, for any $\delta \in (0,1)$ and $H_{n,r}$ as in Thm. 4, each of the following bounds holds (separately) with probability at least $1-\delta$:*

$$\Delta\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(w_{\mathrm{SR}}) = O\left(\sqrt{\tfrac{\|\boldsymbol{k}_{\mathbb{P}}\|_n(\log n \vee \frac{1}{\delta})}{n} + \tfrac{nH_{n,m}}{\delta}}\right) \text{ and}$$

$$\Delta\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(w_{\mathrm{LSR}}) = O\left(\sqrt{\tfrac{\|\boldsymbol{k}_{\mathbb{P}}\|_n(\log n \vee \frac{1}{\delta})}{n} + \tfrac{n(H_{n,m}+H_{n,r})}{\delta}}\right).$$

Example 4. Instantiate the assumptions of Thm. 6 with γ-slow-growing input points (Def. 3), LogGrowth $\boldsymbol{k}_{\mathbb{P}}$, and a heavily compressed coreset size $m = (\log n)^{2(\alpha+\beta\gamma)+\epsilon}$ for any $\epsilon > 0$. Then SR delivers $\widetilde{O}(n^{-1/2})$ excess MMD with high probability in $O(n^2)$ time, and LSR with $r = m$ and $T = \Theta(\sqrt{n})$ achieves the same in $\widetilde{O}(n^{1.5})$ time.

4.2. Constant-preserving coresets via Stein Cholesky

For applications supporting negative weights, we next introduce a constant-preserving compression algorithm, CholeskyThinning (CT, Alg. 6), suitable for summarizing a debiased input sequence. CT first applies WeightedRPCholesky to a *constant-regularized kernel* $\boldsymbol{k}_{\mathbb{P}}(x, y) + c$ to select an initial coreset and then uses a combination of

KT-Swap-LS and closed-form optimal constant-preserving reweighting to greedily refine the support and weights. The regularized kernel ensures that WeightedRPCholesky, originally developed for compression with unconstrained weights, also yields a high-quality coreset when its weights are constrained to sum to 1, and our CT standalone analysis (Prop. H.1) improves upon the runtime and error guarantees of RT. In Alg. 7, we combine CT with SteinThinning or Low-rankDebiasing to obtain Stein Cholesky (SC) and Low-rank SC (LSC), our approaches to debiased constant-preserving compression. Our MMD guarantees for SC and LSC (proved in App. H.2) improve upon the rates of Thm. 6.

Algorithm 6 Cholesky Thinning (CT)

Input: mean-zero kernel $\boldsymbol{k}_{\mathbb{P}}$, points $\mathcal{S}_n = (x_i)_{i=1}^n$, weights $w \in \Delta_{n-1}$, output size m
$c \leftarrow$ AVERAGE(Largest m entries of $(\boldsymbol{k}_{\mathbb{P}}(x_i, x_i))_{i=1}^n$)
$\mathtt{I}, F \leftarrow$ WeightedRPCholesky$(\boldsymbol{k}_{\mathbb{P}}+c, \mathcal{S}_n, w, m)$; $w \leftarrow \mathbf{0}_n$
$w[\mathtt{I}] \leftarrow \operatorname{argmin}_{w' \in \mathbb{R}^{|\mathtt{I}|}:\sum_i w'_i=1} w'^\top \boldsymbol{k}_{\mathbb{P}}(\mathcal{S}_n[\mathtt{I}], \mathcal{S}_n[\mathtt{I}])w'$
$w \leftarrow$ KT-Swap-LS$(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, w, \text{CP})$; $\mathtt{I} \leftarrow \{i : w_i \neq 0\}$
$w[\mathtt{I}] \leftarrow \operatorname{argmin}_{w' \in \mathbb{R}^{|\mathtt{I}|}:\sum_i w'_i=1} w'^\top \boldsymbol{k}_{\mathbb{P}}(\mathcal{S}_n[\mathtt{I}], \mathcal{S}_n[\mathtt{I}])w'$
Return: $w_{\text{CT}} \leftarrow w \in \mathbb{R}^n$ with $\|w_{\text{CT}}\|_0 \leq m$, $\mathbf{1}_n^\top w_{\text{CT}} = 1$

Algorithm 7 (Low-rank) Stein Cholesky (SC / LSC)

Input: mean-zero kernel $\boldsymbol{k}_{\mathbb{P}}$, points $\mathcal{S}_n$, output size m, rank r, AGM steps T, adaptive rounds Q

$$w \leftarrow \begin{cases} \text{Low-rankDebiasing}(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, r, T, Q) & \text{if low-rank} \\ \text{SteinThinning}(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n) & \text{otherwise} \end{cases}$$

$w_{\text{SC}} \leftarrow$ CholeskyThinning$(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, w, m)$
Return: $w_{\text{SC}} \in \mathbb{R}^n$ with $\|w_{\text{SC}}\|_0 \leq m$ and $\mathbf{1}_n^\top w_{\text{SC}} = 1$

Theorem 7 (MMD guarantee for SC / LSC)**.** *Under Assum. $(\boldsymbol{\alpha},\boldsymbol{\beta})$-params, Stein Cholesky (Alg. 7) takes $O(d_{\boldsymbol{k}_{\mathbb{P}}} n^2 + (d_{\boldsymbol{k}_{\mathbb{P}}}+m)nm + m^3)$ time to output w_{SC}, and Low-rank SC takes $O((d_{\boldsymbol{k}_{\mathbb{P}}}+r+T)nr + (d_{\boldsymbol{k}_{\mathbb{P}}}+m)nm + m^3)$ time to output w_{LSC}. Moreover, for any $\delta \in (0,1)$, with probability at least $1-\delta$, each of the following bounds hold:*

$$\Delta\text{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(w_{\text{SC}}) = 2\,\text{MMD}_{\text{OPT}} + O\Big(\sqrt{\tfrac{\|\boldsymbol{k}_{\mathbb{P}}\|_n \log n}{\delta n} + \tfrac{H_{n,m'}}{\delta}}\Big) \text{ and}$$

$$\Delta\text{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(w_{\text{LSC}}) = 2\,\text{MMD}_{\text{OPT}} + O\Big(\sqrt{\tfrac{\|\boldsymbol{k}_{\mathbb{P}}\|_n (\log n \vee 1/\delta)}{\delta n} + \tfrac{H_{n,m'}}{\delta} + \tfrac{n H_{n,r}}{\delta^2}}\Big)$$

for $H_{n,r}$ as in Thm. 4 and $m' \triangleq m + \log 2 - 2\sqrt{m \log 2 + 1}$.

Example 5. Instantiate the assumptions of Thm. 7 with γ-slow-growing input points (Def. 3), LOGGROWTH $\boldsymbol{k}_{\mathbb{P}}$, and a heavily compressed coreset size $m = (\log n)^{2(\alpha+\beta\gamma)+\epsilon}$ for any $\epsilon > 0$. Then SC delivers $\widetilde{O}(n^{-1/2})$ excess MMD with high probability in $O(n^2)$ time, and LSC with $r = m$ and $T = \Theta(\sqrt{n})$ achieves the same in $\widetilde{O}(n^{1.5})$ time.

Remark 3. While we present our results for a target precision of $1/\sqrt{n}$, a coarser target precision of $1/\sqrt{n_0}$ for $n_0 < n$ can be achieved more quickly by standard thinning the input sequence down to size n_0 before running SR, LSR, SC, or LSC.

5. Experiments

We next evaluate the practical utility of our procedures when faced with three common sources of bias: (1) burn-in, (2) approximate MCMC, and (3) tempering. In all experiments, we use a Stein kernel $\boldsymbol{k}_p$ with an inverse multiquadric (IMQ) base kernel $\boldsymbol{k}(x, y) = (1 + \|x-y\|_M^2/\sigma^2)^{-1/2}$ for σ equal to the median pairwise $\|\cdot\|_M$ distance amongst 1000 points standard thinned from the input. To vary output MMD precision, we first standard thin the input to size $n_0 \in \{2^{10}, 2^{12}, 2^{14}, 2^{16}, 2^{18}, 2^{20}\}$ before applying any method, as discussed in Rems. 2 and 3. For low-rank or weighted coreset methods, we show results for $m = r = n^\tau$. When comparing weighted coresets, we optimally reweight every coreset. We report the median over 5 independent runs for all error metrics. We implement our algorithms in JAX (Bradbury et al., 2018) and refer the reader to App. I for additional experiment details.

Correcting for burn-in The initial iterates of a Markov chain are biased by its starting point and need not accurately reflect the target distribution $\mathbb{P}$. Classical burn-in corrections use convergence diagnostics to detect and discard these iterates but typically require running multiple independent Markov chains (Cowles and Carlin, 1996). Alternatively, our proposed debiased compression methods can be used to correct for burn-in given just a single chain.

We test this claim using an experimental setup from Riabiz et al. (2022, Sec. 4.1) and the 6-chain "burn-in oracle" diagnostic of Vats and Knudson (2021). We aim to compress a posterior $\mathbb{P}$ over the parameters in the Goodwin model of oscillatory enzymatic control ($d = 4$) using $n = 2 \times 10^6$ points from a preconditioned Metropolis-adjusted Langevin algorithm (P-MALA) chain. We repeat this experiment with three alternative MCMC algorithms in App. I.3. Our primary metric is $\text{MMD}_{\boldsymbol{k}_{\mathbb{P}}}$ to $\mathbb{P}$ with $M = I$, but, for external validation, we also measure the energy distance (Riabiz et al., 2022, Eq. 11) to an auxiliary MCMC chain of length n. Trajectory plots of the first two coordinates (Fig. 1, left) highlight the substantial burn-in period for the Goodwin chain and the ability of LSKT to mimic the 6-chain burn-in oracle using only a single chain. In Fig. 1 (right), for both the MMD metric and the auxiliary energy distance, our proposed methods consistently outperform Stein thinning and match the quality of 6-chain burn-

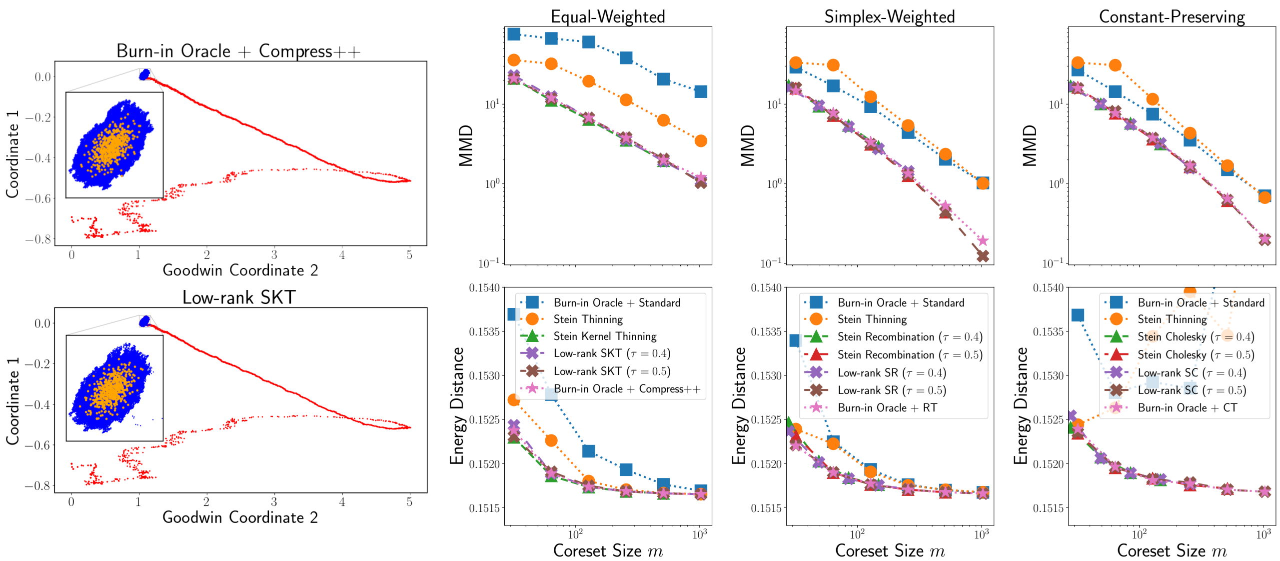

Figure 1: **Correcting for burn-in.** *Left:* Before selecting coresets (orange), the burn-in oracle uses 6 independent Markov chains to discard burn-in (red) while LSKT identifies the same high-density region (blue) with 1 chain. *Right:* Using only one chain, our methods consistently outperform the Stein and standard thinning baselines and match the 6-chain oracle.

in removal paired with unbiased compression. The spike in baseline energy distance for the constant-preserving task can be attributed to the selection of overly large weight values due to poor matrix conditioning; the simplex-weighted task does not suffer from this issue due to its regularizing nonnegativity constraint.

Correcting for approximate MCMC In posterior inference, MCMC algorithms typically require iterating over every datapoint to draw each new sample point. When datasets are large, approximating MCMC using datapoint mini-batches can reduce sampling time at the cost of persistent bias and an unknown stationary distribution that prohibits debiasing via importance sampling. Our proposed methods can correct for these biases during compression by computing full-dataset scores on a small subset of n_0 standard thinned points. To evaluate this protocol, we compress a Bayesian logistic regression posterior conditioned on the Forest Covtype dataset ($d = 54$) using $n = 2^{24}$ approximate MCMC points from the stochastic gradient Fisher scoring sampler (Ahn et al., 2012) with batch size 32. Following Wang et al. (2024), we set $M = -\nabla^2 \log p(x_{\text{mode}})$ at the sample mode x_{mode} and use 2^{20} surrogate ground truth points from the No U-turn Sampler (Hoffman et al., 2014) to evaluate energy distance. We find that our proposals improve upon standard thinning and Stein thinning for each compression task, not just in the optimized MMD metric (Fig. 2, top) but also in the auxiliary energy distance (Fig. 2, middle) and when measuring integration error for the mean (Fig. I.4).

Correcting for tempering Tempering, targeting a less-peaked and more dispersed distribution $\mathbb{Q}$, is a popular technique to improve the speed of MCMC convergence. One can correct for the sample bias using importance sampling, but this requires knowledge of the tempered density and can introduce substantial variance (Gramacy et al., 2010). Alternatively, one can use constructions of this work to correct for tempering during compression; this requires no importance weighting and no knowledge of $\mathbb{Q}$. To test this proposal, we compress the cardiac calcium signaling model posterior ($d = 38$) of Riabiz et al. (2022, Sec. 4.3) with $M = I$ and $n = 3 \times 10^6$ tempered points from a Gaussian random walk Metropolis-Hastings chain. As discussed by Riabiz et al., compression is essential in this setting as the ultimate aim is to propagate posterior uncertainty through a human heart simulator, a feat which requires over 1000 CPU hours for each summary point retained. Our methods perform on par with Stein thinning for equal-weighted compression and yield substantial gains over Stein (and standard) thinning for the two weighted compression tasks.

6. Conclusions and Future Work

We have introduced and analyzed a suite of new procedures for compressing a biased input sequence into an accurate summary of a target distribution. For equal-weighted compression, Stein kernel thinning delivers $\sqrt{n}$ points with $\widetilde{O}(n^{-1/2})$ MMD in $O(n^2)$ time, and low-rank SKT can improve this running time to $\widetilde{O}(n^{3/2})$. For simplex-weighted and constant-preserving compression, Stein recombination

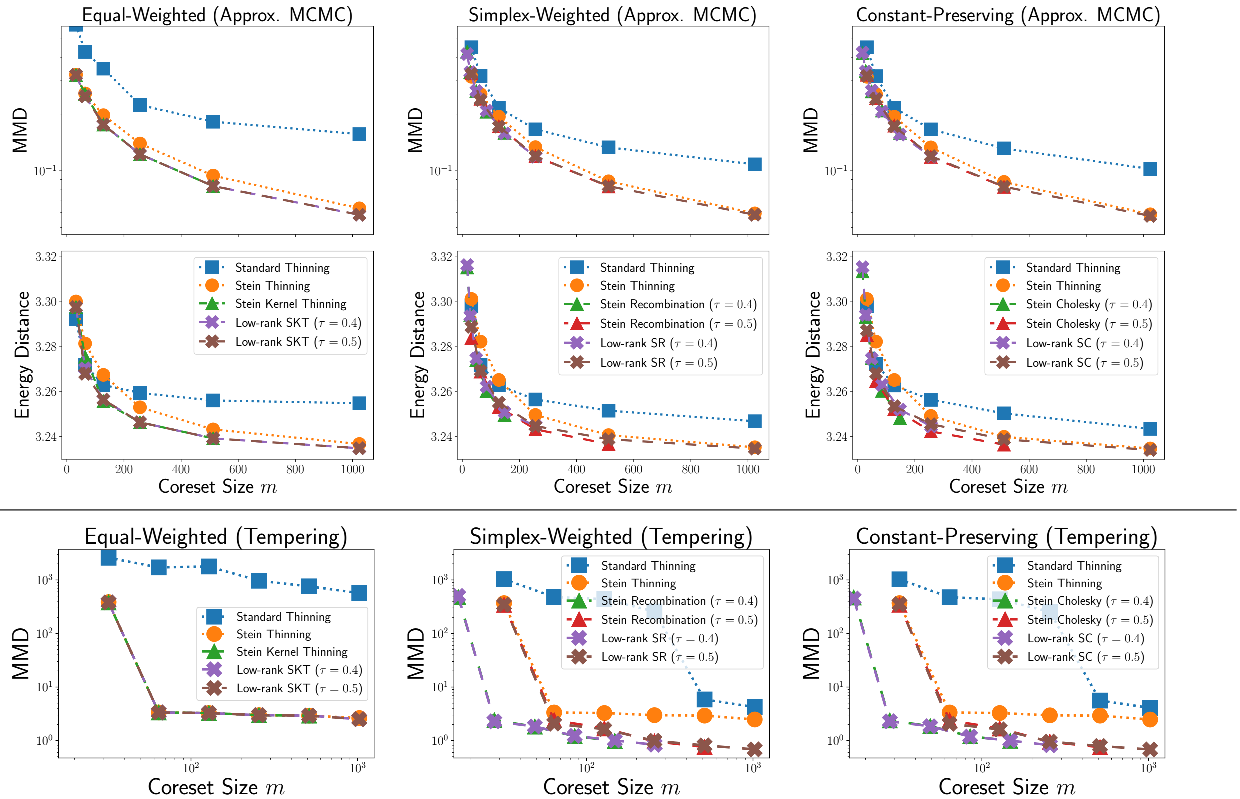

Figure 2: **Correcting for approximate MCMC (top) and tempering (bottom).** For posterior inference over the parameters of Bayesian logistic regression ($d = 54$, top) and a cardiac calcium signaling model ($d = 38$, bottom), our concise coreset constructions correct for approximate MCMC and tempering biases without need for explicit importance sampling.

and Stein Cholesky provide enhanced parsimony, matching these guarantees with as few as poly-log(n) points. Recent work has identified some limitations of score-based discrepancies, like Stein kernel MMDs, and developed modified objectives that are more sensitive to the relative density of isolated modes (Liu et al., 2023; Bénard et al., 2024). A valuable next step would be to extend our constructions to provide compression guarantees for these modified discrepancy measures. Other opportunities for future work include marrying the better-than-i.i.d. guarantees of this work with the non-myopic compression of Teymur et al. (2021), the control-variate compression of Chopin and Ducrocq (2021), and the online compression of Hawkins et al. (2022).

Broader Impact Statement

This paper presents work with the aim of advancing the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

Acknowledgments

We thank Marina Riabiz for making the Markov chain data used in Sec. 5 available and Jeffrey Rosenthal for helpful discussions concerning the geometric ergodicity of Markov chains.

References

E. Afzali and S. Muthukumarana. Gradient-free kernel conditional stein discrepancy goodness of fit testing. *Machine Learning with Applications*, 12:100463, 2023. ISSN 2666-8270. doi: https://doi.org/10.1016/j.mlwa.2023.100463. URL https://www.sciencedirect.com/science/article/pii/S2666827023000166.

S. Ahn, A. Korattikara, and M. Welling. Bayesian posterior sampling via stochastic gradient fisher scoring. *arXiv preprint arXiv:1206.6380*, 2012.

N. Aronszajn. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

Y. F. Atchadé and F. Perron. On the geometric ergodicity of metropolis-hastings algorithms. *Statistics*, 41(1):77–84, 2007.

A. Barp, F.-X. Briol, A. Duncan, M. Girolami, and L. Mackey. Minimum stein discrepancy estimators. *Advances in Neural Information Processing Systems*, 32, 2019.

A. Barp, C.-J. Simon-Gabriel, M. Girolami, and L. Mackey. Targeted separation and convergence with kernel discrepancies. *arXiv preprint arXiv:2209.12835*, 2022.

C. Bénard, B. Staber, and S. Da Veiga. Kernel stein discrepancy thinning: a theoretical perspective of pathologies and a practical fix with regularization. *Advances in Neural Information Processing Systems*, 36, 2024.

P. Billingsley. *Convergence of probability measures*. John Wiley & Sons, 2013.

J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

R. C. Bradley. Basic properties of strong mixing conditions. a survey and some open questions. 2005.

C. Carmeli, E. De Vito, and A. Toigo. Vector valued reproducing kernel hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(04): 377–408, 2006.

T. K. Chandra. De la vallée poussin's theorem, uniform integrability, tightness and moments. *Statistics & Probability Letters*, 107:136–141, 2015.

Y. Chen, E. N. Epperly, J. A. Tropp, and R. J. Webber. Randomly pivoted cholesky: Practical approximation of a kernel matrix with few entry evaluations. *arXiv preprint arXiv:2207.06503*, 2022.

N. Chopin and G. Ducrocq. Fast compression of mcmc output. *Entropy*, 23(8):1017, 2021.

K. Chwialkowski, H. Strathmann, and A. Gretton. A kernel test of goodness of fit. In M. F. Balcan and K. Q. Weinberger, editors, *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2606–2615, New York, New York, USA, 20–22 Jun 2016. PMLR. URL https://proceedings.mlr.press/v48/chwialkowski16.html.

M. K. Cowles and B. P. Carlin. Markov chain monte carlo convergence diagnostics: a comparative review. *Journal of the American Statistical Association*, 91(434):883–904, 1996.

A. Dax et al. Low-rank positive approximants of symmetric matrices. *Advances in Linear Algebra & Matrix Theory*, 4(03):172, 2014.

R. Douc and O. Cappé. Comparison of resampling schemes for particle filtering. In *ISPA 2005. Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis, 2005.*, pages 64–69. IEEE, 2005.

R. Douc, E. Moulines, P. Priouret, and P. Soulier. *Markov chains*, volume 1. Springer, 2018.

A. Durmus and E. Moulines. Nonasymptotic convergence analysis for the unadjusted langevin algorithm. 2017.

A. Durmus and É. Moulines. On the geometric convergence for mala under verifiable conditions. *arXiv preprint arXiv:2201.01951*, 2022.

A. Durmus, É. Moulines, and E. Saksman. Irreducibility and geometric ergodicity of hamiltonian monte carlo. *The Annals of Statistics*, 48(6):3545–3564, 2020.

R. Dwivedi and L. Mackey. Kernel thinning. *arXiv preprint arXiv:2105.05842*, 2021.

R. Dwivedi and L. Mackey. Generalized kernel thinning. In *International Conference on Learning Representations*, 2022.

E. Epperly and E. Moreno. Kernel quadrature with randomly pivoted cholesky. *Advances in Neural Information Processing Systems*, 36, 2024.

M. A. Erdogdu, L. Mackey, and O. Shamir. Global nonconvex optimization with discretized diffusions. *Advances in Neural Information Processing Systems*, 31, 2018.

M. A. Gallegos-Herrada, D. Ledvinka, and J. S. Rosenthal. Equivalences of geometric ergodicity of markov chains, 2023.

B. Ghojogh, A. Ghodsi, F. Karray, and M. Crowley. Kkt conditions, first-order and second-order optimization, and distributed optimization: tutorial and survey. *arXiv preprint arXiv:2110.01858*, 2021.

J. Gorham and L. Mackey. Measuring sample quality with kernels. In *International Conference on Machine Learning*, pages 1292–1301. PMLR, 2017.

J. Gorham, A. B. Duncan, S. J. Vollmer, and L. Mackey. Measuring sample quality with diffusions. *The Annals of Applied Probability*, 29(5):2884–2928, 2019.

R. Gramacy, R. Samworth, and R. King. Importance tempering. *Statistics and Computing*, 20:1–7, 2010.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *The Journal of Machine Learning Research*, 13(1):723–773, 2012.

C. Hawkins, A. Koppel, and Z. Zhang. Online, informative mcmc thinning with kernelized stein discrepancy. *arXiv preprint arXiv:2201.07130*, 2022.

S. Hayakawa, H. Oberhauser, and T. Lyons. Positively weighted kernel quadrature via subsampling. *Advances in Neural Information Processing Systems*, 35:6886–6900, 2022.

S. Hayakawa, H. Oberhauser, and T. Lyons. Sampling-based nyström approximation and kernel quadrature. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

L. Hodgkinson, R. Salomone, and F. Roosta. The reproducing stein kernel approach for post-hoc corrected sampling. *arXiv preprint arXiv:2001.09266*, 2020.

M. D. Hoffman, A. Gelman, et al. The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, 2014.

A. A. Johnson. *Geometric ergodicity of Gibbs samplers*. university of minnesota, 2009.

L. Li, J.-G. Liu, and Y. Wang. Geometric ergodicity of sgld via reflection coupling. *arXiv preprint arXiv:2301.06769*, 2023.

Q. Liu and J. Lee. Black-box importance sampling. In *Artificial Intelligence and Statistics*, pages 952–961. PMLR, 2017.

Q. Liu, J. Lee, and M. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *International conference on machine learning*, pages 276–284. PMLR, 2016.

X. Liu, A. B. Duncan, and A. Gandy. Using perturbation to improve goodness-of-fit tests based on kernelized stein discrepancy. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 21527–21547. PMLR, 23–29 Jul 2023.

F. Merlevède, M. Peligrad, and S. Utev. Sharp conditions for the clt of linear processes in a hilbert space. *Journal of Theoretical Probability*, 10(3):681–693, 1997.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

H. Q. Minh. Some properties of gaussian reproducing kernel hilbert spaces and their implications for function approximation and learning theory. *Constructive Approximation*, 32:307–338, 2010.

S. Niederer, L. Mitchell, N. Smith, and G. Plank. Simulating human cardiac electrophysiology on clinical timescales. *Frontiers in physiology*, 2:14, 2011.

V. I. Paulsen and M. Raghupathi. *An introduction to the theory of reproducing kernel Hilbert spaces*, volume 152. Cambridge university press, 2016.

D. Phan, N. Pradhan, and M. Jankowiak. Composable effects for flexible and accelerated probabilistic programming in numpyro. *arXiv preprint arXiv:1912.11554*, 2019.

I. Pinelis. Exact lower and upper bounds on the incomplete gamma function. *arXiv preprint arXiv:2005.06384*, 2020.

Y. Pitcan. A note on concentration inequalities for u-statistics. *arXiv preprint arXiv:1712.06160*, 2017.

Q. Qin. Geometric ergodicity of trans-dimensional markov chain monte carlo algorithms. *arXiv preprint arXiv:2308.00139*, 2023.

M. Riabiz, W. Y. Chen, J. Cockayne, P. Swietach, S. A. Niederer, L. Mackey, and C. J. Oates. Replication Data for: Optimal Thinning of MCMC Output, 2020. URL https://doi.org/10.7910/DVN/MDKNWM. Accessed on Mar 23, 2021.

M. Riabiz, W. Y. Chen, J. Cockayne, P. Swietach, S. A. Niederer, L. Mackey, and C. J. Oates. Optimal thinning of mcmc output. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 84(4):1059–1081, 2022.

G. O. Roberts and R. L. Tweedie. Geometric convergence and central limit theorems for multidimensional hastings and metropolis algorithms. *Biometrika*, 83(1):95–110, 1996.

A. Shetty, R. Dwivedi, and L. Mackey. Distribution compression in near-linear time. In *International Conference on Learning Representations*, 2022.

I. Steinwart and A. Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

I. Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and rkhss. *Constructive Approximation*, 35:363–417, 2012.

H.-W. Sun and D.-X. Zhou. Reproducing kernel hilbert spaces associated with analytic translation-invariant mercer kernels. *Journal of Fourier Analysis and Applications*, 14(1):89–101, 2008.

M. Tchernychova. *Caratheodory cubature measures*. PhD thesis, University of Oxford, 2016.

O. Teymur, J. Gorham, M. Riabiz, and C. Oates. Optimal quantisation of probability measures using maximum mean discrepancy. In *International Conference on Artificial Intelligence and Statistics*, pages 1027–1035. PMLR, 2021.

D. Vats and C. Knudson. Revisiting the gelman–rubin diagnostic. *Statistical Science*, 36(4):518–529, 2021.

M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge university press, 2019.

C. Wang, Y. Chen, H. Kanagawa, and C. J. Oates. Stein Π-importance sampling. *Advances in Neural Information Processing Systems*, 36, 2024.

J.-K. Wang, J. Abernethy, and K. Y. Levy. No-regret dynamics in the fenchel game: A unified framework for algorithmic convex optimization. *Mathematical Programming*, pages 1–66, 2023.

J. Wellner et al. *Weak convergence and empirical processes: with applications to statistics*. Springer Science & Business Media, 2013.

J. Yang, Q. Liu, V. Rao, and J. Neville. Goodness-of-fit testing for discrete distributions via stein discrepancy. In *International Conference on Machine Learning*, pages 5561–5570. PMLR, 2018.

F. Zhang. *The Schur complement and its applications*, volume 4. Springer Science & Business Media, 2006.

D.-X. Zhou. The covering number in learning theory. *Journal of Complexity*, 18(3):739–767, 2002.

Appendix Contents

A. Appendix Notation

For the point sequence $\mathcal{S}_n = (x_i)_{i\in[n]}$, we define $\mathbb{S}_n \triangleq \frac{1}{n}\sum_{i\in[n]} \delta_{x_i}$. For a weight vector $w \in \mathbb{R}^n$, we define the support $\mathrm{supp}(w) \triangleq \{i \in [n] : w_i \neq 0\}$ and the signed measure $\mathbb{S}_n^w \triangleq \sum_{i\in[n]} w_i\delta_{x_i}$. For a matrix $K \in \mathbb{R}^{n\times n}$ and $w \in \Delta_{n-1}$, we define the weighted matrix $K^w \triangleq \mathrm{diag}(\sqrt{w})K\,\mathrm{diag}(\sqrt{w})$. For positive semidefinite (PSD) matrices (A, B), we use $A \succeq B$ (resp. $A \preceq B$) to mean $A - B$ (resp. $B - A$) is PSD. For a symmetric PSD (SPSD) matrix M, we let $M^{1/2}$ denote a symmetric matrix square root satisfying $M = M^{1/2}M^{1/2}$. For $A \in \mathbb{R}^{n\times m}$, we denote $\|A\|_p \triangleq \sup_{x\neq 0} \frac{\|Ax\|_p}{\|x\|_p}$. We will use $\mathbb{1}_E$ to denote the indicator function for an event E.

Notations used only in a specific section will be introduced within.

B. Spectral Analysis of Kernel Matrices

The goal of this section is to develop spectral bounds for kernel matrices.

In App. B.1, we transfer the bounds on covering numbers from the definition of PolyGrowth or LogGrowth kernels to bounds on the eigenvalues of the kernel matrices. This sets the theoretical foundation for the algorithms in later sections as their error guarantees rely on the fast decay of eigenvalues of kernel matrices.

In App. B.2, we show that Stein kernels are PolyGrowth (resp. LogGrowth) provided that their base kernels are differentiable (resp. radially analytic). Hence we obtain spectral bounds for a wide range of Stein kernels.

Notation For a normed space E, we use $\|\cdot\|_E$ to denote its norm, $\mathcal{B}_E(p, r) \triangleq \{x \in E : \|x - p\|_E \leq r\}$ to denote the closed ball of radius r centered at p in E with the shorthand $\mathcal{B}_E(r) \triangleq \mathcal{B}_E(0, r)$ and $\mathcal{B}_E \triangleq \mathcal{B}_E(1)$. When E is an RKHS with kernel $\boldsymbol{k}$, for brevity we use $\boldsymbol{k}$ in place of E in the subscript. Let $\mathfrak{F}(\mathcal{X}, \mathcal{Y})$ denote the space of functions from $\mathcal{X}$ to $\mathcal{Y}$, and $\mathfrak{B}(E, F)$ denote the space of bounded linear functions between normed spaces E, F. For a set A, we use $\ell_\infty(A)$ to denote the space of bounded $\mathbb{R}$-valued functions on A equipped with the sup-norm $\|f\|_{\infty,A} \triangleq \sup_{x\in A}|f(x)|$. We use $E \hookrightarrow F$ to denote the inclusion map. e use $\lambda_\ell(T)$ to denote the ℓ-th largest eigenvalue of an operator T.

B.1. Bounding the spectrum of kernel matrices

We first introduce the general Mercer representation theorem from Steinwart and Scovel (2012), which shows the existence of a discrete spectrum of the integral operator associated with a continuous square-integrable kernel. The theorem also provides a series expansion of the kernel, i.e., the Mercer representation, in terms of the eigenvalues and eigenfunctions.

Lemma B.1 (General Mercer representation (Steinwart and Scovel, 2012))**.** *Consider a kernel $\boldsymbol{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ that is jointly continuous in both inputs and a probability measure μ such that $\int \boldsymbol{k}(x, x)\mathrm{d}\mu(x) < \infty$. Then the following holds.*

(a) The inclusion $\mathcal{H}_{\boldsymbol{k}} \hookrightarrow \mathcal{L}^2(\mu)$ is a compact operator, i.e., $\mathcal{B}_{\boldsymbol{k}}$ is a compact subset of $\mathcal{L}^2(\mu)$. In particular, this inclusion is continuous.

(b) The Hilbert-space adjoint of the inclusion $\mathcal{H}_{\boldsymbol{k}} \hookrightarrow \mathcal{L}^2(\mu)$ is the compact operator $S_{\boldsymbol{k},\mu} : \mathcal{L}^2(\mu) \to \mathcal{H}_{\boldsymbol{k}}$ defined as

$$S_{\boldsymbol{k},\mu} f \triangleq \int \boldsymbol{k}(\cdot, x)f(x)\mathrm{d}\mu(x). \tag{3}$$

We also have $S^*_{\boldsymbol{k},\mu} \triangleq \mathcal{H}_{\boldsymbol{k}} \hookrightarrow \mathcal{L}^2(\mu)$. Hence the operator

$$T_{\boldsymbol{k},\mu} \triangleq S^*_{\boldsymbol{k},\mu} S_{\boldsymbol{k},\mu} : \mathcal{L}^2(\mu) \to \mathcal{L}^2(\mu) \tag{4}$$

is also compact.

(c) There exist $\{\lambda_\ell\}_{\ell=1}^\infty$ with $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ and $\{\phi_\ell\}_{\ell=1}^\infty \subset \mathcal{H}_{\boldsymbol{k}}$ such that $\{\phi_\ell\}_{\ell=1}^\infty$ is an orthonormal system in $\mathcal{L}^2(\mu)$ and $\{\lambda_\ell\}_{\ell=1}^\infty$ (resp. $\{\phi_\ell\}_{\ell=1}^\infty$) consists of the eigenvalues (resp. eigenfunctions) of $T_{\boldsymbol{k},\mu}$ with eigendecomposition, for $f \in \mathcal{L}^2(\mu)$,

$$T_{\boldsymbol{k},\mu} f = \textstyle\sum_{\ell=1}^\infty \lambda_\ell \langle f, \phi_\ell\rangle_{\mathcal{L}^2(\mu)} \phi_\ell$$

with convergence in $\mathcal{L}^2(\mu)$.

(d) We have the following series expansion

$$\boldsymbol{k}(x, x') = \sum_{\ell=1}^{\infty} \lambda_i \phi_\ell(x)\phi_\ell(x'), \tag{5}$$

where the series convergence is absolute and uniform in x, x' on all $A \times A \subset \operatorname{supp}\mu \times \operatorname{supp}\mu$.

Proof of Lem. B.1. Part (a) and (b) follow respectively from Steinwart and Scovel (2012, Lem. 2.3 and 2.2). Part (c) follows from part (a) and Steinwart and Scovel (2012, Lem. 2.12). Finally, part (d) follows from Steinwart and Scovel (2012, Cor. 3.5). □

We will use the following lemma regarding the restriction of covering numbers.

Lemma B.2 (Covering number is preserved in restriction)**.** *For a kernel $\boldsymbol{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ and a set $A \subset \mathbb{R}^d$, we have $\mathcal{N}_{\boldsymbol{k}}(A, \epsilon) = \mathcal{N}_{\boldsymbol{k}|_A}(A, \epsilon)$, for $\boldsymbol{k}|_A$, the restricted kernel of $\boldsymbol{k}$ to A (Paulsen and Raghupathi, 2016, Sec. 5.4).*

Proof of Lem. B.2. It suffices to show that a $(\boldsymbol{k}, A, \epsilon)$ cover can be converted to a cover of $(\boldsymbol{k}|_A, A, \epsilon)$ of the same cardinality and vice versa.

Let $\mathcal{C} \subset \mathcal{B}_{\boldsymbol{k}|_A}$ be a $(\boldsymbol{k}|_A, A, \epsilon)$ cover. For any $f \in \mathcal{C}$, we have $\|f\|_{\boldsymbol{k}|_A} = \inf\left\{\|\tilde{f}\|_{\boldsymbol{k}} : \tilde{f} \in \mathcal{H}_{\boldsymbol{k}}, \tilde{f}|_A = f\right\} \le 1$ (Paulsen and Raghupathi, 2016, Corollary 5.8). Moreover, the infimum is attained by some $\tilde{f} \in \mathcal{H}_{\boldsymbol{k}}$ such that $\|\tilde{f}\|_{\boldsymbol{k}} = \|f\|_{\boldsymbol{k}|_A} \le 1$ and $\tilde{f}|_A = f$. Now form $\tilde{\mathcal{C}} = \{\tilde{f} : f \in \mathcal{C}\}$. For any $\tilde{h} \in \mathcal{B}_{\boldsymbol{k}}$, there exists $f \in \mathcal{C}$ such that

$$\left\|\tilde{h}|_A - f\right\|_{\infty,A} \le \epsilon \implies = \left\|\tilde{h} - \tilde{f}\right\|_{\infty,A} \le \epsilon,$$

so $\tilde{\mathcal{C}}$ is a $(\boldsymbol{k}|_A, A, \epsilon)$ cover.

For the other direction, let $\tilde{\mathcal{C}} \subset \mathcal{B}_{\boldsymbol{k}}$ be a $(\boldsymbol{k}, A, \epsilon)$ cover. Define $\mathcal{C} = \{\tilde{f}|_A : \tilde{f} \in \tilde{\mathcal{C}}\} \subset \mathcal{H}_{\boldsymbol{k}|_A}$. Since $\|\tilde{f}|_A\|_{\boldsymbol{k}|_A} \le \|\tilde{f}\|_{\boldsymbol{k}}$, we have $\mathcal{C} \subset \mathcal{B}_{\boldsymbol{k}|_A}$. For any $h \in \mathcal{B}_{\boldsymbol{k}_A]}$, again by Paulsen and Raghupathi (2016, Corollary 5.8), there exists $\tilde{h} \in \mathcal{H}_{\boldsymbol{k}}$ such that $\|\tilde{h}\|_{\boldsymbol{k}} = \|h\|_{\boldsymbol{k}|_A} \le 1$, so there exists $\tilde{f} \in \tilde{\mathcal{C}}$ such that

$$\left\|\tilde{h} - \tilde{f}\right\|_{\infty,A} \le \epsilon \implies \left\|h - \tilde{f}|_A\right\|_{\infty,A} \le \epsilon,$$

Hence $\mathcal{C}$ is a $(\boldsymbol{k}, A, \epsilon)$ cover. □

The goal for the rest of this section is to transfer the bounds of the covering number in the definition of a PolyGrowth or LogGrowth kernel from Assum. **$(\boldsymbol{\alpha}, \boldsymbol{\beta})$-kernel** to bounds on entropy numbers (Steinwart and Christmann, 2008, Def. 6.20) that are closely related to eigenvalues of the integral operator (4).

Definition B.1 (Entropy number of a bounded linear map)**.** *For a bounded linear operator $S : E \to F$ between normed spaces E, F, for $\ell \in \mathbb{N}$, the ℓ-th entropy number of S is defined as*

$$e_\ell(S) \triangleq \inf\left\{\epsilon > 0 : \exists s_1, \dots, s_{2^{\ell-1}} \in S(\mathcal{B}_E) \text{ such that } S(\mathcal{B}_E) \subset \bigcup_{i=1}^{2^{\ell-1}} \mathcal{B}_F(s_i, \epsilon)\right\}.$$

The following lemma shows the relation between covering numbers and entropy numbers.

Lemma B.3 (Relation between covering number and entropy number)**.** *Suppose a kernel $\boldsymbol{k}$ is jointly continuous and $A \subset \mathbb{R}^d$ is bounded. Then for any $\epsilon > 0$,*

$$e_{\lceil \log_2 \mathcal{N}_{\boldsymbol{k}}(A,\epsilon)\rceil + 1}(\mathcal{H}_{\boldsymbol{k}|_A} \hookrightarrow \ell_\infty(A)) \le \epsilon.$$

Proof of Lem. B.3. First, the assumption implies $\boldsymbol{k}|_A$ is a bounded kernel, so by Steinwart and Christmann (2008, Lemma 4.23), the inclusion $\mathcal{H}_{\boldsymbol{k}|_A} \hookrightarrow \ell_\infty(A)$ is continuous. By the definition of $\mathcal{N}_{\boldsymbol{k}|_A}(A, \epsilon)$, by adding arbitrary elements into the cover if necessary, there exists a $(\boldsymbol{k}|_A, A, \epsilon)$ cover of $\mathcal{B}_{\boldsymbol{k}|_A}$ of cardinality $2^{\lceil \log_2(\mathcal{N}_{\boldsymbol{k}|_A}(A,\epsilon))\rceil} \ge \mathcal{N}_{\boldsymbol{k}|_A}(A, \epsilon)$. Hence

$$e_{\lceil \log_2 \mathcal{N}_{\boldsymbol{k}|_A}(A,\epsilon)\rceil + 1}(\mathcal{H}_{\boldsymbol{k}|_A} \hookrightarrow \ell_\infty(A)) \le \epsilon.$$

The claim follows since $\mathcal{N}_{\boldsymbol{k}|_A}(A, \epsilon) = \mathcal{N}_{\boldsymbol{k}}(A, \epsilon)$ by Lem. B.2. □

Proposition B.1 (ℓ_∞-entropy number bound for POLYGROWTH or LOGGROWTH $\boldsymbol{k}$)**.** *Suppose a kernel $\boldsymbol{k}$ satisfies Assum. ($\boldsymbol{\alpha}, \boldsymbol{\beta}$)-kernel. Let $\mathfrak{C}_d > 0$ denote the constant that appears in the Assum. ($\boldsymbol{\alpha}, \boldsymbol{\beta}$)-kernel. Define*

$$L_{\boldsymbol{k}}(r) \triangleq \frac{\mathfrak{C}_d}{\log 2} r^\beta. \tag{6}$$

Then for any $r > 0$ and $\ell \in \mathbb{N}$ that satisfies $\ell > L_{\boldsymbol{k}}(r+1) + 1$, we have

$$e_\ell(\mathcal{H}_{\boldsymbol{k}|_{\mathcal{B}_2(r)}} \hookrightarrow \ell_\infty(\mathcal{B}_2(r))) \le \begin{cases} \left(\frac{L_{\boldsymbol{k}}(r+1)}{\ell-1}\right)^{\frac{1}{\alpha}} & \text{if } \boldsymbol{k} \text{ is } \textsc{PolyGrowth}(\alpha,\beta), \text{ and} \\ \exp\left(1 - \left(\frac{\ell-1}{L_{\boldsymbol{k}}(r+1)}\right)^{\frac{1}{\alpha}}\right) & \text{if } \boldsymbol{k} \text{ is } \textsc{LogGrowth}(\alpha,\beta). \end{cases}$$

Proof of Prop. B.1. By Lem. B.3 and the fact that e_ℓ is monotonically decreasing in ℓ by definition, if $\ell \ge \log_2 \mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(r), \epsilon) + 1$ for some $\epsilon > 0$, then

$$e_\ell(\mathcal{H}_{\boldsymbol{k}|_{\mathcal{B}_2(r)}} \hookrightarrow \ell_\infty(\mathcal{B}_2(r))) \le e_{\lceil \log_2 \mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(r),\epsilon)\rceil + 1}(\mathcal{H}_{\boldsymbol{k}|_{\mathcal{B}_2(r)}} \hookrightarrow \ell_\infty(\mathcal{B}_2(r))) \le \epsilon. \tag{7}$$

For the POLYGROWTH case, by its definition, the condition $\ell \ge \log_2 \mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(r), \epsilon) + 1$ is met if $\epsilon \in (0,1)$ and

$$\ell \ge \frac{\mathfrak{C}_d}{\log 2}(1/\epsilon)^\alpha (r+1)^\beta + 1 \iff \epsilon \le \left(\frac{L_{\boldsymbol{k}}(r+1)}{\ell-1}\right)^{\frac{1}{\alpha}}.$$

Hence (7) holds with $\epsilon = \left(\frac{L_{\boldsymbol{k}}(r+1)}{\ell-1}\right)^{\frac{1}{\alpha}}$, as long as $\epsilon \in (0,1)$, so ℓ needs to satisfy

$$1 > \left(\frac{L_{\boldsymbol{k}}(r+1)}{\ell-1}\right)^{\frac{1}{\alpha}} \iff \ell > L_{\boldsymbol{k}}(r+1) + 1.$$

Similarly, for the LOGGROWTH case, the condition $\ell \ge \log_2 \mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(r), \epsilon) + 1$ is met if $\epsilon \in (0,1)$ and

$$\ell \ge \frac{\mathfrak{C}_d}{\log 2}(\log(1/\epsilon) + 1)^\alpha (r+1)^\beta + 1 \iff \epsilon \le \exp\left(1 - \left(\frac{\ell-1}{L_{\boldsymbol{k}}(r+1)}\right)^{\frac{1}{\alpha}}\right).$$

Hence (7) holds with $\epsilon = \exp\left(1 - \left(\frac{\ell-1}{L_{\boldsymbol{k}}(r+1)}\right)^{\frac{1}{\alpha}}\right)$, as long as $\epsilon \in (0,1)$, so ℓ needs to satisfy

$$1 > \exp\left(1 - \left(\frac{\ell-1}{L_{\boldsymbol{k}}(r+1)}\right)^{\frac{1}{\alpha}}\right) \iff \ell > L_{\boldsymbol{k}}(r+1) + 1.$$

□

Next, we show that we can transfer bounds on entropy numbers to obtain bounds for the eigenvalues of kernel matrices, which will become handy when we develop sub-quadratic-time algorithms in Sec. 3. We rely on the following lemma, which summarizes the relevant facts from Steinwart and Christmann (2008, Appendix A).

Lemma B.4 (Eigenvalue is bounded by entropy number)**.** *Let $\boldsymbol{k}$ be a jointly continuous kernel and $\mathbb{P}$ be a distribution such that $\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}(x,x)] < \infty$, and recall that $\lambda_\ell(\cdot)$ denotes the ℓ-th largest eigenvalue of a linear operator. Then, for all $\ell \in \mathbb{N}$,*

$$\lambda_\ell(T_{\boldsymbol{k},\mathbb{P}}) \le 4e_\ell^2(\mathcal{H}_{\boldsymbol{k}} \hookrightarrow \mathcal{L}^2(\mathbb{P})).$$

Proof of Lem. B.4. For any bounded linear operator $S : \mathcal{H}_1 \to \mathcal{H}_2$ between Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, we have $a_\ell(S) \le 2e_\ell(S)$, where a_ℓ is the ℓ-th approximation number defined in Steinwart and Christmann (2008, (A.29)). Recall the operator $S^*_{\boldsymbol{k},\mathbb{P}} = \mathcal{H}_{\boldsymbol{k}} \hookrightarrow \mathcal{L}^2(\mathbb{P})$ from (3), which is compact (in particular bounded) by Lem. B.1(a). Thus

$$s_\ell(S^*_{\boldsymbol{k},\mathbb{P}}) = a_\ell(S^*_{\boldsymbol{k},\mathbb{P}}) \le 2e_\ell(S^*_{\boldsymbol{k},\mathbb{P}}),$$

where the first equality follows from the paragraph below Steinwart and Christmann (2008, (A.29))) and s_ℓ is ℓ-th singular number of an operator (Steinwart and Christmann, 2008, (A.25)). Then using the identities mentioned under Steinwart and Christmann (2008, (A.25)) and Steinwart and Christmann (2008, (A.27)) and that all operators involved are compact by Lem. B.1(b), we have

$$\lambda_\ell(T_{\boldsymbol{k},\mathbb{P}}) = \lambda_\ell(S^*_{\boldsymbol{k},\mathbb{P}} S_{\boldsymbol{k},\mathbb{P}}) = s_\ell(S^*_{\boldsymbol{k},\mathbb{P}} S_{\boldsymbol{k},\mathbb{P}}) = s_\ell^2(S^*_{\boldsymbol{k},\mathbb{P}}) \le 4e_\ell^2(S^*_{\boldsymbol{k},\mathbb{P}}).$$

□

The previous lemma allows us to bound eigenvalues of kernel matrices by ℓ_∞-entropy numbers.

Proposition B.2 (Eigenvalue of kernel matrix is bounded by ℓ_∞-entropy number)**.** *Let $\boldsymbol{k}$ be a jointly continuous kernel. Define $K \triangleq \boldsymbol{k}(\mathcal{S}_n, \mathcal{S}_n)$ for the sequence of points $\mathcal{S}_n = (x_1, \dots, x_n) \subset \mathbb{R}^d$. For any $w \in \Delta_{n-1}$, recall the notation $\mathbb{S}_n^w = \sum_{i\in[n]} w_i \delta_{x_i}$, $K^w = \operatorname{diag}(\sqrt{w}) K \operatorname{diag}(\sqrt{w})$, and $R_n = 1 + \sup_{i\in[n]} \|x_i\|_2$. Then for all $\ell \in \mathbb{N}$,*

$$\lambda_\ell(K^w) \overset{(i)}{=} \lambda_\ell(T_{\boldsymbol{k},\mathbb{S}_n^w}) \overset{(ii)}{\leq} 4e_\ell^2(\mathcal{H}_{\boldsymbol{k}|_{\mathcal{B}_2(R_n-1)}} \hookrightarrow \ell_\infty(\mathcal{B}_2(R_n - 1))). \tag{8}$$

Proof of Prop. B.2. Without loss of generality, we assume $w_i > 0$ for all $i \in [n]$, since otherwise, we can consider a smaller set of points by removing the ones with zero weights.

Proof of equality (i) from display (8) Note that $\mathcal{L}^2(\mathbb{S}_n^w)$ is isometric to $\mathbb{R}^n$. Let $K \triangleq \boldsymbol{k}(\mathcal{S}_n, \mathcal{S}_n)$ denote the kernel matrix. The action of $T_{\boldsymbol{k},\mathbb{S}_n^w}$ is given by, for $i \in [n]$,

$$T_{\boldsymbol{k},\mathbb{S}_n^w} f(x_i) = \textstyle\sum_{j\in[n]} w_j \boldsymbol{k}(x_i, x_j) f(x_j),$$

so in matrix form, $T_{\boldsymbol{k},\mathbb{S}_n^w} f = K \operatorname{diag}(w) f$, and hence $T_{\boldsymbol{k},\mathbb{S}_n^w} = K \operatorname{diag}(w)$. If λ_ℓ is an eigenvalue of $K \operatorname{diag}(w)$ with eigenvector v_ℓ, then

$$\begin{aligned} K \operatorname{diag}(w) v_\ell = \lambda_\ell v_\ell &\iff \operatorname{diag}(\sqrt{w}) K \operatorname{diag}(w) v_\ell = \lambda_\ell \operatorname{diag}(\sqrt{w}) v_\ell \\ &\iff \operatorname{diag}(\sqrt{w}) K \operatorname{diag}(\sqrt{w}) (\operatorname{diag}(\sqrt{w}) v_\ell) = \lambda_\ell \operatorname{diag}(\sqrt{w}) v_\ell, \end{aligned}$$

where we used $w_i > 0$ for all $i \in [n]$. Hence the eigenspectrum of $T_{\boldsymbol{k},\mathbb{S}_n^w}$ agrees with that of $\operatorname{diag}(\sqrt{w}) K \operatorname{diag}(\sqrt{w})$.

Proof of bound (ii) from display (8) By Lem. B.4, we have $\lambda_\ell(T_{\boldsymbol{k},\mathbb{S}_n^w}) \leq 4e_\ell^2(\mathcal{H}_{\boldsymbol{k}|_{\mathcal{B}_2(R_n-1)}} \hookrightarrow \mathcal{L}^2(\mathbb{S}_n^w))$. Finally, using Def. B.1, we have $e_\ell(\mathcal{H}_{\boldsymbol{k}|_{\mathcal{B}_2(R_n-1)}} \hookrightarrow \mathcal{L}^2(\mathbb{S}_n^w)) \leq e_\ell(\mathcal{H}_{\boldsymbol{k}|_{\mathcal{B}_2(R_n-1)}} \hookrightarrow \ell_\infty(\mathcal{B}_2(R_n - 1)))$ because $\mathbb{S}_n^w$ is supported in $\mathcal{B}_2(R_n - 1)$ and the fact that $\|\cdot\|_{\mathcal{L}^2(\mathbb{P})} \leq \|\cdot\|_\infty$ for any $\mathbb{P}$. ☐

Combining the tools developed so far, we have the following corollary for bounding the eigenvalues of POLYGROWTH and LOGGROWTH kernel matrices.

Corollary B.1 (Eigenvalue bound for POLYGROWTH or LOGGROWTH kernel matrix)**.** *Suppose a kernel $\boldsymbol{k}$ satisfies Assum. $(\boldsymbol{\alpha}, \boldsymbol{\beta})$-kernel. Let $\mathcal{S}_n = (x_1, \dots, x_n) \subset \mathbb{R}^d$ be a sequence of points. For any $w \in \Delta_{n-1}$, using the notation $L_{\boldsymbol{k}}$ from (6), for any $\ell > L_{\boldsymbol{k}}(R_n) + 1$, we have*

$$\lambda_\ell(K^w) \leq \begin{cases} 4\left(\frac{L_{\boldsymbol{k}}(R_n)}{\ell-1}\right)^{\frac{2}{\alpha}} & \textsc{PolyGrowth}(\alpha, \beta) \quad \textit{and} \\ 4\exp\left(2 - 2\left(\frac{\ell-1}{L_{\boldsymbol{k}}(R_n)}\right)^{\frac{1}{\alpha}}\right) & \textsc{LogGrowth}(\alpha, \beta). \end{cases}$$

Proof of Cor. B.1. The claim follows by applying Prop. B.2 and Prop. B.1. ☐

B.2. Spectral decay of Stein kernels

The goal of this section is to show that a Stein kernel $\boldsymbol{k}_p$ satisfies Assum. $(\boldsymbol{\alpha}, \boldsymbol{\beta})$-kernel provided that the base kernel is sufficiently smooth and to derive the parameters α, β for POLYGROWTH and LOGGROWTH cases.

For a Stein kernel $\boldsymbol{k}_p$ with preconditioning matrix M, we define

$$\mathfrak{S}_p(r) \triangleq \max\left(1, \sup_{\|x\|_2 \leq r} \left\|M^{1/2} \nabla \log p(x)\right\|_2\right). \tag{9}$$

We start by noting a useful alternative expression for a Stein kernel where we only need access to the density via the score $\nabla \log p$.

Proposition B.3 (Alternative expression for Stein kernel)**.** *The Stein kernel $\boldsymbol{k}_p$ has the following alternative form:*

$$\begin{aligned} \boldsymbol{k}_p(x, y) = \langle \nabla \log p(x), M \nabla \log p(y) \rangle \boldsymbol{k}(x, y) + \langle \nabla \log p(x), M \nabla_y \boldsymbol{k}(x, y) \rangle + \\ \langle \nabla \log p(y), M \nabla_x \boldsymbol{k}(x, y) \rangle + \operatorname{tr}(M \nabla_x \nabla_y \boldsymbol{k}(x, y)), \end{aligned} \tag{10}$$

where $\nabla_x \nabla_y \boldsymbol{k}(x, y)$ denotes the $d \times d$ matrix $(\partial_{x_i} \partial_{y_j} \boldsymbol{k}(x, y))_{i,j\in[d]}$.

Proof of Prop. B.3. We compute

$$\begin{aligned}(\nabla_x \cdot (p(x)M\boldsymbol{k}(x,y)p(y)))_j &= \textstyle\sum_{i\in[d]} M_{ij}\left(\partial_{x_i}p(x)\boldsymbol{k}(x,y)p(y) + p(x)\partial_{x_i}\boldsymbol{k}(x,y)p(y)\right).\\ \nabla_y \cdot \nabla_x \cdot (p(x)M\boldsymbol{k}(x,y)p(y)) &= \textstyle\sum_{i,j\in[d]} M_{ij}\left(\partial_{x_i}p(x)\partial_{y_j}p(y)\boldsymbol{k}(x,y) + \partial_{x_i}p(x)\partial_{y_j}\boldsymbol{k}(x,y)p(y)\right)\\ &\quad + \textstyle\sum_{i,j\in[d]} M_{ij}\left(p(x)\partial_{y_j}p(y)\partial_{x_i}\boldsymbol{k}(x,y) + p(x)\partial_{x_i}\partial_{y_j}\boldsymbol{k}(x,y)p(y)\right).\end{aligned}$$

The four terms in the last equation correspond to the four terms in (10). □

In what follows, we will make use of a matrix-valued kernel $\boldsymbol{K} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^{d\times d}$ which generates an RKHS $\mathcal{H}_{\boldsymbol{K}}$ of vector-valued functions. See Carmeli et al. (2006) for an introduction to vector-valued RKHS theory.

Our next goal is to build a Hilbert-space isometry between the direct sum Hilbert space $\mathcal{H}_{\boldsymbol{k}}{}^{\oplus d}$ and $\mathcal{H}_{\boldsymbol{k}_p}$ to represent functions in $\mathcal{H}_{\boldsymbol{k}_p}$ using functions from $\mathcal{H}_{\boldsymbol{k}}$.

Lemma B.5 (Preconditioned matrix-valued RKHS from a scalar kernel)**.** *Let $\boldsymbol{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be kernel and $\mathcal{H}_{\boldsymbol{k}}$ be the corresponding RKHS. Let $M \in \mathbb{R}^{d\times d}$ be an SPSD matrix. Consider the map $\iota : \mathcal{H}_{\boldsymbol{k}}{}^{\oplus d} \to \mathfrak{F}(\mathbb{R}^d, \mathbb{R}^d)$ defined by $(f_1, \dots, f_d) \mapsto [x \mapsto M^{1/2}(f_1(x), \dots, f_d(x))]$, where $\mathcal{H}_{\boldsymbol{k}}{}^{\oplus d}$ is the direct-sum Hilbert space of d copies of $\mathcal{H}_{\boldsymbol{k}}$ Then ι is a Hilbert-space isometry onto a vector-valued RKHS $\mathcal{H}_{\boldsymbol{K}}$ with matrix-valued reproducing kernel given by $\boldsymbol{K}(x, y) = \boldsymbol{k}(x, y)M$.*

Proof of Lem. B.5. Define $\gamma : \mathbb{R}^d \to \mathfrak{F}(\mathbb{R}^d, \mathcal{H}_{\boldsymbol{k}}{}^{\oplus d})$ via

$$\gamma(x)(\alpha) \triangleq \boldsymbol{k}(x, \cdot)M^{1/2}\alpha.$$

We have

$$\|\gamma(x)(\alpha)\|_{\mathcal{H}_{\boldsymbol{k}}{}^{\oplus d}} \le \|\boldsymbol{k}(x, \cdot)\|_{\boldsymbol{k}} \left\|M^{1/2}\right\|_2 \|\alpha\|_2 ,$$

so $\gamma(x)$ is bounded. Since $\gamma(x)$ is also linear, we have $\gamma(x) \in \mathfrak{B}(\mathbb{R}^d, \mathcal{H}_{\boldsymbol{k}}{}^{\oplus d})$. Let $\gamma(x)^* : \mathcal{H}_{\boldsymbol{k}}{}^{\oplus d} \to \mathbb{R}^d$ denote the Hilbert-space adjoint of $\gamma(x)$. Then for any $(f_1, \dots, f_d) \in \mathcal{H}_{\boldsymbol{k}}{}^{\oplus d}$, $\alpha \in \mathbb{R}^d$, we have

$$\begin{aligned}\langle \gamma(x)^*(f_1, \dots, f_d), \alpha\rangle &= \langle (f_1, \dots, f_d), \gamma(x)(\alpha)\rangle_{\mathcal{H}_{\boldsymbol{k}}{}^{\oplus d}}\\ &= \langle (f_1, \dots, f_d), \boldsymbol{k}(x, \cdot)M^{1/2}\alpha\rangle_{\mathcal{H}_{\boldsymbol{k}}{}^{\oplus d}}\\ &= \langle (f_1(x), \dots, f_d(x)), M^{1/2}\alpha\rangle\\ &= \langle M^{1/2}(f_1(x), \dots, f_d(x)), \alpha\rangle.\end{aligned}$$

Hence $\gamma(x)^*(f_1, \dots, f_d) = M^{1/2}(f_1(x), \dots, f_d(x))$, so $\iota(f_1, \dots, f_d)(x) = \gamma(x)^*(f_1, \dots, f_d)$. By Carmeli et al. (2006, Proposition 2.4), we see that ι is a partial isometry from $\mathcal{H}_{\boldsymbol{k}}{}^{\oplus d}$ onto a vector-valued RKHS space withv reproducing kernel $\boldsymbol{K}(x, y) = \gamma(x)^*\gamma(y) : \mathbb{R}^d \to \mathbb{R}^d$. For $\alpha \in \mathbb{R}^d$, previous calculation implies

$$\gamma(x)^*\gamma(y)(\alpha) = \gamma(x)^*(\boldsymbol{k}(y, \cdot)M^{1/2}\alpha) = M^{1/2}\boldsymbol{k}(y, x)M^{1/2}\alpha = \boldsymbol{k}(x, y)M.$$

□

Lemma B.6 (Stein operator is an isometry)**.** *Consider a Stein kernel $\boldsymbol{k}_p$ with base kernel $\boldsymbol{k}$ and preconditioning matrix M. Then, the Stein operator $\mathcal{S}_p$ defined by $\mathcal{S}_p(v) \triangleq \frac{1}{p}\nabla \cdot (pv)$ is an isometry from $\mathcal{H}_{\boldsymbol{K}}$ with $\boldsymbol{K} \triangleq \boldsymbol{k}M$ to $\mathcal{H}_{\boldsymbol{k}_p}$.*

Proof. This follows from Barp et al. (2022, Theorem 2.6) applied to $\boldsymbol{K}$. □

The previous two lemmas show that $\mathcal{S}_p \circ \iota$ is a Hilbert space isometry from $\mathcal{H}_{\boldsymbol{k}}{}^{\oplus d}$ to $\mathcal{H}_{\boldsymbol{k}_p}$. Note that $\mathcal{S}_p(v) = \langle \nabla \log p, h\rangle + \nabla \cdot h$. Hence, we immediately have

$$\mathcal{H}_{\boldsymbol{k}_p} = \left\{\langle \nabla \log p, M^{1/2}f\rangle + \nabla \cdot (M^{1/2}f) : f = (f_1, \dots, f_d) \in \mathcal{H}_{\boldsymbol{k}}{}^{\oplus d}\right\}. \tag{11}$$

We next build a divergence RKHS which is one of the summands used to form $\mathcal{H}_{\boldsymbol{k}_p}$.

Lemma B.7 (Divergence RKHS)**.** *Let $\boldsymbol{k} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a continuously differentiable kernel. Let M be an SPSD matrix. Define $\nabla^{\otimes 2} \cdot (M\boldsymbol{k}) : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ via*

$$\nabla^{\otimes 2} \cdot (M\boldsymbol{k})(x, y) \triangleq \nabla_y \cdot \nabla_x \cdot (M\boldsymbol{k}(x, y)) = \operatorname{tr}(M\nabla_x\nabla_y\boldsymbol{k}(x, y)). \tag{12}$$

Then $\nabla^{\otimes 2} \cdot (M\boldsymbol{k})$ is a kernel, and its RKHS $\mathcal{H}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}$ has the following explicit form

$$\mathcal{H}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})} = \nabla \cdot \mathcal{H}_{\boldsymbol{K}} = \left\{ \nabla \cdot (M^{1/2} f) : f = (f_1, \dots, f_d) \in \mathcal{H}_{\boldsymbol{k}}{}^{\oplus d} \right\}, \tag{13}$$

where $\boldsymbol{K} = M\boldsymbol{k}$. Moreover, $\nabla\cdot : \mathcal{H}_{\boldsymbol{K}} \to \mathcal{H}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}$ is an isometry.

Proof of Lem. B.7. First of all, by Steinwart and Christmann (2008, Corollary 4.36), every $f \in \mathcal{H}_{\boldsymbol{k}}$ is continuously differentiable, so $\partial_{x_i} f$ exists. By Lem. B.5, $\nabla \cdot \mathcal{H}_{\boldsymbol{K}}$ is well-defined and the right equality in (13) holds.

Define $\gamma : \mathbb{R}^d \to \mathfrak{F}(\mathbb{R}, \mathcal{H}_{\boldsymbol{K}})$ via

$$\gamma(x)(t) \triangleq t \textstyle\sum_{i=1}^d \partial_{x_i} \boldsymbol{K}(x, \cdot) e_i,$$

where $e_i \in \mathbb{R}^d$ is the ith standard basis vector in $\mathbb{R}^d$; by Barp et al. (2022, Lemma C.8) we have $\partial_{x_i} \boldsymbol{K}(x, \cdot)e_i \in \mathcal{H}_{\boldsymbol{K}}$. Note that

$$\|\gamma(x)(t)\|_{\boldsymbol{K}} = |t| \left\| \textstyle\sum_{i=1}^d \partial_{x_i} \boldsymbol{K}(x, \cdot)e_i \right\|_{\boldsymbol{K}},$$

so $\gamma(x) \in \mathfrak{B}(\mathbb{R}, \mathcal{H}_{\boldsymbol{K}})$. The adjoint $\gamma(x)^* \in \mathfrak{B}(\mathcal{H}_{\boldsymbol{K}}, \mathbb{R})$ must satisfy, for any $h \in \mathcal{H}_{\boldsymbol{K}}$,

$$t\gamma(x)^* h = \langle h, \gamma(x)(t) \rangle_{\boldsymbol{K}} = \left\langle h, t \textstyle\sum_{i=1}^d \partial_{x_i} \boldsymbol{K}(x, \cdot)e_i \right\rangle_{\boldsymbol{K}} = t\nabla \cdot h,$$

where we used the fact (Barp et al., 2022, Lemma C.8) that, for $c \in \mathbb{R}^d$, $h \in \mathcal{H}_{\boldsymbol{K}}$, $\langle \partial_{x_i} \boldsymbol{K}(x, \cdot)c, h \rangle = c^\top \partial_{x_i} h(x)$. So we find $\gamma(x)^*(h) = \nabla \cdot h(x)$. By Carmeli et al. (2006, Proposition 2.4), the map $A : \mathcal{H}_{\boldsymbol{K}} \to \mathfrak{F}(\mathbb{R}^d, \mathbb{R})$ defined by $A(h)(x) = \gamma^*(x)(h) = \nabla \cdot h(x)$, i.e., $A = \nabla\cdot$, is a partial isometry from $\mathcal{H}_{\boldsymbol{K}}$ to an RKHS $\mathcal{H}_{\nabla\cdot\boldsymbol{K}}$ with reproducing kernel

$$\gamma(x)^*\gamma(y) = \nabla \cdot \left(\textstyle\sum_{i=1}^d \partial_{x_i} \boldsymbol{K}(x, \cdot)e_i \right)(y) = \nabla_y \cdot \nabla_x \cdot \boldsymbol{K}(x, y) = \nabla^{\otimes 2} \cdot (M\boldsymbol{k})(x, y).$$

□

The following lemma shows that we can project a covering of $\mathcal{B}_{\boldsymbol{k}}$ consisting of arbitrary functions to a covering using functions only in $\mathcal{B}_{\boldsymbol{k}}$ while inflating the covering radius by at most 2.

Lemma B.8 (Projection of coverings into RKHS balls)**.** *Let $\boldsymbol{k}$ be a kernel, $A \subset \mathbb{R}^d$ be a set, and $\epsilon > 0$. Let $\mathcal{C}$ be a set of functions such that for any $f \in \mathcal{B}_{\boldsymbol{k}}$, there exists $g \in \mathcal{C}$ such that $\|f - g\|_{\infty, A} \le \epsilon$. Then*

$$\mathcal{N}_{\boldsymbol{k}}(A, 2\epsilon) \le |\mathcal{C}|.$$

Proof. We will build a $(\boldsymbol{k}, A, 2\epsilon)$ covering $\mathcal{C}'$ as follows. For any $h \in \mathcal{C}$, if there exists $h' \in \mathcal{B}_{\boldsymbol{k}}$ with $\|h' - h\|_{\infty, A} \le \epsilon$, then we include h' in $\mathcal{C}'$. By construction, $|\mathcal{C}'| \le |\mathcal{C}|$. Then, for any $f \in \mathcal{B}_{\boldsymbol{k}}$, by assumption, there exists $g \in \mathcal{C}$ such that $\|f - g\|_{\infty, A} \le \epsilon$. By construction, there exists $g' \in \mathcal{C}'$ such that $\|g' - g\|_{\infty, A} \le \epsilon$. Thus

$$\|f - g'\|_{\infty, A} \le \|f - g\|_{\infty, A} + \|g - g'\|_{\infty, A} \le 2\epsilon.$$

Hence C' is a $(\boldsymbol{k}, A, 2\epsilon)$ covering. □

We are now ready to bound the covering numbers of $\boldsymbol{k}_p$ by those of $\boldsymbol{k}$ and $\nabla^{\otimes 2} \cdot (M\boldsymbol{k})$. Our key insight towards this end is that any element in $\mathcal{H}_{\boldsymbol{k}_p}$ can be decomposed as a sum of functions originated from $\mathcal{H}_{\boldsymbol{k}}$ and a function from the divergence RKHS $\mathcal{H}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}$.

Lemma B.9 (Upper bounding covering number of Stein kernel with that of its base kernel)**.** *Let $\boldsymbol{k}_p$ be a Stein kernel with density p and preconditioning matrix M. For any $A \subset \mathbb{R}^d$, $\epsilon_1, \epsilon_2 > 0$,*

$$\mathcal{N}_{\boldsymbol{k}_p}(A, \epsilon) \leq \mathcal{N}_{\boldsymbol{k}}(A, \epsilon_1)^d \mathcal{N}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}(A, \epsilon_2),$$

for $\epsilon = 2(\sqrt{d}\epsilon_1 \sup_{x\in A} \left\|M^{1/2}\nabla \log p(x)\right\| + \epsilon_2)$.

Proof of Lem. B.9. Let $\mathcal{C}_{\boldsymbol{k}}$ be a $(\boldsymbol{k}, A, \epsilon_1)$ covering and $\mathcal{C}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}$ be a $(\nabla^{\otimes 2}\cdot(M\boldsymbol{k}), A, \epsilon_2)$ covering. Define $b \triangleq M^{1/2}\nabla \log p$. Form

$$\mathcal{C} \triangleq \left\{\langle b, \tilde{f}\rangle + \tilde{g} : \tilde{f} = (\tilde{f}_1, \dots, \tilde{f}_d) \in (\mathcal{C}_{\boldsymbol{k}})^d, \tilde{g} \in \mathcal{C}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}\right\} \subset \mathfrak{F}(\mathbb{R}^d, \mathbb{R}).$$

Then $|\mathcal{C}| \leq |\mathcal{C}_{\boldsymbol{k}}|^d \left|\mathcal{C}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}\right|$. Let $\boldsymbol{K} \triangleq \boldsymbol{k}M$. For any $h \in \mathcal{B}_{\boldsymbol{k}_p}$, by (11), we can find $f = (f_1, \dots, f_d) \in \mathcal{H}_{\boldsymbol{k}}{}^{\oplus d}$ with $f_i \in \mathcal{H}_{\boldsymbol{k}}$ such that

$$h = \mathcal{S}_p \circ \iota(f) = \langle \nabla \log p, M^{1/2} f\rangle + \nabla\cdot(M^{1/2}f) = \langle b, f\rangle + \nabla\cdot(M^{1/2}f).$$

Since ι and $\mathcal{S}_p$ are isometries, we have $f \in \mathcal{B}_{\mathcal{H}_{\boldsymbol{k}}{}^{\oplus d}}$. Since, for each i,

$$\|f_i\|_{\boldsymbol{k}} \leq \sqrt{\textstyle\sum_{j=1}^d \|f_j\|_{\boldsymbol{k}}^2} = \|f\|_{\mathcal{H}_{\boldsymbol{k}}{}^{\oplus d}} \leq 1,$$

we have $f_i \in \mathcal{B}_{\boldsymbol{k}}$. By Lem. B.7, $\nabla\cdot : \mathcal{H}_{\boldsymbol{K}} \to \mathcal{H}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}$ is also an isometry, so $\nabla\cdot(M^{1/2}f) \in \mathcal{B}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}$. Thus there exist $\tilde{f}_i \in \mathcal{C}_{\boldsymbol{k}}$ for each i and $\tilde{g} \in \mathcal{C}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}$ such that

$$\left\|f_i - \tilde{f}_i\right\|_{\infty,A} \leq \epsilon_1, \quad \left\|\nabla\cdot(M^{1/2}f) - \tilde{g}\right\|_{\infty,A} \leq \epsilon_2.$$

Let

$$\tilde{h}(x) \triangleq \langle b, \tilde{f}\rangle + \tilde{g} \in \mathcal{C}.$$

Then for $x \in A$,

$$\begin{aligned}\left|h(x) - \tilde{h}(x)\right| &= \left|\langle b(x), f(x) - \tilde{f}(x)\rangle + \nabla\cdot(M^{1/2}f(x)) - \tilde{g}(x)\right| \\ &\leq \|b(x)\| \sqrt{\textstyle\sum_{i=1}^d (f_i(x) - \tilde{f}_i(x))^2} + \left|\nabla\cdot(M^{1/2}f(x)) - \tilde{g}(x)\right| \\ &\leq \sqrt{d}\epsilon_1 \|b(x)\| + \epsilon_2.\end{aligned}$$

Hence

$$\left\|h - \tilde{h}\right\|_{\infty,A} \leq \sqrt{d}\epsilon_1 \sup_{x\in A} \|b(x)\| + \epsilon_2 \triangleq \epsilon_3.$$

Note that $\mathcal{C}$ that we constructed is not necessarily contained in $\mathcal{B}_{\boldsymbol{k}_p}$. By Lem. B.8, we can get a $(\boldsymbol{k}_p, A, 2\epsilon_3)$ covering and we are done. □

Corollary B.2 (Log-covering number bound for Stein kernel)**.** *Let $\boldsymbol{k}_p$ be a Stein kernel and $A \subset \mathbb{R}^d$. For any $r > 0$, $\epsilon > 0$,*

$$\log \mathcal{N}_{\boldsymbol{k}_p}(A, \epsilon) \leq d \log \mathcal{N}_{\boldsymbol{k}}\left(A, \tfrac{\epsilon}{4\sqrt{d}\mathfrak{S}_p(r)}\right) + \log \mathcal{N}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}\left(A, \tfrac{\epsilon}{4}\right),$$

where $\mathfrak{S}_p$ is defined in (9).

Proof. This is direct from Lem. B.9 with $\epsilon_1 = \frac{\epsilon}{4\sqrt{d}\mathfrak{S}_p(r)}$, $\epsilon_2 = \frac{\epsilon}{4}$. □

B.2.1. CASE OF DIFFERENTIABLE BASE KERNEL

Definition B.2 (s-times continuously differentiable kernel). *A kernel $\boldsymbol{k}$ is* s-times continuously differentiable *for $s \in \mathbb{N}$ if all partial derivatives $\partial^{\alpha,\alpha}\boldsymbol{k}$ exist and are continuous for all multi-indices $\alpha \in \mathbb{N}_0^d$ with $|\alpha| \le s$.*

Proposition B.4 (Covering number bound for $\boldsymbol{k}_p$ with differentiable base kernel). *Suppose $\boldsymbol{k}_p$ is a Stein kernel with an s-times continuously differentiable base kernel $\boldsymbol{k}$ for $s \ge 2$. Then there exist a constant $\mathfrak{C}_d > 0$ depending only on $(s, d, \boldsymbol{k}, M)$ such that for any $r > 0, \epsilon \in (0,1)$,*

$$\log \mathcal{N}_{\boldsymbol{k}_p}(\mathcal{B}_2(r), \epsilon) \le \mathfrak{C}_d r^d \mathfrak{S}_p^{d/s}(r)(1/\epsilon)^{d/(s-1)}.$$

Proof of Prop. B.4. Since $\boldsymbol{k}$ is s-times continuously differentiable, the divergence kernel $\nabla^{\otimes 2} \cdot (M\boldsymbol{k})$ is $(s-1)$-times continuously differentiable. By Dwivedi and Mackey (2022, Proposition 2(b)), there exists constants c_1, c_2 depending only on $(s, d, \boldsymbol{k}, M)$ such that, for any $r > 0, \epsilon_1, \epsilon_2 > 0$,

$$\log \mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(r), \epsilon_1) \le c_1 r^d (1/\epsilon)^{d/s},$$
$$\log \mathcal{N}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}(\mathcal{B}_2(r), \epsilon_2) \le c_2 r^d (1/\epsilon)^{d/(s-1)}.$$

By Cor. B.2 with $A = \mathcal{B}_2(r)$, we have, for any $r > 0$ and $\epsilon \in (0,1)$,

$$\begin{aligned}\log \mathcal{N}_{\boldsymbol{k}_p}(\mathcal{B}_2(r), \epsilon) &\le c_1 d r^d (4\sqrt{d}\mathfrak{S}_p(r))^{d/s}(1/\epsilon)^{d/\epsilon} + c_2 r^d (4/\epsilon)^{d/(s-1)} \\ &\le \mathfrak{C}_d r^d \mathfrak{S}_p^{d/s}(r)(1/\epsilon)^{d/(s-1)}\end{aligned}$$

for some $\mathfrak{C}_d > 0$ depending only on $(s, d, \boldsymbol{k}, M)$. □

B.2.2. CASE OF RADIALLY ANALYTIC BASE KERNEL

For a symmetric positive definite $M \in \mathbb{R}^{d\times d}$, we define, for $x \in \mathbb{R}^d$,

$$\|x\|_M \triangleq \sqrt{x^\top M^{-1} x}.$$

Definition B.3 (Radially analytic kernel). *A kernel $\boldsymbol{k}$ is* radially analytic *if $\boldsymbol{k}$ satisfies $\boldsymbol{k}(x, y) = \kappa(\|x - y\|_M^2)$ for a symmetric positive definite matrix $M \in \mathbb{R}^{d\times d}$ and a function $\kappa : \mathbb{R}_{\ge 0} \to \mathbb{R}$ real-analytic everywhere with convergence radius $R_\kappa > 0$ such that there exists a constant $C_\kappa > 0$ for which*

$$\left|\tfrac{1}{j!}\kappa_+^{(j)}(0)\right| \le C_\kappa (2/R_\kappa)^j, \textit{ for all } j \in \mathbb{N}_0, \tag{14}$$

where $\kappa_+^{(j)}$ indicates the j-th right-sided derivative of κ.

Example B.1 (Gaussian kernel). Consider the Gaussian kernel $\boldsymbol{k}(x, y) = \kappa(\|x - y\|_M^2)$ with $\kappa(t) = e^{-\frac{t}{2\sigma^2}}$ where $\sigma > 0$. Note the exponential function is real-analytic everywhere, and so is κ. Since $\kappa(t) = \sum_{j=0}^\infty \frac{(-t/2\sigma^2)^j}{j}$, we find $\frac{1}{j!}\kappa^{(j)}(0) = \frac{(-1)^j}{j(2\sigma^2)^j}$. Hence (14) holds with $C_\kappa = 1$ and $R_\kappa = 2\inf_{j\ge 0}(j(2\sigma^2)^j)^{1/j} = 4\sigma^2$.

Example B.2 (IMQ kernel). Consider the inverse multiquadric kernel $\boldsymbol{k}(x, y) = \kappa(\|x - y\|_M^2)$ with $\kappa(t) = (c^2 + t)^{-\beta}$ where $c, \beta > 0$. By Sun and Zhou (2008, Example 3), κ is real-analytic everywhere with $C_\kappa = c^{-2\beta}(2\beta + 1)^{\beta+1}$ and $R_\kappa = c^2$.

A simple calculation yields the following lemma.

Proposition B.5 (Expression for $\boldsymbol{k}_p$ with a radially analytic base kernel). *Suppose a Stein kernel $\boldsymbol{k}_p$ has a symmetric positive definite preconditioning matrix and a base kernel $\boldsymbol{k}(x, y) = \kappa(\|x - y\|_M^2)$ where κ is twice-differentiable. Then*

$$\begin{aligned}\boldsymbol{k}_p(x, y) = \langle \nabla \log p(x), M\nabla \log p(y)\rangle \kappa(\|x - y\|_M^2) - \\ 2\kappa_+'(\|x - y\|_M^2)\langle x - y, \nabla \log p(x) - \nabla \log p(y)\rangle - \\ 4\kappa_+''(\|x - y\|_M^2)\,\|x - y\|_M^2 - 2d\kappa_+'(\|x - y\|_M^2).\end{aligned} \tag{15}$$

In particular,

$$\boldsymbol{k}_p(x, x) = \left\|M^{1/2}\nabla \log p(x)\right\|_2^2 \kappa(0) - 2d\kappa_+'(0).$$

Proof of Prop. B.5. From $\boldsymbol{k}(x,y) = \kappa(\|x-y\|_M^2) = \kappa((x-y)^\top M^{-1}(x-y))$, we compute, using (12),

$$\begin{aligned}
\nabla_y \boldsymbol{k}(x,y) &= -2\kappa_+'(\|x-y\|_M^2)M^{-1}(x-y)\\
\nabla_x\nabla_y \boldsymbol{k}(x,y) &= -2\kappa_+'(\|x-y\|_M^2)M^{-1} - 4\kappa_+''(\|x-y\|_M^2)M^{-1}(x-y)((x-y)M^{-1})^\top\\
\nabla^{\otimes 2}\cdot(M\boldsymbol{k})(x,y) &= \mathrm{tr}(M\nabla_x\nabla_y \boldsymbol{k}(x,y)) = -4\kappa_+''(\|x-y\|_M^2)\,\|x-y\|_M^2 - 2d\kappa_+'(\|x-y\|_M^2). \qquad (16)
\end{aligned}$$

The form (15) then follows from applying Prop. B.3. □

We next show that the divergence kernel $\nabla^{\otimes 2}\cdot(M\boldsymbol{k})$ is radially analytic given that $\boldsymbol{k}$ is.

Lemma B.10 (Convergence radius of divergence kernel)**.** *Let $\boldsymbol{k}$ be a radially analytic kernel with the corresponding real-analytic function κ, convergence radius R_κ with constant C_κ, and a symmetric positive definite matrix M. Then*

$$\nabla^{\otimes 2}\cdot(M\boldsymbol{k})(x,y) = \kappa_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}(\|x-y\|_M^2),$$

where $\kappa_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}:\mathbb{R}_{\geq 0}\to\mathbb{R}$ is the real-analytic function defined as

$$\kappa_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}(t) \triangleq -4\kappa_+''(t)t - 2d\kappa_+'(t).$$

Moreover, $\kappa_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}$ has convergence radius with constant

$$R_{\kappa_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}} = \tfrac{R_\kappa}{4d+8}, \quad C_{\kappa_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}} = 4dC_\kappa/R_\kappa.$$

Proof of Lem. B.10. The first statement regarding the form of $\kappa_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}$ directly follows from (16). Next, iterative differentiation yields, for $j\in\mathbb{N}_0$,

$$\kappa^{(j)}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}(t) = -(2d+4j)\kappa_+^{(j+1)}(t) - 4\kappa_+^{(j+2)}(t)t.$$

Thus

$$\begin{aligned}
\left|\tfrac{1}{j!}\kappa^{(j)}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}(0)\right| &= \tfrac{2d+4j}{j!}\kappa_+^{(j+1)}(0)\\
&= \tfrac{(2d+4j)(j+1)}{(j+1)!}\kappa_+^{(j+1)}(0)\\
&\leq (2d+4j)(j+1)C_\kappa(2/R_\kappa)^{j+1}. \qquad (17)
\end{aligned}$$

For $j\geq 1$,

$$\begin{aligned}
\left|\tfrac{1}{j!}\kappa^{(j)}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}(0)\right| &\leq (2C_\kappa/R_\kappa)\left(((2d+4j)(j+1))^{1/j}2/R_\kappa\right)^j\\
&\leq (2C_\kappa/R_\kappa)\left((2(2d+4)\cdot 2/R_\kappa\right)^j.
\end{aligned}$$

where we used the fact that $((2d+4j)(j+1))^{1/j}$ is decreasing in j. For $j=0$, (17) is just $2dC_\kappa\cdot 2/R_\kappa$. Hence $\kappa_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}$ is analytic with $C_{\kappa_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}} = 4dC_\kappa/R_\kappa$ and $R_{\kappa_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}} = \frac{R_\kappa}{4d+8}$. □

We will use the following lemma repeatedly.

Lemma B.11 (Covering number of radially analytic kernel with M-metric)**.** *Let $\boldsymbol{k}_0$ be a radially analytic kernel with $\boldsymbol{k}_0(x,y)=\kappa(\|x-y\|_2^2)$. For any symmetric positive definite $M\in\mathbb{R}^{d\times d}$, consider the radially analytic kernel $\boldsymbol{k}(x,y)\triangleq\kappa(\|x-y\|_M^2)$. Then for any $A\subset\mathbb{R}^d$ and $\epsilon>0$, we have*

$$\mathcal{N}_{\boldsymbol{k}}(M^{-1/2}(A),\epsilon) = \mathcal{N}_{\boldsymbol{k}_0}(A,\epsilon).$$

In particular, for any $r>0$,

$$\mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(r),\epsilon) \leq \mathcal{N}_{\boldsymbol{k}_0}(\mathcal{B}_2(r\|M^{1/2}\|_2),\epsilon).$$

Proof. Note that $\boldsymbol{k}(x,y) = \boldsymbol{k}_0(M^{-1/2}x, M^{-1/2}y)$. By Paulsen and Raghupathi (2016, Theorem 5.7), $\mathcal{H}_{\boldsymbol{k}} = \{f \circ M^{-1/2} : f \in \mathcal{H}_{\boldsymbol{k}_0}\}$, and moreover $\mathcal{B}_{\boldsymbol{k}} = \{f \circ M^{-1} : f \in \mathcal{B}_{\boldsymbol{k}_0}\}$. Let $\mathcal{C}_0$ be a $(\boldsymbol{k}_0, A, \epsilon)$ covering. Form $\mathcal{C} = \{h \circ M^{-1/2} : h \in \mathcal{C}_0\} \subset \mathcal{B}_{\boldsymbol{k}}$. For any element $f \circ M^{-1/2} \in \mathcal{B}_{\boldsymbol{k}}$ where $f \in \mathcal{B}_{\boldsymbol{k}_0}$, there exists $h \in \mathcal{C}_0$ such that $\|f - h\|_{\infty,A} \le \epsilon$. Thus

$$\left\|f \circ M^{-1/2} - h \circ M^{-1/2}\right\|_{\infty, M^{-1/2}(A)} = \|f - h\|_{\infty,A} \le \epsilon.$$

Thus $\mathcal{N}_{\boldsymbol{k}}(M^{-1/2}(A), \epsilon) \le \mathcal{N}_{\boldsymbol{k}_0}(A, \epsilon)$. By considering M^{-1} in place of M, we get our desired equality.

For the second statement, by letting $A = M^{1/2}\mathcal{B}_2(r)$, we have

$$\mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(r), \epsilon) = \mathcal{N}_{\boldsymbol{k}_0}(M^{1/2}\mathcal{B}_2(r), \epsilon) \le \mathcal{N}_{\boldsymbol{k}_0}(\mathcal{B}_2(r\|M^{1/2}\|_2), \epsilon),$$

where we use the fact that $M^{1/2}\mathcal{B}_2(r) \subset \mathcal{B}_2(r\|M^{1/2}\|_2)$. □

In the next lemma, we rephrase the result from Sun and Zhou (2008, Theorem 2) for bounding the covering number of a radially analytic kernel.

Lemma B.12 (Covering number bound for radially analytic kernel)**.** *Let $\boldsymbol{k}$ be a radially analytic kernel with $\boldsymbol{k}(x,y) = \kappa(\|x-y\|_2^2)$. Then, there exist a polynomial $P(r)$ of degree $2d$ and a constant C depending only on (κ, d) such that for any $r > 0$, $\epsilon \in (0, 1/2)$,*

$$\log \mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(r), \epsilon) \le P(r)(\log(1/\epsilon) + C)^{d+1}.$$

Proof of Lem. B.12. Let R_κ, C_κ denote the constants for κ as in (14). By and Sun and Zhou (2008, Theorem 2) with $R = 1$, $D = 2r$, and Lem. B.2, for $\epsilon \in (0, 1/2)$, we have

$$\log \mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(r), \epsilon) \le N_2(\mathcal{B}_2(r), r^\dagger/2)\left(4\log(1/\epsilon) + 2 + 4\log\left(16\sqrt{C_\kappa} + 1\right)\right)^{d+1},$$

where $r^\dagger = \min(\frac{\sqrt{R_\kappa}}{2d}, \sqrt{R_\kappa + (2r)^2} - 2r)$, and $N_2(\mathcal{B}_2(r), r^\dagger/2)$ is the covering number of $\mathcal{B}_2(r)$ as a subset of $\mathbb{R}^d$, which can be further bounded by (Wainwright, 2019, (5.8))

$$N_2(\mathcal{B}_2(r), r^\dagger/2) \le \left(1 + \tfrac{4r}{r^\dagger}\right)^d.$$

If $r^\dagger = \sqrt{R_\kappa + (2r)^2} - 2r$, then $\frac{r}{r^\dagger} = \frac{r}{\sqrt{R_\kappa + (2r)^2} - 2r} = \frac{r(\sqrt{R_\kappa + (2r)^2} + 2r)}{R_\kappa} \le \frac{r(\sqrt{R_\kappa} + 4r)}{R_\kappa}$ which is a quadratic polynomial in r. Hence for a constant $C > 0$ and a polynomial $P(r)$ of degree $2d$ that depend only on (κ, d), we have the claim. □

Proposition B.6 (Covering number bound for $\boldsymbol{k}_p$ with radially analytic base kernel)**.** *Suppose $\boldsymbol{k}_p$ is a Stein kernel with a preconditioning matrix M and a radially analytic base kernel $\boldsymbol{k}$ based on a real-analytic function κ. Then there exist a constant $C > 0$ and a polynomial $P(r)$ of degree $2d$ depending only on (κ, d, M) such that for any $r > 0, \epsilon \in (0,1)$,*

$$\log \mathcal{N}_{\boldsymbol{k}_p}(\mathcal{B}_2(r), \epsilon) \le \left(\log \tfrac{\mathfrak{S}_p(r)}{\epsilon} + C\right)^{d+1} P(r). \tag{18}$$

Proof of Prop. B.6. Recall $\boldsymbol{k}(x,y) = \kappa(\|x-y\|_M^2)$. Consider $\boldsymbol{k}_0(x,y) \triangleq \kappa(\|x-y\|_2^2)$. For $\epsilon_1 \in (0, 1/2)$, by Lem. B.11, we have

$$\log \mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(r/\|M^{1/2}\|_2), \epsilon_1) \le \log \mathcal{N}_{\boldsymbol{k}_0}(\mathcal{B}_2(r), \epsilon_1/2).$$

Thus by Lem. B.12, there exists a polynomial $P_{\boldsymbol{k}}(r)$ of degree $2d$ and a constant $C_{\boldsymbol{k}}$ depending only on (κ, d, M) such that

$$\log \mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(r), \epsilon_1) \le P_{\boldsymbol{k}}(r)(\log(1/\epsilon_1) + C_{\boldsymbol{k}})^{d+1}$$

Similarly, for $\epsilon_2 \in (0, 1/2)$, by Lem. B.10 and Lem. B.12, we have, for a constant $C_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})} > 0$ and a polynomial $P_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}(r)$ of degree $2d$ that depend only on (κ, d, M),

$$\log \mathcal{N}_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}(\mathcal{B}_2(r), \epsilon_2) \le P_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})}(r)(\log(1/\epsilon_2) + C_{\nabla^{\otimes 2}\cdot(M\boldsymbol{k})})^{d+1}.$$

For a given $\epsilon \in (0,1)$, let $\epsilon_1 = \frac{\epsilon}{4\sqrt{d}\mathfrak{S}_p(r)}$ and $\epsilon_2 = \frac{\epsilon}{4}$. Then since $\mathfrak{S}_p \geq 1$, we have $\epsilon_1, \epsilon_2 \in (0, 1/2)$. By Cor. B.2 with $A = \mathcal{B}_2(r)$, we obtain, for a constants $C > 0$ and a polynomial $P(r)$ of degree $2d$ that depend only on (κ, d, M),

$$\log \mathcal{N}_{\boldsymbol{k}_p}(\mathcal{B}_2(r), \epsilon) \leq P(r)(\log(1/\epsilon) + \log \mathfrak{S}_p(r) + C)^{d+1}.$$

Hence (18) is shown. □

When $\log \mathfrak{S}_p(r)$ grows polynomially in r, we apply Prop. B.6 to immediately obtain the following.

Corollary B.3. *Under the assumption of Prop. B.6, suppose $\mathfrak{S}_p(r) = O(\mathrm{poly}(r))$. Then for any $\delta > 0$, there exists $\mathfrak{C}_d > 0$ such that*

$$\log \mathcal{N}_{\boldsymbol{k}_p}(\mathcal{B}_2(r), \epsilon) \leq \mathfrak{C}_d \log(e/\epsilon)^{d+1}(r+1)^{2d+\delta}.$$

Proof of Cor. B.3. This immediately follows from Prop. B.6 by using $\delta > 0$ to absorb the $\log \mathfrak{S}_p(r) = O(r^\delta)$ term. □

B.2.3. Proof of Prop. 1: Stein kernel growth rates

This follows from Prop. B.4 and Cor. B.3, and by noticing that if $\sup_{\|x\|_2 \leq r} \|\nabla \log p(x)\|_2$ is bounded by a degree d_ℓ polynomial, then so is

$$\mathfrak{S}_p(r) = \sup_{\|x\|_2 \leq r} \left\|M^{1/2} \nabla \log p(x)\right\|_2 \leq \left\|M^{1/2}\right\|_2 \sup_{\|x\|_2 \leq r} \|\nabla \log p(x)\|_2 .$$

□

C. A Debiasing Benchmark

C.1. MMD of unbiased i.i.d. sample points

We start by showing that sequence of n points sampled i.i.d. from $\mathbb{P}$ achieves $\Theta(n^{-1})$ squared $\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}$ to $\mathbb{P}$ in expectation.

Proposition C.1 (MMD of unbiased i.i.d. sample points)**.** *Let $\boldsymbol{k}_\mathbb{P}$ be a kernel satisfying Assum. 1 with $\mathfrak{p} \geq 1$. Let $\mathcal{S}_n = (x_i)_{i\in[n]}$ be n i.i.d. samples from $\mathbb{P}$. Then*

$$\mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n, \mathbb{P})^2] = \frac{\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}_\mathbb{P}(x,x)]}{n}.$$

Proof of Prop. C.1. We compute

$$\mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n, \mathbb{P})^2] = \mathbb{E}[\textstyle\sum_{i,j\in[n]} \frac{1}{n^2}\boldsymbol{k}_\mathbb{P}(x_i, x_j)] = \frac{1}{n^2}\sum_{i,j\in[n]} \mathbb{E}[\boldsymbol{k}_\mathbb{P}(x_i, x_j)] = \frac{1}{n}\mathbb{E}[\boldsymbol{k}_\mathbb{P}(x_1, x_1)],$$

where we used the fact that $\boldsymbol{k}_\mathbb{P}$ is mean-zero with respect to $\mathbb{P}$ and the independence of x_i, x_j for $i \neq j$. □

C.2. Proof of Thm. 1: Debiasing via simplex reweighting

We make use of the self-normalized importance sampling weights $w_j^{\mathrm{SNIS}} = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x_j)/\sum_{i\in[n]} \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x_i)$ for $j \in [n]$ in our proofs. Notice that $(w_1^{\mathrm{SNIS}}, \ldots, w_n^{\mathrm{SNIS}})^\top \in \Delta_{n-1}$ and hence

$$\mathrm{MMD}_{\mathrm{OPT}} \leq \mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(w_i^{\mathrm{SNIS}}\delta_{x_i}, \mathbb{P}) = \frac{\|\sum_{i=1}^n \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x_i)\boldsymbol{k}_\mathbb{P}(x_i,\cdot)\|_{\boldsymbol{k}_\mathbb{P}}}{\sum_{i=1}^n \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x_i)} = \frac{\|\frac{1}{n}\sum_{i=1}^n \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x_i)\boldsymbol{k}_\mathbb{P}(x_i,\cdot)\|_{\boldsymbol{k}_\mathbb{P}}}{\frac{1}{n}\sum_{i=1}^n \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x_i)}.$$

Introduce the bounded in probability notation $X_n = O_p(g_n)$ to mean $\Pr(|X_n/g_n| > C_\epsilon) \leq \epsilon$ for all $n \geq N_\epsilon$ for any $\epsilon > 0$. Then we claim that under the conditions assumed in Thm. 1,

$$\|\tfrac{1}{n}\textstyle\sum_{i=1}^n \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x_i)\boldsymbol{k}_\mathbb{P}(x_i,\cdot)\|_{\boldsymbol{k}_\mathbb{P}} = O_p(n^{-\frac{1}{2}}) \quad \text{and} \quad \tfrac{1}{n}\textstyle\sum_{i=1}^n \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x_i) \to 1 \text{ almost surely,} \tag{19}$$

so that by Slutsky's theorem (Wellner et al., 2013, Ex. 1.4.7), we have $\mathrm{MMD}_{\mathrm{OPT}} = O_p(n^{-\frac{1}{2}})$ as desired. We prove the claims in (19) in two main steps: (a) first, we construct a weighted RKHS and then (b) establish a central limit theorem (CLT) that allows us to conclude both claims from (19)

Constructing a weighted and separable RKHS Define the kernel $\boldsymbol{k}_{\mathbb{Q}}(x,y) \triangleq \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x)\boldsymbol{k}_{\mathbb{P}}(x,y)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(y)$ with Hilbert space $\mathcal{H}_{\boldsymbol{k}_{\mathbb{Q}}} = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$ and the elements $\xi_i \triangleq \boldsymbol{k}_{\mathbb{Q}}(x_i,\cdot) = \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x_i)\boldsymbol{k}_{\mathbb{P}}(x_i,\cdot)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\cdot) \in \mathcal{H}_{\boldsymbol{k}_{\mathbb{Q}}}$ for each $i \in \mathbb{N}$. By Paulsen and Raghupathi (2016, Prop. 5.20), any element in $\mathcal{H}_{\boldsymbol{k}_{\mathbb{Q}}}$ is represented by $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}f$ for some $f \in \mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$ and moreover, $f \mapsto \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}f$ preserves inner product between the two RKHSs, i.e., $\langle f,g\rangle_{\boldsymbol{k}_{\mathbb{P}}} = \langle \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}f, \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}g\rangle_{\boldsymbol{k}_{\mathbb{Q}}}$ for $f,g \in \mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$, which in turn implies $\|f\|_{\boldsymbol{k}_{\mathbb{P}}} = \|\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}f\|_{\boldsymbol{k}_{\mathbb{Q}}}$. As a result, we also have that

$$\|\tfrac{1}{n}\textstyle\sum_{i=1}^n \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x_i)\boldsymbol{k}_{\mathbb{P}}(x_i,\cdot)\|_{\boldsymbol{k}_{\mathbb{P}}} = \|\tfrac{1}{n}\sum_{i=1}^n \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x_i)\boldsymbol{k}_{\mathbb{P}}(x_i,\cdot)\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(\cdot)\|_{\boldsymbol{k}_{\mathbb{Q}}} = \|\tfrac{1}{n}\sum_{i=1}^n \xi_i\|_{\boldsymbol{k}_{\mathbb{Q}}}. \tag{20}$$

Since $\mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$ is separable, there exists a dense countable subset $(f_n)_{n\in\mathbb{N}} \subset \mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$. For any $\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}f \in \mathcal{H}_{\boldsymbol{k}_{\mathbb{Q}}}$, there exists $\{n_k\}_{k\in\mathbb{N}}$ such that $\lim_{k\to\infty}\|f_{n_k} - f\|_{\boldsymbol{k}_{\mathbb{P}}} = 0$. Since $\|\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}f_{n_k} - \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}f\|_{\boldsymbol{k}_{\mathbb{Q}}} = \|\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(f_{n_k} - f)\|_{\boldsymbol{k}_{\mathbb{Q}}} = \|f_{n_k} - f\|_{\boldsymbol{k}_{\mathbb{P}}}$ due to inner-product preservation, we thus have $\lim_{k\to\infty}\|\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}f_{n_k} - \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}f\|_{\boldsymbol{k}_{\mathbb{Q}}} = \lim_{k\to\infty}\|f_{n_k} - f\|_{\boldsymbol{k}_{\mathbb{P}}}0$, so $(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}f_n)_{n\in\mathbb{N}}$ is dense in $\mathcal{H}_{\boldsymbol{k}_{\mathbb{Q}}}$, showing that $\mathcal{H}_{\boldsymbol{k}_{\mathbb{Q}}}$ is separable.

Harris recurrence of the chain $(x_i)_{i\in\mathbb{N}}$ Let μ_1 denote the distribution of x_1. Since $\mathcal{S}_\infty = (x_i)_{i=1}^\infty$ is a homogeneous ϕ-irreducible geometrically ergodic Markov chain with stationary distribution $\mathbb{Q}$, it is also positive (Meyn and Tweedie, 2012, Ch. 10) by definition and aperiodic by Douc et al. (2018, Lem. 9.3.9). Moreover, since $\mathcal{S}_\infty$ is ϕ-irreducible, aperiodic, and geometrically ergodic in the sense of Gallegos-Herrada et al. (2023, Thm. 1) and μ_1 is absolutely continuous with respect to $\mathbb{P}$, we will assume, without loss of generality, that $\mathcal{S}_\infty$ is Harris recurrent (Meyn and Tweedie, 2012, Ch. 9), since, by Qin (2023, Lem. 9), $\mathcal{S}_\infty$ is equal to a geometrically ergodic Harris chain with probability 1.

CLT for $\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i$ We now show that $\frac{1}{\sqrt{n}}\sum_{i=1}^n \xi_i$ converges to a Gaussian random element taking values in $\mathcal{H}_{\boldsymbol{k}_{\mathbb{Q}}}$. We separate the proof in two parts: first when the initial distribution $\mu_1 = \mathbb{Q}$ and next when $\mu_1 \neq \mathbb{Q}$.

Case 1: $\mu_1 = \mathbb{Q}$ When $\mu_1 = \mathbb{Q}$, $\mathcal{S}_\infty$ is a strictly stationary chain. By Bradley (2005, Thm. 3.7 and (1.11)), since $\mathcal{S}_\infty$ is geometrically ergodic, its strong mixing coefficients $(\tilde{\alpha}_i)_{i\in\mathbb{N}}$ satisfy $\tilde{\alpha}_i \le C\rho^i$ for some $C > 0$ and $\rho \in [0,1)$ and all $i \in \mathbb{N}$. Since each ξ_i is a measurable function of x_i, the strong mixing coefficients $(\alpha_i)_{i\in\mathbb{N}}$ of $(\xi_i)_{i\in\mathbb{N}}$ satisfy $\alpha_i \le \tilde{\alpha}_i \le C\rho^i$ for each $i \in \mathbb{N}$. Consequently, $\sum_{i\in\mathbb{N}} i^{2/\delta}\alpha_i < \infty$ for $\delta = 2\mathfrak{q} - 2 > 0$. Note that we also have

$$\mathbb{E}_{z\sim\mathbb{Q}}[\|\boldsymbol{k}_{\mathbb{Q}}(z,\cdot)\|_{\boldsymbol{k}_{\mathbb{Q}}}^{2+\delta}] = \mathbb{E}_{z\sim\mathbb{Q}}[\boldsymbol{k}_{\mathbb{Q}}(z,z)^{\mathfrak{q}}] = \mathbb{E}_{z\sim\mathbb{Q}}[\tfrac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(z)^{2\mathfrak{q}}\boldsymbol{k}_{\mathbb{P}}(z,z)^{\mathfrak{q}}] = \mathbb{E}_{x\sim\mathbb{P}}[\tfrac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x)^{2\mathfrak{q}-1}\boldsymbol{k}_{\mathbb{P}}(x,x)^{\mathfrak{q}}] < \infty,$$

$\mathbb{E}_{x_i\sim\mathbb{Q}}[\xi_i] = \mathbb{E}_{x_i\sim\mathbb{P}}[\boldsymbol{k}_{\mathbb{P}}(x_i,\cdot)] = 0$ and that $\mathcal{H}_{\boldsymbol{k}_{\mathbb{Q}}}$ is separable. Since $\mathcal{S}_\infty$ is a strictly stationary chain, we conclude that $(\xi_i)_{i\in\mathbb{N}}$ is a strictly stationary centered sequence of $\mathcal{H}_{\boldsymbol{k}_{\mathbb{Q}}}$-valued random variables satisfying the conditions needed to invoke Merlevède et al. (1997, Cor. 1), and hence $\sum_{i=1}^n \xi_i/\sqrt{n}$ converges in distribution to a Gaussian random element taking values in $\mathcal{H}_{\boldsymbol{k}_{\mathbb{Q}}}$.

Case 2: $\mu_1 \neq \mathbb{Q}$ Since $\mathcal{S}_\infty$ is positive Harris and $\sum_{i=1}^n \xi_i/\sqrt{n}$ satisfies a CLT for the initial distribution $\mu_1 = \mathbb{Q}$, Meyn and Tweedie (2012, Prop. 17.1.6) implies that $\sum_{i=1}^n \xi_i/\sqrt{n}$ also satisfies the same CLT for any initial distribution μ_1.

Putting the pieces together for (19) Since, for any initial distribution for x_1, the sequence $(\sum_{i=1}^n \xi_i/\sqrt{n})_{n\in\mathbb{N}}$ converges in distribution and that $\mathcal{H}_{\boldsymbol{k}_{\mathbb{Q}}}$ is separable and (by virtue of being a Hilbert space) complete, Prokhorov's theorem (Billingsley, 2013, Thm. 5.2) implies that the sequence is also tight, i.e., $\|\sum_{i=1}^n \xi_i\|_{\boldsymbol{k}_{\mathbb{Q}}}/\sqrt{n} = O_p(1)$. Consequently,

$$\|\tfrac{1}{n}\textstyle\sum_{i=1}^n \frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x_i)\boldsymbol{k}_{\mathbb{P}}(x_i,\cdot)\|_{\boldsymbol{k}_{\mathbb{P}}} \overset{(20)}{=} \|\tfrac{1}{n}\sum_{i=1}^n \xi_i\|_{\boldsymbol{k}_{\mathbb{Q}}} = \tfrac{1}{\sqrt{n}}\cdot\|\tfrac{\sum_{i=1}^n \xi_i}{\sqrt{n}}\|_{\boldsymbol{k}_{\mathbb{Q}}} = O_p(n^{-\frac{1}{2}}),$$

as desired for the first claim in (19). Moreover, the strong law of large numbers for positive Harris chains (Meyn and Tweedie, 2012, Thm. 17.0.1(i)) implies that $\frac{1}{n}\sum_{i\in[n]}\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(x_i)$ converges almost surely to $\mathbb{E}_{z\sim\mathbb{Q}}[\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}(z)] = 1$ as desired for the second claim in (19). □

C.3. Proof of Thm. 2: Better-than-i.i.d. debiasing via simplex reweighting

We start with Thm. C.1, proved in App. C.4, that bounds $\mathrm{MMD}_{\mathrm{OPT}}$ in terms of the eigenvalues of the integral operator of the kernel $\boldsymbol{k}_{\mathbb{P}}$. Our proof makes use of a weight construction from Liu and Lee (2017, Theorem 3.2), but is a non-trivial generalization of their proof as we no longer assume uniform bounds on the eigenfunctions, and instead leverage truncated variations of Bernstein's inequality (Lems. C.2 and C.3) to establish suitable concentration bounds.

Theorem C.1 (Debiasing via i.i.d. simplex reweighting)**.** *Consider a kernel $\boldsymbol{k}_{\mathbb{P}}$ satisfying Assum. 1 with $\mathfrak{p} = 2$. Let $(\lambda_\ell)_{\ell=1}^\infty$ be the decreasing sequence of eigenvalues of $T_{\boldsymbol{k}_{\mathbb{P}},\mathbb{P}}$ defined in (4). Let $\mathcal{S}_n$ be a sequence of $n \in 2\mathbb{N}$ i.i.d. random variables with law $\mathbb{Q}$ such that $\mathbb{P}$ is absolutely continuous with respect to $\mathbb{Q}$ and $\|\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\|_\infty \leq M$ for some $M > 0$. Futhermore, assume there exist constants $\delta_n, B_n > 0$ such that $\Pr\left(\|\boldsymbol{k}_{\mathbb{P}}\|_n > B_n\right) < \delta_n$. Then for all $L \in \mathbb{N}$ such that $\lambda_L > 0$, we have*

$$\mathbb{E}[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w_{\mathrm{OPT}}}, \mathbb{P})] \leq \frac{8M}{n}\left(\frac{2M}{n}\frac{\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}^2_{\mathbb{P}}(x,x)]}{\lambda_L} + \textstyle\sum_{\ell>L}\lambda_\ell\right) + \epsilon_n \mathbb{E}[\boldsymbol{k}^2_{\mathbb{P}}(x_1, x_1)], \tag{21}$$

where

$$\epsilon_n^2 \triangleq n\exp\left(\frac{-3n}{16MB_n/\lambda_L}\right) + 2\exp\left(\frac{-n}{16M^2}\right) + 2\exp\left(-\frac{n}{64M^2(\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}_{\mathbb{P}}(x,x)]+B_n/12)/\lambda_L}\right) + \delta_n. \tag{22}$$

Given Thm. C.1, Thm. 2 follows, i.e., we have $\mathbb{E}[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w_{\mathrm{OPT}}}, \mathbb{P})] = o(n^{-1})$, as long as we can show (i) $\mathbb{E}[\boldsymbol{k}^2_{\mathbb{P}}(x_1, x_1)] < \infty$, which in turn holds when $\mathfrak{q} > 3$ as assumed in Thm. 2, and (ii) find sequences $(B_n)_{n=1}^\infty$, $(\delta_n)_{n=1}^\infty$, and $(L_n)_{n=1}^\infty$ such that $\Pr(\|\boldsymbol{k}_{\mathbb{P}}\|_n > B_n) < \delta_n$ for all n and the following conditions are met:

(a) $\frac{\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}^2_{\mathbb{P}}(x,x)]}{\lambda_{L_n}} = o(n)$;

(b) $\frac{B_n}{\lambda_{L_n}} = O(n^\beta)$, for some $\beta < 1$;

(c) $\sum_{\ell>L_n}\lambda_\ell = o(1)$;

(d) $\delta_n = o(n^{-2})$.

We now proceed to establish these conditions under the assumptions of Thm. 2.

Condition (d) By the de La Vallée Poussin Theorem (Chandra, 2015, Thm. 1.3) applied to the $\mathbb{Q}$-integrable function $x \mapsto \boldsymbol{k}_{\mathbb{P}}(x,x)^{\mathfrak{q}}$ (which is a uniformly integrable family with one function), there exists a convex increasing function G such that $\lim_{t\to\infty}\frac{G(t)}{t} = \infty$ and $\mathbb{E}[G(\boldsymbol{k}_{\mathbb{P}}(x_1,x_1)^{\mathfrak{q}})] < \infty$. Thus,

$$\begin{aligned}\Pr\left(\boldsymbol{k}_{\mathbb{P}}(x_1,x_1) > n^{3/\mathfrak{q}}\right) = \Pr\left(\boldsymbol{k}_{\mathbb{P}}(x_1,x_1)^{\mathfrak{q}} > n^3\right) &= \Pr\left(G(\boldsymbol{k}_{\mathbb{P}}(x_1,x_1))^{\mathfrak{q}} > G(n^3)\right)\\ &\leq \frac{\mathbb{E}[G(\boldsymbol{k}_{\mathbb{P}}(x_1,x_1))^{\mathfrak{q}}]}{G(n^3)} = o(n^{-3}),\end{aligned}$$

where the last step uses $\lim_{t\to\infty}\frac{G(t)}{t} = \infty$. Hence by the union bound,

$$\Pr\left(\|\boldsymbol{k}_{\mathbb{P}}\|_n > n^{3/\mathfrak{q}}\right) = \Pr\left(\max_{i\in[n]}\boldsymbol{k}_{\mathbb{P}}(x_i,x_i) > n^{3/\mathfrak{q}}\right) \leq n\Pr\left(\boldsymbol{k}_{\mathbb{P}}(x_1,x_1) > n^{3/\mathfrak{q}}\right) = o(n^{-2}).$$

Hence if we set $B_n = n^\tau$ for $\tau \triangleq 3/\mathfrak{q} < 1$, there exists $(\delta_n)_{n=1}^\infty$ such that $\delta_n = o(n^{-2})$. This fulfills (d) and that $\Pr(\|\boldsymbol{k}_{\mathbb{P}}\|_n > B_n) < \delta_n$.

To prove remaining conditions, without loss of generality, we assume that $\lambda_\ell > 0$ for all $\ell \in \mathbb{N}$, since otherwise we can choose L_n to be, for all n, the largest ℓ such that $\lambda_\ell > 0$. Then $\sum_{\ell>L_n}\lambda_{L_n} = 0$ and all other conditions are met.

Condition (c) If $L_n \to \infty$, then (c) is fulfilled since $\sum_\ell \lambda_\ell < \infty$, which follows from Lem. B.1(d) and that

$$\textstyle\sum_\ell \lambda_\ell = \sum_{\ell=1}^\infty \lambda_i \mathbb{E}_{x\sim\mathbb{P}}[\phi_\ell(x)^2] = \sum_{\ell=1}^\infty \lambda_i \mathbb{E}_{x\sim\mathbb{P}}[\phi_\ell(x)^2] = \mathbb{E}_{x\sim\mathbb{P}}[\sum_{\ell=1}^\infty \lambda_i\phi_\ell(x)^2] = \mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}_{\mathbb{P}}(x,x)] < \infty.$$

Conditions (a) and (b) Note that the condition (a) is subsumed by (b) since $\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}^2_{\mathbb{P}}(x,x)] < \infty$. It remains to choose $(L_n)_{n=1}^\infty$ to satisfy (b) such that $\lim_{n\to\infty} L_n = \infty$. Define $L_n \triangleq \max\{\ell \in \mathbb{N} : \lambda_\ell \geq n^{\frac{\tau-1}{2}}\}$. Then L_n is well-defined for $n \geq (\frac{1}{\lambda_1})^{\frac{2}{1-\tau}}$, since for such n we have $\lambda_1 \geq n^{\frac{\tau-1}{2}}$. Hence for $n \geq (\frac{1}{\lambda_1})^{\frac{2}{1-\tau}}$, we have

$$\frac{B_n}{\lambda_{L_n}} \leq \frac{n^\tau}{n^{\frac{\tau-1}{2}}} = n^{\frac{\tau+1}{2}},$$

so (b) is satisfied with $\beta = \frac{\tau+1}{2} < 1$. Since $\tau < 1$, L_n is non-decreasing in n and $n^{\frac{\tau-1}{2}}$ decreases to 0. Since each $\lambda_\ell > 0$, we therefore have $\lim_{n\to\infty} L_n = \infty$. $\square$

C.4. Proof of Thm. C.1: Debiasing via i.i.d. simplex reweighting

We will slowly build up towards proving Thm. C.1. First notice $\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}_{\mathbb{P}}^2(x,x)] < \infty$ implies $\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}_{\mathbb{P}}(x,x)] < \infty$, so Lem. B.1 holds. Fix any $L \in \mathbb{N}$ satisfying $\lambda_L > 0$. Since n is even, we can define $\mathcal{D}_0 \triangleq [n/2]$ and $\mathcal{D}_1 \triangleq [n] \setminus \mathcal{D}_0$. We will use $\mathcal{S}_{\mathcal{D}_0}$ and $\mathcal{S}_{\mathcal{D}_1}$ to denote the subsets of $\mathcal{S}_n$ with indices in $\mathcal{D}_0$ and $\mathcal{D}_1$ respectively. Let $(\phi_\ell)_{\ell=1}^\infty \subset \mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$ be eigenfunctions corresponding to the eigenvalues $(\lambda_\ell)_{\ell=1}^\infty$ by Lem. B.1(c), so that $(\phi_\ell)_{\ell=1}^\infty$ is an orthonormal system of $\mathcal{L}^2(\mathbb{P})$.

We start with a useful lemma.

Lemma C.1 ($\mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$ consists of mean-zero functions)**.** *Let $\boldsymbol{k}_{\mathbb{P}}$ be a kernel satisfying Assum. 1. Then for any $f \in \mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$, we have $\mathbb{P}f = 0$.*

Proof. Fix $f \in \mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$. By Steinwart and Christmann (2008, Thm 4.26), f is $\mathbb{P}$ integrable. Consider the linear operator I that maps $f \mapsto \mathbb{P}f$. Since

$$|I(f)| = |\mathbb{P}f| \le \mathbb{P}|f| = \int |\langle f, \boldsymbol{k}_{\mathbb{P}}(x,\cdot)\rangle_{\boldsymbol{k}_{\mathbb{P}}}| \mathrm{d}\mathbb{P} \le \int \|f\|_{\boldsymbol{k}_{\mathbb{P}}} \sqrt{\boldsymbol{k}_{\mathbb{P}}(x,x)}\mathrm{d}\mathbb{P} = \|f\|_{\boldsymbol{k}_{\mathbb{P}}} \mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}_{\mathbb{P}}(x,x)^{\frac{1}{2}}].$$

Hence I is a continuous linear operator, so by the Riez representation theorem (Steinwart and Christmann, 2008, Thm. A.5.12), there exists $g \in \mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$ such that $I(h) = \langle h, g\rangle_{\boldsymbol{k}_{\mathbb{P}}}$ for any $h \in \mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$.

By Steinwart and Christmann (2008, Thm. 4.21), the set

$$H_{\text{pre}} \triangleq \left\{\sum_{i=1}^n \alpha_i \boldsymbol{k}_{\mathbb{P}}(\cdot, x_i) : n \in \mathbb{N}, (\alpha_i)_{i\in[n]} \subset \mathbb{R}, (x_i)_{i\in[n]} \subset \mathbb{R}^d\right\}$$

is dense in $\mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$. Note that H_{pre} consists of mean zero functions under $\mathbb{P}$ by linearity. So there exists f_n converging to f in $\mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$ where each f_n has $\mathbb{P}f_n = I(f_n) = \langle f_n, g\rangle_{\boldsymbol{k}_{\mathbb{P}}} = 0$. Since

$$\lim_{n\to\infty} |\langle f, g\rangle_{\boldsymbol{k}_{\mathbb{P}}} - \langle f_n, g\rangle_{\boldsymbol{k}_{\mathbb{P}}}| = \lim_{n\to\infty} |\langle f - f_n, g\rangle_{\boldsymbol{k}_{\mathbb{P}}}| \le \lim_{n\to\infty} \|f - f_n\|_{\boldsymbol{k}_{\mathbb{P}}} \|g\|_{\boldsymbol{k}_{\mathbb{P}}} = 0,$$

we have $\mathbb{P}f = \langle f, g\rangle_{\boldsymbol{k}_{\mathbb{P}}} = 0$. □

In particular, the assumption $\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}_{\mathbb{P}}^2(x,x)] < \infty$ of Thm. C.1 implies $\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}_{\mathbb{P}}(x,x)^{\frac{1}{2}}] < \infty$, so Lem. C.1 holds.

Step 1. Build control variate weights

Fix any $L \ge 1$ and $h \in \mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$, and let $\hat{h}_{\mathcal{D}_0}$ denote the eigen-expansion truncated approximation of h based on $\mathcal{D}_0$,

$$\hat{h}_{\mathcal{D}_0}(x) \triangleq \sum_{\ell=1}^L \hat{\beta}_{\ell,0}\phi_\ell(x) \quad \text{for} \quad \hat{\beta}_{\ell,0} \triangleq \frac{2}{n}\sum_{i\in\mathcal{D}_0} h(x_i)\phi_\ell(x_i)\xi(x_i).$$

Then

$$\mathbb{E}[\hat{\beta}_{\ell,0}] = \mathbb{E}\left[\frac{2}{n}\sum_{i\in\mathcal{D}_0} h(x_i)\phi_\ell(x_i)\xi(x_i)\right] = \langle h, \phi_\ell\rangle_{\mathcal{L}^2(\mathbb{P})}. \tag{23}$$

Next, define the control variate

$$\hat{Z}_0[h] = \frac{2}{n}\sum_{i\in D_1}\left(\xi(x_i)(h(x_i) - \hat{h}_{\mathcal{D}_0}(x_i))\right). \tag{24}$$

which satisfies

$$\mathbb{E}[\hat{Z}_0[h]] = \mathbb{E}_{x\sim\mathbb{P}}\left[h(x) - \sum_{\ell=1}^L \mathbb{E}[\hat{\beta}_{\ell,0}]\phi_\ell(x)\right] = 0, \tag{25}$$

since functions in $\mathcal{H}_{\boldsymbol{k}_{\mathbb{P}}}$ have mean 0 with respect to $\mathbb{P}$ (Lem. C.1). Similarly, we define $\hat{Z}_1[h]$ by swapping $\mathcal{D}_0$ and $\mathcal{D}_1$. Then we form $\hat{Z}[h] \triangleq \frac{\hat{Z}_0[h]+\hat{Z}_1[h]}{2}$. We can rewrite $\hat{Z}[h]$ as a quadrature rule over $\mathcal{S}_n$ (Liu and Lee, 2017, Lemma B.6)

$$\hat{Z}[h] = \sum_{i\in[n]} w_i h(x_i), \tag{26}$$

where w_i is defined as (whose randomness depends on the randomness in $\mathcal{S}_n$)

$$w_i \triangleq \begin{cases} \frac{1}{n}\xi(x_i) - \frac{2}{n^2}\sum_{j\in\mathcal{D}_1}\xi(x_i)\xi(x_j)\langle\Phi_L(x_i), \Phi_L(x_j)\rangle, \forall i \in \mathcal{D}_0, \\ \frac{1}{n}\xi(x_i) - \frac{2}{n^2}\sum_{j\in\mathcal{D}_0}\xi(x_i)\xi(x_j)\langle\Phi_L(x_i), \Phi_L(x_j)\rangle, \forall i \in \mathcal{D}_1, \end{cases} \tag{27}$$

and $\Phi_L(x) \triangleq (\phi_1(x), \ldots, \phi_L(x))$.

Step 2. Show $\mathbb{E}[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^w, \mathbb{P})] = o(n^{-1})$

We first bound the variance of the control variate $\hat{Z}_0[h]$ for $h = \phi_{\ell'}$ for $\ell' \in \mathbb{N}$. Let us fix $\ell' \in \mathbb{N}$. From (24), we compute

$$\begin{aligned}\mathbb{E}[\hat{Z}_0[h]^2] = \tfrac{4}{n^2}\mathbb{E}\left[\left(\sum_{i\in\mathcal{D}_1}\xi(x_i)(h(x_i)-\hat{h}_{\mathcal{D}_0}(x_i))\right)^2\right] &= \tfrac{4}{n^2}\mathbb{E}\left[\sum_{i\in\mathcal{D}_1}\xi(x_i)^2(h(x_i)-\hat{h}_{\mathcal{D}_0}(x_i))^2\right]\\ &= \tfrac{2}{n}\mathbb{E}[\mathbb{E}_{x\sim\mathbb{Q}}[\xi(x)^2(h(x)-\hat{h}_{\mathcal{D}_0}(x))^2|\mathcal{S}_{\mathcal{D}_0}]]\\ &= \tfrac{2}{n}\mathbb{E}[\mathbb{E}_{x\sim\mathbb{P}}[\xi(x)(h(x)-\hat{h}_{\mathcal{D}_0}(x))^2|\mathcal{S}_{\mathcal{D}_0}]]\\ &\le \tfrac{2M}{n}\mathbb{E}[\mathbb{E}_{x\sim\mathbb{P}}[(h(x)-\hat{h}_{\mathcal{D}_0}(x))^2|\mathcal{S}_{\mathcal{D}_0}]],\end{aligned}$$

where in the second equality, the cross terms are zero due to the independence of points x_i and the equality (25). By the definition of $\hat{h}_{\mathcal{D}_0}$, we compute

$$\begin{aligned}\mathbb{E}_{x\sim\mathbb{P}}[(h(x)-\hat{h}_{\mathcal{D}_0}(x))^2|\mathcal{S}_{\mathcal{D}_0}] &= \mathbb{E}_{x\sim\mathbb{P}}\left[\left(\phi_{\ell'}(x)-\sum_{\ell\le L}\hat{\beta}_{\ell,0}\phi_\ell(x)\right)^2\middle|\mathcal{S}_{\mathcal{D}_0}\right]\\ &= \mathbb{E}_{x\sim\mathbb{P}}\left[\phi_{\ell'}^2(x)+\sum_{\ell\le L}\hat{\beta}_{\ell,0}^2\phi_\ell^2(x)-2\phi_{\ell'}(x)\sum_{\ell\le L}\hat{\beta}_{\ell,0}\phi_\ell(x)\middle|\mathcal{S}_{\mathcal{D}_0}\right]\\ &= 1+\sum_{\ell\le L}\hat{\beta}_{\ell,0}^2-2\sum_{\ell\le L}\hat{\beta}_{\ell,0}\mathbb{1}_{\ell'=\ell}\\ &= 1+\sum_{\ell\le L}\hat{\beta}_{\ell,0}^2-2\hat{\beta}_{\ell',0}\mathbb{1}_{\ell'\le L},\end{aligned}$$

where we use the fact that $(\phi_\ell)_{\ell=1}^\infty$ is an orthonormal system in $\mathcal{L}^2(\mathbb{P})$. By (23) with $h = \phi_{\ell'}$, we have $\mathbb{E}[\hat{\beta}_{\ell',0}] = 1$. On the other hand, we can bound, again using the orthonormality of $(\phi_\ell)_{\ell=1}^\infty$,

$$\mathbb{E}[\hat{\beta}_{\ell,0}^2] = \mathbb{E}\left[\left(\tfrac{2}{n}\sum_{i\in\mathcal{D}_0}\phi_\ell(x_i)\phi_{\ell'}(x_i)\xi(x_i)\right)^2\right] = \tfrac{4}{n^2}\mathbb{E}\left[\sum_{i\in\mathcal{D}_0}(\phi_\ell(x_i)\phi_{\ell'}(x_i)\xi(x_i))^2\right] \le \tfrac{2M}{n}\mathbb{E}_{x\sim\mathbb{P}}[(\phi_\ell(x)\phi_{\ell'}(x))^2].$$

Thus for all $\ell' \in \mathbb{N}$,

$$\mathbb{E}[\hat{Z}_0[\phi_{\ell'}]^2] \le \tfrac{2M}{n}\left(1+\tfrac{2M}{n}\sum_{\ell\le L}\mathbb{E}_{x\sim\mathbb{P}}[(\phi_\ell(x)\phi_{\ell'}(x))^2]-2\mathbb{1}_{\ell'\le L}\right) \le \tfrac{2M}{n}\left(\tfrac{2M}{n}\sum_{\ell\le L}\mathbb{E}_{x\sim\mathbb{P}}[(\phi_\ell(x)\phi_{\ell'}(x))^2]+\mathbb{1}_{\ell'>L}\right).$$

Since $\hat{Z}[h] = \frac{\hat{Z}_0[h]+\hat{Z}_1[h]}{2}$ and $(\frac{a+b}{2})^2 \le \frac{a^2+b^2}{2}$ for $a, b \in \mathbb{R}$, and, by symmetry, $\mathbb{E}[\hat{Z}_0[h]^2] = \mathbb{E}[\hat{Z}_1[h]^2]$, we have

$$\mathbb{E}[\hat{Z}[\phi_{\ell'}]^2] \le \tfrac{2M}{n}\left(\tfrac{2M}{n}\sum_{\ell\le L}\mathbb{E}_{x\sim\mathbb{P}}[(\phi_\ell(x)\phi_{\ell'}(x))^2]+\mathbb{1}_{\ell'>L}\right). \tag{28}$$

Now we have

$$\begin{aligned}\mathbb{E}[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^w,\mathbb{P})] = \mathbb{E}\left[\sum_{i,j\in[n]}w_iw_j\boldsymbol{k}_{\mathbb{P}}(x_i,x_j)\right] &= \mathbb{E}\left[\sum_{i,j\in[n]}w_iw_j\sum_{\ell'=1}^\infty\lambda_{\ell'}\phi_{\ell'}(x_i)\phi_{\ell'}(x_j)\right]\\ &= \mathbb{E}\left[\sum_{\ell'=1}^\infty\sum_{i,j\in[n]}w_iw_j\lambda_{\ell'}\phi_{\ell'}(x_i)\phi_{\ell'}(x_j)\right]\\ &= \mathbb{E}\left[\sum_{\ell'=1}^\infty\lambda_{\ell'}\left(\sum_{i\in[n]}w_i\phi_{\ell'}(x_i)\right)^2\right]\\ &= \sum_{\ell'=1}^\infty\lambda_{\ell'}\mathbb{E}\left[\left(\sum_{i\in[n]}w_i\phi_{\ell'}(x_i)\right)^2\right] = \sum_{\ell'=1}^\infty\lambda_{\ell'}\mathbb{E}[\hat{Z}[\phi_{\ell'}]^2],\end{aligned}$$

where the second and third equalities are due to the absolute convergence of the Mercer series (Lem. B.1(d)), the fourth equality follows from Tonelli's theorem (Steinwart and Christmann, 2008, Thm. A.3.10), and the last step is due to (26). Plugging in (28), we have

$$\mathbb{E}[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^w,\mathbb{P})] \le \tfrac{2M}{n}\left(\tfrac{2M}{n}\sum_{\ell'=1}^\infty\sum_{\ell\le L}\lambda_{\ell'}\mathbb{E}_{x\sim\mathbb{P}}[(\phi_\ell(x)\phi_{\ell'}(x))^2]+\sum_{\ell>L}\lambda_\ell\right).$$

Since the eigenvalues are nonnegative and non-increasing, we can write, by (5),

$$\boldsymbol{k}_{\mathbb{P}}^2(x,x) = \left(\sum_{\ell=1}^\infty\lambda_\ell\phi_\ell(x)^2\right)^2 \ge \sum_{\ell'=1}^\infty\sum_{\ell\le L}\lambda_{\ell'}\lambda_\ell(\phi_\ell(x)\phi_{\ell'}(x))^2 \ge \lambda_L\sum_{\ell'=1}^\infty\sum_{\ell\le L}\lambda_{\ell'}(\phi_\ell(x)\phi_{\ell'}(x))^2.$$

Thus by Tonelli's theorem (Steinwart and Christmann, 2008, Thm. A.3.10),

$$\sum_{\ell'=1}^{\infty} \sum_{\ell \le L} \lambda_{\ell'} \mathbb{E}_{x\sim\mathbb{P}}[(\phi_\ell(x)\phi_{\ell'}(x))^2] = \mathbb{E}_{x\sim\mathbb{P}}\left[\sum_{\ell'=1}^{\infty} \sum_{\ell \le L} \lambda_{\ell'} (\phi_\ell(x)\phi_{\ell'}(x))^2\right] \le \frac{\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}_{\mathbb{P}}^2(x,x)]}{\lambda_L}.$$

Finally, we have

$$\mathbb{E}[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^w, \mathbb{P})] \le \frac{2M}{n}\left(\frac{2M}{n}\frac{\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}_{\mathbb{P}}^2(x,x)]}{\lambda_L} + \sum_{\ell > L} \lambda_\ell\right). \tag{29}$$

Step 3. Meet the non-negative constraint

We now show that the weights (27) are nonnegative and sum close to one with high probability. For $i \in \mathcal{D}_0$, we have

$$w_i = \tfrac{1}{n}\xi(x_i)\,(1 - T_i) \quad \text{for} \quad T_i \triangleq \tfrac{2}{n}\sum_{j\in\mathcal{D}_1} \xi(x_j)\langle \Phi_L(x_i), \Phi_L(x_j)\rangle.$$

Our first goal is to derive an upper bound for $\Pr(\min_{i\in\mathcal{D}_0} w_i < 0)$. Define the event

$$E_n \triangleq \{\|\boldsymbol{k}_{\mathbb{P}}\|_n \le B_n\}, \tag{30}$$

so $\Pr(E_n^c) < \delta_n$ by the assumption on $\|\boldsymbol{k}_{\mathbb{P}}\|_n$. Then

$$\Pr\left(\min_{i\in[n]} w_i < 0, E_n\right) = \Pr\left(\max_{i\in[n]} T_i > 1, E_n\right) \le n\Pr(T_1 \mathbb{1}_{E_n} > 1), \tag{31}$$

where we applied the union bound and used the fact that $T_i \mathbb{1}_{E_n}$ has the same law for different i. To further bound $\Pr(T_1 \mathbb{1}_{E_n} > 1)$, we will use the following lemma.

Lemma C.2 (Truncated Bernstein inequality)**.** *Let $X_1, \dots, X_n$ be i.i.d. random variables with $\mathbb{E}[X_1] = 0$ and $\mathbb{E}[X_1^2] < \infty$. For any $B > 0$, $t > 0$,*

$$\Pr\left(\frac{1}{n}\sum_{i\in[n]} X_i \mathbb{1}_{X_i \le B} > t\right) \le \exp\left(\frac{-nt^2}{2(\mathbb{E}[X_1^2] + \frac{Bt}{3})}\right).$$

Proof of Lem. C.2. Fix any $B > 0$ and $t > 0$ and define, for each $i \in [n]$, $Y_i \triangleq X_i \mathbb{1}_{X_i \le B}$. Then $Y_i \le B$,

$$\mathbb{E}[Y_i] = \mathbb{E}[X_i \mathbb{1}_{X_i \le B}] \le \mathbb{E}[X_i \mathbb{1}_{X_i \le B}] + \mathbb{E}[X_i \mathbb{1}_{X_i > B}] = \mathbb{E}[X_i] = 0, \quad \text{and}$$
$$\mathbb{E}[Y_i^2] = \mathbb{E}[X_i^2 \mathbb{1}_{X_i \le B}] \le \mathbb{E}[X_i^2] = \mathbb{E}[X_1^2].$$

Now we can invoke the non-positivity of $\mathbb{E}[Y_i]$, the one-sided Bernstein inequality (Wainwright, 2019, Prop. 2.14), and the relation $\mathbb{E}[Y_i^2] \le \mathbb{E}[X_1^2]$ to conclude that

$$\Pr\left(\frac{1}{n}\sum_{i\in[n]} Y_i > t\right) \le \Pr\left(\frac{1}{n}\sum_{i\in[n]} (Y_i - \mathbb{E}[Y_i]) > t\right) \le \exp\left(\frac{-nt^2}{2(\frac{1}{n}\sum_{i\in[n]}\mathbb{E}[Y_i^2] + \frac{Bt}{3})}\right) \le \exp\left(\frac{-nt^2}{2(\mathbb{E}[X_1^2] + \frac{Bt}{3})}\right).$$

□

For $j \in \mathcal{D}_1$, define $X_j \triangleq \xi(x_j)\langle \Phi_L(x_1), \Phi_L(x_j)\rangle$ and note that

$$\begin{aligned}
\mathbb{E}[X_j|x_1] &= \mathbb{E}_{x\sim\mathbb{Q}}[\xi(x)\langle \Phi_L(x_1), \Phi_L(x)\rangle|x_1] = \mathbb{E}_{x\sim\mathbb{P}}[\langle \Phi_L(x_1), \Phi_L(x)\rangle|x_1] = 0 \\
\mathbb{E}[X_j^2|x_1] &= \mathbb{E}[\xi(x_j)^2\langle \Phi_L(x_1), \Phi_L(x_j)\rangle^2|x_1] \le M\mathbb{E}_{x\sim\mathbb{P}}[\langle \Phi_L(x_1), \Phi_L(x)\rangle^2|x_1] \\
&= M\mathbb{E}_{x\sim\mathbb{P}}\left[\sum_{\ell,\ell'\le L} \phi_\ell(x_1)\phi_{\ell'}(x_1)\phi_\ell(x)\phi_{\ell'}(x)\Big| x_1\right] \\
&= M\sum_{\ell,\ell'\le L} \phi_\ell(x_1)\phi_{\ell'}(x_1)\mathbb{E}_{x\sim\mathbb{P}}\left[\phi_\ell(x)\phi_{\ell'}(x)\right] \\
&= M\,\|\Phi_L(x_1)\|_2^2.
\end{aligned}$$

Since $\lambda_1 \ge \lambda_2 \ge \cdots \ge 0$, for any $x \in \mathbb{R}^d$, we can bound $\|\Phi_L(x)\|_2^2$ via

$$\|\Phi_L(x)\|_2^2 = \sum_{\ell\le L} \phi_\ell(x)^2 \le \frac{\sum_{\ell\le L}\lambda_\ell\phi_\ell(x)^2}{\lambda_L} \le \frac{\sum_{\ell=1}^{\infty}\lambda_\ell\phi_\ell(x)^2}{\lambda_L} = \frac{\boldsymbol{k}_{\mathbb{P}}(x,x)}{\lambda_L}, \tag{32}$$

where we applied Lem. B.1(d) for the last equality. Thus

$$|X_j| \le M \left\|\Phi_L(x_1)\right\|_2 \left\|\Phi_L(x_j)\right\|_2 \le M \left\|\Phi_L(x_1)\right\|_2 \sqrt{\tfrac{\boldsymbol{k}_{\mathbb{P}}(x_j,x_j)}{\lambda_L}},$$

so if we let $B \triangleq \sqrt{\frac{B_n}{\lambda_L}} M \left\|\Phi_L(x_1)\right\|_2$, then

$$E_n = \left\{\sup_{i\in[n]} \boldsymbol{k}_{\mathbb{P}}(x_i,x_i) \le B_n\right\} \subset \bigcap_{j\in\mathcal{D}_1}\{|X_j| \le B\}.$$

Since $T_1 = \frac{2}{n}\sum_{j\in\mathcal{D}_1} X_j$, we have inclusions of events

$$\{T_1\mathbb{1}_{E_n} > 1\} = \{T_1 > 1\} \cap E_n \subset \left\{\tfrac{2}{n}\sum_{j\in\mathcal{D}_1} X_j \mathbb{1}_{X_j\le B} > 1\right\}.$$

Thus Lem. C.2 with $t = 1$ and conditioned on x_1 implies

$$\begin{aligned}\Pr\left(T_1\mathbb{1}_{E_n} > 1 | x_1\right) &\le \Pr\left(\tfrac{2}{n}\sum_{j\in\mathcal{D}_1} X_j\mathbb{1}_{X_j\le B} > 1 \Big| x_1\right) \\ &\le \exp\left(\frac{-n}{4(M\|\Phi_L(x_1)\|_2^2 + \sqrt{\frac{B_n}{\lambda_L}}M\|\Phi_L(x_1)\|_2/3)}\right).\end{aligned}$$

On event $\{\boldsymbol{k}_{\mathbb{P}}(x_1,x_1) \le B_n\}$, by (32), we have

$$\left\|\Phi_L(x_1)\right\|_2 \le \sqrt{\tfrac{B_n}{\lambda_L}}.$$

Hence

$$\Pr\left(T_1\mathbb{1}_{E_n} > 1|x_1\right)\mathbb{1}_{\boldsymbol{k}_{\mathbb{P}}(x_1,x_1)\le B_n} \le \exp\left(\frac{-n}{\frac{16}{3}M\frac{B_n}{\lambda_L}}\right).$$

On the other hand, $\{\boldsymbol{k}_{\mathbb{P}}(x_1,x_1) > B_n\} \notin E_n$, so

$$\Pr\left(T_1\mathbb{1}_{E_n} > 1|x_1\right)\mathbb{1}_{\boldsymbol{k}_{\mathbb{P}}(x_1,x_1)>B_n} = 0$$

Thus

$$\Pr(T_1\mathbb{1}_{E_n} > 1) = \mathbb{E}[\Pr(T_1\mathbb{1}_{E_n} > 1|x_1)] \le \exp\left(\frac{-n}{\frac{16}{3}M\frac{B_n}{\lambda_L}}\right).$$

Combining the last inequality with (31), we have:

$$\Pr\left(\min_{i\in[n]} w_i < 0, E_n\right) \le n\exp\left(\frac{-n}{\frac{16}{3}M\frac{B_n}{\lambda_L}}\right). \tag{33}$$

Step 4. Meet the sum-to-one constraint

Let

$$S \triangleq \sum_{i\in\mathcal{D}_0} w_i = \sum_{i\in\mathcal{D}_0} \tfrac{1}{n}\xi(x_i)\left(1 - \tfrac{2}{n}\sum_{j\in\mathcal{D}_1}\xi(x_j)\langle\Phi_L(x_i),\Phi_L(x_j)\rangle\right).$$

We now derive a bound for $\Pr(S < 1/2 - t/2)$ for $t \in (0,1)$. Let

$$S_1 \triangleq \tfrac{1}{n}\sum_{i\in\mathcal{D}_0}\xi(x_i), \quad S_2 \triangleq -\tfrac{2}{n^2}\sum_{i\in\mathcal{D}_0}\sum_{j\in\mathcal{D}_1}\xi(x_i)\xi(x_j)\langle\Phi_L(x_i),\Phi_L(x_j)\rangle,$$

so $S = S_1 + S_2$. Note that $\mathbb{E}[S_1] = 1/2$ and $\mathbb{E}[S_2] = 0$ since $\mathcal{D}_0$ and $\mathcal{D}_1$ are disjoint. Let E_n be the same event defined as in (30). For $t_1 \in (0, t/2)$ to be determined later and $t_2 \triangleq t/2 - t_1$, we have, by the union bound

$$\Pr(S < 1/2 - t/2, E_n) \le \Pr(S_1 < 1/2 - t_1, E_n) + \Pr(S_2 < -t_2, E_n).$$

By Hoeffding's inequality and the assumption $\xi(x) \le M$, we have

$$\Pr(S_1 < 1/2 - t_1, E_n) \le \Pr\left(\tfrac{2}{n}\sum_{i\in\mathcal{D}_0}\tfrac{\xi(x_i)}{2} - 1/2 < -t_1\right) \le \exp\left(\tfrac{-2(n/2)t_1^2}{(M/2)^2}\right) = \exp\left(\tfrac{-4nt_1^2}{M^2}\right). \tag{34}$$

To give a concentration bound for $\Pr(S_2 < -t_2, E_n)$, we will use the following lemma.

Lemma C.3 (U-statistic Bernstein's inequality)**.** *Let $h : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a function bounded above by $b > 0$. Assume $n \in 2\mathbb{N}$ and let $x_1, \dots, x_n$ be i.i.d. random variables taking values in $\mathcal{X}$. Denote $m_h \triangleq \mathbb{E}[h(x_1, x_2)]$ and $\sigma_h^2 \triangleq \mathrm{Var}[h(x_1, x_2)]$. Let $\mathcal{D}_0 = [n/2]$ and $\mathcal{D}_1 = [n] \setminus [n/2]$. Define*

$$U \triangleq \frac{1}{(n/2)^2} \sum_{i\in\mathcal{D}_0} \sum_{j\in\mathcal{D}_1} h(x_i, x_j).$$

Then

$$\Pr(U - m_h > t) \le \exp\left(\frac{-nt^2}{4(\sigma_h^2 + \frac{bt}{3})}\right).$$

Proof of Lem. C.3. We adapt the proof from Pitcan (2017, Section 3) as follows. Let $k \triangleq n/2$. Define $V : \mathcal{X}^n \to \mathbb{R}$ as

$$V(x_1, \dots, x_n) \triangleq \frac{1}{k} \sum_{i\in[k]} h(x_i, x_{i+k}).$$

Then note that

$$U = \frac{1}{k!} \sum_{\sigma\in\mathrm{perm}(k)} V_\sigma,$$
$$V_\sigma \triangleq V(x_{\sigma_1}, \dots, x_{\sigma_k}),$$

where $\mathrm{perm}(k)$ is the set of all permutations of $[k]$; this is because every $h(x_i, x_j)$ term for $i \in \mathcal{D}_0, j \in \mathcal{D}_1$ will appear in the summation an equal number of times. For a fixed $\sigma \in \mathrm{perm}(k)$, the random variable $V(x_{\sigma_1}, \dots, x_{\sigma_k}, x_{k+1}, \dots, x_n)$ is a sum of k i.i.d. terms $h(x_{\sigma_i}, x_{i+k})$. Denote $V = V(x_1, \dots, x_n)$. For any $s > 0$, we have, by independence,

$$\begin{aligned}\mathbb{E}[e^{s(V-m_h)}] &= \mathbb{E}\left[\exp\left(\frac{s}{k} \sum_{i\in[k]} (h(x_i, x_{i+[k]}) - m_h)\right)\right] \\ &= \left(\mathbb{E}\left[\exp\left(\frac{s}{k}(h(x_1, x_2) - m_h)\right)\right]\right)^k\end{aligned}$$

By the one-sided Bernstein's lemma Wainwright (2019, Prop. 2.14) applied to $\frac{h(x_1,x_2)}{k}$ which is upper bounded by $\frac{b}{k}$ with variance $\frac{\sigma_h^2}{k^2}$, we have

$$\mathbb{E}\left[\exp\left(s\frac{h(x_1,x_2)-m_h}{k}\right)\right] \le \exp\left(\frac{s^2\sigma_h^2/2}{k(k-\frac{bs}{3})}\right),$$

for $s \in [0, 3k/b)$. Next, by Markov's inequality and Jensen's inequality,

$$\begin{aligned}\Pr(U - m_h > t) &= \Pr\left(e^{s(U-m_h)} > e^{st}\right) \le \mathbb{E}[e^{s(U-m_h)}]e^{-st} \\ &= \mathbb{E}\left[\exp\left(\frac{1}{(n/2)!} \sum_{\sigma\in\mathrm{perm}(n/2)} s(V_\sigma - m_h)\right)\right] e^{-st} \\ &\le \mathbb{E}\left[\frac{1}{(n/2)!} \sum_{\sigma\in\mathrm{perm}(n/2)} \exp(s(V_\sigma - m_h))\right] e^{-st} \\ &= \mathbb{E}[e^{s(V-m_h)}]e^{-st}.\end{aligned}$$

Therefore,

$$\Pr(U - m_h > t) \le \exp\left(\frac{s^2\sigma_h^2}{2(k-\frac{bs}{3})} - st\right).$$

Now, we get the desired bound if we pick $s = \frac{k^2 t}{k\sigma_h^2 + \frac{ktb}{3}} \in [0, 3k/b)$ and simplify. □

Let

$$h(x, x') \triangleq \xi(x)\xi(x')\langle \Phi_L(x), \Phi_L(x')\rangle$$
$$\bar{h}(x, x') \triangleq h(x, x')\mathbb{1}_{h(x,x')\le M^2\frac{B_n}{\lambda_L}}.$$

Then

$$\Pr(S_2 < -t_2, E_n) = \Pr\left(\tfrac{1}{(n/2)^2}\sum_{i\in\mathcal{D}_0}\sum_{j\in\mathcal{D}_1} h(x_i,x_j) > 2t_2, E_n\right)$$
$$\le \Pr\left(\tfrac{1}{(n/2)^2}\sum_{i\in\mathcal{D}_0}\sum_{j\in\mathcal{D}_1} \bar{h}(x_i,x_j) > 2t_2\right), \tag{35}$$

where the last inequality used the fact that, for $i \in \mathcal{D}_0, j \in \mathcal{D}_1$,

$$E_n \subset \{\max(\boldsymbol{k}_{\mathbb{P}}(x_i,x_i), \boldsymbol{k}_{\mathbb{P}}(x_j,x_j)) \le B_n\} \subset \left\{h(x_i,x_j) \le M^2\tfrac{B_n}{\lambda_L}\right\},$$

using (32). We further compute

$$\begin{aligned}
m_{\bar{h}} &= \mathbb{E}[\bar{h}(x_1,x_2)] \le \mathbb{E}[h(x_1,x_2)] = \mathbb{E}[\xi(x_1)\xi(x_2)\langle\Phi_L(x_1),\Phi_L(x_2)\rangle] \\
&= \textstyle\sum_{\ell\le L}\mathbb{E}[\xi(x_1)\xi(x_2)\phi_\ell(x_1)\phi_\ell(x_2)] \\
&= \textstyle\sum_{\ell\le L}(\mathbb{E}_{x\sim\mathbb{P}}[\phi_\ell(x)])^2 = 0,
\end{aligned}$$

and

$$\begin{aligned}
\sigma_{\bar{h}}^2 &= \mathrm{Var}[\bar{h}(x_1,x_2)] \le \mathbb{E}[\bar{h}(x_1,x_2)^2] \le \mathbb{E}[h(x_1,x_2)^2] \\
&= \mathbb{E}\left[(\xi(x_1)\xi(x_2)\langle\Phi_L(x_1),\Phi_L(x_2)\rangle)^2\right] \\
&\le M^2\mathbb{E}_{(x,x')\sim\mathbb{P}\times\mathbb{P}}[\langle\Phi_L(x),\Phi_L(x')\rangle^2] \\
&= M^2\mathbb{E}_{(x,x')\sim\mathbb{P}\times\mathbb{P}}\left[\textstyle\sum_{\ell,\ell'\le L}\phi_\ell(x)\phi_{\ell'}(x)\phi_\ell(x')\phi_{\ell'}(x')\right] \\
&= M^2\textstyle\sum_{\ell,\ell'\le L}(\mathbb{E}[\phi_\ell(x)\phi_{\ell'}(x)])^2 = LM^2.
\end{aligned}$$

Since $\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}_{\mathbb{P}}(x,x)] = \sum_\ell \lambda_\ell \ge L\lambda_L$, we have $L \le \frac{\|\boldsymbol{k}_{\mathbb{P}}\|^2_{\mathcal{L}^2(\mathbb{P})}}{\lambda_L}$, so that $\sigma^2_{\bar{h}} \le \frac{M^2\|\boldsymbol{k}_{\mathbb{P}}\|^2_{\mathcal{L}^2(\mathbb{P})}}{\lambda_L}$. Applying Lem. C.3 to $\bar{h}$, which is bounded by $M^2\frac{B_n}{\lambda_L}$ and using the fact that $m_{\bar{h}} \le 0$, we have

$$\Pr\left(\tfrac{1}{(n/2)^2}\sum_{i\in\mathcal{D}_0}\sum_{j\in\mathcal{D}_1}\bar{h}(x_i,x_j) > 2t_2\right) \le \Pr\left(\tfrac{1}{(n/2)^2}\sum_{i\in\mathcal{D}_0}\sum_{j\in\mathcal{D}_1}\bar{h}(x_i,x_j) - m_{\bar{h}} > 2t_2\right)$$
$$\le \exp\left(\frac{-n(2t_2)^2}{4\left(\frac{M^2\|\boldsymbol{k}_{\mathbb{P}}\|^2_{\mathcal{L}^2(\mathbb{P})}}{\lambda_L} + 2M^2\frac{B_n}{\lambda_L}t_2/3\right)}\right). \tag{36}$$

Thus combining (34), (35), (36), we get

$$\Pr(S < 1/2 - t/2, E_n) \le \exp\left(\frac{-4nt_1^2}{M^2}\right) + \exp\left(\frac{-nt_2^2}{\left(\frac{M^2\|\boldsymbol{k}_{\mathbb{P}}\|^2_{\mathcal{L}^2(\mathbb{P})}}{\lambda_L} + 2M^2\frac{B_n}{\lambda_L}t_2/3\right)}\right).$$

Finally, by symmetry and the union bound, for $t \in (0,1)$, $t \in (0, t/2)$ and $t_2 = t/2 - t_1$, we have

$$\begin{aligned}
\Pr\left(\textstyle\sum_{i\in[n]} w_i < 1 - t, E_n\right) &\le \Pr\left(\textstyle\sum_{i\in\mathcal{D}_0} w_i < 1/2 - t/2, E_n\right) + \Pr\left(\textstyle\sum_{i\in\mathcal{D}_1} w_i < 1/2 - t/2, E_n\right) \\
&= 2\Pr(S < 1/2 - t/2, E_n) \\
&\le 2\left(\exp\left(\frac{-4nt_1^2}{M^2}\right) + \exp\left(\frac{-nt_2^2}{\left(\frac{M^2\|\boldsymbol{k}_{\mathbb{P}}\|^2_{\mathcal{L}^2(\mathbb{P})}}{\lambda_L} + 2M^2\frac{B_n}{\lambda_L}t_2/3\right)}\right)\right).
\end{aligned} \tag{37}$$

Step 5. Putting it all together

Define the event

$$F_n = \left\{ \min_{i\in[n]} w_i \geq 0, \sum_{i\in[n]} w_i \geq \tfrac{1}{2} \right\}.$$

Then, by the union bound,

$$\Pr(F_n^c) \leq \Pr\big(\min_{i\in[n]} w_i < 0, E_n\big) + \Pr\Big(\sum_{i\in[n]} w_i < \tfrac{1}{2}, E_n\Big) + \Pr(E_n^c).$$

Applying (33) and (37) to bound the last expression with $t = 1/2$, $t_1 = t_2 = 1/8$, we have $\Pr(F_n^c) \leq \epsilon_n^2$ for ϵ_n defined in (22). On the event F_n, if we define $w^+ \in \Delta_{n-1}$ via

$$w_i^+ \triangleq \tfrac{w_i}{\sum_{i\in[n]} w_i},$$

then $w_i^+ = \alpha w_i$ for $i \in [n]$ and $\alpha \triangleq \frac{1}{\sum_{i\in[n]} w_i} \leq 2$. Let $\tilde{w} \in \Delta_{n-1}$ be the weight defined by $\tilde{w}_1 = 1$ and $\tilde{w}_i = 0$ for $i > 1$. Since w_{OPT} is the best simplex weight, we have $\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w_{\mathrm{OPT}}}, \mathbb{P}) \leq \min(\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^+}, \mathbb{P}), \mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{\tilde{w}}, \mathbb{P}))$. Hence

$$\begin{aligned}\mathbb{E}\left[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w_{\mathrm{OPT}}}, \mathbb{P})\right] &= \mathbb{E}\left[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w_{\mathrm{OPT}}}, \mathbb{P})\mathbb{1}_{F_n}\right] + \mathbb{E}\left[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w_{\mathrm{OPT}}}, \mathbb{P})\mathbb{1}_{F_n^c}\right] \\ &\leq \mathbb{E}\left[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^+}, \mathbb{P})\mathbb{1}_{F_n}\right] + \mathbb{E}\left[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{\tilde{w}}, \mathbb{P})\mathbb{1}_{F_n^c}\right].\end{aligned}$$

For the first term, we have the bound

$$\begin{aligned}\mathbb{E}\left[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^+}, \mathbb{P})\mathbb{1}_{F_n}\right] &= \mathbb{E}\left[\sum_{i,j\in[n]} w_i^+ w_j^+ \boldsymbol{k}_{\mathbb{P}}(x_i, x_j)\mathbb{1}_{F_n}\right] = \mathbb{E}\left[\alpha^2 \sum_{i,j\in[n]} w_i w_j \boldsymbol{k}_{\mathbb{P}}(x_i, x_j)\mathbb{1}_{F_n}\right] \\ &\leq 4\mathbb{E}\left[\sum_{i,j\in[n]} w_i w_j \boldsymbol{k}_{\mathbb{P}}(x_i, x_j)\right] \leq \tfrac{8M}{n}\left(\tfrac{2M}{n}\tfrac{\mathbb{E}_{x\sim\mathbb{P}}[\boldsymbol{k}_{\mathbb{P}}^2(x,x)]}{\lambda_L} + \sum_{\ell>L} \lambda_\ell\right),\end{aligned}$$

where we applied (29) for the last inequality. For the second term, by the Cauchy-Schwartz inequality,

$$\begin{aligned}\mathbb{E}\left[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{\tilde{w}}, \mathbb{P})\mathbb{1}_{F_n^c}\right] &\leq \sqrt{\Pr(F_n^c)}\sqrt{\mathbb{E}\left[\left(\sum_{i,j\in[n]} \boldsymbol{k}_{\mathbb{P}}(x_i, x_j)\tilde{w}_i\tilde{w}_j\right)^2\right]} \\ &\leq \sqrt{\Pr(F_n^c)}\sqrt{\mathbb{E}[\boldsymbol{k}_{\mathbb{P}}(x_1, x_1)^2]}.\end{aligned}$$

Putting everything together we obtain (21). □

D. Stein Kernel Thinning

In this section, we detail our Stein thinning implementation in App. D.1, our kernel thinning implementation and analysis in App. D.2, and our proof of Thm. 3 in App. D.3.

D.1. Stein Thinning with sufficient statistics

For an input point set of size n, the original implementation of Stein Thinning of Riabiz et al. (2022) takes $O(nm^2)$ time to output a coreset of size m. In Alg. D.1, we show that this runtime can be improved to $O(nm)$ using sufficient statistics. The idea is to maintain a vector $g \in \mathbb{R}^n$ such that $g = 2\boldsymbol{k}_{\mathbb{P}}(\mathbb{S}_n, \mathbb{S}_n)w$ where w is the weight representing the current coreset.

D.2. Kernel Thinning targeting $\mathbb{P}$

Our KernelThinning implementation is detailed in Alg. D.2. Since we are able to directly compute $\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^w, \mathbb{P})$, we use KT-Swap (Alg. D.3) in place of the standard KT-SWAP subroutine (Dwivedi and Mackey, 2022, Algorithm 1b) to choose candidate points to swap in so as to greedily minimize $\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^w, \mathbb{P})$. To facilitate our subsequent SKT analysis, we restate the guarantees of KT-SPLIT (Dwivedi and Mackey, 2022, Theorem 2) in the sub-Gaussian format of (Shetty et al., 2022, Definition 3).

Algorithm D.1 SteinThinning (ST) with sufficient statistics

Input: kernel $\boldsymbol{k}_{\mathbb{P}}$ with zero-mean under $\mathbb{P}$, input points $\mathcal{S}_n = (x_i)_{i\in[n]}$, output size m
$w \leftarrow \mathbf{0} \in \mathbb{R}^n$
$j \leftarrow \operatorname{argmin}_{i\in[n]} \boldsymbol{k}_{\mathbb{P}}(x_i, x_i)$
$w_j \leftarrow 1$
$g \leftarrow 2\boldsymbol{k}_{\mathbb{P}}(\mathcal{S}_n, x_j)$ ▷ *maintain sufficient statistics* $g = 2\boldsymbol{k}_{\mathbb{P}}(\mathcal{S}_n, \mathcal{S}_n)w$
for $t = 1$ **to** $m-1$ **do**
  $j \leftarrow \operatorname{argmin}_{i\in[n]}\{tg_i + \boldsymbol{k}_{\mathbb{P}}(x_i, x_i)\}$
  $w \leftarrow \frac{t}{t+1}w + \frac{1}{t+1}e_j$
  $g \leftarrow \frac{t}{t+1}g + \frac{2}{t+1}\boldsymbol{k}_{\mathbb{P}}(\mathcal{S}_n, x_j)$
end for
Return: w

Algorithm D.2 KernelThinning (KT) (adapted from Dwivedi and Mackey (2022, Alg. 1))

Input: kernel $\boldsymbol{k}_{\mathbb{P}}$ with zero-mean under $\mathbb{P}$, input points $\mathcal{S}_n = (x_i)_{i\in[n]}$, multiplicity n' with $\log_2 \frac{n'}{m} \in \mathbb{N}$, weight $w \in \Delta_{n-1} \cap (\frac{\mathbb{N}_0}{n'})^n$, output size m with $\frac{n'}{m} \in 2^{\mathbb{N}}$, failure probability δ
$\mathtt{S} \leftarrow$ index sequence where $k \in [n]$ appears $n'w_k$ times
$\mathfrak{t} \leftarrow \log_2 \frac{n'}{m} \in \mathbb{N}$
$(\mathtt{I}^{(\ell)})_{\ell\in[2^{\mathfrak{t}}]} \leftarrow$ KT-SPLIT$(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n[\mathtt{S}], \mathfrak{t}, \delta/n')$ ▷ KT-SPLIT *is from Dwivedi and Mackey (2022, Algorithm 1a) and we set* $\delta_i = \delta$ *for all* i
$\mathtt{I}^{(\ell)} \leftarrow \mathtt{S}[\mathtt{I}^{(\ell)}]$ for each $\ell \in [2^{\mathfrak{t}}]$
$\mathtt{I} \leftarrow$ KT-Swap$(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, (\mathtt{I}^{(\ell)})_{\ell\in[2^{\mathfrak{t}}]})$
$w_{\mathrm{KT}} \leftarrow$ simplex weight corresponding to $\mathtt{I}$ ▷ $w_i = \frac{\text{number of occurrences of } i \text{ in } \mathtt{I}}{|\mathtt{I}|}$
Return: $w_{\mathrm{KT}} \in \Delta_{n-1} \cap (\frac{\mathbb{N}_0}{m})^n$ ▷ *Hence* $\|w_{\mathrm{KT}}\|_0 \le m$

Lemma D.1 (Sub-Gaussian guarantee for KT-SPLIT)**.** *Let $\mathcal{S}_n$ be a sequence of n points and $\boldsymbol{k}$ a kernel. For any $\delta \in (0,1)$ and $m \in \mathbb{N}$ such that $\log_2 \frac{n}{m} \in \mathbb{N}$, consider the* KT-SPLIT *algorithm (Dwivedi and Mackey, 2022, Algorithm 1a) with $\boldsymbol{k}_{\mathrm{split}} = \boldsymbol{k}$, thinning parameter $\mathfrak{t} = \log_2 \frac{n}{m}$, and $\delta_i = \frac{\delta}{n}$ to compress $\mathcal{S}_n$ to $2^{\mathfrak{t}}$ coresets $\{\mathcal{S}_{\mathrm{out}}^{(i)}\}_{i\in[2^{\mathfrak{t}}]}$ where each $\mathcal{S}_{\mathrm{out}}^{(i)}$ has m points. Denote the signed measure $\phi^{(i)} \triangleq \frac{1}{n}\sum_{x\in\mathcal{S}_n}\delta_x - \frac{2^{\mathfrak{t}}}{n}\sum_{x\in\mathcal{S}_{\mathrm{out}}^{(i)}}\delta_x$. Then for each $i \in [2^{\mathfrak{t}}]$, on an event $\mathcal{E}_{\mathrm{equi}}^{(i)}$ with $\mathbb{P}(\mathcal{E}_{\mathrm{equi}}^{(i)}) \ge 1 - \frac{\delta}{2}$, $\phi^{(i)} = \tilde{\phi}^{(i)}$ for a random signed measure $\tilde{\phi}^{(i)}$[5] such that, for any $\delta' \in (0,1)$,*

$$\Pr\left(\left\|\tilde{\phi}^{(i)}\boldsymbol{k}\right\|_{\mathcal{H}_{\boldsymbol{k}}} \ge a_{n,m} + v_{n,m}\sqrt{\log\left(\tfrac{1}{\delta'}\right)}\right) \le \delta',$$

where

$$a_{n,m} \triangleq \frac{1}{m}\left(2 + \sqrt{\frac{8}{3}\|\boldsymbol{k}\|_n \log\left(\frac{6(\log_2\frac{n}{m})m}{\delta}\right)\log(4\mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(R_n), m^{-1}))}\right),$$
$$v_{n,m} \triangleq \frac{1}{m}\sqrt{\frac{8}{3}\|\boldsymbol{k}\|_n \log\left(\frac{6(\log_2\frac{n}{m})m}{\delta}\right)}.$$

Proof of Lem. D.1. Fix $i \in [2^{\mathfrak{t}}]$, $\delta \in (0,1)$ and $n, m \in \mathbb{N}$ such that $\mathfrak{t} = \log_2 \frac{n}{m} \in \mathbb{N}$. Define $\phi \triangleq \phi^{(i)}$. By the proof of Dwivedi and Mackey (2022, Thms. 1 and 2), there exists an event $\mathcal{E}_{\mathrm{equi}}$ with $\Pr\left(\mathcal{E}_{\mathrm{equi}}^c\right) \le \frac{\delta}{2}$ such that, on this event, $\phi = \tilde{\phi}$ where $\tilde{\phi}$ is a signed measure such that, for any $\delta' \in (0,1)$, with probability at least $1-\delta'$,

$$\left\|\tilde{\phi}\boldsymbol{k}\right\|_{\mathcal{H}_{\boldsymbol{k}}} \le \inf_{\epsilon\in(0,1), A:\mathcal{S}_n\subset A} 2\epsilon + \frac{2^{\mathfrak{t}}}{n}\sqrt{\frac{8}{3}\|\boldsymbol{k}\|_n \log\left(\frac{6tn}{2^{\mathfrak{t}}\delta}\right)\left[\log\frac{4}{\delta'} + \log\mathcal{N}_{\boldsymbol{k}}(A,\epsilon)\right]}.$$

[5]This is the signed measure returned by repeated applications of self-balancing Hilbert walk (SBHW) (Dwivedi and Mackey, 2021, Algorithm 3). Although SBHW returns an element of $\mathcal{H}_{\boldsymbol{k}}$, by tracing the algorithm, the returned output is equivalent to a signed measure via the correspondence $\sum_{i\in[n]} c_i\boldsymbol{k}(x_i,\cdot) \Leftrightarrow \sum_{i\in[n]} c_i\delta_{x_i}$. The usage of signed measures is consistent with Shetty et al. (2022).

Algorithm D.3 KT-Swap (modified Dwivedi and Mackey (2022, Alg. 1b) to minimize MMD to $\mathbb{P}$)

Input: kernel $\boldsymbol{k}_{\mathbb{P}}$ with zero-mean under $\mathbb{P}$, input points $\mathcal{S}_n = (x_i)_{i\in[n]}$, candidate coreset indices $(\mathtt{I}^{(\ell)})_{\ell\in[L]}$
$m \leftarrow |\mathtt{I}^{(0)}|$ ▷ *all candidate coresets are of the same size*
$\mathtt{I} \leftarrow \mathtt{I}^{(\ell^*)}$ for $\ell^* \in \operatorname{argmin}_{\ell\in[L]} \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathcal{S}_n[\mathtt{I}^{(\ell)}], \mathbb{P})$ ▷ *select the best* KT-SPLIT *coreset*
$\mathtt{I}_{\mathrm{ST}} \leftarrow$ index sequence of SteinThinning$(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, m)$▷ *add Stein thinning baseline*
$\mathtt{C} = \{\mathtt{I}, \mathtt{I}_{\mathrm{ST}}\}$ ▷ *shortlisted candidates*
for $\mathtt{I} \in \mathtt{C}$ **do**
  $g \leftarrow \mathbf{0} \in \mathbb{R}^n$ ▷ *maintain sufficient statistics* $g = \sum_{j\in[m]} \boldsymbol{k}_{\mathbb{P}}(x_{\mathtt{I}_j}, \mathcal{S}_n)$
  $\mathtt{Kdiag} \leftarrow (\boldsymbol{k}_{\mathbb{P}}(x_i, x_i))_{i\in[n]}$
  for $j = 1$ **to** m **do**
    $g \leftarrow g + \boldsymbol{k}_{\mathbb{P}}(x_{\mathtt{I}_j}, \mathcal{S}_n)$
  end for
  for $j = 1$ **to** m **do**
    $\Delta = 2(g - \boldsymbol{k}_{\mathbb{P}}(x_{\mathtt{I}_j}, \mathcal{S}_n)) + \mathtt{Kdiag}$ ▷ *this is the change in* $\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathcal{S}_n[\mathtt{I}], \mathbb{P})$ *if we were to replace* $\mathtt{I}_j$
    $k \leftarrow \operatorname{argmin}_{i\in[n]} \Delta_i$
    $g = g - \boldsymbol{k}_{\mathbb{P}}(x_{\mathtt{I}_j}, \mathcal{S}_n) + \boldsymbol{k}_{\mathbb{P}}(x_k, \mathcal{S}_n)$
    $\mathtt{I}_j \leftarrow k$
  end for
end for
Return: $\mathtt{I} = \operatorname{argmin}_{\mathtt{I}\in\mathtt{C}} \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathcal{S}_n[\mathtt{I}], \mathbb{P})$

Note that on $\mathcal{E}_{\text{equi}}$, $\left\|\tilde{\phi}\boldsymbol{k}\right\|_{\mathcal{H}_{\boldsymbol{k}}} = \|\phi\boldsymbol{k}\|_{\mathcal{H}_{\boldsymbol{k}}}$. We choose $A = \mathcal{B}_2(R_n)$ and $\epsilon = \frac{2^{\mathfrak{t}}}{n} = m^{-1}$, so that, with probability at least $1-\delta'$, using the fact that $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$,

$$\begin{aligned}\left\|\tilde{\phi}\boldsymbol{k}\right\|_{\mathcal{H}_{\boldsymbol{k}}} &\le \tfrac{2^{\mathfrak{t}+1}}{n} + \tfrac{2^{\mathfrak{t}}}{n}\sqrt{\tfrac{8}{3}\|\boldsymbol{k}\|_n \log\left(\tfrac{6\mathfrak{t}n}{2^{\mathfrak{t}}\delta}\right)\left[\log\tfrac{4}{\delta'} + \log\mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(R_n), m^{-1})\right]} \\ &\le \tfrac{2^{\mathfrak{t}+1}}{n} + \tfrac{2^{\mathfrak{t}}}{n}\sqrt{\tfrac{8}{3}\|\boldsymbol{k}\|_n \log\left(\tfrac{6\mathfrak{t}n}{2^{\mathfrak{t}}\delta}\right)}\left[\sqrt{\log\tfrac{1}{\delta'}} + \sqrt{\log 4\mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(R_n), m^{-1})}\right] \\ &\le a_{n,m} + v_{n,m}\sqrt{\log\left(\tfrac{1}{\delta'}\right)},\end{aligned} \tag{38}$$

for $a_{n,m}, v_{n,m}$ in Lem. D.1. ◻

Corollary D.1 (MMD guarantee for KT-SPLIT)**.** *Let $\mathcal{S}_\infty$ be an infinite sequence of points in $\mathbb{R}^d$ and $\boldsymbol{k}$ a kernel. For any $\delta \in (0,1)$ and $n, m \in \mathbb{N}$ such that $\log_2 \frac{n}{m} \in \mathbb{N}$, consider the* KT-SPLIT *algorithm (Dwivedi and Mackey, 2022, Algorithm 1a) with parameters $\boldsymbol{k}_{\text{split}} = \boldsymbol{k}$ and $\delta_i = \frac{\delta}{n}$ to compress $\mathcal{S}_n$ to $2^{\mathfrak{t}}$ coresets $\{\mathcal{S}_{\text{out}}^{(i)}\}_{i\in[2^{\mathfrak{t}}]}$ where $\mathfrak{t} = \log_2 \frac{n}{m}$, each with m points. Then for any $i \in [2^{\mathfrak{t}}]$, with probability at least $1-\delta$,*

$$\mathrm{MMD}_{\boldsymbol{k}}(\mathbb{S}_n, \mathbb{S}_{\text{out}}^{(i)}) \le \tfrac{1}{m}\left(2 + \sqrt{\tfrac{8}{3}\|\boldsymbol{k}\|_n \log\left(\tfrac{6(\log_2\frac{n}{m})m}{\delta}\right)\left(\log\mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(R_n), m^{-1}) + \log\tfrac{8}{\delta}\right)}\right). \tag{39}$$

Proof of Cor. D.1. Fix $i \in [2^{\mathfrak{t}}]$. By taking $\delta' = \frac{\delta}{2}$ in (38), we obtain (39). This occurs with probability

$$\begin{aligned}&\Pr\left(\mathrm{MMD}_{\boldsymbol{k}}(\mathbb{S}_n, \mathbb{S}_{\text{out}}^{(i)}) < a_{n,m} + v_{n,m}\sqrt{\log\left(\tfrac{1}{\delta'}\right)}\right) \\ =&1 - \Pr\left(\mathrm{MMD}_{\boldsymbol{k}}(\mathbb{S}_n, \mathbb{S}_{\text{out}}^{(i)}) \ge a_{n,m} + v_{n,m}\sqrt{\log\left(\tfrac{1}{\delta'}\right)}\right) \\ \ge&1 - \Pr\left(\mathcal{E}_{\text{equi}}^{(i)}, \mathrm{MMD}_{\boldsymbol{k}}(\mathbb{S}_n, \mathbb{S}_{\text{out}}^{(i)}) \ge a_{n,m} + v_{n,m}\sqrt{\log\left(\tfrac{1}{\delta'}\right)}\right) - \Pr\left({\mathcal{E}_{\text{equi}}^{(i)}}^c\right) \\ \ge&1 - \Pr\left(\left\|\tilde{\phi}^{(i)}\boldsymbol{k}\right\|_{\mathcal{H}_{\boldsymbol{k}}} \ge a_{n,m} + v_{n,m}\sqrt{\log\left(\tfrac{1}{\delta'}\right)}\right) - \Pr\left({\mathcal{E}_{\text{equi}}^{(i)}}^c\right) \\ \ge&1 - \tfrac{\delta}{2} - \tfrac{\delta}{2} = 1 - \delta.\end{aligned}$$

◻

D.3. Proof of Thm. 3: MMD guarantee for SKT

Thm. 3 will follows directly from Assum. (α, β)-kernel and the following statement for a generic covering number.

Theorem D.1. *Let $\boldsymbol{k}_{\mathbb{P}}$ be a kernel satisfying Assum. 1. Let $\mathcal{S}_\infty$ be an infinite sequence of points. Then for a prefix sequence $\mathcal{S}_n$ of n points, $m \in [n]$, and $n' \triangleq m2^{\lceil \log_2 \frac{n}{m} \rceil}$, SKT outputs w_{SKT} in $O(n^2 d_{\boldsymbol{k}_{\mathbb{P}}})$ time that satisfies, with probability at least $1-\delta$,*

$$\Delta\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(w_{\mathrm{SKT}}) \le \sqrt{\left(\tfrac{1+\log n'}{n'}\right)\|\boldsymbol{k}_{\mathbb{P}}\|_n} + \tfrac{1}{m}\left(2 + \sqrt{\tfrac{8}{3}\|\boldsymbol{k}\|_n \log\left(\tfrac{6(\log_2 \frac{n'}{m})m}{\delta}\right)\left(\log \mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(R_n), m^{-1}) + \log \tfrac{8}{\delta}\right)}\right).$$

Proof of Thm. D.1. The runtime of SKT comes from the fact that all of SteinThinning (with output size n), KT-SPLIT, and KT-Swap take $O(d_{\boldsymbol{k}_{\mathbb{P}}} n^2)$ time.

By Riabiz et al. (2022, Theorem 1), SteinThinning (which is a deterministic algorithm) from n points to n' points has the following guarantee

$$\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^\dagger}, \mathbb{P}) \le \mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w_{\mathrm{OPT}}}, \mathbb{P}) + \left(\tfrac{1+\log n'}{n'}\right)\|\boldsymbol{k}_{\mathbb{P}}\|_n,$$

where we denote the output weight of SteinThinning as $w^\dagger$. Using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$, we have

$$\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^\dagger}, \mathbb{P}) \le \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w_{\mathrm{OPT}}}, \mathbb{P}) + \sqrt{\left(\tfrac{1+\log n'}{n'}\right)\|\boldsymbol{k}_{\mathbb{P}}\|_n}.$$

Fix $\delta \in (0,1)$. By Cor. D.1 with $\boldsymbol{k} = \boldsymbol{k}_{\mathbb{P}}$ and $\mathfrak{t} = \log_2 \frac{n'}{m}$, with probability at least $1-\delta$, we have, for any $i \in [2^{\mathfrak{t}}]$,

$$\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^\dagger}, \mathbb{S}_{\mathrm{out}}^{(i)}) \le \tfrac{1}{m}\left(2 + \sqrt{\tfrac{8}{3}\|\boldsymbol{k}\|_n \log\left(\tfrac{6(\log_2 \frac{n'}{m})m}{\delta}\right)\left(\log \mathcal{N}_{\boldsymbol{k}}(\mathcal{B}_2(R_n), m^{-1}) + \log \tfrac{8}{\delta}\right)}\right),$$

where $\mathbb{S}_{\mathrm{out}}^{(i)}$ is the i-th coreset output by KT-SPLIT. Since KT-Swap can only decrease the MMD to $\mathbb{P}$, we have, by the triangle inequality of $\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}$,

$$\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w_{\mathrm{SKT}}}, \mathbb{P}) \le \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_{\mathrm{out}}^{(1)}, \mathbb{P}) \le \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_{\mathrm{out}}^{(1)}, \mathbb{S}_n^{w^\dagger}) + \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^\dagger}, \mathbb{P}),$$

which gives the desired bound. □

Thm. 3 now follows from Thm. D.1, the kernel growth definitions in Assum. (α, β)-kernel, $n \le n' \le 2n$, and that $\log_2(\frac{n'}{m})m \le n'$. □

E. Resampling of Simplex Weights

Integral to many of our algorithms is a resampling procedure that turns a simplex-weighted point set of size n into an equal-weighted point set of size m while incurring at most $O(1/\sqrt{m})$ MMD error. The motivation for wanting an equal-weighted point set is two-fold: First, in LSKT, we need to provide an equal-weighted point set to KT-Compress++, but the output of LD is a simplex weight. Secondly, we can exploit the fact that non-zero weights are bounded away from zero in equal-weighted point sets to provide a tighter analysis of WeightedRPCholesky. While i.i.d. resampling also achieves the $O(1/\sqrt{m})$ goal, we choose Resample (Alg. E.3), a stratified residual resampling algorithm (Douc and Cappé, 2005, Sec. 3.2, 3.3). In this section, we derive an MMD bound for Resample and show that it is better in expectation than using i.i.d. resampling or residual resampling alone.

Let D_w^{inv} be the inverse of the cumulative distribution function of the multinomial distribution with weight w, i.e.,

$$D_w^{\mathrm{inv}}(u) \triangleq \min\left\{i \in [n] : u \le \textstyle\sum_{j=1}^i w_j\right\}.$$

Algorithm E.1 i.i.d. resampling

Input: Weights $w \in \Delta_{n-1}$, output size m
$w' \leftarrow \mathbf{0} \in \mathbb{R}^n$
for $j = 1$ **to** m **do**
  Draw $U_j \sim \text{Uniform}([0, 1))$
  $\mathtt{I}_j \leftarrow D_w^{\text{inv}}(U_j)$
  $w'_{\mathtt{I}_j} \leftarrow w'_{\mathtt{I}_j} + \frac{1}{m}$
end for
Return: $w' \in \Delta_{n-1} \cap (\frac{\mathbb{N}_0}{m})^n$

Algorithm E.2 Residual resampling

Input: Weights $w \in \Delta_{n-1}$, output size m
$w'_i \leftarrow \frac{\lfloor m w_i \rfloor}{m}, \forall i \in [n]$
$r \leftarrow m - \sum_{i \in [n]} \lfloor m w_i \rfloor \in \mathbb{N}$
$\eta_i \leftarrow \frac{m w_i - \lfloor m w_i \rfloor}{r}, \forall i \in [n]$
for $j = 1$ **to** r **do**
  Draw $U_j \sim \text{Uniform}([0, 1))$
  $\mathtt{I}_j \leftarrow D_\eta^{\text{inv}}(U_j)$
  $w'_{\mathtt{I}_j} \leftarrow w'_{\mathtt{I}_j} + \frac{1}{m}$
end for
Return: $w' \in \Delta_{n-1} \cap (\frac{\mathbb{N}_0}{m})^n$

Proposition E.1 (MMD guarantee of resampling algorithms)**.** *Consider any kernel $\boldsymbol{k}$, points $\mathcal{S}_n = (x_1, \ldots, x_n) \subset \mathbb{R}^d$, and a weight vector $w \in \Delta_{n-1}$.*

(a) *Using the notation from Alg. E.1, let X, X' be independent random variables with law $\mathbb{S}_n^w$. Then, the output weight vector $w^{\text{i.i.d.}} \triangleq w'$ of Alg. E.1 satisfies*

$$\mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{i.i.d.}}}, \mathbb{S}_n^w)] = \frac{\mathbb{E}\boldsymbol{k}(X,X) - \mathbb{E}\boldsymbol{k}(X,X')}{m}. \tag{40}$$

(b) *Using the notation from Alg. E.2, let R, R' be independent random variables with law $\mathbb{S}_n^\eta$. Then, the output weight vector $w^{\text{resid}} \triangleq w'$ of Alg. E.2 satisfies*

$$\mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{resid}}}, \mathbb{S}_n^w)] = \frac{r(\mathbb{E}\boldsymbol{k}(R,R) - \mathbb{E}\boldsymbol{k}(R,R'))}{m^2}. \tag{41}$$

(c) *Using the notation from Alg. E.3, let $R_j \triangleq x_{\mathtt{I}_j}$ and R'_j be an independent copy of R_j. Let R be an independent random variable with law $\mathbb{S}_n^\eta$. Then, the output weight vector $w^{\text{sr}} \triangleq w'$ of Alg. E.3 satisfies*

$$\mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{sr}}}, \mathbb{S}_n^w)] = \frac{r\mathbb{E}\boldsymbol{k}(R,R) - \sum_{j \in [r]} \mathbb{E}\boldsymbol{k}(R_j, R'_j)}{m^2}. \tag{42}$$

Algorithm E.3 Stratified residual resampling (Resample)

Input: Weights $w \in \Delta_{n-1}$, output size m
$w'_i \leftarrow \frac{\lfloor m w_i \rfloor}{m}, \forall i \in [n]$
$r \leftarrow m - \sum_{i \in [n]} \lfloor m w_i \rfloor \in \mathbb{N}$
$\eta_i \leftarrow \frac{m w_i - \lfloor m w_i \rfloor}{r}, \forall i \in [n]$
for $j = 1$ **to** r **do**
  Draw $U_j \sim \text{Uniform}([\frac{j}{r}, \frac{j+1}{r}))$
  $\mathtt{I}_j \leftarrow D_\eta^{\text{inv}}(U_j)$
  $w'_{\mathtt{I}_j} \leftarrow w'_{\mathtt{I}_j} + \frac{1}{m}$
end for
Return: $w' \in \Delta_{n-1} \cap (\frac{\mathbb{N}_0}{m})^n$

Proof of Prop. E.1(a). Let $X_i \triangleq x_{\mathtt{I}_i}$. As random signed measures, we have

$$\mathbb{S}_n^{w'} - \mathbb{S}_n^{w} = \frac{1}{m}\sum_{i\in[m]} \delta_{X_i} - \sum_{i\in[n]} w_i\delta_{x_i}.$$

Hence

$$\begin{aligned}\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w'}, \mathbb{S}_n^{w}) &= ((\mathbb{S}_n^{w'} - \mathbb{S}_n^{w}) \times (\mathbb{S}_n^{w'} - \mathbb{S}_n^{w}))\boldsymbol{k} \\ &= \frac{1}{m^2}\sum_{i,i'\in[m]} \boldsymbol{k}(X_i, X_{i'}) - \frac{2}{m}\sum_{i\in[m],i'\in[n]} w_{i'}\boldsymbol{k}(X_i, x_{i'}) + \sum_{i,i'\in[n]} w_i w_{i'}\boldsymbol{k}(x_i, x_{i'}).\end{aligned}$$

Since each X_i is distributed to $\mathbb{S}_n^w$ and X_i and $X_{i'}$ are independent for $i \neq i'$, taking expectation, we have

$$\mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w'}, \mathbb{S}_n^{w})] = \frac{1}{m}\mathbb{E}\boldsymbol{k}(X, X) + \frac{m-1}{m}\mathbb{E}\boldsymbol{k}(X, X') - 2\mathbb{E}\boldsymbol{k}(X, X') + \mathbb{E}\boldsymbol{k}(X, X').$$

This gives the bound (40). □

Proof of Prop. E.1(b). Let $R_j \triangleq x_{\mathtt{I}_j}$. As random signed measures, we have

$$\begin{aligned}\mathbb{S}_n^{w'} - \mathbb{S}_n^{w} &= \left(\sum_{i\in[n]} \frac{\lfloor mw_i\rfloor}{m}\delta_{x_i} + \frac{1}{m}\sum_{j\in[r]} \delta_{R_j}\right) - \sum_{i\in[n]} w_i\delta_{x_i} \\ &= \frac{1}{m}\sum_{j\in[r]} \delta_{R_j} - \sum_{i\in[n]} \left(w_i - \frac{\lfloor mw_i\rfloor}{m}\right)\delta_{x_i} \\ &= \frac{1}{m}\sum_{j\in[r]} \delta_{R_j} - \frac{r}{m}\sum_{i\in[n]} \eta_i\delta_{x_i}.\end{aligned}$$

Hence

$$\begin{aligned}\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w'}, \mathbb{S}_n^{w}) &= ((\mathbb{S}_n^{w'} - \mathbb{S}_n^{w}) \times (\mathbb{S}_n^{w'} - \mathbb{S}_n^{w}))\boldsymbol{k} \\ &= \frac{1}{m^2}\sum_{j,j'\in[r]} \boldsymbol{k}(R_j, R_{j'}) - \frac{2r}{m^2}\sum_{j\in[r],i\in[n]} \eta_i\boldsymbol{k}(R_j, x_i) + \frac{r^2}{m^2}\sum_{i,i'\in[n]} \eta_i\eta_j\boldsymbol{k}(x_i, x_j).\end{aligned} \tag{43}$$

Since each R_j is distributed to $\mathbb{S}_n^\eta$ and R_j and $R_{j'}$ are independent for $j \neq j'$, taking expectation, we have

$$\mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w'}, \mathbb{S}_n^{w})] = \frac{r}{m^2}\mathbb{E}\boldsymbol{k}(R, R) + \frac{r(r-1)}{m^2}\mathbb{E}\boldsymbol{k}(R, R') - \frac{2r^2}{m^2}\mathbb{E}\boldsymbol{k}(R, R') + \frac{r^2}{m^2}\mathbb{E}\boldsymbol{k}(R, R').$$

This gives the bound (41). □

Proof of Prop. E.1(c). We repeat the same steps from the previous part of the proof to get (43). In the case of (c), R_j's are not identically distributed so the analysis is different. Let R' be an independent copy of R. Taking expectation of (43), we have

$$m^2\mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w'}, \mathbb{S}_n^{w})] = \sum_{j\in[r]} \mathbb{E}\boldsymbol{k}(R_j, R_j) + \sum_{j\in[r]}\sum_{j'\in[r]\setminus\{j\}} \mathbb{E}\boldsymbol{k}(R_j, R_{j'}) - 2r\sum_{j\in[r]} \mathbb{E}\boldsymbol{k}(R_j, R) + r^2\mathbb{E}\boldsymbol{k}(R, R').$$

Note

$$\sum_{j\in[r]} \mathbb{E}\boldsymbol{k}(R_j, R_j) = \sum_{j\in[r]} r\int_{[\frac{j}{r}, \frac{j+1}{r})} \boldsymbol{k}(x_{D_\eta^{\mathrm{inv}}(u)}, x_{D_\eta^{\mathrm{inv}}(u)})\mathrm{d}u = r\int_0^1 \boldsymbol{k}(x_{D_\eta^{\mathrm{inv}}(u)}, x_{D_\eta^{\mathrm{inv}}(u)})\mathrm{d}u = r\mathbb{E}\boldsymbol{k}(R, R),$$

where we used the fact that $x_{D_\eta^{\mathrm{inv}}(U)} \overset{D}{=} R$ for $U \sim \mathrm{Uniform}([0, 1])$. Similarly, we deduce

$$\begin{aligned}\sum_{j\in[r]}\sum_{j'\in[r]\setminus\{j\}} \mathbb{E}\boldsymbol{k}(R_j, R_{j'}) &= \sum_{j\in[r]} \left(\sum_{j'\in[r]} \mathbb{E}\boldsymbol{k}(R_j, R'_{j'}) - \mathbb{E}\boldsymbol{k}(R_j, R'_j)\right) \\ &= \sum_{j\in[r]}(r\mathbb{E}\boldsymbol{k}(R_j, R') - \mathbb{E}\boldsymbol{k}(R_j, R'_j)) \\ &= r^2\mathbb{E}\boldsymbol{k}(R, R') - \sum_{j\in[r]} \mathbb{E}\boldsymbol{k}(R_j, R'_j),\end{aligned}$$

and also

$$\sum_{j\in[r]} \mathbb{E}\boldsymbol{k}(R_j, R) = r\mathbb{E}\boldsymbol{k}(R, R').$$

Combining terms, we get

$$\begin{aligned}m^2\mathbb{E}\,\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w'}, \mathbb{S}_n^{w}) &= r\mathbb{E}\boldsymbol{k}(R, R) + r^2\mathbb{E}\boldsymbol{k}(R, R') - \sum_{j\in[r]} \mathbb{E}\boldsymbol{k}(R_j, R'_j) - 2r^2\mathbb{E}\boldsymbol{k}(R, R') + r^2\mathbb{E}\boldsymbol{k}(R, R') \\ &= r\mathbb{E}\boldsymbol{k}(R, R) - \sum_{j\in[r]} \mathbb{E}\boldsymbol{k}(R_j, R'_j),\end{aligned}$$

which yields the desired bound (42). □

The next proposition shows that stratifying the residuals always improves upon using i.i.d. sampling or residual resampling alone. We need the following convexity lemma.

Lemma E.1 (Convexity of squared MMD)**.** *Let $\boldsymbol{k}$ be a kernel. Let $\mathcal{S}_n = (x_1, \dots, x_n)$ be an arbitrary set of points. The function $E_{\boldsymbol{k}} : \mathbb{R}^n \to \mathbb{R}$ defined by*

$$E_{\boldsymbol{k}}(w) \triangleq \|\mathbb{S}_n^w \boldsymbol{k}\|_{\mathcal{H}_{\boldsymbol{k}}}^2 = \textstyle\sum_{i,j\in[n]} w_i w_j \boldsymbol{k}(x_i, x_j)$$

is convex on $\mathbb{R}^n$.

Proof of Lem. E.1. Since $\boldsymbol{k}$ is a kernel, the Hessian $\nabla^2 E_{\boldsymbol{k}} = 2\boldsymbol{k}(\mathcal{S}_n, \mathcal{S}_n)$ is PSD, and hence $E_{\boldsymbol{k}}$ is convex. $\square$

Proposition E.2 (Stratified residual resampling improves MMD)**.** *Under the assumptions of Prop. E.1, we have*

$$\mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{i.i.d.}}}, \mathbb{S}_n^w)] \geq \mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{resid}}}, \mathbb{S}_n^w)] \geq \mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{sr}}}, \mathbb{S}_n^w)].$$

Proof of Prop. E.2. Let $K \triangleq \boldsymbol{k}(\mathcal{S}_n, \mathcal{S}_n)$. To show the first inequality, note that since $\eta = \frac{mw - \lfloor mw \rfloor}{r}$, by Prop. E.1,

$$\begin{aligned}
\mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{resid}}}, \mathbb{S}_n^w)] &= \tfrac{r(\mathbb{E}\boldsymbol{k}(R,R) - \mathbb{E}\boldsymbol{k}(R,R'))}{m^2} \\
&= \tfrac{r(\sum_{i\in[n]} K_{ii}\eta_i - \sum_{i,j\in[n]} K_{ij}\eta_i\eta_j)}{m^2} \\
&= \tfrac{1}{m}\left(\textstyle\sum_{i\in[n]} K_{ii}\left(w_i - \frac{\lfloor mw_i \rfloor}{m}\right) - \frac{m}{r}\left(w - \frac{\lfloor mw \rfloor}{m}\right)^\top K \left(w - \frac{\lfloor mw \rfloor}{m}\right)\right).
\end{aligned}$$

Hence

$$\begin{aligned}
&\mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{i.i.d.}}}, \mathbb{S}_n^w)] - \mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{resid}}}, \mathbb{S}_n^w)] \\
=&\tfrac{1}{m}\left(\textstyle\sum_{i\in[n]} K_{ii}\frac{\lfloor mw_i \rfloor}{m} + \frac{m}{r}\left(w - \frac{\lfloor mw \rfloor}{m}\right)^\top K\left(w - \frac{\lfloor mw \rfloor}{m}\right) - w^\top K w\right) \\
=&\tfrac{1}{m}\left((1-\theta)\textstyle\sum_{i\in[n]} K_{ii}\xi_i + \theta\eta^\top K\eta - w^\top K w\right),
\end{aligned}$$

where we let $\xi \triangleq \frac{m}{m-r}\frac{\lfloor mw \rfloor}{m}$ and $\theta \triangleq \frac{r}{m}$. Note that $w = \theta\eta + (1-\theta)\xi$. By Lem. E.1 and Jensen's inequality, we have

$$w^\top K w = E_{\boldsymbol{k}}(w) \leq \theta E_{\boldsymbol{k}}(\eta) + (1-\theta)E_{\boldsymbol{k}}(\xi) = \theta\eta^\top K\eta + (1-\theta)\xi^\top K\xi \leq \theta\eta^\top K\eta + (1-\theta)\textstyle\sum_{i\in[n]} K_{ii}\xi_i,$$

where the last inequality follows from Prop. E.1(a) with $w = \xi$ and the fact that MMD is nonnegative. Hence we have shown

$$\mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{i.i.d.}}}, \mathbb{S}_n^w)] - \mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{resid}}}, \mathbb{S}_n^w)] \geq 0,$$

as desired.

For the second inequality, by Prop. E.1, we compute

$$\mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{resid}}}, \mathbb{S}_n^w)] - \mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{sr}}}, \mathbb{S}_n^w)] = \tfrac{r}{m^2}\left(\tfrac{1}{r}\textstyle\sum_{j\in[r]}\mathbb{E}\boldsymbol{k}(R_j, R_j') - \mathbb{E}\boldsymbol{k}(R, R')\right).$$

Note that

$$\mathbb{E}\boldsymbol{k}(R, R') = \textstyle\int_{[0,1)}\int_{[0,1)} k(x_{D_\eta^{\text{inv}}(u)}, x_{D_\eta^{\text{inv}}(v)})\mathrm{d}u\mathrm{d}v = E_{\boldsymbol{k}}\left((D_\eta^{\text{inv}})_\#\text{Uniform}[0,1)\right),$$

where we used $T_\#\mu$ to denote the pushforward measure of μ by T. Similarly,

$$\begin{aligned}
\tfrac{1}{r}\textstyle\sum_{j\in[r]}\mathbb{E}\boldsymbol{k}(R_j, R_j') &= \tfrac{1}{r}\textstyle\sum_{j\in[r]}\int_{[\frac{j}{r},\frac{j+1}{r})}\int_{[\frac{j}{r},\frac{j+1}{r})} k(x_{D_\eta^{\text{inv}}(u)}, x_{D_\eta^{\text{inv}}(v)})\mathrm{d}u\mathrm{d}v \\
&= \tfrac{1}{r}\textstyle\sum_{j\in[r]} E_{\boldsymbol{k}}\left((D_\eta^{\text{inv}})_\#\text{Uniform}\left[\frac{j}{r},\frac{j+1}{r}\right)\right) \\
&\leq E_{\boldsymbol{k}}\left((D_\eta^{\text{inv}})_\#\text{Uniform}[0,1)\right) = \mathbb{E}\boldsymbol{k}(R, R'),
\end{aligned}$$

where in the last inequality we applied Jensen's inequality since $E_{\boldsymbol{k}}$ is convex by Lem. E.1. Hence we have shown

$$\mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{resid}}}, \mathbb{S}_n^w)] - \mathbb{E}[\mathrm{MMD}_{\boldsymbol{k}}^2(\mathbb{S}_n^{w^{\text{sr}}}, \mathbb{S}_n^w)] \geq 0$$

and the proof is complete. $\square$

F. Accelerated Debiased Compression

In this section, we provide supplementary algorithmic details and deferred analyses for LSKT (Alg. 3). In WeightedRPCholesky (Alg. F.1), we provide details for the weighted extension of Chen et al. (2022, Alg. 2.1) that is used extensively in our algorithms. The details of AMD (Wang et al., 2023, Alg. 14) are provided in Alg. F.2. In App. F.1, we give the proof of Thm. 4 for the MMD error guarantee of LD (Alg. 2). In App. F.2, we provide details on KT-Compress++ modified from Compress++ (Shetty et al., 2022) to minimize MMD to $\mathbb{P}$. Finally, Thm. 5 is proved in App. F.3.

Algorithm F.1 Weighted Randomly Pivoted Cholesky (WeightedRPCholesky) (extension of Chen et al. (2022, Alg. 2.1))

Input: kernel $\boldsymbol{k}$, points $\mathcal{S}_n = (x_i)_{i=1}^n$, simplex weights $w \in \Delta_{n-1}$, rank r
$\tilde{\boldsymbol{k}}(i,j) \triangleq \boldsymbol{k}(x_i, x_j)\sqrt{w_i}\sqrt{w_j}$ ▷ *reweighted kernel matrix function*
$F \leftarrow \mathbf{0}_{n\times r}, \mathtt{S} \leftarrow \{\}, d \leftarrow (\tilde{\boldsymbol{k}}(i,i))_{i\in[n]}$
for $i = 1$ **to** r **do**
  Sample $s \sim d/\sum_{j\in[n]} d_j$
  $\mathtt{S} \leftarrow \mathtt{S} \cup \{s\}$
  $g \leftarrow \tilde{\boldsymbol{k}}(:, s) - F(:, 1:i-1)F(s, 1:i-1)^\top$
  $F(:, i) \leftarrow g/\sqrt{g_s}$
  $d \leftarrow d - F(:, i)^2$ ▷ *$F(:,i)^2$ denotes a vector with entry-wise squared values of $F(:,i)$*
  $d \leftarrow \max(d, 0)$ ▷ *numerical stability fix, helpful in practice*
end for
$F \leftarrow \mathrm{diag}((1/\sqrt{w_i})_{i\in[n]})F$ ▷ *undo weighting; treat $1/\sqrt{w_i} = 0$ if $w_i = 0$*
Return: $\mathtt{S} \subset [n]$ with $|\mathtt{S}| = r$ and $F \in \mathbb{R}^{n\times r}$

Algorithm F.2 Accelerated Entropic Mirror Descent (AMD) (modification of Wang et al. (2023, Alg. 14))

Input: kernel matrix $K \in \mathbb{R}^{n\times n}$, number of steps T, initial weight $w_0 \in \Delta_{n-1}$, aggressive flag $\mathtt{AGG}$
$\eta \leftarrow \frac{1}{8w_0^\top \mathrm{diag}(K)}$ if $\mathtt{AGG}$ else $\frac{1}{8\max_{i\in[n]} K_{ii}}$
$v_0 \leftarrow w_0$
for $t = 1$ **to** T **do**
  $\beta_t \leftarrow \frac{2}{t+1}$
  $z_t \leftarrow (1-\beta_t)w_{t-1} + \beta_t v_{t-1}$
  $g \leftarrow 2t\eta K z_t$ ▷ *this is $\gamma_t \nabla f(z_t)$ in Wang et al. (2023, Alg. 14) for $f(w) = w^\top K w$*
  $v_t \leftarrow v_{t-1} \cdot \exp(-g)$ ▷ *component-wise exponentiation and multiplication*
  $v_t \leftarrow v_t / \left\|v_t\right\|_1$ ▷ *$v_t = \mathrm{argmin}_{w\in\Delta_{n-1}} \langle g, w\rangle + D^\phi_{v_{t-1}}(w)$ for $\phi(w) = \sum_{i\in[n]} w_i \log w_i$*
  $w_t \leftarrow (1-\beta_t)w_{t-1} + \beta_t v_t$
end for
Return: $w_T \in \Delta_{n-1}$

F.1. Proof of Thm. 4: Debiasing guarantee for LD

We start with a useful lemma that bounds $w^\top(K - \hat{K})w$ by $\mathrm{tr}\left(K - \hat{K}\right)$ for any simplex weights w.

Lemma F.1. *For any PSD matrix $A \in \mathbb{R}^{n\times n}$ and $w \in \Delta_{n-1}$, we have*

$$w^\top A w \le \mathrm{tr}(A^w) \le \max_{i\in[n]} A_{ii} \le \lambda_1(A),$$

where $\lambda_1(A)$ denotes the largest eigenvalue of A.

Proof of Lem. F.1. Note that

$$w^\top A w = \sqrt{w}^\top \mathrm{diag}(\sqrt{w}) A\, \mathrm{diag}(\sqrt{w})\sqrt{w} = \sqrt{w}^\top A^w \sqrt{w}.$$

The condition that $w \in \Delta_{n-1}$ implies $\left\|\sqrt{w}\right\|_2 = 1$, so that

$$\sqrt{w}^\top A^w \sqrt{w} \le \lambda_1(A^w) \le \mathrm{tr}(A^w).$$

To see $\mathrm{tr}(A^w) \le \max i \in [n] A_{ii}$, note that $\mathrm{tr}(A^w) = \sum_{i\in[n]} A_{ii} w_i \le \max_{i\in[n]} A_{ii}$ since $w \in \Delta_{n-1}$..

Since $\lambda_1(A) = \sup_{x:\|x\|_2=1} x^\top A x$, if we let $i^* \triangleq \mathrm{argmin}_{i\in[n]} A_{ii}$, then the simplex weight with 1 on the i^*-th entry has two-norm 1, so we see that $\max_{i\in[n]} A_{ii} \le \lambda_1(A)$. □

Our next lemma bounds the suboptimality of surrogate optimization of a low-rank plus diagonal approximation of K.

Lemma F.2 (Suboptimality of surrogate optimization). *Let $\boldsymbol{k}_{\mathbb{P}}$ be a kernel satisfying Assum. 1. Let $\mathcal{S}_n = (x_1, \dots, x_n) \subset \mathbb{R}^d$ be a sequence of points. Define $K \triangleq \boldsymbol{k}_{\mathbb{P}}(\mathcal{S}_n, \mathcal{S}_n) \in \mathbb{R}^{n\times n}$. Suppose $\widehat{K} \in \mathbb{R}^{n\times n}$ is another PSD matrix such that $K \succeq \widehat{K}$. Define $D \triangleq \mathrm{diag}(K - \widehat{K})$, the diagonal part of $K - \widehat{K}$, and form $K' \triangleq \widehat{K} + D$. Let $w' \in \mathrm{argmin}_{w\in\Delta_{n-1}} w'^\top K' w'$. Then for any $w \in \Delta_{n-1}$,*

$$\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^w, \mathbb{P}) \le \mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w_{\mathrm{OPT}}}, \mathbb{P}) + \mathrm{tr}\Big((K - \widehat{K})^w\Big) + \max_{i\in[n]}(K - \widehat{K})_{ii} + (w^\top K' w - w'^\top K' w'). \tag{44}$$

Proof of Lem. F.2. Since $K = K' + (K - \widehat{K}) - D$ by construction, we have

$$\begin{aligned} w^\top K w &= w^\top K' w + w^\top (K - \widehat{K}) w - w^\top D w \\ &\le w^\top K' w + w^\top (K - \widehat{K}) w \\ &= (w^\top K' w - w'^\top K' w') + w'^\top K' w' + w^\top (K - \widehat{K}) w \\ &\le (w^\top K' w - w'^\top K' w') + w'^\top K' w' + \mathrm{tr}\Big((K - \widehat{K})^w\Big), \end{aligned}$$

where we used the fact that $D \succeq 0$ and Lem. F.1. Next, by the definition of w', we have

$$\begin{aligned} w'^\top K' w' \le (w_{\mathrm{OPT}})^\top K' w_{\mathrm{OPT}} &= (w_{\mathrm{OPT}})^\top (K' - K) w_{\mathrm{OPT}} + (w_{\mathrm{OPT}})^\top K w_{\mathrm{OPT}} \\ &= (w_{\mathrm{OPT}})^\top (D - (K - \widehat{K})) w_{\mathrm{OPT}} + (w_{\mathrm{OPT}})^\top K w_{\mathrm{OPT}} \\ &\le (w_{\mathrm{OPT}})^\top D w_{\mathrm{OPT}} + (w_{\mathrm{OPT}})^\top K w_{\mathrm{OPT}} \\ &\le \max_{i\in[n]}(K - \widehat{K})_{ii} + (w_{\mathrm{OPT}})^\top K w_{\mathrm{OPT}}, \end{aligned}$$

where we used the fact $K \succeq \widehat{K}$ in the penultimate step and Lem. F.1 in the last step. Hence we have shown our claim. □

Lem. F.2 shows that to control $\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^w, \mathbb{P})$, it suffices to separately control the approximation error in terms of $\mathrm{tr}\Big(K - \hat{K}\Big)$ and the optimization error $(w^\top K' w - w'^\top K' w')$. The next result establishes that using WeightedRPCholesky, we can obtain polynomial and exponential decay bounds for $\mathrm{tr}\Big(K - \hat{K}\Big)$ in expectation depending on the kernel growth of $\boldsymbol{k}_{\mathbb{P}}$.

Proposition F.1 (Approximation error of WeightedRPCholesky). *Let $\boldsymbol{k}$ be a kernel satisfying Assum. $(\boldsymbol{\alpha}, \boldsymbol{\beta})$-kernel. Let $\mathcal{S}_\infty$ be an infinite sequence of points in $\mathbb{R}^d$. For any $w \in \Delta_{n-1}$, let F be the low-rank approximation factor output by WeightedRPCholesky$(\boldsymbol{k}, \mathcal{S}_n, w, r)$. Define $K \triangleq \boldsymbol{k}(\mathcal{S}_n, \mathcal{S}_n)$. If $r \ge (\frac{\mathfrak{C}_d R_n^\beta + 1}{\sqrt{\log 2}} + \sqrt{\log 2})^2 - \frac{1}{\log 2}$, then, with the expectation taken over the randomness in WeightedRPCholesky,*

$$\mathbb{E}\left[\mathrm{tr}\left((K - FF^\top)^w\right)\right] \le H_{n,r}, \tag{45}$$

where $H_{n,r}$ is defined as

$$H_{n,r} \triangleq \begin{cases} 8\sum_{\ell=\mathfrak{U}(r)}^n \left(\frac{L_{\boldsymbol{k}}(R_n)}{\ell}\right)^{\frac{2}{\alpha}} & \textsc{PolyGrowth}(\alpha,\beta), \\ 8\sum_{\ell=\mathfrak{U}(r)}^n \exp\Big(1 - \big(\frac{\ell}{L_{\boldsymbol{k}}(R_n)}\big)^{\frac{1}{\alpha}}\Big) & \textsc{LogGrowth}(\alpha,\beta), \end{cases} \tag{46}$$

for $L_{\boldsymbol{k}}$ defined in (6) and

$$\mathfrak{U}(r) \triangleq \left\lfloor \sqrt{\frac{r + \frac{1}{\log 2}}{\log 2}} - \frac{1}{\log 2} \right\rfloor. \tag{47}$$

Moreover, $H_{n,r}$ satisfies the bounds in Thm. 4.

Proof of Prop. F.1. Recall the notation $L_{\boldsymbol{k}}(R_n) = \frac{\mathfrak{C}_d R_n^\beta}{\log 2}$ from (6). Define $q \triangleq \mathfrak{U}(r)$ so that q is the biggest integer for which $r \geq 2q + q^2 \log 2$. The lower bound assumption of r is chosen such that $q > L_{\boldsymbol{k}}(R_n) > 0$. By Chen et al. (2022, Theorem 3.1) with $\epsilon = 1$, we have

$$\mathbb{E}\left[\operatorname{tr}\left((K - FF^\top)^w\right)\right] \leq 2\sum_{\ell=q+1}^n \lambda_\ell(K^w). \tag{48}$$

Since $q > L_k(R_n)$, we can apply Cor. B.1 to bound $\lambda_\ell(K^w)$ for $\ell \geq q+1$ and obtain (45) since $H_{n,r}$ (46) is constructed to match the bounds when applying Cor. B.1 to (48). It remains to justify the bounds for $H_{n,r}$ in Thm. 4.

If $\boldsymbol{k}$ is PolyGrowth(α, β), by Assum. **($\boldsymbol{\alpha}, \boldsymbol{\beta}$)-kernel** we have $\alpha < 2$. Hence

$$H_{n,r} = 8\sum_{\ell=q}^n \left(\frac{L_{\boldsymbol{k}}(R_n)}{\ell}\right)^{\frac{2}{\alpha}} \leq 8L_{\boldsymbol{k}}(R_n)^{\frac{2}{\alpha}} \int_{q-1}^\infty \ell^{-\frac{2}{\alpha}} \mathrm{d}\ell = 8L_{\boldsymbol{k}}(R_n)^{\frac{2}{\alpha}}(q-1)^{1-\frac{2}{\alpha}} = O\left(\sqrt{r}(\tfrac{R_n^{2\beta}}{r})^{\frac{1}{\alpha}}\right),$$

where we used the fact that $\int_{q-1}^\infty \ell^{-\frac{2}{\alpha}} \mathrm{d}\ell = (q-1)^{1-\frac{2}{\alpha}}$ for $\alpha < 2$, $L_{\boldsymbol{k}}(R_n) = O(R_n^\beta)$, and $q = \Theta(\sqrt{r})$.

If $\boldsymbol{k}$ is LogGrowth(α, β), then

$$H_{n,r} = 8\sum_{\ell=q}^n \exp\left(1 - \left(\frac{\ell}{L_{\boldsymbol{k}}(R_n)}\right)^{\frac{1}{\alpha}}\right) = 8e\sum_{\ell=q}^n c^{\ell^{1/\alpha}} \leq 8e\int_{\ell=q-1}^\infty c^{\ell^{1/\alpha}},$$

where $c \triangleq \exp\left(-L_{\boldsymbol{k}}(R_n)^{-1/\alpha}\right) \in (0,1)$. Defining $m \triangleq -\log c > 0$ and $q' = q-1$, we have

$$\int_{x=q'}^\infty c^{x^{1/\alpha}} \mathrm{d}x = \int_{x=q'}^\infty \exp\left(-mx^{1/\alpha}\right)\mathrm{d}x = \alpha q'(mq'^{1/\alpha})^{-\alpha}\Gamma(\alpha, mq'^{1/\alpha}) = \alpha m^{-\alpha}\Gamma(\alpha, mq'^{1/\alpha}), \tag{49}$$

where $\Gamma(\alpha, x) \triangleq \int_x^\infty t^{\alpha-1}e^{-t}\mathrm{d}t$ is the incomplete gamma function. Since $\alpha > 0$, by Pinelis (2020, Thm. 1.1), we have

$$\Gamma(\alpha, mq'^{1/\alpha}) \leq \frac{(mq'^{1/\alpha}+b)^\alpha - (mq'^{1/\alpha})^\alpha}{\alpha b} e^{-mq'^{1/\alpha}},$$

where b is a known constant depending only on α. By the equivalence of norms on $\mathbb{R}^2$, there exists $C_\alpha > 0$ such that $(x+y)^\alpha \leq C_\alpha(x^\alpha + y^\alpha)$ for any $x, y > 0$. Hence

$$\Gamma(\alpha, mq'^{1/\alpha}) \leq \frac{(mq'^{1/\alpha}+b)^\alpha}{\alpha b} e^{-mq'^{1/\alpha}} \leq \frac{C_\alpha(m^\alpha q' + b^\alpha)}{\alpha b} e^{-mq'^{1/\alpha}}.$$

Hence from (49) we deduce

$$\sum_{\ell=q'}^\infty c^{\ell^{1/\alpha}} \leq C_\alpha(q'b^{-1} + b^{\alpha-1}m^{-\alpha})e^{-mq'^{1/\alpha}}. \tag{50}$$

Since $m = -\log c = L_{\boldsymbol{k}}(R_n)^{-1/\alpha}$, we can bound the exponent by

$$-mq'^{1/\alpha} = -(L_k(R_n)^{-1}q')^{1/\alpha} = -(\tfrac{q'\log 2}{\mathfrak{C}_d R_n^\beta})^{1/\alpha} \leq -(\tfrac{0.83\sqrt{r}-2.39}{\mathfrak{C}_d R_n^\beta})^{1/\alpha},$$

where we used the fact that $q'\log 2 = (q-1)\log 2 \geq (\sqrt{\frac{r+\frac{1}{\log 2}}{\log 2}} - \frac{1}{\log 2} - 2)\log 2 \geq 0.83\sqrt{r} - 2.39$. On the other hand, since $q' = q - 1 \geq L(R_n) = m^{-\alpha}$, we can absorb the $b^{\alpha-1}m^{-1}$ term in (50) into q and finally obtain the bounds for $H_{n,r}$ in Thm. 4. □

The last piece of our analysis involves bounding the optimization error $(w^\top K'w - w'^\top K'w')$ in (44).

Lemma F.3 (AMD guarantee for debiasing)**.** *Let $K \in \mathbb{R}^{n\times n}$ be an SPSD matrix. Let $f(w) \triangleq w^\top Kw$. Then the final iterate x_T of Nesterov's 1-memory method (Wang et al., 2023, Algorithm 14) after T steps with objective function $f(w)$, norm $\|\cdot\| = \|\cdot\|_1$, distance-generating function $\phi(x) = \sum_{i=1}^n x_i \log x_i$, and initial point $w_0 = (\frac{1}{n}, \ldots, \frac{1}{n}) \in \Delta_{n-1}$ satisfies*

$$f(w_T) - f(w_{\mathrm{OPT}}) \leq \frac{16\log n \max_{i\in[n]} K_{ii}}{T^2},$$

where $w_{\mathrm{OPT}} \in \operatorname{argmin}_{x\in\mathbb{R}^n} f(x)$.

Proof of Lem. F.3. We apply Wang et al. (2023, Theorem 14). Hence it remains to determine the smoothness constant $L > 0$ such that, for all $x, y \in \Delta_{n-1}$,

$$\|\nabla f(x) - \nabla f(y)\|_\infty \le L \|x - y\|_1 ,$$

and an upper bound for the Bregman divergence $D^\phi_{w_0}(w_{\mathrm{OPT}}) = \sum_{i=1}^n w_{\mathrm{OPT}\,i} \log \frac{w_{\mathrm{OPT}\,i}}{(w_0)_i} = \sum_{i=1}^n w_{\mathrm{OPT}\,i} \log n w_{\mathrm{OPT}\,i}$. To determine L, note $\nabla f(w) = 2Kw$, so we have, for any $x, y \in \Delta_{n-1}$,

$$\begin{aligned}\|\nabla f(x) - \nabla f(y)\|_\infty &= 2\|K(x-y)\|_\infty = 2\max_{i\in[n]} |K_{i,:}(x-y)| \\ &\le 2\max_{i\in[n]} \|K_{i,:}\|_\infty \|x-y\|_1 = 2\left(\max_{i\in[n]} K_{ii}\right)\|x-y\|_1 = 2\left(\max_{i\in[n]} K_{ii}\right)\|x-y\|_1 ,\end{aligned}$$

where we used the fact that the largest entry in an SPSD matrix appears on its diagonal. Thus we can take the smoothness constant to be

$$L = 2\max_{i\in[n]} K_{ii}.$$

To bound $D^\phi_{w_0}(w_{\mathrm{OPT}})$, note that by Jensen's inequality,

$$D^\phi_{w_0}(w) = \sum_{i=1}^n w_i \log n w_i \le \log\left(\sum_{i=1}^n n w_i^2\right) = \log n + \log \|w\|_2^2 \le \log n,$$

where we used the fact that $\|w\|_2^2 \le \|w\|_1 = 1$ for $w \in \Delta_{n-1}$. □

With these tools in hand, we turn to the proof of Thm. 4. For the runtime of LD, it follows from the fact that WeightedRPCholesky takes $O((d_{\boldsymbol{k}_\mathbb{P}} + r)nr)$ time and one step of AMD takes $O(nr)$ time.

The error analysis is different for the first adaptive iteration and the ensuing adaptive iterations. Roughly speaking, we will show that the first adaptive iteration brings the MMD gap to the desired level, while the ensuing iterations do not introduce an excessive amount of error.

Step 1. Bound $\Delta\,\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(w^{(1)})$

Let $K \triangleq \boldsymbol{k}_\mathbb{P}(\mathcal{S}_n, \mathcal{S}_n)$ and F denote the low-rank approximation factor generated by WeightedRPCholesky. Denote $\widehat{K} \triangleq FF^\top$. Then $K' = \widehat{K} + \mathrm{diag}(K - \widehat{K})$. First, note that since $w^{(0)} = (\frac{1}{n}, \ldots, \frac{1}{n})$, Resample returns $\tilde{w} = w^{(0)}$ with probability one. By Lem. F.2, we have, using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$ repeatedly and Lem. F.1 that $\mathrm{tr}\Big((K - \widehat{K})^w\Big) \le \lambda_1(K - \widehat{K})$ and $\max_{i\in[n]}(K - \widehat{K})_{ii} \le \lambda_1(K - \widehat{K})$,

$$\begin{aligned}\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w^{(1)}}, \mathbb{P}) &\le \mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w_{\mathrm{OPT}}}, \mathbb{P}) + \sqrt{2\lambda_1(K - \widehat{K})} + \sqrt{{w^{(1)}}^\top K' w^{(1)} - w'^\top K' w'} \\ &\le \mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w_{\mathrm{OPT}}}, \mathbb{P}) + \sqrt{2\lambda_1(K - \widehat{K})} + \sqrt{\tfrac{16\log n \|\boldsymbol{k}_\mathbb{P}\|_n}{T^2}},\end{aligned}$$

where we applied Lem. F.1 and Lem. F.3 in the last inequality. Fix $\delta \in (0, 1)$. By Markov's inequality, we have

$$\Pr\left(\sqrt{\lambda_1(K - \widehat{K})} > \sqrt{\tfrac{\mathbb{E}\left[\lambda_1(K - \widehat{K})\right]}{\delta}}\right) \le \delta.$$

This means that with probability at least $1 - \delta$, we have

$$\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w^{(1)}}, \mathbb{P}) \le \mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w_{\mathrm{OPT}}}, \mathbb{P}) + \sqrt{\tfrac{2\mathbb{E}\left[\lambda_1(K - \widehat{K})\right]}{\delta}} + \sqrt{\tfrac{16\log n \|\boldsymbol{k}_\mathbb{P}\|_n}{T^2}}.$$

Note that the lower bound condition on r in Assum. $(\boldsymbol{\alpha},\boldsymbol{\beta})$-params implies the lower bound condition in Prop. F.1. Hence, by Prop. F.1 with $w = (\frac{1}{n}, \ldots, \frac{1}{n})$ and using the identity $\lambda_1(K - \widehat{K}) \le \mathrm{tr}\Big(K - \widehat{K}\Big)$ while noting that a factor of n appears, we have

$$\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w^{(1)}}, \mathbb{P}) \le \mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w_{\mathrm{OPT}}}, \mathbb{P}) + \sqrt{\tfrac{2nH_{n,r}}{\delta}} + \sqrt{\tfrac{16\|\boldsymbol{k}_\mathbb{P}\|_n \log n}{T^2}}.$$

Step 2. Bound the error of the remaining iterations

Fix $\delta > 0$. The previous step shows that, with probability at least $1 - \frac{\delta}{2}$,

$$\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^{(1)}}, \mathbb{P}) \le \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w_{\mathrm{OPT}}}, \mathbb{P}) + \sqrt{\tfrac{4nH_{n,r}}{\delta}} + \sqrt{\tfrac{16\|\boldsymbol{k}_{\mathbb{P}}\|_n \log n}{T^2}}.$$

Fix $q > 1$, and let $\tilde{w}$ be the resampled weight defined in the q-th iteration in Alg. 2. Without loss of generality, we assume $\tilde{w}_i > 0$ for all $i > 0$, since if $w_i = 0$ then index i is irrelevant for the rest of the algorithm. Thus, thanks to Resample, we have $\tilde{w}_i \ge 1/n$ for all $i \in [n]$. Let a/b denote the entry-wise division between two vectors. As in the previous step of the proof, we let $K \triangleq \boldsymbol{k}_{\mathbb{P}}(\mathcal{S}_n, \mathcal{S}_n)$, F be the low-rank factor output by WeightedRPCholesky$(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, \tilde{w}, r)$, and $\widehat{K} = FF^\top$. For any $w \in \Delta_{n-1}$, recall the notation $K^w \triangleq \mathrm{diag}(\sqrt{w}) K \,\mathrm{diag}(\sqrt{w})$. Then we have

$$\begin{aligned}
w^\top K w &= (w/\sqrt{\tilde{w}})^\top \mathrm{diag}(K^{\tilde{w}})(w/\sqrt{\tilde{w}}) \\
&= (w/\sqrt{\tilde{w}})^\top (\mathrm{diag}(\sqrt{\tilde{w}})\widehat{K}\,\mathrm{diag}(\sqrt{\tilde{w}})) + \mathrm{diag}(\sqrt{\tilde{w}})(K - \widehat{K})\,\mathrm{diag}(\sqrt{\tilde{w}})))(w/\sqrt{\tilde{w}}) \\
&= w^\top \widehat{K} w + (w/\sqrt{\tilde{w}})^\top (\mathrm{diag}(\sqrt{\tilde{w}})(K - \widehat{K})\,\mathrm{diag}(\sqrt{\tilde{w}})))(w/\sqrt{\tilde{w}}) \\
&\le w^\top \widehat{K} w + \max\nolimits_{i\in[n]}(1/\tilde{w}_i)\,\mathrm{tr}\Big(\mathrm{diag}(\sqrt{\tilde{w}})(K - \widehat{K})\,\mathrm{diag}(\sqrt{\tilde{w}})\Big) \\
&\le w^\top \widehat{K} w + n\,\mathrm{tr}\Big((K - \widehat{K})^{\tilde{w}}\Big).
\end{aligned} \tag{51}$$

Note that

$$K' = \widehat{K} + \mathrm{diag}(K - \widehat{K}) = K + (\widehat{K} - K) + \mathrm{diag}(K - \widehat{K}).$$

Since $K' \succeq \widehat{K}$, we have

$$w^{(q)\top} \widehat{K} w^{(q)} \le w^{(q)\top} K' w^{(q)} \le \tilde{w}^\top K' \tilde{w}, \tag{52}$$

where the last inequality follows from the if conditioning at the end of Alg. 2. In addition,

$$\begin{aligned}
\tilde{w}^\top K' \tilde{w} &= \tilde{w}^\top (K + (\widehat{K} - K) + \mathrm{diag}(K - \widehat{K}))\tilde{w} \\
&\le \tilde{w}^\top K \tilde{w} + \tilde{w}^\top \mathrm{diag}(K - \widehat{K})\tilde{w} \\
&= \tilde{w}^\top K \tilde{w} + \sqrt{\tilde{w}}^\top \mathrm{diag}((K - \widehat{K})^{\tilde{w}})\sqrt{\tilde{w}} \\
&\le w^\top K \tilde{w} + \mathrm{tr}\Big((K - \widehat{K})^{\tilde{w}}\Big),
\end{aligned}$$

where we used the fact that $K \succeq \widehat{K}$ and $\left\|\sqrt{\tilde{w}}\right\|_2 = 1$. Plugging the previous inequality into (52) and then into (51) with $w = w^{(q)}$, we get

$$w^{(q)\top} K w^{(q)} \le \tilde{w}^\top K \tilde{w} + (n+1)\,\mathrm{tr}\Big((K - \widehat{K})^{\tilde{w}}\Big). \tag{53}$$

Taking square-root on both sides using $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$ for $a, b \ge 0$ and the triangle inequality, we get

$$\begin{aligned}
\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^{(q)}}, \mathbb{P}) &\le \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{\tilde{w}}, \mathbb{P}) + \sqrt{(n+1)\,\mathrm{tr}\Big((K - \widehat{K})^{\tilde{w}}\Big)} \\
&\le \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^{(q-1)}}, \mathbb{P}) + \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^{(q-1)}}, \mathbb{S}_n^{\tilde{w}}) + \sqrt{(n+1)\,\mathrm{tr}\Big((K - \widehat{K})^{\tilde{w}}\Big)}.
\end{aligned}$$

By Markov's inequality, we have

$$\begin{aligned}
\Pr\left(\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^{(q-1)}}, \mathbb{S}_n^{\tilde{w}}) > \sqrt{\tfrac{4Q\mathbb{E}\left[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^{(q-1)}}, \mathbb{S}_n^{\tilde{w}})\right]}{\delta}}\right) &\le \tfrac{\delta}{4Q} \\
\Pr\left(\sqrt{\mathrm{tr}\Big((K - \widehat{K})^{\tilde{w}}\Big)} > \sqrt{\tfrac{4Q\mathbb{E}\left[\mathrm{tr}\left((K-\widehat{K})^{\tilde{w}}\right)\right]}{\delta}}\right) &\le \tfrac{\delta}{4Q}.
\end{aligned}$$

By Prop. E.1(c), we have

$$\mathbb{E}\left[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^{(q-1)}}, \mathbb{S}_n^{\tilde{w}})\right] = \mathbb{E}\left[\mathbb{E}\left[\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^{(q-1)}}, \mathbb{S}_n^{\tilde{w}})\middle| w^{(q-1)}\right]\right] \le \frac{\|\boldsymbol{k}_{\mathbb{P}}\|_n}{n}.$$

Thus by the union bound, with probability at least $1 - \frac{\delta}{2Q}$, using Prop. F.1 (recall low-rank approximation $\widehat{K}$ is obtained using $\tilde{w}$), we have

$$\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^{(q)}}, \mathbb{P}) \le \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^{(q-1)}}, \mathbb{P}) + \sqrt{\frac{4Q\|\boldsymbol{k}_{\mathbb{P}}\|_n}{n\delta}} + \sqrt{\frac{4Q(n+1)H_{n,r}}{\delta}}. \tag{54}$$

Finally, applying union bound and summing up the bounds for $q = 1, \ldots, Q$, we get, with probability at least $1 - \delta$,

$$\Delta\,\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(w^{(q)}) \le \sqrt{\frac{2nH_{n,r}}{\delta}} + \sqrt{\frac{16\|\boldsymbol{k}_{\mathbb{P}}\|_n \log n}{T^2}} + (Q-1)\left(\sqrt{\frac{4Q\|\boldsymbol{k}_{\mathbb{P}}\|_n}{n\delta}} + \sqrt{\frac{4Q(n+1)H_{n,r}}{\delta}}\right).$$

This matches the stated asymptotic bound in Thm. 4. □

F.2. Thinning with KT-Compress++

For compression with target distribution $\mathbb{P}$, we modify the original KT-Compress++ algorithm of (Shetty et al., 2022, Ex. 6): in HALVE and THIN of Compress++, we use KT-SPLIT with kernel $\boldsymbol{k}_{\mathbb{P}}$ without KT-SWAP (so our version of Compress++ outputs $2^{\mathfrak{g}}$ coresets, each of size $\sqrt{n}$), followed by KT-Swap to obtain a size $\sqrt{n}$ coreset. We call the resulting thinning algorithm KT-Compress++. We show in Lem. F.4 and Cor. F.1 that KT-Compress++ satisfies an MMD guarantee similar to that of quadratic-time kernel thinning.

Algorithm F.3 KT-Compress++ (modified Shetty et al. (2022, Alg. 2) to minimize MMD to $\mathbb{P}$)

Input: kernel $\boldsymbol{k}_{\mathbb{P}}$ with zero-mean under $\mathbb{P}$, input points $\mathcal{S}_n = (x_i)_{i\in[n]}$, multiplicity n' with $n' \in 4^{\mathbb{N}}$, weight $w \in \Delta_{n-1} \cap (\frac{\mathbb{N}_0}{n'})^n$, thinning parameter $\mathfrak{g}$, failure probability δ
$\mathtt{S} \leftarrow$ index sequence where $k \in [n]$ appears $n'w_k$ times
$(\mathtt{I}^{(\ell)})_{\ell\in[2^{\mathfrak{g}}]} \leftarrow$ Compress++$(\mathfrak{g}, \mathcal{S}_n[\mathtt{S}])$ ▷ *Shetty et al. (2022, Ex. 6) with KT substituted with* KT-SPLIT *in* HALVE *and* THIN.
$\mathtt{I}^{(\ell)} \leftarrow \mathtt{S}[\mathtt{I}^{(\ell)}]$ for each $\ell \in [2^{\mathfrak{g}}]$
$\mathtt{I} \leftarrow$ KT-Swap$(\boldsymbol{k}_{\mathbb{P}}, \mathcal{S}_n, (\mathtt{I}^{(\ell)})_{\ell\in[2^{\mathfrak{g}}]})$
$w_{\text{C++}} \leftarrow$ simplex weights corresponding to $\mathtt{I}$ ▷ $w_i = \frac{\text{number of occurrences of } i \text{ in } \mathtt{I}}{|\mathtt{I}|}$
Return: $w_{\text{C++}} \in \Delta_{n-1} \cap (\frac{\mathbb{N}_0}{\sqrt{n}})^n$ ▷ *Hence* $\|w_{\text{C++}}\|_0 \le \sqrt{n}$

Lemma F.4 (Sub-gaussian guarantee for Compress++)**.** *Let $\mathcal{S}_n$ be a sequence of n points with $n \in 4^{\mathbb{N}}$. For any $\delta \in (0,1)$ and integer $\mathfrak{g} \ge \lceil \log_2 \log(n+1) + 3.1 \rceil$, consider the* Compress++ *algorithm (Shetty et al., 2022, Algorithm 2) with thinning parameter $\mathfrak{g}$, halving algorithm* HALVE$^{(k)} \triangleq$ symmetrized[6](KT-SPLIT$(\boldsymbol{k}, \cdot, 1, \frac{n_k^2}{4n2^{\mathfrak{g}}(\mathfrak{g}+(\beta_n+1)2^{\mathfrak{g}})}\delta)$) *for an input of $n_k \triangleq 2^{\mathfrak{g}+1+k}\sqrt{n}$ points and $\beta_n \triangleq \log_2\left(\frac{n}{n_0}\right)$, and with thinning algorithm* THIN $\triangleq$ KT-SPLIT$(\boldsymbol{k}, \cdot, \mathfrak{g}, \frac{\mathfrak{g}}{\mathfrak{g}+(\beta_n+1)2^{\mathfrak{g}}}\delta)$. *Then this instantiation of* Compress++ *compresses $\mathcal{S}_n$ to $2^{\mathfrak{g}}$ coresets $(\mathcal{S}_{\text{out}}^{(i)})_{i\in[2^{\mathfrak{g}}]}$ of $\sqrt{n}$ points each. Denote the signed measure $\phi^{(i)} \triangleq \frac{1}{n}\sum_{x\in\mathcal{S}_n}\delta_x - \frac{1}{\sqrt{n}}\sum_{x\in\mathcal{S}_{\text{out}}^{(i)}}\delta_x$. Then for each $i \in [2^{\mathfrak{g}}]$, on an event $\mathcal{E}_{\text{equi}}^{(i)}$ with $\Pr\left(\mathcal{E}_{\text{equi}}^{(i)}\right) \ge 1 - \frac{\delta}{2}$, $\phi^{(i)} = \tilde{\phi}^{(i)}$ for a random signed measure $\tilde{\phi}^{(i)}$ such that, for any $\delta' \in (0,1)$,*

$$\Pr\left(\left\|\tilde{\phi}^{(i)}\boldsymbol{k}\right\|_{\mathcal{H}_{\boldsymbol{k}}} \ge a'_n\left(1 + \sqrt{\log\left(\tfrac{1}{\delta'}\right)}\right)\right) \le \delta',$$

where

$$a'_n = \frac{4}{\sqrt{n}}\left(2 + \sqrt{\tfrac{8}{3}\|\boldsymbol{k}\|_n \log\left(\frac{6\sqrt{n}(\mathfrak{g}+(\frac{\log_2 n}{2}-\mathfrak{g})2^{\mathfrak{g}})}{\delta}\right)\log\left(4\mathcal{N}_{\boldsymbol{k}}\left(\mathcal{B}_2(R_n), n^{-1/2}\right)\right)}\right).$$

[6] Any halving algorithm can be converted into an unbiased one by symmetrization, i.e., returning either the output half or its complement with equal probability (Shetty et al., 2022, Remark 3).

Proof of Lem. F.4. This proof is similar to the one for Shetty et al. (2022, Ex. 6) but with explicit constant tracking and is self-contained, invoking only Shetty et al. (2022, Thm. 4) which gives MMD guarantees for Compress++ given the sub-Gaussian parameters of HALVE and THIN.

Recall that n_k is the number of input points for the halving subroutine at recursion level k in Compress++, and β_n is the total number of recursion levels. Let $\mathcal{S}_{\mathrm{C}}$ denote the output of COMPRESS (Shetty et al., 2022, Alg. 1) of size $2^{\mathfrak{g}}\sqrt{n}$. Fix $\delta, \delta' \in (0,1)$. Suppose we use $\textsc{Halve}^{(k)} \triangleq \text{symmetrized}(\textsc{kt-split}(\boldsymbol{k}, \cdot, 1, \gamma_k\delta))$ for an input of n_k points for γ_k to be determined. Suppose we use $\textsc{Thin} \triangleq \textsc{kt-split}(\boldsymbol{k}, \cdot, \mathfrak{g}, \gamma'\delta)$ for γ' to be determined; this is the kernel thinning stage that thins $2^{\mathfrak{g}}\sqrt{n}$ points to $2^{\mathfrak{g}}$ coresets, each with $\sqrt{n}$ points. Since the analysis is the same for all coresets, we will fix an arbitrary coreset without superscript in the notation.

By Lem. D.1, with notation $t \triangleq \log\frac{1}{\delta'}$, there exist events $\mathcal{E}_{k,j}$, $\mathcal{E}_{\mathrm{T}}$, and random signed measures $\phi_{k,j}$, $\tilde{\phi}_{k,j}$, ϕ_{T}, $\tilde{\phi}_{\mathrm{T}}$ for $0 \le k \le \beta_n$ and $j \in [4^k]$ such that

(a) $\Pr\left(\mathcal{E}_{k,j}^c\right) \le \frac{\gamma_k\delta}{2}$ and $\Pr(\mathcal{E}_{\mathrm{T}}^c) \le \frac{\gamma'\delta}{2}$,

(b) $\mathbb{1}_{\mathcal{E}_{k,j}}\phi_{k,j} = \mathbb{1}_{\mathcal{E}_{k,j}}\tilde{\phi}_{k,j}$ and $\mathbb{1}_{\mathcal{E}_{\mathrm{T}}}\phi_{\mathrm{T}} = \mathbb{1}_{\mathcal{E}_{\mathrm{T}}}\tilde{\phi}_{\mathrm{T}}$,

(c) We have

$$\Pr\left(\left\|\tilde{\phi}_{k,j}\boldsymbol{k}\right\|_{\mathcal{H}_{\boldsymbol{k}}} \ge a_{n_k} + v_{n_k}\sqrt{t}\,\middle|\,\{\tilde{\phi}_{k',j'}\}_{k'>k,j'\ge 1}, \{\tilde{\phi}_{k',j'}\}_{k',j'<j}\right) \le e^{-t}$$
$$\Pr\left(\left\|\tilde{\phi}_{\mathrm{T}}\boldsymbol{k}\right\|_{\mathcal{H}_{\boldsymbol{k}}} \ge a'_n + v'_n\sqrt{t}\,\middle|\,\mathcal{S}_{\mathrm{C}}\right) \le e^{-t},$$

where, by Lem. D.1, and by increasing the sub-Gaussian constants if necessary, we have

$$a_{n_k} \triangleq v_{n_k} \triangleq a_{n_k, n_k/2} = \tfrac{2}{n_k}\left(2 + \sqrt{\tfrac{8}{3}\|\boldsymbol{k}\|_n \log\left(\tfrac{3n_k}{\gamma_k\delta}\right)\log\left(4\mathcal{N}_{\boldsymbol{k}}\left(\mathcal{B}_2(R_n), \tfrac{2}{n_k}\right)\right)}\right),$$
$$a'_n \triangleq v'_n \triangleq a_{2^{\mathfrak{g}}\sqrt{n},\sqrt{n}} = \tfrac{1}{\sqrt{n}}\left(2 + \sqrt{\tfrac{8}{3}\|\boldsymbol{k}\|_n \log\left(\tfrac{6\mathfrak{g}\sqrt{n}}{\gamma'\delta}\right)\log\left(4\mathcal{N}_{\boldsymbol{k}}\left(\mathcal{B}_2(R_n), n^{-1/2}\right)\right)}\right), \quad \text{and}$$

(d) $\mathbb{E}\left[\tilde{\phi}_{k,j}\boldsymbol{k}\,\middle|\,\{\tilde{\phi}_{k',j'}\}_{k'>k,j'\ge 1}, \{\tilde{\phi}_{k',j'}\}_{k',j'<j}\right] = 0$.

Hence on the event $\mathcal{E} = \bigcap_{k,j}\mathcal{E}_{k,j} \cap \mathcal{E}_{\mathrm{T}}$, these properties hold simultaneously. We will choose $\{\gamma_k\}_k$ and γ' such that $\Pr(\mathcal{E}^c) \le \frac{\delta}{2}$. By the union bound,

$$\Pr(\mathcal{E}^c) \le \Pr(\mathcal{E}_{\mathrm{T}}^c) + \textstyle\sum_{k=0}^{\beta_n}\sum_{j=1}^{4^k}\Pr\left(\mathcal{E}_{k,j}^c\right) \le \frac{\gamma'\delta}{2} + \textstyle\sum_{k=0}^{\beta_n} 4^k\frac{\gamma_k\delta}{2}. \tag{55}$$

On the event $\mathcal{E}$, we apply Shetty et al. (2022, Thm. 4, Rmk. 7) to get a sub-Gaussian guarantee for $\mathrm{MMD}_{\boldsymbol{k}}(\mathbb{S}_n, \mathbb{S}_{\mathrm{out}})$. We want to choose γ_k, γ' such that the rescaled quantities $\tilde{\zeta}_{\mathrm{H}} \triangleq \frac{n_0}{2}a_{n_0}$ and $\tilde{\zeta}_{\mathrm{T}} \triangleq \sqrt{n}a'_n$ satisfy $\tilde{\zeta}_{\mathrm{H}} = \tilde{\zeta}_{\mathrm{T}}$ (Shetty et al., 2022, Eq. (13)), which implies that we need

$$\frac{3n_0}{\gamma_0\delta} = \frac{6\mathfrak{g}\sqrt{n}}{\gamma'\delta} \iff \frac{\gamma_0}{\gamma'} = \frac{2^{\mathfrak{g}}}{\mathfrak{g}}. \tag{56}$$

Hence if we take $\gamma' = \frac{\mathfrak{g}}{\mathfrak{g}+(\beta_n+1)2^{\mathfrak{g}}}$ and $\gamma_k = \frac{n_k^2}{4n2^{\mathfrak{g}}(\mathfrak{g}+(\beta_n+1)2^{\mathfrak{g}})}$, then (56) holds and the upper bound in (55) becomes $\frac{\delta}{2}$. Note that $n_k a_{n_k}$ is non-decreasing in n_k, so by Shetty et al. (2022, Theorem 4, Remark 7), Compress++$(\delta, \mathfrak{g})$ outputs a signed measure ϕ that, on the event $\mathcal{E}$ with $\Pr(\mathcal{E}^c) \le \frac{\delta}{2}$, equals another signed measure $\tilde{\phi}$ that satisfies, for any $\delta' \in (0,1)$,

$$\Pr\left(\left\|\tilde{\phi}\boldsymbol{k}\right\|_{\mathcal{H}_{\boldsymbol{k}}} \ge \hat{a}_n + \hat{v}_n\sqrt{\log\left(\tfrac{1}{\delta'}\right)}\right) \le \delta',$$

where $\hat{a}_n, \hat{v}_n$ satisfy $\max(\hat{a}_n, \hat{v}_n) \le 4a'_n$ whenever $\mathfrak{g} \ge \lceil\log_2\log(n+1) + 3.1\rceil$. □

Corollary F.1 (MMD guarantee for Compress++). *Let $\mathcal{S}_\infty$ be an infinite sequence of points in $\mathbb{R}^d$ and $\boldsymbol{k}$ a kernel. For any $\delta \in (0,1)$ and $n \in \mathbb{N}$ such that $n \in 4^{\mathbb{N}}$, consider the* Compress++ *with the same parameters as in Lem. F.4 with $\mathfrak{g} \geq \lceil \log_2 \log(n+1) + 3.1 \rceil$. Then for any $i \in [\sqrt{n}]$, with probability at least $1-\delta$,*

$$\mathrm{MMD}_{\boldsymbol{k}}(\mathbb{S}_n, \mathbb{S}_{\text{out}}^{(i)}) \leq \frac{4}{\sqrt{n}} \left(2 + \sqrt{\frac{8}{3}\|\boldsymbol{k}\|_n \log\left(\frac{6\sqrt{n}(\mathfrak{g}+(\frac{\log_2 n}{2}-\mathfrak{g})2^{\mathfrak{g}})}{\delta}\right) \log\left(4\mathcal{N}_{\boldsymbol{k}}\left(\mathcal{B}_2(R_n), n^{-1/2}\right)\right)}\right) \left(1 + \sqrt{\log\frac{2}{\delta}}\right).$$

Proof. After applying Lem. F.4 with $\delta' = \frac{\delta}{2}$ and following the same argument as in the proof of Cor. D.1, we have, with probability at least $1-\delta$,

$$\mathrm{MMD}_{\boldsymbol{k}}(\mathbb{S}_n, \mathbb{S}_{\text{out}}^{(i)}) \leq a'_n \left(1 + \sqrt{\log\frac{2}{\delta}}\right).$$

Plugging in the expression of a'_n from Lem. F.4 gives the claimed bound. □

F.3. Proof of Thm. 5: MMD guarantee for LSKT

First of all, the claimed runtime follows from the runtime of LD (Thm. 4), the $O(d_{\boldsymbol{k}_\mathbb{P}} 4^{\mathfrak{g}} n \log n)$ runtime of Compress++, and the $O(d_{\boldsymbol{k}_\mathbb{P}} n^{1.5})$ runtime of KT-Swap.

Without loss of generality assume $n \in 4^{\mathbb{N}}$. Fix $\delta \in (0,1)$. Let $w^\diamond$ denote the output of LD, and w^{sr} denote the output of Resample, both regarded as random variables. By Thm. 4, we have, with probability at least $1-\frac{\delta}{3}$,

$$\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w^\diamond}, \mathbb{P}) = \mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w_{\text{OPT}}}, \mathbb{P}) + O\left(\sqrt{\frac{nH_{n,r}}{\delta}}\right) + O\left(\sqrt{\frac{\|\boldsymbol{k}_\mathbb{P}\|_n \max(\log n, 1/\delta)}{n}}\right). \tag{57}$$

By Prop. E.1(c) with $\boldsymbol{k} = \boldsymbol{k}_\mathbb{P}$, we have the upper bound

$$\mathbb{E}\left[\mathrm{MMD}^2_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w^{\text{sr}}}, \mathbb{S}_n^{w^\diamond})\right] = \mathbb{E}\left[\mathbb{E}\left[\mathrm{MMD}^2_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w^{\text{sr}}}, \mathbb{S}_n^{w^\diamond}) \middle| w^\diamond\right]\right] \leq \frac{\|\boldsymbol{k}_\mathbb{P}\|_n}{n}.$$

Thus, by Markov's inequality,

$$\Pr\left(\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w^{\text{sr}}}, \mathbb{S}_n^{w^\diamond}) \geq \sqrt{\frac{3\|\boldsymbol{k}_\mathbb{P}\|_n}{n\delta}}\right) \leq \frac{\delta}{3}.$$

Hence, with probability at least $1-\frac{\delta}{3}$, we have

$$\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w^{\text{sr}}}, \mathbb{S}_n^{w^\diamond}) \leq \sqrt{\frac{3\|\boldsymbol{k}_\mathbb{P}\|_n}{n\delta}}. \tag{58}$$

Let $\mathbb{S}_{\text{out}}^{(i)}$ denote the i-th coreset output by THIN in KT-Compress++ (Alg. F.3). By Cor. F.1 with $\boldsymbol{k} = \boldsymbol{k}_\mathbb{P}$, we have, with probability at least $1-\frac{\delta}{3}$,

$$\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w^{\text{sr}}}, \mathbb{S}_{\text{out}}^{(i)}) = O\left(\sqrt{\frac{\|\boldsymbol{k}_\mathbb{P}\|_n \log n \log\left(e\mathcal{N}_{\boldsymbol{k}_\mathbb{P}}(\mathcal{B}_2(R_n), n^{-1/2})\right)}{n}\log\frac{e}{\delta}}\right).$$

Since KT-Swap can never increase $\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\cdot, \mathbb{P})$, we have, by the triangle inequality,

$$\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w_{\text{LSKT}}}, \mathbb{P}) \leq \mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_{\text{out}}^{(1)}, \mathbb{P}) \leq \mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_{\text{out}}^{(1)}, \mathbb{S}_n^{w^{\text{sr}}}) + \mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w^{\text{sr}}}, \mathbb{S}_n^{w^\diamond}) + \mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w^\diamond}, \mathbb{P}). \tag{59}$$

By the union bound, with probability at least $1-\delta$, the bounds (57), (58), (59) hold, so that the claim is shown by adding together the right-hand sides of these bounds and applying Assum. $(\boldsymbol{\alpha}, \boldsymbol{\beta})$-kernel. □

G. Simplex-Weighted Debiased Compression

In this section, we provide deferred analyses for RT and SR/LSR, as well as the algorithmic details of Recombination (Alg. G.1) and KT-Swap-LS (Alg. G.2).

Algorithm G.1 Recombination (rephrasing of Tchernychova (2016, Alg. 1) that takes $O(m^3 \log n)$ time)

Input: matrix $A \in \mathbb{R}^{m\times n}$ with $m < n$ and one row of A all positive, a nonnegative vector $x_0 \in \mathbb{R}^n_{\geq 0}$.
function FindBFS(A, x_0)
 $\triangleright$ *The requirement of A and x_0 are the same as the input. This subroutine takes $O(n^3)$ time.*
 $x \leftarrow x_0$
 $U, S, V^\top \leftarrow$ SVD(A) $\triangleright$ *any $O(n^3)$-time SVD algorithm that gives $USV^\top = A$*
 $V \leftarrow (V^\top)_{m+1:n}$ $\triangleright$ *$V \in \mathbb{R}^{(n-m)\times n}$ so that the null space of A is spanned by the rows of V*
 for $i = 1$ **to** $n - m$ **do**
  $v \leftarrow V_i$
  $k \leftarrow \operatorname{argmin}_{j\in[n]:v_j>0} \frac{x_j}{v_j}$ $\triangleright$ *This must succeed because $Av = 0$ and A has an all-positive row, so one of the coordinates of v must be positive.*
  $x \leftarrow x - \frac{x_k}{v_k} v$ $\triangleright$ *This zeros out the k-th coordinate of x while still ensuring x is nonnegative.*
  for $j = i + 1$ **to** $n - m$ **do**
   $V_j \leftarrow V_j - \frac{V_{j,k}}{v_k} v$ $\triangleright$ *$\{V_j\}_{j=i+1}^{n-m}$ remain independent and have 0 on the k-th coordinate.*
  end for
 end for
 return: $x \in \mathbb{R}^n_{\geq 0}$ such that $Ax = Ax_0$ and $\|x\|_0 \leq m$.
end function
$x \leftarrow x_0$
while $\|x\|_0 > 2m$ **do**
 Divide $\{i \in [n] : x_i > 0\}$ into $2m$ index blocks $\mathtt{I}_1, \ldots, \mathtt{I}_{2m}$, each of size at most $\left\lfloor \frac{\|x\|_0}{2m} \right\rfloor$.
 $A_i \leftarrow A_{:,\mathtt{I}_i} x_{\mathtt{I}_i} \in \mathbb{R}^m, \forall i \in [2m]$
 Form $\hat{A}$ to be the $m \times 2m$ matrix with columns A_i $\triangleright$ *Hence, one row of A contains all positive entries.*
 $\hat{x} \leftarrow$ FindFBS$(\hat{A}, \mathbf{1}_{2m})$ $\triangleright$ *$\|\hat{x}\|_0 \leq n$ and $\hat{A}\hat{x} = \sum A_i \hat{x}_i = \sum A_i = \sum A_{:,\mathtt{I}_i} x_{\mathtt{I}_i} = Ax$.*
 for $i = 1$ **to** $2m$ **do**
  $x_{\mathtt{I}_i} \leftarrow \hat{x}_i \cdot x_{\mathtt{I}_i}$ **if** $\hat{x}_i > 0$ **else** 0
 end for
 $\triangleright$ *After the update, the support of x shrinks by 2 while it maintains that $Ax = Ax_0$.*
end while
if $\|x\|_0 \geq m + 1$ **then**
 $\mathtt{I} \leftarrow \{i \in [n] : x_i > 0\}$
 $x_{\mathtt{I}} = $ FindBFS$(A_{:,\mathtt{I}}, x_{\mathtt{I}})$
end if
Return: $x \in \mathbb{R}^n_{\geq 0}$ such that $Ax = Ax_0$ and $\|x\|_0 \leq m$.

G.1. MMD guarantee for RT

Proposition G.1 (RT guarantee)**.** *Under Assums. 1 and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$-kernel, given $w \in \Delta_{n-1}$ and that $m \geq (\frac{\mathfrak{C}_d R_n^\beta + 1}{\sqrt{\log 2}} + \sqrt{\log 2})^2 - \frac{1}{\log 2} + 1$, RecombinationThinning (Alg. 4) outputs $w_{\mathrm{RT}} \in \Delta_{n-1}$ with $\|w_{\mathrm{RT}}\|_0 \leq m$ in $O((d_{\boldsymbol{k}_\mathbb{P}} + m)nm + m^3 \log n)$ time such that with probability at least $1 - \delta$,*

$$\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w_{\mathrm{RT}}}, \mathbb{P}) \leq \mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w}, \mathbb{P}) + \sqrt{\tfrac{2\|\boldsymbol{k}_\mathbb{P}\|_n}{n\delta}} + \sqrt{\tfrac{2nH_{n,m-1}}{\delta}}, \tag{60}$$

where $H_{n,r}$ is defined in (46).

Proof of Prop. G.1. The runtime follows from the $O((d_{\boldsymbol{k}_\mathbb{P}} + m)nm)$ runtime of WeightedRPCholesky, the $O(d_{\boldsymbol{k}_\mathbb{P}} nm)$ runtime of KT-Swap-LS, and the $O(m^3 \log n)$ runtime of Recombination (Tchernychova, 2016) which dominates the $O(m^3)$ weight optimization step.

Recall $w' \in \Delta_{n-1}$ from RT. The formation of F in Alg. 4 is identical to the formation of F (with $r = m - 1$) in Alg. 2 for $q > 1$. Thus by (51) with $w = w'$, $K = \boldsymbol{k}_\mathbb{P}(\mathcal{S}_n, \mathcal{S}_n)$,

$$w'^\top K w' \leq w'^\top F F^\top w' + n \operatorname{tr}\big((K - FF^\top)^{\tilde{w}}\big),$$

Algorithm G.2 KT-Swap with Linear Search (KT-Swap-LS)

Input: kernel $\boldsymbol{k}_{\mathbb{P}}$ with zero-mean under $\mathbb{P}$, input points $\mathcal{S}_n = (x_i)_{i\in[n]}$, weights $w \in \Delta_{n-1}$, $\texttt{fmt} \in \{\text{SPLX}, \text{CP}\}$
$\texttt{S} \leftarrow \{i \in [n] : w_i \neq 0\}$
▷ *Maintain two sufficient statistics:* $g = Kw$ *and* $D = w^\top Kw$.
function Add(g, D, i, t)
  $g \leftarrow g + t\boldsymbol{k}_{\mathbb{P}}(\mathcal{S}_n, x_i)$
  $D \leftarrow D + 2tg_i + t^2\boldsymbol{k}_{\mathbb{P}}(x_i, x_i)$
  return: (g, D)
end function
function Scale(g, D, α)
  $g \leftarrow \alpha g$
  $D \leftarrow \alpha^2 D$
  return: (g, D)
end function
$\texttt{Kdiag} \leftarrow \boldsymbol{k}_{\mathbb{P}}(\mathcal{S}_n, \mathcal{S}_n)$
$g \leftarrow \mathbf{0} \in \mathbb{R}^n$
$D \leftarrow 0$
for i **in** $\texttt{S}$ **do**
  $(g, D) \leftarrow \text{Add}(g, D, i, w_i)$
end for
for i **in** $\texttt{S}$ **do**
  if $w_i = 1$ **then continue**; ▷ *We cannot swap* i *out if* $\sum_{j\neq i} w_j = 0$*!*
  ▷ *First zero out* w_i.
  $(g, D) \leftarrow \text{Add}(g, D, i, -w_i)$
  $(g, D) \leftarrow \text{Scale}(g, D, \frac{1}{1-w_i})$
  $w_i = 0$
  ▷ *Next perform line search to add back a point.*
  $\alpha = (D - g)./(D - 2g + \texttt{Kdiag})$; ▷ $\alpha_i = \operatorname{argmin}_t \text{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{te_i+(1-t)w}, \mathbb{P}) = \operatorname{argmin}_t (1-t)^2 D + 2t(1-t)g + t^2 \mathit{Kdiag}$
  if $\texttt{fmt} = \text{SPLX}$ **then**
    $\alpha = \texttt{clip}(\alpha, 0, 1)$; ▷ *Clipping* α *to* $[0, 1]$*. This corresponds to* $\operatorname{argmin}_{t\in[0,1]} \text{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{te_i+(1-t)w}, \mathbb{P})$.
  end if
  $D' \leftarrow (1-\alpha)^2 D + 2\alpha(1-\alpha)g + \alpha^2\texttt{Kdiag}$ ▷ *multiplications are element-wise*
  $k \leftarrow \operatorname{argmin}_i D'_i$
  $(g, D) \leftarrow \text{Scale}(g, D, 1 - \alpha_k)$
  $(g, D) \leftarrow \text{Add}(g, D, k, \alpha_k)$
end for
Return: $w \in \Delta_{n-1}$

where $K = \boldsymbol{k}_{\mathbb{P}}(\mathcal{S}_n, \mathcal{S}_n)$. By construction of w' using Recombination, we have $F^\top \tilde{w} = F^\top w'$. Since $K \succeq FF^\top$, we have

$$w'^\top K w' \leq \tilde{w}^\top FF^\top \tilde{w} + n\,\text{tr}\big((K - FF^\top)^{\tilde{w}}\big) \leq \tilde{w}^\top K\tilde{w} + n\,\text{tr}\big((K - FF^\top)^{\tilde{w}}\big).$$

We recognize the right-hand side is precisely the right-hand side of (53) aside from having a multiplier of n instead of $n + 1$ in front of the trace and that F is rank $m - 1$. Now applying (54) with $Q = \frac{1}{2}$, $w^{(q)} = w'$, $w^{(q-1)} = w$, $r = m - 1$, and noticing that KT-Swap-LS and the quadratic-programming solve at the end cannot decrease the objective, we obtain (60) with probability at least $1 - \delta$. Note that the lower bound of m in Assum. $(\boldsymbol{\alpha},\boldsymbol{\beta})$-params makes $r = m - 1$ satisfy the lower bound for r in Prop. F.1. □

G.2. Proof of Thm. 6: MMD guarantee for SR/ LSR

The claimed runtime follows from the runtime of SteinThinning (Alg. D.1) or LD (Thm. 4) plus the runtime of RT (Prop. G.1).

Note the lower bound for m in Assum. ($\boldsymbol{\alpha}$,$\boldsymbol{\beta}$)-params implies the lower bound condition in Prop. G.1. For the case of SR, we proceed as in the proof of Thm. 3 and use Prop. G.1. For the case of LSR, we proceed as in the proof of Thm. 5 and use Thm. 4 and Prop. G.1. $\square$

H. Constant-Preserving Debiased Compression

In this section, we provide deferred analyses for CT and SC/LSC.

H.1. MMD guarantee for CT

Proposition H.1 (CT guarantee). *Under Assums. 1 and ($\boldsymbol{\alpha}, \boldsymbol{\beta}$)-kernel, given $w \in \Delta_{n-1}$ and $m \geq (\frac{\mathfrak{C}_d R_n^\beta + 1}{\sqrt{\log 2}} + \frac{2}{\sqrt{\log 2}})^2 - \frac{1}{\log 2}$, CT outputs $w_{\mathrm{CT}} \in \mathbb{R}^n$ with $\mathbf{1}_n^\top w_{\mathrm{CT}} = 1$ and $\|w_{\mathrm{CT}}\|_0 \leq m$ in $O((d_{\boldsymbol{k}_\mathbb{P}} + m)nm + m^3)$ time such that, for any $\delta \in (0, 1)$, with probability $1 - \delta$,*

$$\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w_{\mathrm{CT}}}, \mathbb{P}) \leq 2\,\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w}, \mathbb{P}) + \sqrt{\frac{4H_{n,m'}}{\delta}},$$

where $H_{n,m}$ is defined in (46) and $m' \triangleq m + \log 2 - 2\sqrt{m \log 2 + 1}$.

Proof of Prop. H.1. The runtime follows from the $O((d_{\boldsymbol{k}_\mathbb{P}} + m)nm)$ runtime of WeightedRPCholesky, the $O(nm)$ runtime of KT-Swap-LS, and the $O(m^3)$ runtime of matrix inversion in solving the two minimization problems using (64).

To improve the clarity of notation, we will use $w^\diamond$ to denote the input weight w to CT. For index sequences $\mathtt{I}, \mathtt{J} \subset [n]$ and a kernel $\boldsymbol{k}$, we use $\boldsymbol{k}(\mathtt{I}, \mathtt{J})$ to indicate the matrix $\boldsymbol{k}(\mathcal{S}_n[\mathtt{I}], \mathcal{S}_n[\mathtt{J}]) = [\boldsymbol{k}(x_i, x_j)]_{i \in \mathtt{I}, j \in \mathtt{J}}$, and similarly for a function $f : \mathbb{R}^n \to \mathbb{R}$, we use $f(\mathtt{I})$ to denote the vector $(f(x_i))_{i \in \mathtt{I}}$.

Recall the regularized kernel is $\boldsymbol{k}_c \triangleq \boldsymbol{k}_\mathbb{P} + c$. Suppose for now that $c > 0$ is an arbitrary constant. Let $\mathtt{I}$ denote the indices output by WeightedRPCholesky in CT. Let

$$w^c \triangleq \mathrm{argmin}_{w: \mathrm{supp}(w) \subset \mathtt{I}}\, \mathrm{MMD}^2_{\boldsymbol{k}_c}(\mathbb{S}_n^w, \mathbb{S}_n^{w^\diamond}).$$

Note that w^c is not a probability vector and may not sum to 1.

Step 1. Bound $\mathrm{MMD}^2_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^c}, \mathbb{S}_n^{w^\diamond})$ in terms of WeightedRPCholesky approximation error

We start by using an argument similar to that of Epperly and Moreno (2024, Prop. 3) to exploit the optimality condition of w^c. Since

$$\mathrm{argmin}_{w: \mathrm{supp}(w) \subset \mathtt{I}}\, \mathrm{MMD}^2_{\boldsymbol{k}_c}(\mathbb{S}_n^w, \mathbb{S}_n^{w^\diamond}) = \mathrm{argmin}_{w: \mathrm{supp}(w) \subset \mathtt{I}}\, w_\mathtt{I}^\top \boldsymbol{k}_c(\mathtt{I}, \mathtt{I}) w_\mathtt{I} - 2 w^{\diamond\top} \boldsymbol{k}_c(\mathcal{S}_n, \mathtt{I}) w_\mathtt{I},$$

by optimality, w^c satisfies,

$$\boldsymbol{k}_c(\mathtt{I}, \mathtt{I}) w^c_\mathtt{I} = \mathbb{S}_n^{w^\diamond} \boldsymbol{k}_c(\mathtt{I}).$$

We comment that the index sequence $\mathtt{I}$ returned by WeightedRPCholesky makes $\boldsymbol{k}_c(\mathtt{I}, \mathtt{I})$ invertible with probability 1: by the Guttman rank additivity formula of Schur complement (Zhang, 2006, Eq. (6.0.4)), each iteration of WeightedRPCholesky chooses a pivot with a non-zero diagonal and thus increases the rank of the low-rank approximation matrix, which is spanned by the columns of pivots, by 1. Hence

$$\begin{aligned}\mathbb{S}_n^{w^c} \boldsymbol{k}_c(\cdot) = \boldsymbol{k}_c(\cdot, \mathcal{S}_n) w^c = \boldsymbol{k}_c(\cdot, \mathtt{I}) w^c_\mathtt{I} = \boldsymbol{k}_c(\cdot, \mathtt{I}) \boldsymbol{k}_c(\mathtt{I}, \mathtt{I})^{-1} \boldsymbol{k}_c(\mathtt{I}, \mathtt{I}) w^c_\mathtt{I} \\ = \boldsymbol{k}_c(\cdot, \mathtt{I}) \boldsymbol{k}_c(\mathtt{I}, \mathtt{I})^{-1} \mathbb{S}_n^{w^\diamond} \boldsymbol{k}_c(\mathtt{I}) = \mathbb{S}_n^{w^\diamond} \boldsymbol{k}_{c\mathtt{I}}(\cdot),\end{aligned}$$

where $\boldsymbol{k}_{c\mathtt{I}}(x, y) \triangleq \boldsymbol{k}_c(x, \mathtt{I}) \boldsymbol{k}_c(\mathtt{I}, \mathtt{I})^{-1} \boldsymbol{k}_c(\mathtt{I}, y)$. Then

$$\mathrm{MMD}^2_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^c}, \mathbb{S}_n^{w^\diamond}) = \left\|\mathbb{S}_n^{w^\diamond} \boldsymbol{k}_c - \mathbb{S}_n^{w^c} \boldsymbol{k}_c\right\|^2_{\boldsymbol{k}_c} = \left\|\mathbb{S}_n^{w^\diamond} \boldsymbol{k}_c - \mathbb{S}_n^{w^\diamond} \boldsymbol{k}_{c\mathtt{I}}\right\|^2_{\boldsymbol{k}_c} = w^{\diamond\top} (\boldsymbol{k}_c - \boldsymbol{k}_{c\mathtt{I}})(\mathcal{S}_n, \mathcal{S}_n) w^\diamond.$$

Recall the index set $\mathtt{I}$ consists of the m pivots selected by WeightedRPCholesky on the input matrix

$$K_c^\diamond \triangleq \boldsymbol{k}_c(\mathcal{S}_n, \mathcal{S}_n)^{w^\diamond}.$$

Define

$$\widehat{K}_c^\diamond \triangleq \boldsymbol{k}_{c\mathtt{I}}(\mathcal{S}_n,\mathcal{S}_n)^{w^\diamond}.$$

Thus, by Lem. F.1,

$$\mathrm{MMD}^2_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^c},\mathbb{S}_n^{w^\diamond}) = w^{\diamond\top}(\boldsymbol{k}_c-\boldsymbol{k}_{c\mathtt{I}})(\mathcal{S}_n,\mathcal{S}_n)w^\diamond = \sqrt{w^\diamond}^\top(K_c^\diamond-\widehat{K}_c^\diamond)\sqrt{w^\diamond} \le \lambda_1(K_c^\diamond-\widehat{K}_c^\diamond) \le \mathrm{tr}\Big(K_c^\diamond-\widehat{K}_c^\diamond\Big).$$

Step 2. Bound $\mathrm{tr}\Big(K_c^\diamond-\widehat{K}_c^\diamond\Big)$ using the trace bound of the unregularized kernel

Let $[\![A]\!]_r$ denote the best rank-r approximation of an SPSD matrix $A\in\mathbb{R}^{n\times n}$ in the sense that

$$[\![A]\!]_r \triangleq \mathrm{argmin}_{\substack{X\in\mathbb{R}^{n\times n}\\ X=X^\top\\ A\succeq X\succeq 0\\ \mathrm{rank}(X)\le r}} \ \mathrm{tr}(A-X). \tag{61}$$

By the Eckart-Young-Mirsky theorem applied to symmetric matrices (Dax et al., 2014, Theorem 19), the solution to (61) is given by r-truncated eigenvalue decomposition of A, so that

$$\mathrm{tr}(A-[\![A]\!]_r) = \sum_{\ell=r+1}^n \lambda_\ell(A).$$

Let $q\triangleq\mathfrak{U}(m)$ where $\mathfrak{U}$ is defined in (47), so that by Chen et al. (2022, Thm. 3.1) with $\epsilon=1$, we have

$$\mathbb{E}\Big[\mathrm{tr}\Big(K_c^\diamond-\widehat{K}_c^\diamond\Big)\Big] \le 2\,\mathrm{tr}\Big(K_c^\diamond-[\![K_c^\diamond]\!]_q\Big).$$

We know one specific rank-q approximation of $K_c^\diamond$:

$$\widetilde{K}_c^\diamond \triangleq [\![K^\diamond]\!]_{q-1} + \mathrm{diag}(\sqrt{w^\diamond})c\mathbf{1}_n\mathbf{1}_n^\top\,\mathrm{diag}(\sqrt{w^\diamond}),$$

which satisfies

$$K_c^\diamond-\widetilde{K}_c^\diamond = K^\diamond + \mathrm{diag}(\sqrt{w^\diamond})c\mathbf{1}_n\mathbf{1}_n^\top\,\mathrm{diag}(\sqrt{w^\diamond}) - \widetilde{K}_c^\diamond = K^\diamond - [\![K^\diamond]\!]_{q-1}.$$

Thus by the variational definition in (61), we have

$$\mathrm{tr}\Big(K_c^\diamond-[\![K_c^\diamond]\!]_q\Big) \le \mathrm{tr}\Big(K_c^\diamond-\widetilde{K}_c^\diamond\Big) = \mathrm{tr}\Big(K^\diamond-[\![K^\diamond]\!]_{q-1}\Big) = \sum_{\ell=q}^n \lambda_\ell(K^\diamond).$$

Note the last bound does not depend on c. The tail sum of eigenvalues in the last expression is the same (up to a constant multiplier) as the one in (48) except for an off-by-1 difference in the summation index. A simple calculation shows that for $m'\triangleq m+\log 2-2\sqrt{m\log 2+1}$, we have $\mathfrak{U}(m')=\mathfrak{U}(m)-1$. Another simple calculation shows that $m\ge(\frac{\mathfrak{C}_d R_n^\beta+1}{\sqrt{\log 2}}+\frac{2}{\sqrt{\log 2}})^2-\frac{1}{\log 2}$ implies that m' satisfies the lower bound requirement of r in Prop. F.1. Thus, arguing as in the proof that follows (48), we get

$$\mathbb{E}\Big[\mathrm{tr}\Big(K_c^\diamond-\widehat{K}_c^\diamond\Big)\Big] \le H_{n,m'}.$$

Thus so far we have shown

$$\mathbb{E}[\mathrm{MMD}^2_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^c},\mathbb{S}_n^{w^\diamond})] \le \mathbb{E}\Big[\mathrm{tr}\Big(K_c^\diamond-\widehat{K}_c^\diamond\Big)\Big] \le H_{n,m'}.$$

By Markov's inequality, with probability at least $1-\delta$, we have

$$\mathrm{MMD}_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^c},\mathbb{S}_n^{w^\diamond}) \le \sqrt{\tfrac{H_{n,m'}}{\delta}}.$$

Recall that $\mathrm{MMD}_{\boldsymbol{k}}(\mu,\nu)=\|(\mu-\nu)\boldsymbol{k}\|_{\boldsymbol{k}}$ for signed measures μ,ν. By the triangle inequality, we have

$$\begin{aligned}\mathrm{MMD}_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^c},\mathbb{P}) &\le \mathrm{MMD}_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^c},\mathbb{S}_n^{w^\diamond}) + \mathrm{MMD}_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^\diamond},\mathbb{P})\\ &= \mathrm{MMD}_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^c},\mathbb{S}_n^{w^\diamond}) + \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^\diamond},\mathbb{P}),\end{aligned}$$

where we used that fact that $\sum_{i\in[n]} w_i^{\diamond} = 1$ to get the identity $\mathrm{MMD}_{\boldsymbol{k}_c}(\mathbb{S}^{w^\diamond}, \mathbb{P}) = \mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}^{w^\diamond}, \mathbb{P})$. Hence, with probability at least $1-\delta$,

$$\mathrm{MMD}_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^c}, \mathbb{P}) \le \mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w^\diamond}, \mathbb{P}) + \sqrt{\tfrac{H_{n,m'}}{\delta}}. \tag{62}$$

Step 3. Incorporating sum-to-one constraint

We now turn w^c into a constant-preserving weight while not inflating the MMD by much. Define

$$w^1 \triangleq \mathrm{argmin}_{w:\mathrm{supp}(w)\subset \mathtt{I}, \sum_{i\in[n]} w_i = 1} \mathrm{MMD}^2_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^w, \mathbb{P}). \tag{63}$$

Note w^1 is the weight right before KT-Swap-LS step in CT. Let $K_\mathtt{I} = \boldsymbol{k}_\mathbb{P}(\mathtt{I}, \mathtt{I})$. Let $\mathbf{1}_\mathtt{I}$ denote the $|\mathtt{I}|$-dimensional all-one veector. The Karush-Kuhn-Tucker condition (Ghojogh et al., 2021, Sec. 4.7) applied to (63) implies that, the solution w^1 is a stationary point of the Lagrangian function

$$L(w_\mathtt{I}, \lambda) \triangleq w_\mathtt{I}^\top K_\mathtt{I} w_\mathtt{I} + \lambda(\mathbf{1}_\mathtt{I}^\top w_\mathtt{I} - 1).$$

Then $\nabla_{w_\mathtt{I}} L(w_\mathtt{I}^1, \lambda) = 0$ implies $2K_\mathtt{I} w_\mathtt{I}^1 - \lambda \mathbf{1}_\mathtt{I} = 0$, so $w_\mathtt{I}^1 = \frac{\lambda K_\mathtt{I}^{-1}\mathbf{1}_\mathtt{I}}{2}$. The Lagrangian multiplier λ is determined by the constraint $\mathbf{1}_\mathtt{I}^\top w_\mathtt{I} = 1$, so we find

$$w_\mathtt{I}^1 = \tfrac{K_\mathtt{I}^{-1}\mathbf{1}_\mathtt{I}}{\mathbf{1}_\mathtt{I}^\top K_\mathtt{I}^{-1}\mathbf{1}_\mathtt{I}}. \tag{64}$$

Define

$$w^{c,\mathbb{P}} \triangleq \mathrm{argmin}_{w:\mathrm{supp}(w)\subset \mathtt{I}} \mathrm{MMD}^2_{\boldsymbol{k}_c}(\mathbb{S}_n^w, \mathbb{P}).$$

Since $w^{c,\mathbb{P}}$ is optimized to minimize $\mathrm{MMD}_{\boldsymbol{k}_c}$ to $\mathbb{P}$ on the same support as w^c, we have

$$\mathrm{MMD}_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^{c,\mathbb{P}}}, \mathbb{P}) \le \mathrm{MMD}_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^c}, \mathbb{P}).$$

The optimality condition for $w^{c,\mathbb{P}}$ is

$$(K_\mathtt{I} + c\mathbf{1}_\mathtt{I}\mathbf{1}_\mathtt{I}^\top)w - c\mathbf{1}_\mathtt{I} = 0,$$

and hence by the Sherman–Morrison formula,

$$w_\mathtt{I}^{c,\mathbb{P}} = (K_\mathtt{I} + c\mathbf{1}_\mathtt{I}\mathbf{1}_\mathtt{I}^\top)^{-1} c\mathbf{1}_\mathtt{I} = \left(K_\mathtt{I}^{-1} - \tfrac{cK_\mathtt{I}^{-1}\mathbf{1}_\mathtt{I}\mathbf{1}_\mathtt{I}^\top K_\mathtt{I}^{-1}}{1+c\mathbf{1}_\mathtt{I}^\top K_\mathtt{I}^{-1}\mathbf{1}_\mathtt{I}}\right) c\mathbf{1}_\mathtt{I} = \tfrac{K_\mathtt{I}^{-1}\mathbf{1}_\mathtt{I}}{1/c + \mathbf{1}_\mathtt{I}^\top K_\mathtt{I}^{-1}\mathbf{1}_\mathtt{I}}.$$

Let $\rho_c \triangleq \frac{\mathbf{1}_\mathtt{I}^\top K_\mathtt{I}^{-1}\mathbf{1}_\mathtt{I}}{1/c + \mathbf{1}_\mathtt{I}^\top K_\mathtt{I}^{-1}\mathbf{1}_\mathtt{I}}$, so that $w_\mathtt{I}^{c,\mathbb{P}} = \rho_c w_\mathtt{I}^1$. In particular, w^1 and $w^{c,\mathbb{P}}$ are scalar multiples of one another. To relate $\mathrm{MMD}_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w^1}, \mathbb{P})$ and $\mathrm{MMD}_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^{c,\mathbb{P}}}, \mathbb{P})$, note that

$$\begin{aligned}\mathrm{MMD}^2_{\boldsymbol{k}_\mathbb{P}}(\mathbb{S}_n^{w^1}, \mathbb{P}) &= {w_\mathtt{I}^1}^\top K_\mathtt{I} w_\mathtt{I}^1 = \tfrac{{w_\mathtt{I}^{c,\mathbb{P}}}^\top K_\mathtt{I} w_\mathtt{I}^{c,\mathbb{P}}}{\rho_c^2} = \tfrac{{w_\mathtt{I}^{c,\mathbb{P}}}^\top (K_\mathtt{I} + c\mathbf{1}_\mathtt{I}\mathbf{1}_\mathtt{I}^\top) w_\mathtt{I}^{c,\mathbb{P}} - c(\mathbf{1}_\mathtt{I}^\top w_\mathtt{I}^c)^2}{\rho_c^2} \\ &= \tfrac{\mathrm{MMD}^2_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^{c,\mathbb{P}}}, \mathbb{P}) + 2c\mathbf{1}_\mathtt{I}^\top w_\mathtt{I}^c - c - c(\mathbf{1}_\mathtt{I}^\top w_\mathtt{I}^c)^2}{\rho_c^2} = \tfrac{\mathrm{MMD}^2_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^{c,\mathbb{P}}}, \mathbb{P}) - c(\rho_c - 1)^2}{\rho_c^2}.\end{aligned}$$

So far the argument does not depend on any particular choice of $c > 0$. Let us now discuss how to choose c. Note that

$$\mathbf{1}_\mathtt{I}^\top K_\mathtt{I}^{-1}\mathbf{1}_\mathtt{I} = m \tfrac{\mathbf{1}_\mathtt{I}}{\sqrt{m}}^\top K_\mathtt{I}^{-1} \tfrac{\mathbf{1}_\mathtt{I}}{\sqrt{m}} \ge m\lambda_m(K_\mathtt{I}^{-1}) \ge \tfrac{m}{\lambda_1(K_\mathtt{I})} \ge \tfrac{m}{\mathrm{tr}(K_\mathtt{I})} \ge \tfrac{m}{\sum_{i\in[m]} \mathrm{diag}(K)_i^{\downarrow}},$$

where $\mathrm{diag}(K)^{\downarrow}$ denote the diagonal entries of $K = \boldsymbol{k}_\mathbb{P}(\mathcal{S}_n, \mathcal{S}_n)$ sorted in descending order. Thus

$$\rho_c = \tfrac{1}{\frac{1}{c\mathbf{1}_\mathtt{I}^\top K_\mathtt{I}^{-1}\mathbf{1}_\mathtt{I}} + 1} \ge \tfrac{1}{\frac{\sum_{i\in[m]} \mathrm{diag}(K)_i^{\downarrow}}{mc} + 1}.$$

Hence we can choose c to make sure ρ_c is bounded from below by a positive value. Recall in CT, we take

$$c = \frac{\sum_{i\in[m]} \mathrm{diag}(K)_i^{\downarrow}}{m},$$

so that $\rho_c \ge \frac{1}{2}$ and

$$\mathrm{MMD}^2_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^1}, \mathbb{P}) = \frac{\mathrm{MMD}^2_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^{c,\mathbb{P}}}, \mathbb{P}) - c(\rho_c - 1)^2}{\rho_c^2} \le 4\,\mathrm{MMD}^2_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^{c,\mathbb{P}}}, \mathbb{P}).$$

Hence by (62) and the fact that KT-Swap-LS and the final reweighting in CT only improves MMD, we have, with probability at least $1-\delta$,

$$\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w_{\mathrm{CT}}}, \mathbb{P}) \le \mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^1}, \mathbb{P}) \le 2\,\mathrm{MMD}_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^{c,\mathbb{P}}}, \mathbb{P}) \le 2\,\mathrm{MMD}_{\boldsymbol{k}_c}(\mathbb{S}_n^{w^{c}}, \mathbb{P}) \le 2\,\mathrm{MMD}_{\boldsymbol{k}_{\mathbb{P}}}(\mathbb{S}_n^{w^{\diamond}}, \mathbb{P}) + 2\sqrt{\frac{H_{n,m'}}{\delta}},$$

where we use (62) in the last inequality. □

H.2. Proof of Thm. 7: MMD guarantee for SC / LSC

The claimed runtime follows from the runtime of SteinThinning (Alg. D.1) or LD (Thm. 4) plus the runtime of CT (Prop. H.1).

Note the lower bound for m in Assum. $(\boldsymbol{\alpha},\boldsymbol{\beta})$-params implies the lower bound condition in Prop. H.1. For the case of SC, we proceed as in the proof of Thm. 3 and use Prop. H.1. For the case of LSC, we proceed as in the proof of Thm. 5 by invoking Thm. 4 and Prop. H.1. □

I. Implementation and Experimental Details

I.1. $O(d)$-time Stein kernel evaluation

In this section, we show that for $\mathcal{S}_n = (x_i)_{i\in[n]}$, each Stein kernel evaluation $\boldsymbol{k}_p(x_i, x_j)$ for a radially analytic base kernel (Def. B.3) can be done in $O(d)$ time after computing certain sufficient statistics in $O(nd^2 + d^3)$ time. Let $M \in \mathbb{R}^{d\times d}$ be a positive definite preconditioning matrix for $\boldsymbol{k}_p$. Let L be the Cholesky decomposition of M which can be done in $O(d^3)$ time so that $M = LL^\top$. From the expression (15), we can achieve $O(d)$ time evaluation if we can compute $\|x-y\|_M^2$ and $M\nabla \log p(x)$ in $O(d)$ time. For $M\nabla\log p(x)$, we can simply precompute $M\nabla \log p(x_i)$ for all $i \in [n]$. For $\|x-y\|_M^2$, we have

$$\|x-y\|_M^2 = (x-y)^\top M^{-1}(x-y) = (x-y)^\top (LL^\top)^{-1}(x-y) = \left\|L^{-1}(x-y)\right\|_2^2.$$

Hence it suffices to precompute $L^{-1}x_i$ for all $i \in [n]$, and we can precompute the inverse L^{-1} in $O(d^3)$ time.

I.2. Default parameters for algorithms

For LD, we always use $Q = 3$. To ensure that the guarantees of Lem. F.3 and Thm. 4 hold while achieving fast convergence in practice, we take the step size of AMD to be $1/(8\|\boldsymbol{k}_{\mathbb{P}}\|_n)$ in the first adaptive round and $1/(8\sum_{i\in[n]} w_i^{(q-1)} \boldsymbol{k}_{\mathbb{P}}(x_i, x_i))$ in subsequent adaptive rounds. We use $T = 7\sqrt{n_0}$ for AMD in all experiments.

We implemented our modified versions of KernelThinning and KT-Compress++ in JAX (Bradbury et al., 2018) so that certain subroutines can achieve a speedup using just-in-time compilation and the parallel computation power of GPUs. For Compress++, we use $\mathfrak{g} = 4$ in all experiments as in Shetty et al. (2022). For both KernelThinning and KT-Compress++, we use choose $\delta = 1/2$ as in the `goodpoints` library.

Each experiment was run with a single NVIDIA RTX 6000 GPU and an AMD EPYC 7513 32-Core CPU.

I.3. Correcting for burn-in details

We use the four MCMC chains provided by Riabiz et al. (2022) that include both the sample points and their scores. The reference chain used to compute the energy distance is the same one used in Riabiz et al. (2022) for the energy distance and was kindly provided by the authors.

In Tab. I.1, we collect the runtime for the burn-in correction experiments.

Fig. I.1, Fig. I.2, Fig. I.3, display the results of the burn-in correction experiment of Sec. 5 repeated with three other MCMC algorithms: MALA without preconditioning, random walk (RW), and adaptive random walk (ADA-RW). The results of P-MALA from Sec. 5 are also included for completeness. For all four chains, our methods reliably achieve better quality coresets when compared with the baseline methods.

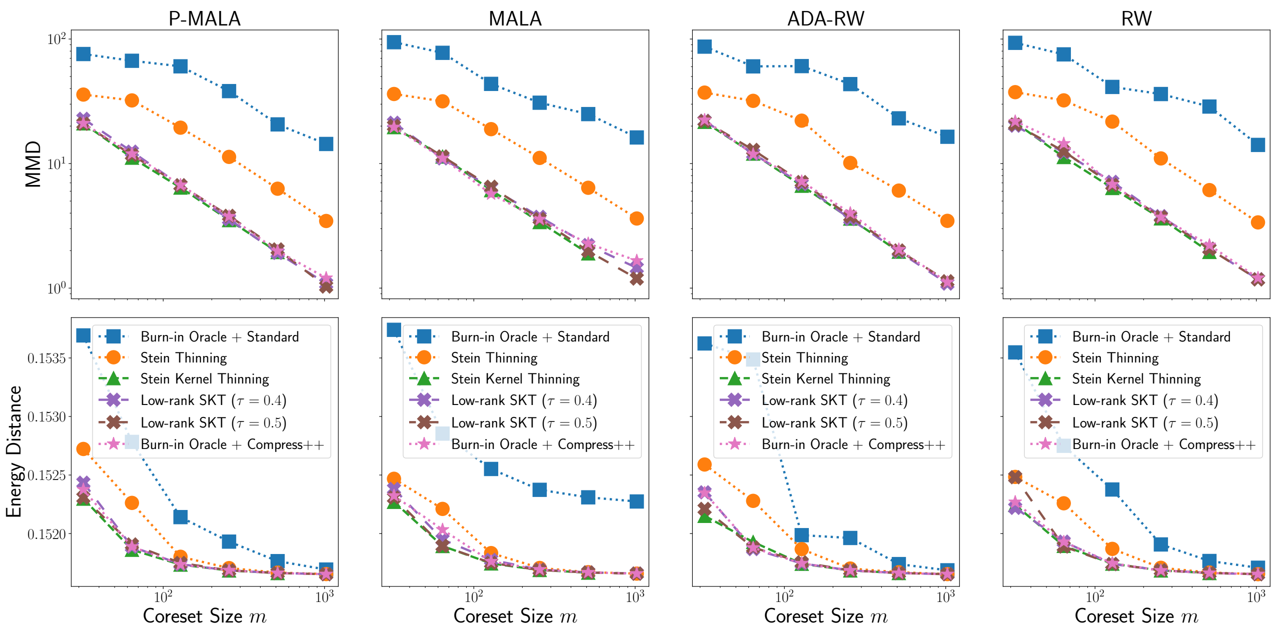

Figure I.1: **Correcting for burn-in with equal-weighted compression.** For each of four MCMC algorithms and using only one chain, our methods consistently outperform the Stein and standard thinning baselines and match the 6-chain oracle.

n_0	**ST**	**LD** (0.5)	**LD** (0.4)	**KT**	**KT-Compress++**	**RT** (0.5)	**RT** (0.4)	**CT** (0.5)	**CT** (0.4)
2^{14}	2.50	13.22	12.88	7.31	3.49	0.79	0.60	2.06	1.96
2^{16}	8.48	16.15	15.82	20.77	5.90	2.59	1.68	3.66	3.04
2^{18}	111.06	32.14	20.60	193.03	11.73	11.16	2.63	6.48	3.67
2^{20}	-	314.67	131.31	-	35.99	113.71	11.06	51.14	8.42

Table I.1: **Breakdown of runtime (in seconds)** for the burn-in correction experiment ($d = 4$) of Sec. 5. n_0 is the input size after standard thinning from the length $n = 2 \times 10^6$ chain (Rem. 2). Each runtime is the median of 3 runs. KT and KT-Compress++ output $m = \sqrt{n_0}$ equal-weighted points. RT and CT respectively output $m = n_0^\tau$ points with simplex or constant-preserving weights for τ shown in parentheses. In addition, LD, RT, and CT use the rank n_0^τ. ST and KT took longer than 30 minutes for $n_0 = 2^{20}$ and hence their numbers are not reported.

I.4. Correcting for approximate MCMC details

Surrogate ground truth Following Liu and Lee (2017), we took the first 10,000 data points and generated 2^{20} surrogate ground truth sample points using NUTS (Hoffman et al., 2014) for the evaluation. To generate the surrogate ground truth using NUTS, we used `numpyro` (Phan et al., 2019). It took 12 hours to generate the surrogate ground truth points using the GPU implementation, and we estimate it would have taken 200 hours using the CPU implementation.

SGFS For SGFS, we used batch size 32 and the step size schedule $\eta/(1+t)^{0.55}$ where t is the step count and η is the initial step size. We chose η from $\{10.0, 5.0, 1.0, 0.5, 0.1, 0.05, 0.01\}$, found $\eta = 1.0$ gave the best standard thinning

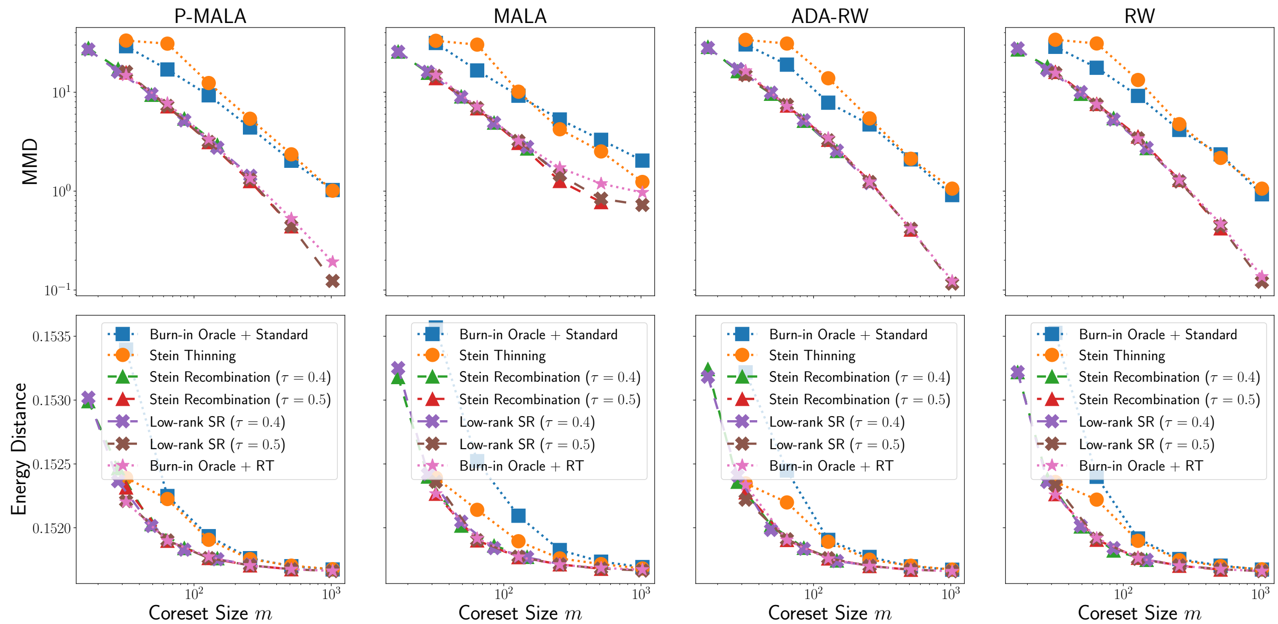

Figure I.2: **Correcting for burn-in with simplex-weighted compression.** For each of four MCMC algorithms and using only one chain, our methods consistently outperform the Stein and standard thinning baselines and match the 6-chain oracle.

MMD to get a coreset size of $m = 2^{10}$, and hence we fixed $\eta = 1.0$ in all experiments. We used the version of SGFS (Ahn et al., 2012, SGFS-f) that involves inversion of $d \times d$ matrices — we found the faster version (SGFS-d) that inverts only the diagonal resulted in significantly worse mixing. We implemented SGFS in `numpy` and ran it on the CPU.

Runtime The SGFS chain of length 2^{24} took approximately 2 hours to generate using the CPU. Remarkably, all of our low-rank methods finish within 10 minutes for $n_0 = 2^{20}$, which is orders of magnitude faster than the time taken to generate the NUTS surrogate ground truth.

Additional results In Fig. I.4, we plot the posterior mean mean-squared error (MSE) for each compression method in the approximate MCMC experiment of Sec. 5.

I.5. Correcting for tempering details

In the data release of Riabiz et al. (2020), we noticed there were 349 sample points for which the provided scores were NaNs, so we removed those points at the recommendation of the authors.

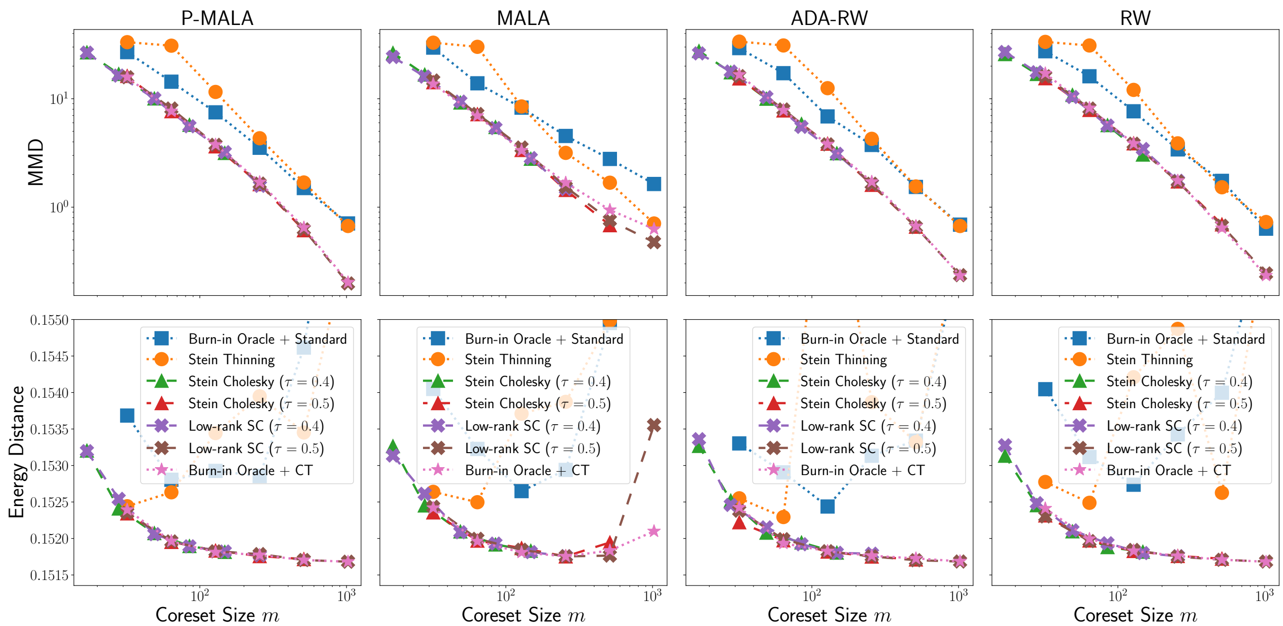

Figure I.3: **Correcting for burn-in with constant-preserving compression.** For each of four MCMC algorithms and using only one chain, our methods consistently outperform the Stein and standard thinning baselines and match the 6-chain oracle.

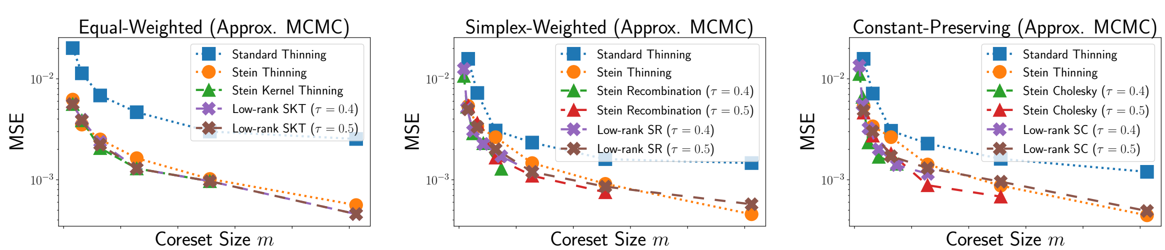

Figure I.4: **Posterior mean mean-squared error (MSE)** for the approximate MCMC compression experiment of Sec. 5. MSE is computed as $\|\hat{\mathbb{E}}_{\mathbb{P}} Z - \sum_{i \in [n_0]} w_i x_i\|_M^2 / d$ where $\hat{\mathbb{E}}_{\mathbb{P}} Z$ is the mean of the surrogate ground truth NUTS sample.