Optimal Quantisation of Probability Measures Using Maximum Mean Discrepancy

Onur Teymur
Newcastle University
Alan Turing Institute

Jackson Gorham
Whisper.ai, Inc.

Marina Riabiz
King's College London
Alan Turing Institute

Chris. J. Oates
Newcastle University
Alan Turing Institute

Abstract

Several researchers have proposed minimisation of maximum mean discrepancy (MMD) as a method to quantise probability measures, i.e., to approximate a distribution by a representative point set. We consider sequential algorithms that greedily minimise MMD over a discrete candidate set. We propose a novel non-myopic algorithm and, in order to both improve statistical efficiency and reduce computational cost, we investigate a variant that applies this technique to a mini-batch of the candidate set at each iteration. When the candidate points are sampled from the target, the consistency of these new algorithms—and their mini-batch variants—is established. We demonstrate the algorithms on a range of important computational problems, including optimisation of nodes in Bayesian cubature and the thinning of Markov chain output.

1 Introduction

This paper considers the approximation of a probability distribution μ, defined on a set $\mathcal{X}$, by a discrete distribution $\nu = \frac{1}{n}\sum_{i=1}^{n}\delta(x_i)$, for some representative points x_i, where $\delta(x)$ denotes a point mass located at $x \in \mathcal{X}$. This *quantisation* task arises in many areas including numerical cubature (Karvonen, 2019), experimental design (Chaloner and Verdinelli, 1995) and Bayesian computation (Riabiz et al., 2020). To solve the quantisation task one first identifies an optimality criterion, typically a notion of *discrepancy* between μ and ν, and then develops an algorithm to approximately minimise it. Classical optimal quantisation picks the x_i to minimise a Wasserstein distance between ν and μ, which leads to an elegant connection with Voronoi partitions whose centres are the x_i (Graf and Luschgy, 2007). Several other discrepancies exist but are less well-studied for the quantisation task. In this paper we study quantisation with *maximum mean discrepancy* (MMD), as well as a specific version called *kernel Stein discrepancy* (KSD), each of which admit a closed-form expression for a wide class of target distributions μ (e.g. Rustamov, 2019).

Despite several interesting results, optimal quantisation with MMD remains largely unsolved. Quasi Monte Carlo (QMC) provides representative point sets that asymptotically minimise MMD (Hickernell, 1998; Dick and Pillichshammer, 2010); however, these results are typically limited to specific instances of μ and MMD.[1] The use of greedy sequential algorithms, in which x_n is selected conditional on the $x_1, \ldots, x_{n-1}$ already chosen, has received some attention in the MMD context—see Santin and Haasdonk (2017) and the surveys in Oettershagen (2017) and Pronzato and Zhigljavsky (2018). Greedy sequential algorithms have also been proposed for KSD (Chen et al., 2018, 2019), as well as a non-greedy sequential algorithm for minimising MMD, called *kernel herding* (Chen et al., 2010).

In certain situations,[2] the greedy and herding algorithms produce the same sequence of points, with the latter theoretically understood due to its interpretation as a Frank–Wolfe algorithm (Bach et al., 2012; Lacoste-Julien et al., 2015). Outside the translation-invariant context, empirical studies have shown that greedy algorithms tend to outperform kernel herding (Chen et al., 2018). Information-theoretic lower bounds on MMD have been derived in the literature on information-based complexity (Novak and Woźniakowski, 2008) and in Mak et al. (2018), who

Proceedings of the 24th International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130. Copyright 2021 by the author(s).

[1] In Section 2.1 we explain how MMD is parametrised by a *kernel*; the QMC literature typically focuses on μ uniform on $[0,1]^d$, and d-dim tensor products of kernels over $[0,1]$.

[2] Specifically, the algorithms coincide when the kernel on which they are based is translation-invariant.

studied representative points that minimise an *energy distance*; the relationship between energy distances and MMD is clarified in Sejdinovic et al. (2013).

The aforementioned sequential algorithms require that, to select the next point x_n, one has to search over the whole set $\mathcal{X}$. This is often impractical, since $\mathcal{X}$ will typically be an infinite set and may not have useful structure (e.g. a vector space) that can be exploited by a numerical optimisation method.

Extended notions of greedy optimisation, where at each step one seeks to add a point that is merely 'sufficiently close' to optimal, were studied for KSD in Chen et al. (2018). Chen et al. (2019) proposed a stochastic optimisation approach for this purpose. However, the global non-convex optimisation problem that must be solved to find the next point x_n becomes increasingly difficult as more points are selected. This manifests, for example, in the increasing number of iterations required in the approach of Chen et al. (2019).

This paper studies sequential minimisation of MMD over a *finite* candidate set, instead of over the whole of $\mathcal{X}$. This obviates the need to use a numerical optimisation routine, requiring only that a suitable candidate set can be produced. Such an approach was recently described in Riabiz et al. (2020), where an algorithm termed *Stein thinning* was proposed for greedy minimisation of KSD. Discrete candidate sets in the context of kernel herding were discussed in Chen et al. (2010) and Lacoste-Julien et al. (2015), and in Paige et al. (2016) for the case that the subset is chosen from the support of a discrete target. Mak et al. (2018) proposed sequential selection from a discrete candidate set to approximately minimise energy distance, but theoretical analysis of this algorithm was not attempted.

The novel contributions of this paper are as follows:

- We study greedy algorithms for sequential minimisation of MMD, including novel non-myopic algorithms in which multiple points are selected simultaneously. These algorithms are also extended to allow for mini-batching of the candidate set. Consistency is established and a finite-sample-size error bound is provided.
- We show how non-myopic algorithms can be cast as integer quadratic programmes (IQP) that can be exactly solved using standard libraries.
- A detailed empirical assessment is presented, including a study varying the extent of non-myopic selection, up to and including the limiting case in which all points are selected simultaneously. Such non-sequential algorithms require high computational expenditure, and so a semi-definite relaxation of the IQP is considered in the supplement.

The remainder of the paper is structured thus. In Section 2 we provide background on MMD and KSD. In Section 3 our novel methods for optimal quantisation are presented. Our empirical assessment, including comparisons with existing methods, is in Section 4 and our theoretical assessment is in Section 5. The paper concludes with a discussion in Section 6.

2 Background

Let $\mathcal{X}$ be a measurable space and let $\mathcal{P}(\mathcal{X})$ denote the set of probability distributions on $\mathcal{X}$. First we introduce a notion of discrepancy between two measures $\mu, \nu \in \mathcal{P}(\mathcal{X})$, and then specialise this definition to MMD (Section 2.1) and KSD (Section 2.2).

For any $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and set $\mathcal{F}$ consisting of real-valued measurable functions on $\mathcal{X}$, we define a *discrepancy* to be a quantity of the form

$$D_{\mathcal{F}}(\mu, \nu) = \sup_{f \in \mathcal{F}} \left| \int f \,\mathrm{d}\mu - \int f \,\mathrm{d}\nu \right|, \qquad (1)$$

assuming $\mathcal{F}$ was chosen so that all integrals in (1) exist. The set $\mathcal{F}$ is called *measure-determining* if $D_{\mathcal{F}}(\mu, \nu) = 0$ implies $\mu = \nu$, and in this case $D_{\mathcal{F}}$ is called an *integral probability metric* (Müller, 1997). An example is the Wasserstein metric—induced by choosing $\mathcal{F}$ as the set of 1-Lipschitz functions defined on $\mathcal{X}$—that is used in classical quantisation (Dudley, 2018, Thm. 11.8.2). Next we describe how MMD and KSD are induced from specific choices of $\mathcal{F}$.

2.1 Maximum Mean Discrepancy

Consider a symmetric and positive-definite function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, which we call a *kernel*. A kernel *reproduces* a Hilbert space of functions $\mathcal{H}$ from $\mathcal{X} \to \mathbb{R}$ if (i) for all $x \in \mathcal{X}$ we have $k(\cdot, x) \in \mathcal{H}$, and (ii) for all $x \in \mathcal{X}$ and $f \in \mathcal{H}$ we have $\langle k(\cdot, x), f \rangle_{\mathcal{H}} = f(x)$, where $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ denotes the inner product in $\mathcal{H}$. By the Moore–Aronszajn theorem (Aronszajn, 1950), there is a one-to-one mapping between the kernel k and the *reproducing kernel Hilbert space* (RKHS) $\mathcal{H}$, which we make explicit by writing $\mathcal{H}(k)$. A prototypical example of a kernel on $\mathcal{X} \subseteq \mathbb{R}^d$ is the squared-exponential kernel $k(x, y; \ell) = \exp(-\frac{1}{2}\ell^{-2}\|x - y\|^2)$, where $\|\cdot\|$ in this paper denotes the Euclidean norm and $\ell > 0$ is a positive scaling constant.

Choosing the set $\mathcal{F}$ in (1) to be the unit ball $\mathcal{B}(k) := \{f \in \mathcal{H}(k) : \langle f, f \rangle_{\mathcal{H}(k)} \leq 1\}$ of the RKHS $\mathcal{H}(k)$ enables the supremum in (1) to be written in closed form and defines the MMD (Song, 2008):

$$\begin{aligned} \mathrm{MMD}_{\mu,k}(\nu)^2 &:= D_{\mathcal{B}(k)}(\mu, \nu) \\ &= \iint k(x, y) \,\mathrm{d}\nu(x) \,\mathrm{d}\nu(y) - 2 \iint k(x, y) \,\mathrm{d}\nu(x) \,\mathrm{d}\mu(y) \\ &\qquad + \iint k(x, y) \,\mathrm{d}\mu(x) \,\mathrm{d}\mu(y) \qquad (2) \end{aligned}$$

Our notation emphasises ν as the variable of interest, since in this paper we aim to minimise MMD over possible ν for a fixed kernel k and a fixed target μ. Under suitable conditions on k and $\mathcal{X}$ it can be shown that MMD is a metric on $\mathcal{P}(\mathcal{X})$ (in which case the kernel is called *characteristic*); for sufficient conditions see Section 3 of Sriperumbudur et al. (2010). Furthermore, under stronger conditions on k, MMD metrises the weak topology on $\mathcal{P}(\mathcal{X})$ (Sriperumbudur et al., 2010, Thms. 23, 24). This provides theoretical justification for minimisation of MMD: if $\mathrm{MMD}_{\mu,k}(\nu) \to 0$ then $\nu \Rightarrow \mu$, where $\Rightarrow$ denotes weak convergence in $\mathcal{P}(\mathcal{X})$.

Evaluation of MMD requires that μ and ν are either explicit or can be easily approximated (e.g. by sampling), so as to compute the integrals appearing in (2). This is the case in many applications and MMD has been widely used (Arbel et al., 2019; Briol et al., 2019a; Chérief-Abdellatif and Alquier, 2020). In cases where μ is not explicit, such as when it arises as an intractable posterior in a Bayesian context, KSD can be a useful specialisation of MMD that circumvents integration with respect to μ. We describe this next.

2.2 Kernel Stein Discrepancy

While originally proposed as a means of proving distributional convergence, Stein's method (Stein, 1972) can be used to circumvent the integration against μ required in (2) to calculate the MMD. Suppose we have an operator $\mathcal{A}_\mu$ defined on a set of functions $\mathcal{G}$ such that $\int \mathcal{A}_\mu g \,\mathrm{d}\mu = 0$ holds for all $g \in \mathcal{G}$. Choosing $\mathcal{F} = \mathcal{A}_\mu \mathcal{G} := \{\mathcal{A}_\mu g : g \in \mathcal{G}\}$ in (1), we would then have $D_{\mathcal{A}_\mu \mathcal{G}}(\mu,\nu) = \sup_{g\in\mathcal{G}} \left|\int \mathcal{A}_\mu g \,\mathrm{d}\nu\right|$, an expression which no longer involves integrals with respect to μ. Appropriate choices for $\mathcal{A}_\mu$ and $\mathcal{G}$ were studied in Gorham and Mackey (2015), who termed $D_{\mathcal{A}_\mu \mathcal{G}}$ the *Stein discrepancy*, and these will now be described.

Assume μ admits a positive and continuously differentiable density p_μ on $\mathcal{X} = \mathbb{R}^d$; let ∇ and $\nabla\cdot$ denote the gradient and divergence operators respectively; and let $k : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ be a kernel that is continuously differentiable in each argument. Then take

$$\mathcal{A}_\mu g := \nabla \cdot g + u_\mu \cdot g, \qquad u_\mu := \nabla \log p_\mu,$$
$$\mathcal{G} := \{g : \mathbb{R}^d \to \mathbb{R}^d : \textstyle\sum_{i=1}^d \langle g_i, g_i\rangle_{\mathcal{H}(k)} \le 1\}.$$

Note that $\mathcal{G}$ is the unit ball in the d-dimensional tensor product of $\mathcal{H}(k)$. Under mild conditions on k and μ (Gorham and Mackey, 2017, Prop. 1), it holds that $\int \mathcal{A}_\mu g \,\mathrm{d}\mu = 0$ for all $g \in \mathcal{G}$. The set $\mathcal{A}_\mu \mathcal{G}$ can then be shown (Oates et al., 2017) to be the unit ball $\mathcal{B}(k_\mu)$ in a different RKHS $\mathcal{H}(k_\mu)$ with reproducing kernel

$$\begin{aligned} k_\mu(x,y) := {} & \nabla_x \cdot \nabla_y k(x,y) + \nabla_x k(x,y) \cdot u_\mu(y) \\ & + \nabla_y k(x,y) \cdot u_\mu(x) + k(x,y) u_\mu(x) \cdot u_\mu(y), \end{aligned} \tag{3}$$

where subscripts are used to denote the argument upon which a differential operator acts. Since $k_\mu(x,\cdot) \in \mathcal{H}(k_\mu)$, it follows that $\int k_\mu(x,\cdot)\,\mathrm{d}\mu(x) = 0$ and from (2) we arrive at the *kernel Stein discrepancy* (KSD)

$$\mathrm{MMD}_{\mu,k_\mu}(\nu)^2 = \iint k_\mu(x,y)\,\mathrm{d}\nu(x)\,\mathrm{d}\nu(y).$$

Under stronger conditions on μ and k it can be shown that KSD controls weak convergence to μ in $\mathcal{P}(\mathbb{R}^d)$, meaning that if $\mathrm{MMD}_{\mu,k_\mu}(\nu) \to 0$ then $\nu \Rightarrow \mu$ (Gorham and Mackey, 2017, Thm. 8). The description of KSD here is limited to $\mathbb{R}^d$, but constructions also exist for discrete spaces (Yang et al., 2018) and more general Riemannian manifolds (Barp et al., 2018; Xu and Matsuda, 2020; Le et al., 2020). Extensions that use other operators $\mathcal{A}_\mu$ (Gorham et al., 2019; Barp et al., 2019) have also been studied.

3 Methods

In this section we propose novel algorithms for minimisation of MMD over a finite candidate set. The simplest algorithm is described in Section 3.1, and from this we generalise to consider both non-myopic selection of representative points and mini-batching in Section 3.2. A discussion of non-sequential algorithms, as the limit of non-myopic algorithms where all points are simultaneously selected, is given in Section 3.3.

3.1 A Simple Algorithm for Quantisation

In what follows we assume that we are provided with a finite candidate set $\{x_i\}_{i=1}^n \subset \mathcal{X}$ from which representative points are to be selected. Ideally, these candidates should be in regions where μ is supported, but we defer making any assumptions on this set until the theoretical analysis in Section 5. The simplest algorithm that we consider greedily minimises MMD over the candidate set; for each i, pick

$$\pi(i) \in \operatorname*{argmin}_{j\in\{1,\dots,n\}} \mathrm{MMD}_{\mu,k}\left(\tfrac{1}{i}\textstyle\sum_{i'=1}^{i-1} \delta(x_{\pi(i')}) + \tfrac{1}{i}\delta(x_j)\right),$$

to obtain, after m steps, an index sequence $\pi \in \{1,\dots,n\}^m$ and associated empirical distribution $\nu = \frac{1}{m}\sum_{i=1}^m \delta(x_{\pi(i)})$. (The convention $\sum_{i=1}^0 = 0$ is used.) Explicit formulae are contained in Algorithm 1. The computational complexity of selecting m points in this manner is $O(m^2 n)$, provided that the integrals appearing in Algorithm 1 can be evaluated in $O(1)$. Note that candidate points can be selected more than once.

Theorems 1, 2 and 4 in Section 5 provide novel finite-sample-size error bounds for Algorithm 1 (as a special case of Algorithm 2). The two main shortcomings of Algorithm 1 are that (i) the myopic nature of the optimisation may be *statistically inefficient*, and (ii) the

Algorithm 1: Myopic minimisation of MMD

```
Data: A set $\{x_i\}_{i=1}^n$, a distribution $\mu$, a kernel $k$ and a number $m \in \mathbb{N}$ of output points
Result: An index sequence $\pi \in \{1, \dots, n\}^m$
for $i = 1, \dots, m$ do
    $\pi(i) \in \underset{j \in \{1,\dots,n\}}{\operatorname{argmin}} \Big[ \frac{1}{2} k(x_j, x_j) + \sum_{i'=1}^{i-1} k(x_{\pi(i')}, x_j) - i \int k(x, x_j) \mathrm{d}\mu(x) \Big]$
end
```

requirement to scan through a large candidate set during each iteration may lead to unacceptable *computational cost*. In Section 3.2 we propose non-myopic and mini-batch extensions to address these issues.

3.2 Generalised Sequential Algorithms

In Section 3.2.1 we describe a non-myopic extension of Algorithm 1, where multiple points are simultaneously selected at each step. The use of non-myopic optimisation is impractical when a large candidate set is used, and therefore we explain how mini-batches from the candidate set can be employed in Section 3.2.2.

3.2.1 Non-Myopic Minimisation

Now we consider the simultaneous selection of $s > 1$ representative points from the candidate set at each step. This leads to the non-myopic algorithm

$$\pi(i, \cdot) \in \underset{S \in \{1,\dots,n\}^s}{\operatorname{argmin}} \operatorname{MMD}_{\mu,k}\Big(\tfrac{1}{is} \textstyle\sum_{i'=1}^{i-1} \sum_{j=1}^{s} \delta(x_{\pi(i',j)}) + \tfrac{1}{is} \sum_{j \in S} \delta(x_j) \Big),$$

whose output is a bivariate index $\pi \in \{1, \dots, n\}^{m \times s}$, together with the associated empirical distribution $\nu = \frac{1}{ms} \sum_{i=1}^{m} \sum_{j=1}^{s} \delta(x_{\pi(i,j)})$. Explicit formulae are contained in Algorithm 2. The computational complexity of selecting ms points in this manner is $O(m^2 s^2 n^s)$, which is larger than Algorithm 1 when $s > 1$. Theorems 1, 2 and 4 in Section 5 provide novel finite-sample-size error bounds for Algorithm 2.

Despite its daunting computational complexity, we have found that it is practical to exactly implement Algorithm 2 for moderate values of s and n by casting each iteration of the algorithm as an instance of a constrained *integer quadratic programme* (IQP) (e.g. Wolsey, 2020), so that state-of-the-art discrete optimisation methods can be employed. To this end, we represent the indices $S \subset \{1, \dots, n\}^s$ of the s points to be selected at iteration i as a vector $v \in \{0, \dots, s\}^n$ whose jth element indicates the number of copies of x_j that are selected, and where v is constrained to satisfy $\sum_{j=1}^{n} v_j = s$. It is then an algebraic exercise to recast an optimal subset $\pi(i, \cdot)$ as the solution to a constrained IQP:

Algorithm 2: Non-myopic minimisation of MMD

```
Data: A set $\{x_i\}_{i=1}^n$, a distribution $\mu$, a kernel $k$, a number of points to select per iteration $s \in \mathbb{N}$ and a total number of iterations $m \in \mathbb{N}$
Result: An index sequence $\pi \in \{1, \dots, n\}^{m \times s}$
for $i = 1, \dots, m$ do
    $\pi(i, \cdot) \in \underset{S \in \{1,\dots,n\}^s}{\operatorname{argmin}} \Big[ \frac{1}{2} \sum_{j,j' \in S} k(x_j, x_{j'}) + \sum_{i'=1}^{i-1} \sum_{j=1}^{s} \sum_{j' \in S} k(x_{\pi(i',j)}, x_{j'}) - is \sum_{j \in S} \int k(x, x_j) \mathrm{d}\mu(x) \Big]$
end
```

$$\underset{v \in \mathbb{N}_0^s}{\operatorname{argmin}} \ \tfrac{1}{2} v^\top K v + c^{i\top} v \quad \text{s.t.} \quad \mathbf{1}^\top v = s \tag{4}$$

$$K_{j,j'} := k(x_j, x_{j'}), \qquad \mathbf{1}_j := 1 \text{ for } j = 1, \dots, n,$$

$$c_j^i := \textstyle\sum_{i'=1}^{i-1} \sum_{j'=1}^{s} k(x_{\pi(i',j')}, x_j) - is \int k(x, x_j) \, \mathrm{d}\mu(x)$$

Remark 1. *If one further imposes the constraint $v_i \in \{0, 1\}$ for all i, so that each candidate may be selected at most once, then the resulting* binary quadratic programme *(BQP) is equivalent to the* cardinality constrained k-partition *problem from discrete optimisation, which is known to be NP-hard (Rendl, 2016). (The results we present do* not *impose this constraint.)*

3.2.2 Mini-Batching

The exact solution of (4) is practical only for moderate values of s and n. This motivates the idea of considering only a subset of the n candidates at each iteration, a procedure we call *mini-batching* and inspired by the similar idea from stochastic optimisation. There are several ways that mini-batching can be performed, but here we simply state that candidates denoted $\{x_j^i\}_{j=1}^b$ are considered during the ith iteration, with the mini-batch size denoted by $b \in \mathbb{N}$. The non-myopic algorithm for minimisation of MMD with mini-batching is then

$$\pi(i, \cdot) \in \underset{S \in \{1,\dots,b\}^s}{\operatorname{argmin}} \operatorname{MMD}_{\mu,k}\Big(\tfrac{1}{is} \textstyle\sum_{i'=1}^{i-1} \sum_{j=1}^{s} \delta(x_{\pi(i',j)}^{i'}) + \tfrac{1}{is} \sum_{j \in S} \delta(x_j^i) \Big)$$

Explicit formulae are contained in Algorithm 3. The complexity of selecting ms points in this manner is $O(m^2 s^2 b^s)$, which is smaller than Algorithm 2 when $b < n$. As with Algorithm 2, an exact IQP formulation can be employed. Theorem 3 provides a novel finite-sample-size error bound for Algorithm 3.

3.3 Non-Sequential Algorithms

Finally we consider the limit of the non-myopic Algorithm 2, in which all m representative points are simultaneously selected in a single step:

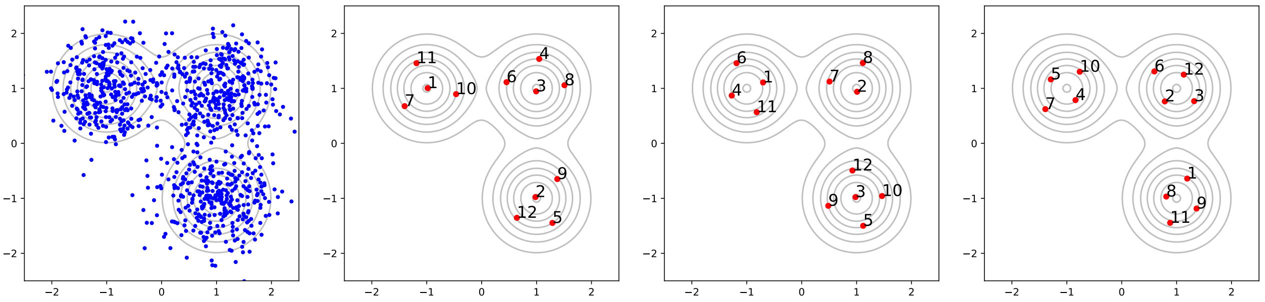

Figure 1: Quantisation of a Gaussian mixture model using MMD. A candidate set of 1000 independent samples (left), from which 12 representative points were selected using: the myopic method (centre-left); non-myopic selection, picking 4 points at a time (centre-right), and by simultaneous selection of all 12 points (right). Simulations were conducted using Algorithms 1 and 2, with a Gaussian kernel whose length-scale was $\ell = 0.25$.

Algorithm 3: Non-myopic minimisation of MMD with mini-batching

```
Data: A set $\{\{x_j^i\}_{j=1}^b\}_{i=1}^m$ of mini-batches, each of
      size $b \in \mathbb{N}$, a distribution $\mu$, a kernel $k$, a
      number of points to select per iteration $s \in \mathbb{N}$,
      and a number of iterations $m \in \mathbb{N}$
Result: An index sequence $\pi \in \{1, \dots, n\}^{m \times s}$
for $i = 1, \dots, m$ do
    $\pi(i, \cdot) \in \underset{S \in \{1,\dots,b\}^s}{\operatorname{argmin}} \Big[ \frac{1}{2} \sum_{j,j' \in S} k(x_j^i, x_{j'}^i)$
        $+ \sum_{i'=1}^{i-1} \sum_{j=1}^{s} \sum_{j' \in S} k(x_{\pi(i',j)}^{i'}, x_{j'}^i)$
        $- is \sum_{j \in S} \int k(x, x_j^i) \mathrm{d}\mu(x) \Big]$
end
```

$$\pi \in \underset{S \in \{1,\dots,n\}^m}{\operatorname{argmin}} \; \mathrm{MMD}_{\mu,k}\left(\tfrac{1}{m} \textstyle\sum_{i \in S} \delta(x_{\pi(i)}) \right) \tag{5}$$

The index set π can again be recast as the solution to an IQP and the associated empirical measure $\nu = \frac{1}{m} \sum_{i=1}^m \delta(x_{\pi(i)})$ provides, by definition, a value for $\mathrm{MMD}_{\mu,k}(\nu)$ that is at least as small as any of the methods so far described (thus satisfying the same error bounds derived in Theorems 1–4).

However, it is only practical to exactly solve (5) for small m and thus, to arrive at a practical algorithm, we consider approximation of (5). There are at least two natural ways to do this. Firstly, one could run a numerical solver for the IQP formulation of (5) and terminate after a fixed computational limit is reached; the solver will return a feasible, but not necessarily optimal, solution to the IQP. An advantage of this approach is that no further methodological work is required. Alternatively, one could employ a convex relaxation of the intractable IQP, which introduces an approximation error that is hard to quantify but leads to a convex problem that may be exactly soluble at reasonable cost. We expand on the latter approach in Appendix B, with preliminary empirical comparisons.

4 Empirical Assessment

This section presents an empirical assessment[3] of Algorithms 1–3. Two regimes are considered, corresponding to high compression (small sm/n; Section 4.1) and low compression (large sm/n; Section 4.2) of the target. These occur, respectively, in applications to Bayesian cubature and thinning of Markov chain output. In Section 4.3, we compare our method to a variety of others based on optimisation in continuous spaces, augmenting a study in Chen et al. (2019). For details of the kernels used, and a sensitivity analysis for the kernel parameters, see Appendix C.

Figure 1 illustrates how a non-myopic algorithm may outperform a myopic one. A candidate set was constructed using 1000 independent samples from a test measure, and 12 representative points selected using the myopic (Alg. 1 with $m = 12$), non-myopic (Alg. 2 with $m = 3$ and $s = 4$), and non-sequential (Alg. 2 with $m = 1$ and $s = 12$) approaches. After choosing the first three samples close to the three modes, the myopic method then selects points that temporarily worsen the overall approximation; note in particular the placement of the fourth point. The non-myopic methods do not suffer to the same extent: choosing 4 points together gives better approximations after each of 4, 8 and 12 samples have been chosen ($s = 4$ was chosen deliberately so as to be co-prime to the number of mixture components, 3). Choosing all 12 points at once gives an even better approximation.

4.1 Bayesian Cubature

Larkin (1972) and subsequent authors proposed to cast numerical cubature in the Bayesian framework, such that an integrand f is *a priori* modelled as a Gaussian process with covariance k, then conditioned on data

[3] Our code is written in `Python` and is available at `https://github.com/oteym/OptQuantMMD`

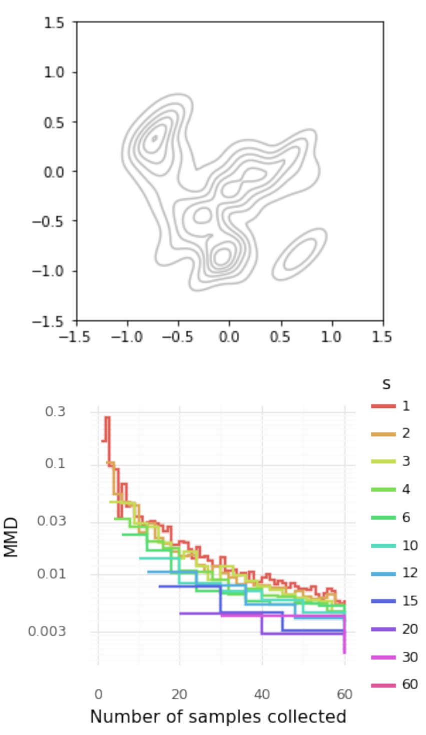

Figure 2: Synthetic data model formed of a mixture of 20 bivariate Gaussians (top). Effect of varying the number s of simultaneously-chosen points on MMD when choosing 60 from 1000 independently sampled points (bottom).

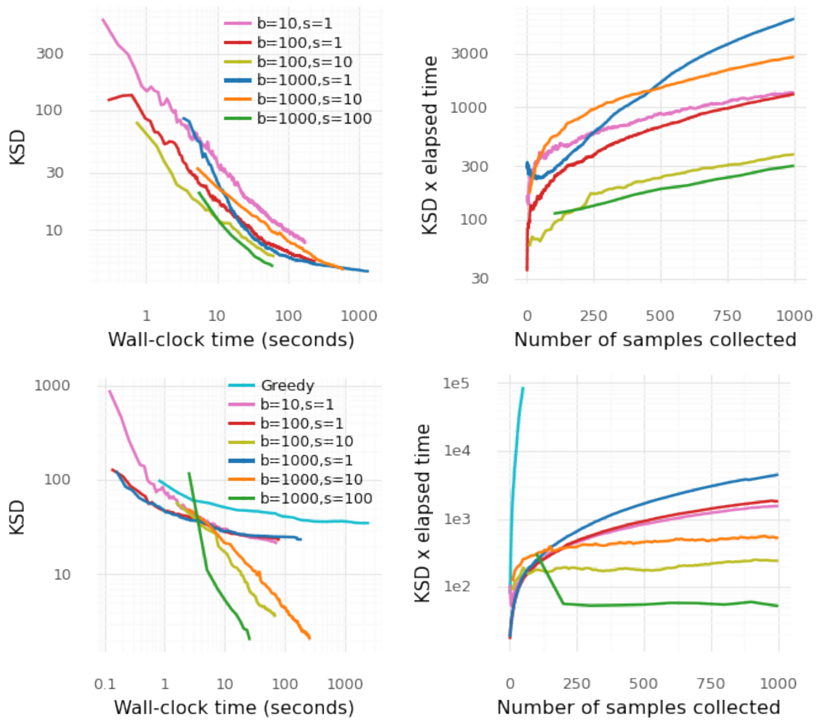

Figure 3: KSD vs. wall-clock time, and KSD × time vs. number of selected samples, shown for the 38-dim calcium signalling model (top two panes) and 4-dim Lotka–Volterra model (bottom two panes). The kernel length-scale in each case was set using the median heuristic (Garreau et al., 2017), and estimated in practice using a uniform subsample of 1000 points for each model. The myopic algorithm of Riabiz et al. (2020) is included in the Lotka–Volterra plots—see main text for details.

$\mathcal{D} = \{f(x_i)\}_{i=1}^n$ as the integrand is evaluated. The posterior standard deviation is (Briol et al., 2019b)

$$\mathrm{Std}\big[\textstyle\int f \,\mathrm{d}\mu|\mathcal{D}\big] = \min_{w_1,\ldots,w_m\in\mathbb{R}} \mathrm{MMD}_{\mu,k}\Big(\sum_{i=1}^m w_i\delta(x_i)\Big). \quad (6)$$

The selection of x_i to minimise (6) is impractical since evaluation of (6) has complexity $O(m^3)$. Huszár and Duvenaud (2012) and Briol et al. (2015) noted that (6) can be bounded above by fixing $w_i = \frac{1}{m}$, and that quantisation methods give a practical means to minimise this bound. All results in Section 5 can therefore be applied and, moreover, the bound is expected to be quite tight—$w_i \ll \frac{1}{m}$ implies that x_i was not optimally placed, thus for optimal x_i we anticipate $w_i \approx \frac{1}{m}$.

BC is most often used when evaluation of f has a high computational cost, and one is prepared to expend resources in the optimisation of the point set $\{x_i\}_{i=1}^m$. Our focus in this section is therefore on the quality of the point set obtained, irrespective of computational cost. Figure 1 suggested that the approximation quality of non-myopic methods depends on s. Figure 2 compares the selection of 60 from 1000 independently sampled points from a mixture of 20 Gaussians, varying s. This gives a set of step functions. Less myopic selections are seen to outperform more myopic ones. Note in particular that MMD of the myopic method ($s = 1$) is observed to decrease non-monotonically. This is a manifestation of the phenomenon also seen in Figure 1, where a particular selection may temporarily worsen the quality of the overall approximation.

Next we consider applications in Bayesian statistics, where both approximation quality and computation time are important. In what follows the density p_μ will be available only up to an unknown normalisation constant and thus KSD—which requires only that $u_\mu = \nabla \log p_\mu$ can be evaluated—will be used.

4.2 Thinning of Markov Chain Output

The use of quantisation to ‘thin’ Markov chain output was proposed in Riabiz et al. (2020), who studied greedy myopic algorithms based on KSD. We revisit the applications from that work to determine whether our methods offer a performance improvement. Unlike Section 4.1, the cost of our algorithms must now be assessed, since their runtime may be comparable to the time required to produce Markov chain output itself.

The datasets[4] consist of (i) 4×10^6 samples from the 38-parameter intracellular calcium signalling model of Hinch et al. (2004), and (ii) 2×10^6 samples from a 4-parameter Lotka–Volterra predator-prey model. The MCMC chains are both highly auto-correlated and start far from any mode. The greedy KSD approach of Riabiz et al. (2020) was found to slow down dramatically after selecting 10^3 samples, due to the need to compute the kernel between selected points and *all* points in the candidate set at each iteration. We employ mini-batching ($b < n$) to ameliorate this, and also investigate the effectiveness of non-myopic selection.

Figure 3 plots KSD against time, with the number of collected samples m fixed at 1000, as well as time-adjusted KSD against m for both models. This acknowledges that both approximation quality and computational time are important. In both cases, larger mini-batches were able to perform better *provided* that s was large enough to realise their potential. Non-myopic selection shows a significant improvement over batch-myopic in the Lotka–Volterra model, and a less significant (though still visible) improvement in the calcium signalling model. The practical upper limit on b for non-myopic methods (due to the requirement to optimise over all b points) may make performance for larger s poorer relatively; in a larger and more complex model, there may be fewer than s 'good' samples to choose from given moderate b. This suggests that control of the ratio s/b may be important; the best results we observed occurred when $s/b = 10^{-1}$.

A comparison to the original myopic algorithm (i.e. $b = n$) of Riabiz et al. (2020) is incuded for the Lotka–Volterra model. This is implemented using the same code and machine as the other simulations. The cyan line shown in the bottom two panes of Figure 3 represents only 50 points (not 1000); collecting just these took 42 minutes. This algorithm is slower still for the calcium signalling model, so it was omitted.

4.3 Comparison with Previous Approaches

Here we compare against approaches based on continuous optimisation, reproducing a 10-dimensional ODE inference task due to Chen et al. (2019). The aim is to minimise KSD whilst controlling the number of evaluations n_{eval} of either the (un-normalised) target μ or its log-gradient u_μ. Figure 4 reports results for *random walk Metropolis* (RWM), the *Metropolis-adjusted Langevin algorithm* (MALA), *Stein variational gradient descent* (SVGD), *minimum energy designs* (MED), *Stein points* (SP), and *Stein point MCMC* (four flavours, denoted SP-$*$, described in Chen et al., 2019). The method from Algorithm 3, shown as a black solid line (OPT MB 100-10; $b = 100$, $s = 10$) and dashed line (OPT MB 10-1; $b = 10$, $s = 1$), was applied to select 100 states from the first mb states visited in the RWM sample path, with m increasing and b fixed. The resulting quantisations are competitive with those produced by existing methods, at comparable computational cost. We additionally include a comparison with batch-uniform selection (B-UNIF 100-10; $b = 100$, $s = 10$, drawn uniformly from the RWM output) in light grey.

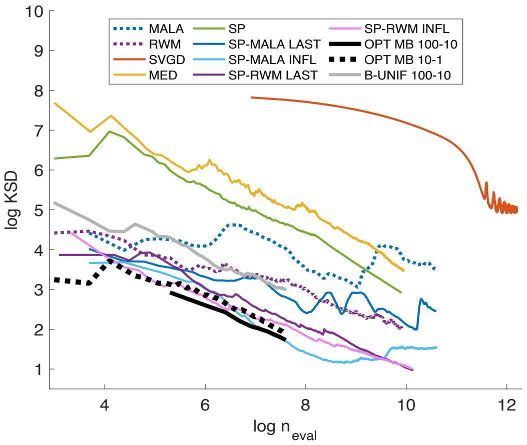

Figure 4: Comparison of quality of approximation for various methods, adjusted by the number of evaluations n_{eval} of μ or u_μ. For details of line labels, refer to the main text, and to Fig. 4 of Chen et al. (2019).

5 Theoretical Assessment

This section presents a theoretical assessment of Algorithms 1–3. Once stated, a standing assumption is understood to hold for the remainder of the main text.

Standing Assumption 1. *Let $\mathcal{X}$ be a measurable space equipped with a probability measure μ. Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be symmetric positive definite and satisfy $C_{\mu,k}^2 := \iint k(x,y)\mathrm{d}\mu(x)\mathrm{d}\mu(y) < \infty$.*

For the kernels k_μ described in Section 2.2, $C_{\mu,k_\mu} = 0$ and the assumption is trivially satisfied. Our first result is a finite-sample-size error bound for non-myopic algorithms when the candidate set is fixed:

Theorem 1. *Let $\{x_i\}_{i=1}^n \subset \mathcal{X}$ be fixed and let $C_{n,k}^2 := \max_{i=1,\dots,n} k(x_i,x_i)$. Consider an index sequence π of length m and with selection size s produced by Algorithm 2. Then for all $m \geq 1$ it holds that*

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

with $C := C_{\mu,k} + C_{n,k}$ an m-independent constant.

[4] Available at `https://doi.org/10.7910/DVN/MDKNWM`.

The proof is provided in Appendix A.1. Aside from providing an explicit error bound, we see that the output of Algorithm 2 converges in MMD to the optimal (weighted) quantisation of μ that is achievable using the candidate point set. Interestingly, all bounds we present are independent of s.

Remark 2. *Theorems 1–4 are stated for general MMD and apply in particular to KSD, for which we set $k = k_\mu$. These results extend the work of Chen et al. (2018, 2019) and Riabiz et al. (2020), who considered only myopic algorithms (i.e., $s = 1$).*

Remark 3. *The theoretical bounds being independent of s do not necessarily imply that $s = 1$ is optimal in practice; indeed our empirical results in Section 4 suggest that it is not.*

Our remaining results explore the cases where the candidate points are randomly sampled. Independent and dependent sampling is considered and, in each case, the following moment bound will be assumed:

Standing Assumption 2. *For some $\gamma > 0$, $C_1 := \sup_{i\in\mathbb{N}} \mathbb{E}\left[e^{\gamma k(x_i,x_i)}\right] < \infty$, where the expectation is taken over $x_1, \ldots, x_n$.*

In the first randomised setting, the x_i are independently sampled from μ, as would typically be possible when μ is explicit:

Theorem 2. *Let $\{x_i\}_{i=1}^n \subset \mathcal{X}$ be independently sampled from μ. Consider an index sequence π of length m produced by Algorithm 2. Then for all $s \in \mathbb{N}$ and all $m, n \geq 1$ it holds that*

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

The proof is provided in Appendix A.2. It is seen that n must be asymptotically increased with m in order for the approximation provided by Algorithm 2 to be consistent. No smoothness assumptions were placed on k; such assumptions can be used to improve the $O(n^{-1})$ term (as in Thm. 1 of Ehler et al., 2019), but we did not consider this useful since the $O(m^{-1})$ term is the bottleneck in the bound.

An analogous (but more technically involved) argument leads to a finite-sample-size error bound when mini-batches are used:

Theorem 3. *Let each mini-batch $\{x_j^i\}_{j=1}^b \subset \mathcal{X}$ be independently sampled from μ. Consider an index sequence π of length m produced by Algorithm 3. Then $\forall\ m, n \geq 1$*

$$\mathbb{E}\Big[\mathrm{MMD}_{\mu,k}\left(\tfrac{1}{ms}\textstyle\sum_{i=1}^m \sum_{j=1}^s \delta(x^i_{\pi(i,j)})\right)^2\Big] \leq \tfrac{\log(C_1)}{b\gamma} + 2\left(C_{\mu,k}^2 + \tfrac{\log(bC_1)}{\gamma}\right)\left(\tfrac{1+\log m}{m}\right).$$

The proof is provided in Appendix A.3. The mini-batch size b plays an analogous role to n in Theorem 2 and must be asymptotically increased with m in order for Algorithm 3 to be consistent.

In our second randomised setting the candidate set arises as a Markov chain sample path. Let V be a function $V : \mathcal{X} \to [1, \infty)$ and, for a function $f : \mathcal{X} \to \mathbb{R}$ and a measure μ on $\mathcal{X}$, let

$$\|f\|_V := \sup_{x\in\mathcal{X}} \tfrac{|f(x)|}{V(x)}, \quad \|\mu\|_V := \sup_{\|f\|_V \leq 1} \left|\textstyle\int f \mathrm{d}\mu\right|.$$

A ψ-irreducible and aperiodic Markov chain with nth step transition kernel P^n is *V-uniformly ergodic* if and only if there exists $R \in [0, \infty)$ and $\rho \in (0, 1)$ such that

$$\|\mathrm{P}^n(x, \cdot) - P\|_V \leq RV(x)\rho^n \tag{7}$$

for all initial states $x \in \mathcal{X}$ and all $n \in \mathbb{N}$ (see Thm. 16.0.1 of Meyn and Tweedie, 2012).

Theorem 4. *Assume that $\int k(x, \cdot)\mathrm{d}\mu(x) = 0$ for all $x \in \mathcal{X}$. Consider a μ-invariant, time-homogeneous, reversible Markov chain $\{x_i\}_{i\in\mathbb{N}} \subset \mathcal{X}$ generated using a V-uniformly ergodic transition kernel, such that (7) is satisfied with $V(x) \geq \sqrt{k(x, x)}$ for all $x \in \mathcal{X}$. Suppose that $C_2 := \sup_{i\in\mathbb{N}} \mathbb{E}[\sqrt{k(x_i, x_i)}V(x_i)] < \infty$. Consider an index sequence π of length m and selection subset size s produced by Algorithm 2. Then, with $C_3 = \frac{2R\rho}{1-\rho}$, we have that*

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

The proof is provided in Appendix A.4. Analysis of mini-batching in the dependent sampling context appears to be more challenging and was not attempted.

6 Discussion

This paper focused on quantisation using MMD, proposing and analysing novel algorithms for this task, but other integral probability metrics could be considered. More generally, if one is interested in compression by means other than quantisation then other approaches may be useful, such as Gaussian mixture models and related approaches from the literature on density estimation (Silverman, 1986).

Some avenues for further research include: (i) extending symmetric structure in μ to the set of representative points (Karvonen et al., 2019); (ii) characterising an *optimal* sampling distribution from which elements of the candidate set can be obtained (Bach, 2017); (iii) further applications of our method, for example to Bayesian neural networks, where quantisation of the posterior provides a promising route to reduce the cost of predicting each label in the test dataset. •

Acknowledgments

The authors are grateful to Lester Mackey and Luc Pronzato for helpful feedback on an earlier draft, and to Wilson Chen for sharing the code used to produce Figure 4. This work was supported by the Lloyd's Register Foundation programme on data-centric engineering at the Alan Turing Institute, UK. MR was supported by the British Heart Foundation–Alan Turing Institute cardiovascular data science award (BHF; SP/18/6/33805).

References

Arbel, M., Korba, A., Salim, A., and Gretton, A. (2019). Maximum mean discrepancy gradient flow. *NeurIPS 32*, pages 6484–6494.

Aronszajn, N. (1950). Theory of reproducing kernels. *Trans. Am. Math. Soc.*, 68(3):337–404.

Bach, F. (2017). On the equivalence between kernel quadrature rules and random feature expansions. *J. Mach. Learn. Res.*, 18(1):714–751.

Bach, F., Lacoste-Julien, S., and Obozinski, G. (2012). On the equivalence between herding and conditional gradient algorithms. *ICML 29*.

Barp, A., Briol, F.-X., Duncan, A., Girolami, M., and Mackey, L. (2019). Minimum Stein discrepancy estimators. *NeurIPS 32*, pages 12964–12976.

Barp, A., Oates, C., Porcu, E., Girolami, M., et al. (2018). A Riemannian-Stein kernel method. *arXiv:1810.04946*.

Briol, F.-X., Barp, A., Duncan, A. B., and Girolami, M. (2019a). Statistical inference for generative models with maximum mean discrepancy. *arXiv:1906.05944*.

Briol, F.-X., Oates, C., Girolami, M., and Osborne, M. A. (2015). Frank-Wolfe Bayesian quadrature: Probabilistic integration with theoretical guarantees. *NeurIPS 28*, pages 1162–1170.

Briol, F.-X., Oates, C. J., Girolami, M., Osborne, M. A., Sejdinovic, D., et al. (2019b). Probabilistic integration: A role in statistical computation? *Stat. Sci.*, 34(1):1–22.

Chaloner, K. and Verdinelli, I. (1995). Bayesian experimental design: a review. *Stat. Sci.*, 10(3):273–304.

Chen, W. Y., Barp, A., Briol, F.-X., Gorham, J., Girolami, M., Mackey, L., and Oates, C. J. (2019). Stein point Markov chain Monte Carlo. *ICML*, 36.

Chen, W. Y., Mackey, L., Gorham, J., Briol, F.-X., and Oates, C. J. (2018). Stein points. *ICML*, 35.

Chen, Y., Welling, M., and Smola, A. (2010). Super-samples from kernel herding. *UAI*, 26:109–116.

Chérief-Abdellatif, B.-E. and Alquier, P. (2020). MMD-Bayes: robust Bayesian estimation via maximum mean discrepancy. *Proc. Mach. Learn. Res.*, 18:1–21.

Dick, J. and Pillichshammer, F. (2010). *Digital nets and sequences: discrepancy theory and quasi–Monte Carlo integration.* Cambridge University Press.

Dudley, R. M. (2018). *Real analysis and probability.* CRC Press.

Ehler, M., Gräf, M., and Oates, C. J. (2019). Optimal Monte Carlo integration on closed manifolds. *Stat. Comput.*, 29(6):1203–1214.

Garreau, D., Jitkrittum, W., and Kanagawa, M. (2017). Large sample analysis of the median heuristic. *arXiv:1707.07269.*

Goemans, M. X. and Williamson, D. P. (1995). Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *J. ACM*, 42(6):1115–1145.

Gorham, J., Duncan, A., Mackey, L., and Vollmer, S. (2019). Measuring sample quality with diffusions. *Ann. Appl. Probab.*, 29(5):2884–2928.

Gorham, J. and Mackey, L. (2015). Measuring sample quality with Stein's method. *NeurIPS 28*, pages 226–234.

Gorham, J. and Mackey, L. (2017). Measuring sample quality with kernels. *ICML 34*, 70:1292–1301.

Graf, S. and Luschgy, H. (2007). *Foundations of quantization for probability distributions.* Springer.

Gurobi Optimization, LLC (2020). *Gurobi Optimizer Reference Manual.* http://www.gurobi.com.

Hickernell, F. (1998). A generalized discrepancy and quadrature error bound. *Math. Comput*, 67(221):299–322.

Hinch, R., Greenstein, J., Tanskanen, A., Xu, L., and Winslow, R. (2004). A simplified local control model of calcium-induced calcium release in cardiac ventricular myocytes. *Biophys. J*, 87(6):3723–3736.

Huszár, F. and Duvenaud, D. (2012). Optimally-weighted herding is Bayesian quadrature. *UAI*, 28:377–386.

Karvonen, T. (2019). Kernel-based and Bayesian methods for numerical integration. *PhD Thesis, Aalto University.*

Karvonen, T., Särkkä, S., and Oates, C. (2019). Symmetry exploits for Bayesian cubature methods. *Stat. Comput.*, 29:1231–1248.

Lacoste-Julien, S., Lindsten, F., and Bach, F. (2015). Sequential kernel herding: Frank-Wolfe optimization for particle filtering. *Proc. Mach. Learn. Res.*, 38:544–552.

Larkin, F. M. (1972). Gaussian measure in Hilbert space and applications in numerical analysis. *Rocky Mt. J. Math.*, 3:379–421.

Le, H., Lewis, A., Bharath, K., and Fallaize, C. (2020). A diffusion approach to Stein's method on Riemannian manifolds. *arXiv:2003.11497.*

Mak, S., Joseph, V. R., et al. (2018). Support points. *Ann. Stat.*, 46(6A):2562–2592.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability.* Springer.

MOSEK ApS (2020). *The MOSEK Optimizer API for Python 9.2.26.* http://www.mosek.com.

Müller, A. (1997). Integral probability metrics and their generating classes of functions. *Adv. App. Prob.*, 29(2):429–443.

Novak, E. and Woźniakowski, H. (2008). *Tractability of Multivariate Problems: Standard information for functionals.* European Mathematical Society.

Oates, C. J., Girolami, M., and Chopin, N. (2017). Control functionals for Monte Carlo integration. *J. R. Statist. Soc. B*, 3(79):695–718.

Oettershagen, J. (2017). Construction of optimal cubature algorithms with applications to econometrics and uncertainty quantification. *PhD thesis, University of Bonn.*

Paige, B., Sejdinovic, D., and Wood, F. (2016). Supersampling with a reservoir. *UAI*, 32:567–576.

Pronzato, L. and Zhigljavsky, A. (2018). Bayesian quadrature, energy minimization, and space-filling design. *SIAM-ASA J. Uncert. Quant.*, 8(3):959–1011.

Rendl, F. (2016). Semidefinite relaxations for partitioning, assignment and ordering problems. *Ann. Oper. Res.*, 240:119–140.

Riabiz, M., Chen, W., Cockayne, J., Swietach, P., Niederer, S. A., Mackey, L., and Oates, C. (2020). Optimal thinning of MCMC output. *arXiv:2005.03952.*

Rustamov, R. M. (2019). Closed-form expressions for maximum mean discrepancy with applications to Wasserstein auto-encoders. *arXiv:1901.03227.*

Santin, G. and Haasdonk, B. (2017). Convergence rate of the data-independent P-greedy algorithm in kernel-based approximation. *Dolomites Research Notes on Approximation*, 10:68–78.

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013). Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *Ann. Stat.*, 41(5):2263–2291.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis.* CRC Press.

Song, L. (2008). Learning via Hilbert space embedding of distributions. *PhD thesis, University of Sydney.*

Sriperumbudur, B. K., Gretton, A., Fukumizu, K., Schölkopf, B., and Lanckriet, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *J. Mach. Learn. Res.*, 11:1517–1561.

Stein, C. (1972). A bound for the error in the normal approximation to the distribution of a sum of dependent random variables. In *Proc. Sixth Berkeley Symp. on Math. Statist. and Prob., Vol. 2*, pages 583–602.

Wolsey, L. A. (2020). *Integer Programming: 2nd Edition.* John Wiley & Sons, Ltd.

Xu, W. and Matsuda, T. (2020). A Stein goodness-of-fit test for directional distributions. *Proc. Mach. Learn Res.*, 108:320–330.

Yang, J., Liu, Q., Rao, V., and Neville, J. (2018). Goodness-of-fit testing for discrete distributions via Stein discrepancy. *Proc. Mach. Learn. Res.*, 80:5561–5570.