Geometric Rates of Convergence for Kernel-based Sampling Algorithms

Rajiv Khanna[1] **Liam Hodgkinson**[1,2] **Michael W. Mahoney**[1,2]

[1]Department of Statistics, UC Berkeley
[2]International Computer Science Institute (ICSI), Berkeley

Abstract

The rate of convergence of weighted kernel herding (WKH) and sequential Bayesian quadrature (SBQ), two kernel-based sampling algorithms for estimating integrals with respect to some target probability measure, is investigated. Under verifiable conditions on the chosen kernel and target measure, we establish a near-geometric rate of convergence for target measures that are nearly atomic. Furthermore, we show these algorithms perform comparably to the theoretical best possible sampling algorithm under the maximum mean discrepancy. An analysis is also conducted in a distributed setting. Our theoretical developments are supported by empirical observations on simulated data as well as a real world application.

1 INTRODUCTION

Estimating expectations is a common problem fundamental to many applications in statistics and machine learning, such as the estimation of sufficient statistics, prediction after marginalization of latent variables, and calculation of risk. The goal is the computation of integrals of the form

$$\mathbb{E}_\pi f(\mathbf{x}) = \int_{\mathcal{X}} f(\mathbf{x}) d\pi(\mathbf{x}), \tag{1}$$

where f is a given function f and π a probability measure over $\mathbb{R}^d$. In the vast majority of cases, the integral is not analytically computable, necessitating numerical approximation. If π is absolutely continuous with density p, classical Gaussian quadrature techniques achieve exponential rates of convergence, but these are known to suffer from the curse of dimensionality. Instead, one often adopts probabilistic techniques. The simplest of these is crude Monte Carlo (MC) integration: generating independent samples $\{\mathbf{x}_i\}_{i=1}^m$ from π, and returning the empirical average of $\{f(\mathbf{x}_i)\}_{i=1}^m$. This method converges at a prohibitively slow rate of $\mathcal{O}(m^{-1/2})$ in L^2. If one cannot easily generate independent samples from π, Markov Chain Monte Carlo (MCMC) is often used instead. This method generates approximate samples from π as iterates of an ergodic Markov chain. Under certain assumptions on the chain, this approach will also converge at rate $\mathcal{O}(m^{-1/2})$.

To achieve "super-root-m" rates of convergence, alternative non-random sampling techniques have been proposed. Quasi Monte Carlo is a classic example with dimension-dependent $\mathcal{O}(m^{-1}\log m)$ convergence. More recently, the herding algorithm [Welling, 2009a,b, Welling and Chen, 2010] was proposed to learn Markov Random Fields (MRFs). First applicable to discrete finite-dimensional spaces, it was later extended to continuous spaces and infinite dimensions through the kernel trick by Chen et al. [2010], who called the resulting algorithm *Kernel Herding* (KH). A general convergence rate of $\mathcal{O}(m^{-1/2})$ was also provided, showing that KH performs, asymptotically, at least as well as crude Monte Carlo methods. However, in practice, KH typically exhibits faster convergence, suggesting these rates can be improved.

Chen et al. [2010] also suggested a *moment matching* interpretation of the algorithm: kernel herding is equivalent to choosing samples $\{\mathbf{x}_i\}_{i=1}^m$ that minimize the Maximum Mean Discrepancy (MMD) metric (see Chen et al. [2010], Huszar and Duvenaud [2012]) between π and the empirical measure of $\{\mathbf{x}_i\}_{i=1}^m$. In an alternative Bayesian approach, O'Hagan [1991] and Ghahramani and Rasmussen [2003] assume a GP prior on f. Each sample $\mathbf{x}_i$ is generated by minimizing the mean squared error of the posterior of f, conditioned on the points $\{\mathbf{x}_j\}_{j<i}$ already sampled. This was shown to be equivalent to KH (with kernel dictated by the covariance of the GP), but with an additional step of also minimizing the weights attached to the sampled points (that is, instead of using uniform weights of $1/m$ [Huszar and Duvenaud, 2012]). Empirically, it was observed that this new algorithm—called *Sequential Bayesian Quadrature* (SBQ)—converges faster than KH due to the additional

Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence (UAI 2021), PMLR 161:2156–2164.

weight optimization step.

A partial explanation for faster convergence was given by Bach et al. [2012]. While KH and SBQ both choose the next sample point to minimize MMD, SBQ also optimizes for the weights. Alternatively, once a sample point is selected, the weights themselves can be optimized. We refer to this procedure as *Weighted Kernel Herding* (WKH). For the case when the weights are constrained to lie in the unit simplex, Bach et al. [2012] noted that WKH is equivalent to the classic algorithm of Frank and Wolfe [1956] on the marginal polytope. Exploiting this connection, they were able to use existing convergence results to analyze constrained WKH. Specifically, if the optimum lies in the relative interior of the marginal polytope with distance b away from the boundary, the convergence rate is $\mathcal{O}(e^{-bm})$. For infinite dimensional kernels, $b = 0$ (i.e., exponential convergence does not hold). On the other hand, $b > 0$ for finite dimensional kernels, but it could be so close to zero that the global $\mathcal{O}(m^{-1/2})$ bound proves tighter. Bach et al. [2012] also point out these issues, suggesting that another approach is required to fully justify the improved empirical performance of WKH.

Here, we attempt to address this deficiency by providing an analysis for near-exponential convergence of unconstrained WKH and SBQ. As noted by Huszar and Duvenaud [2012], the weights in WKH do not lie on the unit simplex in many applications, and so the correspondence to Frank–Wolfe does not hold. Instead, we study the convergence behavior of WKH and SBQ with respect to the *best possible algorithm* under MMD for generating m samples from the target measure — a feat that is almost certainly unachievable in practice. Our analysis effectively says that WKH and SBQ are "good enough," in the sense that one only requires to pick a few more atoms using WKH or SBQ than any other possible algorithm to get close to the performance of the latter. This result is encapsulated in Theorem 1. Within this understanding, we find that the relevant assumption for investigating convergence is *realizability*: that the mean embedding in the kernel space can be exactly reproduced by a linear combination of r samples $\{\mathbf{x}_i\}_{i=1}^r$. This amounts to assuming that π is comprised of finitely many atoms — for any distribution that can be closely approximated by such a measure, this may well be reasonable to consider. In this case, we show that realizability guarantees $\mathcal{O}(e^{-rm})$ convergence.

1.1 CONTRIBUTIONS

Our contributions in this work include:

- An analysis of two algorithms for approximating expectations — WKH and SBQ — highlighting *near exponential* decay with respect to the best possible sampling algorithm under MMD. Our results provide theoretical justifications for empirical observations already made in the literature.
- The introduction of the assumption of realizability in the context of KH algorithms, enabling tighter analysis for distributions close to a finitely atomic measure.
- A *distributed* algorithm for approximating expectations for large scale computations, together with a short analysis of its convergence properties. To the best of our knowledge, herding has not yet been applied in large-scale distributed settings.
- Finally, we present *empirical studies* to validate rapid convergence. Since there is ample empirical evidence supporting the good performance of the KH/SBQ algorithms in earlier works [Huszar and Duvenaud, 2012, Chen et al., 2010, Bach et al., 2012], we focus instead on demonstrating the empirical performance of the distributed algorithm.

1.2 NOTATION

We represent vectors as small letter bolds, e.g., $\mathbf{u}$. Matrices are represented by capital bolds, e.g., $\mathbf{X}, \mathbf{T}$. Sets are represented by sans serif fonts, e.g., S; and the complement of a set S is S^c. A dot product in a reproducing kernel Hilbert space with kernel $k(\cdot,\cdot)$ is represented as $\langle\cdot,\cdot\rangle_k$, and the corresponding norm is $\|\cdot\|_k$. The dual norm is written as $\|\cdot\|_{k*}$. We denote $\{1, 2, \ldots, d\}$ by $[d]$.

2 BACKGROUND

In this section, we discuss some relevant background for the algorithms and methods at hand.

Maximum Mean Discrepancy (MMD). MMD measures the worst-case error between two probability measures π and ν over the unit ball of a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$ with kernel k:

$$\begin{aligned}\mathrm{MMD}(\pi,\nu) &:= \sup_{\substack{f\in\mathcal{H}\\ \|f\|\le 1}} \left| \int f(x)d\pi(x) - \int f(x)d\nu(x) \right| \\ &= \left| \iint k(x,y)d\pi(x)d\pi(y) + \iint k(x,y)d\nu(x)d\nu(y) \right. \\ &\qquad \left. - 2\iint k(x,y)d\pi(x)d\nu(y) \right| \qquad (2) \\ &= \|\mu_\pi - \mu_\nu\|_k \qquad (3)\end{aligned}$$

where μ_π and μ_ν are the respective mean kernel embeddings of π and ν. $\mathrm{MMD}(\pi,\nu) \ge 0$; and if $\mathcal{H}$ is universal, then $\mathrm{MMD}(\pi,\nu) = 0$ if and only if $\pi \equiv \nu$. We refer to Sriperumbudur et al. [2010], Gretton et al. [2007] for further details. The sampling algorithms we consider approximate π by constructing an empirical measure ν that is close to π under MMD.

Algorithm 1 Weighted Kernel Herding : WKH($\mathcal{X}$, m), or Sequential Bayesian Quadrature : SBQ($\mathcal{X}$, m)

1: **INPUT:** kernel function $k(\cdot,\cdot)$, number of iterations m
2: $\mathsf{S} = \emptyset$. // Build solution set S greedily.
3: **for** $n = 1 \dots m$ **do**
4: Use (5) for WKH, or Use (7) for SBQ, to get $\mathbf{x}_n$. $\mathsf{S} = \mathsf{S} \cup \{\mathbf{x}_n\}$
5: Update weights $\mathbf{w} = \mathbf{K}^{-1}\mathbf{z}$, where $\mathbf{K}_{rs} = k(\mathbf{x}_r, \mathbf{x}_s)$ for $r, s \in [1, n]$
6: **end for**
7: return S, $\mathbf{w}$

Weighted Kernel Herding. Recall that our goal is to approximate the expectation of a function f over some probability measure π using a weighted empirical measure:

$$\mathbb{E}_\pi f(\mathbf{x}) = \int_{\mathcal{X}} f(\mathbf{x}) d\pi(\mathbf{x}) \approx \sum_{i=1}^{n} w_i f(\mathbf{x}_i), \qquad (4)$$

where w_i are the weights associated with function evaluations at $\mathbf{x}_i$. For example, taking weights $w_i = 1/n$ and samples $\mathbf{x}_i$ to be independent recovers crude MC integration. Both Kernel Herding [Chen et al., 2010] and Quasi Monte Carlo [Dick and Pillichshammer, 2010] use $w_i = 1/n$ with dependent samples $\mathbf{x}_i$. For brevity, in the sequel, we let $\mathsf{S}_j := \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_j\}$ be the collection of the first j samples, and $z(\mathsf{S}_j) := \sum_j w_j f(\mathbf{x}_j)$.

Here, we present a brief overview of WKH and SBQ, and point the reader to Huszar and Duvenaud [2012], Chen et al. [2010] for further details. As mentioned, these algorithms construct a weighted empirical measure by minimizing MMD to the target measure π. Greedy algorithms for constructing approximations of the form (4) — including WKH and SBQ — typically involve two alternating steps: (1) generate a sample $\mathbf{x}_i$; and then (2) compute the weights w_j across all drawn samples $\{\mathbf{x}_j\}_{j \le i}$. KH chooses the next sample by minimizing MMD, taking all weights to be uniform, that is, $w_i = 1/n$ for $i = 1, \dots, n$. More precisely, given n samples $\{\mathbf{x}_i\}_{i=1}^n$, the next sample $\mathbf{x}_{n+1}^{KH}$ is generated according to

$$\mathbf{x}_{n+1}^{KH} = \arg\min_{\mathbf{x} \in \mathcal{X}} \frac{n}{n+1} \sum_{i=1}^{n} w_i k(\mathbf{x}, \mathbf{x}_i) - 2\mathbb{E}_{\mathbf{x}' \sim \pi}[k(\mathbf{x}, \mathbf{x}')]. \qquad (5)$$

The SBQ algorithm is slightly more involved. Consider imposing a functional Gaussian Process (GP) on f with covariance kernel k, that is, $f \sim GP(0, k)$. Doing so, the quantities in (4) become random variables. The sample $\mathbf{x}_i$ is then chosen to minimize posterior variance, while the corresponding weights are calculated from the resulting posterior mean. More precisely, suppose $\mathbf{x}_1, \dots, \mathbf{x}_n$ are previously drawn samples. A standard result in the theory of GPs asserts that the posterior of f, conditioned on the evaluations $\{f(\mathbf{x}_i)\}_{i=1}^n$, has expectation

$$\hat{f}(\mathbf{x}) = \mathbf{k}^\top \mathbf{K}^{-1} \mathbf{f},$$

where $\mathbf{f}$ is the vector of function evaluations $(f(\mathbf{x}_i))_{i=1}^n$, $\mathbf{k}$ is the vector of kernel evaluations $(k(\mathbf{x}, \mathbf{x}_i))_{i=1}^n$, and $\mathbf{K} := (k(\mathbf{x}_i, \mathbf{x}_j))_{i,j=1}^n$ is the kernel Gram matrix. Consequently, inserting $\hat{f}$ into (4), it becomes clear that $\mathbb{E}z(\mathsf{S}_n) = \mathbf{z}^\top \mathbf{K}^{-1} \mathbf{f}$, where $\mathbf{z}_i := \int k(\mathbf{x}, \mathbf{x}_i) d\pi(\mathbf{x})$. In particular, observe that the weights in (4) are given by $w_i = \sum_j \mathbf{z}_j [\mathbf{K}^{-1}]_{ij}$. The posterior variance becomes

$$\mathrm{cov}(\mathbf{x}, \mathbf{y}) = k(\mathbf{x}, \mathbf{y}) - \mathbf{k}^\top \mathbf{K}^{-1} \mathbf{k}.$$

Therefore, we can write the variance of $z(\mathsf{S}_n)$ as

$$\mathrm{var}(z(\mathsf{S}_n)) = \iint k(\mathbf{x}, \mathbf{y}) d\pi(\mathbf{x}) d\pi(\mathbf{y}) - \mathbf{z}^\top \mathbf{K}^{-1} \mathbf{z}. \quad (6)$$

Given n samples $\{\mathbf{x}_i\}_{i=1}^n$, the SBQ algorithm generates the next sample $\mathbf{x}_{n+1}^{SBQ}$ by minimizing the posterior variance of the approximated integral $z(\mathsf{S}_n)$:

$$\mathbf{x}_{n+1}^{SBQ} = \arg\min_{\mathbf{x} \in \mathcal{X}} \mathrm{var}(z(\mathsf{S}_n \cup \{\mathbf{x}\})). \qquad (7)$$

It has been shown that SBQ is equivalent to minimizing MMD with respect to both samples and weights [Huszar and Duvenaud, 2012].

In this work, we first analyze WKH (Algorithm 1) which performs the update (5) for the samples while also updating weights using SBQ's $w_i = \sum_j \mathbf{z}_j [\mathbf{K}^{-1}]_{ij}$. The same objective $\mathbf{x}_{n+1}^{KH}$ was also considered by Bach et al. [2012], with the additional constraint that the weights w_i are positive and sum to unity. They use the connections to the classic algorithm of Frank and Wolfe [1956] to establish convergence rates under this additional constraint.

3 RELATED WORKS

The connection between WKH and the Frank–Wolfe algorithm was further studied in Briol et al. [2015], providing new variants and rates. Other variants of the Frank–Wolfe algorithm [Jaggi, 2013, Lacoste-Julien and Jaggi, 2015] enjoy faster convergence at the cost of additional memory requirements. Specifically, instead of just selecting sample points, one can think of removing bad points from the set already selected. This variant of Frank–Wolfe, known as *FW with away steps*, is one of the more commonly used in practice, because it is known to converge faster. If the weights are not restricted to lie on the simplex, the analogy to matching pursuit algorithms is obvious — we refer to Locatello et al. [2017] for corresponding convergence rates. However, the linear rates for these algorithms require bounding certain geometric properties of the constraint set, and this may not be straightforward for RKHS-based applications that

usually employ KH and/or SBQ. There have been some recent works that provide sufficient conditions for fast convergence of SBQ under additional assumptions on sufficient exploration of point sets [Kanagawa and Hennig, 2019], or special cases of kernels (e.g. the ones generating Sobolev spaces [Santin and Haasdonk, 2016]). Our setting is more general. Other studies discuss dimension-dependent convergence rates [Briol et al., 2019] for infinitely smooth functions. Such rates often take the form $\mathcal{O}(n^{-1/d})$ for dimension d — our results have no explicit dimension-dependence. More recently, Kanagawa et al. [2020] study convergence properties in misspecified settings. Applying our ideas to these settings may provide an interesting direction for future work.

With the goal of interpreting blackbox models, Khanna et al. [2019] recently exploited connections with submodular optimization to provide the weaker forms of convergence we discuss (in Section 4.1), in the form of an approximation guarantee for SBQ for discrete π. Our result is more general, allowing for arbitrary probability measures. In fact, we do not make use of submodular optimization results. A similar proof technique was also used by Khanna et al. [2017] for proving approximation guarantees of low rank optimization. The proof idea for the distributed algorithm was inspired from tracking the optimum set, which is a common theme in analysis of distributed algorithms in discrete optimization (see, e.g., Altschuler et al. [2016] and references therein).

4 CONVERGENCE RESULTS

In this section, we present our convergence results. The starting point of our analysis is the following re-interpretation of the posterior variance minimization (6) as a variational optimization of MMD [Huszar and Duvenaud, 2012]. We can re-write (6) as a function of a set of chosen sample indices S and weights $\mathbf{w}$:

$$g(\mathsf{S}, \mathbf{w}) := \|\mu_\pi - \sum_{i \in \mathsf{S}} w_i \phi(\mathbf{x}_i)\|_k^2, \tag{8}$$

where ϕ is the respective feature mapping, i.e., $k(\mathbf{x}, \mathbf{y}) = \langle \phi(\mathbf{x}), \phi(\mathbf{y}) \rangle$, and where $\mu_\pi := \int_{\mathbf{y} \sim \pi} \phi(\mathbf{x}) d\pi(\mathbf{x})$ is the mean embedding in the kernel space.

Before presenting our results, we delineate the assumptions we make on the cost function (8). There is an implicit assumption that f lies within the chosen RKHS corresponding to the kernel k. Otherwise, there is no guarantee that any algorithm can approximate f within ϵ error, and the discussion of convergence rates is meaningless.

Assumption 1 (Convexity/Smoothness)**.** *We assume that the loss function $g(\cdot)$ is m_ω-restricted strongly convex and M_Ω-smooth over $\mathsf{S}_\perp^\star \cup \mathsf{S}_n$, with $m_\omega > 0$ and $M_\Omega < \infty$. In other words, for $\mathsf{S}_1, \mathsf{S}_2 \subset \mathsf{S}_r^\star \cup \mathsf{S}_n$, with $Z(\mathsf{S}_j) = \sum_j w_j \phi(\mathbf{x}_j)$, $Z_{12} = Z(\mathsf{S}_1) - Z(\mathsf{S}_2)$, and $D(\mathsf{S}_1, \mathsf{S}_2) = g(\mathsf{S}_1) - g(\mathsf{S}_2) - \langle \nabla g(\mathsf{S}_2), Z(\mathsf{S}_1) - Z(\mathsf{S}_2) \rangle_k$,*

$$\frac{m_\omega}{2} \|Z_{12}\|_k^2 \leq D(\mathsf{S}_1, \mathsf{S}_2) \leq \frac{M_\Omega}{2} \|Z_{12}\|_k^2.$$

Assumption 1 is standard for characterizing geometric rates in optimization; e.g., see Wainwright [2019, Section 9.3.1], and is often implicitly assumed, e.g., by Bach et al. [2012], when drawing upon the Frank–Wolfe connection. This assumption or a slight variation of it, is not only sufficient but also necessary to bound geometric convergence rates. For our case, one can equivalently view the assumption as enforcing that the kernel matrix of the atoms $\mathsf{S}_\perp^\star \cup \mathsf{S}_n$ has a minimum eigenvalue bounded away from zero for any fixed n.

Assumption 2 (Standardization)**.** *We assume that the feature mapping is standardized, that is, $k(\mathbf{x}, \mathbf{x}) = 1$ for all $\mathbf{x} \in \mathcal{X}$.*

Assumption 2 is not restrictive. Rather, as in previous works [Huszar and Duvenaud, 2012], it is enforced to avoid otherwise unnecessary terms for ease of exposition. Any kernel can be standardized over its support, and remain a kernel. Assumptions 1 and 2 are satisfied by many commonly used kernels. Any finite dimensional normalized kernel satisfy them, including the cosine kernel or polynomial kernels; and general normalized RBF kernels on compact supports. We are now ready to present our main result regarding the near linear convergence of the discrepancy metric g.

Theorem 1 (Approximation Guarantee)**.** *Suppose that Assumptions 1 and 2 are satisfied. Let $\mathsf{T}_r := \arg\min_{|\mathsf{S}| \leq r} \min_{\mathbf{w}} g(\mathsf{S}, \mathbf{w})$. For $0 < \epsilon < 1$, consider $s = (r \frac{M_\Omega}{m_\omega} \log \frac{1}{\epsilon})$ iterations of Algorithms 1 (WKH or SBQ), returning the set S_s. Then, $\min_{\mathbf{w}} g(\mathsf{S}_s, \mathbf{w})) \leq (1 - \epsilon)[\min_{\mathbf{w}} g(\mathsf{T}_r, \mathbf{w})] + \epsilon \|\mu_\pi\|_k^2$.*

The proof for WKH is presented in the supplement. Theorem 1 is general, and it is applicable to kernels of any dimension, including infinite dimensional kernels. Intuitively, it may be hard to claim that an infinite dimensional embedding can be closely approximated by a finite number of atoms without additional strong assumptions. However, our analysis gets around this complexity by comparing the performance of the algorithm at hand with best possible finite set of r atoms. Indeed, Theorem 1 is somewhat weaker than a classical convergence result. Instead of providing a rate on closeness to the optimum after k iterations, it provides a contraction factor on how close the algorithm gets to the best possible r steps that *any* algorithm could have taken. The theoretical best possible r samples could even come from an exhaustive combinatorial and computationally hard search over the entire space. However, by choosing only a multiplicative $\mathcal{O}(\log \frac{1}{\epsilon})$ number of extra atoms through the greedy selection processes in WKH or SBQ, we can get provably close to best-case performance.

The SBQ algorithm is equivalent to Algorithm 1, after replacing (5) in Step 4 with (7). Thus, WKH solves a linear program every iteration, while SBQ solves a kernel ℓ_2 minimization problem. The decrease in the cost function (8) per iteration of SBQ is more than its decrease per iteration of WKH. Thus, Theorem 1 also recovers the special case of SBQ for discrete densities studied by Khanna et al. [2019] by exploiting connections to weak submodularity. Their result is also an approximation guarantee (not global convergence guarantee), and only applies for SBQ. Our result is also valid for WKH and provides an alternative proof for SBQ without exploiting weak submodularity.

4.1 REALIZABILITY

Theorem 1 in itself is quite general – it holds for any kernel of arbitrary dimensionality. Here, we further specialize Theorem 1 and provide a sufficient condition under which the convergence rate is geometric (instead of being near-geometric). We encapsulate this sufficient condition as the assumption of realizability.

Assumption 3 (Realizability)**.** *There exists a set $\mathsf{S}_r^\star$ of r atoms, and weights $\mathbf{w}^*$, such that $g(\mathsf{S}_r^\star, \mathbf{w}^*) = 0$.*

Assumption 3 posits that there exists a set of atoms in the mapped domain $\phi(\mathcal{X})$ whose weighted average exactly evaluates to the expectation μ_π so that the discrepancy g is 0. If the RKHS is universal, this is equivalent to assuming that π is finitely atomic. Otherwise, realizability can be achieved with finite-dimensional kernels. By considering the assumption of realizability, we are able to investigate convergence rates independently from the capacity of the target measure π to be approximated (under MMD) by an atomic distribution. Furthermore, in some common use-cases, such as the data summarization task in Section 6, the target measure π is finitely atomic, in which case, realizability is automatically satisfied.

Theorem 2. *Under Assumptions 1 through 3, if S_i is the sequence of iterates produced by Algorithm 1 (WKH or SBQ), the function g converges as* $\min_w g(\mathsf{S}_i, \mathbf{w}) \leq \exp(-\frac{im_\omega}{rM_\Omega})g(\emptyset, 0)$.

Proof Outline. The central idea of the proof is to track and bound the selection of a sample at each iteration, compared to the ideal selection $\mathsf{S}_r^\star$ that could have provided the optimum solution. For this purpose, the properties of the selection subproblem (5) and the assumptions are used. The detailed proof is presented in the supplement.

Theorem 2 provides a linear convergence rate for WKH under the conditions specified in Assumptions 1 through 3. Recall that Bach et al. [2012] also provided a linear rate for finite dimensional kernels by drawing on the equivalence of the herding algorithm to the Frank–Wolfe algorithm. Their rate is $\mathcal{O}(\exp(-b^2m/R^2))$, where b is the distance of the optimum from the boundary of the marginal polytope and R is the width of the marginal polytope. Our result is independent of these constants. As long as Assumption 3 is satisfied, our work shows that exponential convergence is guaranteed. To the best of our knowledge, such a sufficient condition for these harder cases has not been previously been established.

Let us also briefly discuss the case where Assumption 3 is not satisfied, but π is ϵ-close under MMD to a finitely atomic measure $\tilde{\pi}$ that *does* satisfy Assumption 3. Using the triangle inequality for MMD, it is straightforward to show that $\min_w g(\mathsf{S}_i, \mathbf{w}) \leq \exp(-\frac{im_\omega}{rM_\Omega})g(\emptyset, 0) + 2\epsilon$, suggesting near-exponential convergence in these settings. Indeed, in cases where ϵ is small relative to $g(\emptyset, 0)$, this could explain the excellent empirical performance seen for WKH and SBQ.

5 DISTRIBUTED KERNEL HERDING

In many cases, the search over the domain $\mathcal{X}$ can be a severe computational bottleneck to practical use. In this section, we develop a new herding algorithm that can be distributed over multiple machines and run in a streaming map-reduce fashion. We also provide a quick convergence analysis, using techniques presented in Section 4.

The algorithm proceeds as follows. The domain $\mathcal{X}$ is split onto s machines uniformly at random. Each of the s machines has access to only $\mathcal{X}_i \subset \mathcal{X}$, such that $\bigcup_i \mathcal{X}_i = \mathcal{X}$ and $\mathcal{X}_i \cap \mathcal{X}_j = \emptyset$ for $i \neq j$. Each machine runs its own herding algorithm (Algorithm 1) independently, by restricting the search space in (5) to $\mathcal{X}_i$, instead of $\mathcal{X}$. Finally, the iterates generated by each machine are sent to a central server machine. The server *collates* the samples by running another copy of the same algorithm, with $\mathcal{X}$ replaced by the discrete set of samples received. Finally, the best solution out of all the $s+1$ solutions is returned. The pseudo-code is illustrated in Algorithm 2. In what follows, we provide a convergence guarantee in the case of realizability for a single atom.

Algorithm 2 Distributed Kernel Herding: Dist($\mathcal{X}$, k)

1: **INPUT:**kernel function $k(\cdot, \cdot)$, number of iterations k
2: Partition $\mathcal{X} \supset \mathcal{X}_i, i \in [s]$ uniformly at random and transmit $\mathcal{X}_i$ to machine i
3: //Receive solution sets
4: $\mathsf{G}_i, \mathbf{w}_i \leftarrow$ WKH($\mathcal{X}_i$, k) // run in parallel $\forall i \in [s]$
5: $\mathcal{Y} = \cup_i \mathsf{G}_i$
6: $\mathsf{G}_{s+1}, \mathbf{w}_{s+1} \leftarrow$ WKH($\mathcal{Y}$, k)
7: $i^\star \leftarrow \arg\min_{i \in [s+1]} g(\mathsf{G}_i, \mathbf{w}_i)$
8: return $\mathsf{G}_{i^\star}$, $\mathbf{w}_{i^\star}$

Theorem 3. *Under Assumption 1 with $r = 1$, and Assumptions 2 and 3, if S_i is the sequence of iterates produced by Algorithm 2, the function $g(\cdot)$ converges as*

$\mathbb{E}\min_{\mathbf{w}} g(\mathsf{S}_i, \mathbf{w}) \leq \exp(-\frac{im_\omega}{rM_\Omega})g(\emptyset, 0)$.

Proof Sketch: Note that the final set of filtered iterates outputted are the *best* out of $(s+1)$ possible sequences. The proof tracks the possibilities for $\mathsf{S}_1^\star$ when $\mathcal{X}$ is split. The goal is to then show that under all possible scenarios, at least one of the sequences converges linearly. The convergence of individual sequences is based on the proof techniques used in proof of Theorem 2.

We remark that, for the more general case of $r > 1$, our proof technique does not give a non-trivial convergence rate. This is likely an artifact of our proof technique, and improving it is an interesting open question for future research. Nevertheless, as we shall see in the following section, the algorithm displays improved performance in practice.

6 EXPERIMENTS

We refer the readers to earlier works for empirical evidence of the relative performance of the Herding and SBQ algorithms [Bach et al., 2012, Chen et al., 2010, Welling and Chen, 2010, Welling, 2009a,b, Huszar and Duvenaud, 2012, Ghahramani and Rasmussen, 2003]. In this section, we focus on studying the distributed versions of these algorithms to illustrate the speed / accuracy tradeoff.

6.1 MATCHING A DISTRIBUTION

In this study, our goal is to show the tradeoff between performance and scalability when WKH/SBQ are distributed over multiple machines. Towards this end, we extend an experiment considered in Chen et al. [2010], Huszar and Duvenaud [2012]. In this experiment, the target density is a mixture of 20 two-dimensional Gaussian distributions. Samples are chosen by the contending algorithms, and the MMD distance of the sample distribution to the target distribution is reported for different number of samples.

The sampling subroutine (step 4 in Algorithm 1) requires solving an optimization problem over a continuous domain. To make the problem easier, Chen et al. [2010] and Huszar and Duvenaud [2012] first select 10000 points uniformly at random as the set of potential candidates. They note that this does not affect the performance of chosen samples by much. We also adopt the same methodology with the additional step of arbitrarily partitioning these points over s machines for $s = 1, 5$. Our objective is to illustrate the degradation of performance due to partitioning in Algorithm 2.

The results are reported in Figure 1. SBQ-5 and WKH-5 are distributed versions of SBQ and WKH respectively, run over 5 machines. The labels SBQ and WKH are for their respective single machine versions. Since the search space is split and the search step is parallelized, we receive a five-fold speedup in the algorithms, with a graceful degradation in the reported MMD. We notice a similar pattern of degradation for larger number of machines, but this is omitted from the graph to avoid overcrowding.

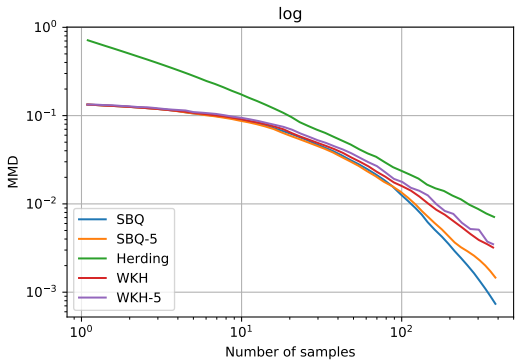
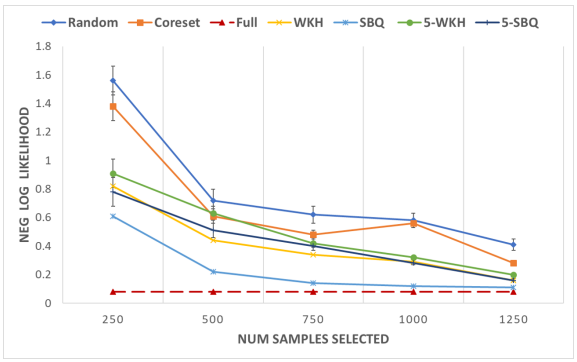


Figure 1: Simulated data results for distributed SBQ/WKH. Herding is WKH with uniform weights. SBQ/WKH are single machine algorithms, while SBQ-5/WKH-5 are their respective distributed versions.

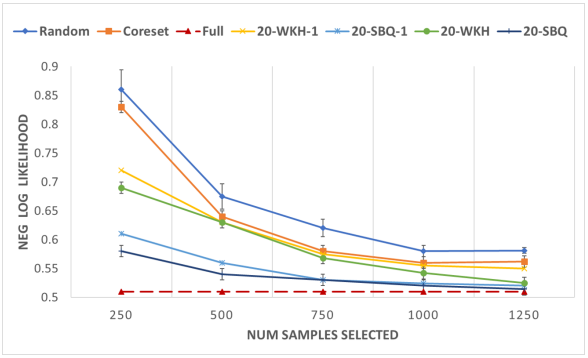
6.2 DATA SUMMARIZATION

In this study, we apply Algorithm 2 to the task of data summarization under logistic regression, as considered by Huggins et al. [2016]. The task of data summarization is as follows. The goal is to select a few data samples that represent the data distribution *sufficiently* well, so that a model built on the selected subsample of the training data does not degrade too much in performance on unseen test data. More specifically, we are interested in approximating the test distribution (i.e., discrete π) using a few samples from the training set. Hence, algorithms such as SBQ and WKH are applicable, provided we have a reasonable kernel function. Recently, Khanna et al. [2019] used SBQ with Fisher Kernels [Jaakkola and Haussler, 1999] for this task. By using the distributed SBQ/WKH over s machines, we obtain a roughly s-times speedup on the run time with minimal loss in the log-likelihood of the selected sample when compared to the results of Khanna et al. [2019] for this task across three different datasets namely `ChemReact`, `CovType` and `WebSpam`.

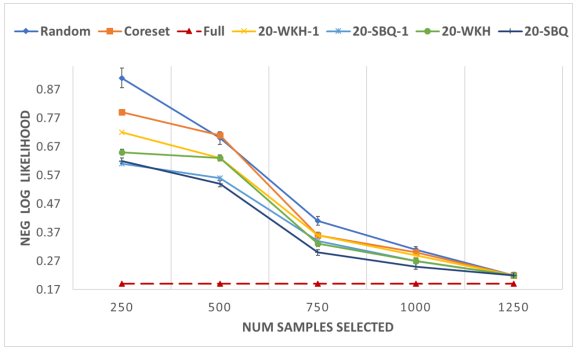
Fisher kernels: For completion, we provide a brief overview of constructing the Fisher kernels. Suppose that we have a parametric model that we learn using maximum likelihood estimation, i.e., $\hat{\theta} := \arg\max \log p(\mathbf{X}|\theta)$, where θ represents the model parameters and $\mathbf{X}$ represents the data. The notion of similarity that Fisher kernels employ is that if two objects are *structurally* similar as the model *sees them*, then slight perturbations in the neighborhood of the fitted parameters $\hat{\theta}$ would impact the fit of the two objects similarly. In other words, the feature embedding $\mathbf{f}_i := \frac{\partial \log p(\mathbf{X}_i|\theta)}{\partial \theta}|_{\theta=\hat{\theta}}$, for an object $\mathbf{X}_i \to \mathbf{f}_i$ can be interpreted as a *feature map-*



(a) ChemReact



(b) CovType



(c) WebSpam

Figure 2: Performance for logistic regression over three datasets for variants of sampling methods. 'Full' reports the numbers for training with the entire training set. For the algorithms s-SBQ and s-WKH, s represents the number of machines used to select the samples. WKH and SBQ represent single machine versions of the algorithms for the smallest dataset, the other two datasets were too big to run on a single machine. The s-WKH-1 experiment is obtained by using only the output of a single split out of s of the dataset run on only one of the machines. Across the three different datasets, the distributed versions of SBQ/WKH proposed in this paper show minimal loss in accuracy while achieving almost linear speedup.

ping which can then be used to define a similarity kernel by a weighted dot product:

$$\kappa(\mathbf{X}_i, \mathbf{X}_j) := \mathbf{f}_i^\top \mathcal{I}^{-1} \mathbf{f}_j,$$

where the matrix $\mathcal{I} := \mathbb{E}_{p(\mathbf{X})}\left[\frac{\partial \log p(\mathbf{X}|\theta)}{\partial \theta}^\top \frac{\partial \log p(\mathbf{X}|\theta)}{\partial \theta}\right]$ is the Fisher information matrix. The information matrix serves to re-scale the dot product, and it is often taken as identity as it loses significance in limit [Jaakkola and Haussler, 1999]. The corresponding kernel is then called the *practical* Fisher kernel, and it is often used in practice.

Another method that also aims to do training data summarization is that of coreset selection [Huggins et al., 2016], albeit with a different goal of reducing the training data size for optimization speedup while still maintaining guaranteed approximation to the training likelihood. Since the goal itself is optimization speedup, coreset selection algorithms typically employ fast methods, while still trying to capture the data distribution by proxy of the training likelihood. Moreover, the coreset selection algorithm is usually closely tied with the respective model, as opposed to being a model-agnostic.

We employ different variants of WKH/SBQ to the problem of training data summarization under logistic regression, as considered by Huggins et al. [2016] using coreset construction. We experiment using three datasets ChemReact, CovType and WebSpam. ChemReact consists of $26,733$ chemicals each of feature size 100. Out of these, 2500 are test data points. The prediction variable is $0/1$ and signifies if a chemical is reactive. CovType has $581,012$ examples each of feature size 54. Out of these, $29,000$ are test points. The task is to predict whether a type of tree is present in each location or not. WebSpam has 350,000 webpages each having 127 features. Out of these, $50,000$ are test data points. The task here is to predict whether a webpage is spam or not. We refer to Huggins et al. [2016] for source of the datasets.

In each of the datasets, we further randomly split the training data into 10% validation and 90% training. We train the logistic regression model on the new training data, and we use the validation set as a proxy to the unseen test set distribution. We build the kernel matrix $\mathbf{K}$ and the affinity vector $\mathbf{z}$, and we run different variants of sampling algorithms to choose samples from the training set to approximate the discrete validation set distribution in the Fisher kernel space. Once the training set samples are extracted, we rebuild the logistic regression model only on the selected samples, and we report negative test likelihood on unseen test data to show how well has the respective algorithm built a model specific dataset summary.

ChemReact is small enough to fit on a single machine, so we run WKH and SBQ on a single machine. To present the tradeoff, we also run 5-WKH and 5-SBQ. These are about 5 times faster than their single machine counterparts, but they degrade in predictive performance. Our aim is to compare

distributed WKH/SBQ against their single machine counterparts. For completeness, we also include the results of coreset selection algorithm and random data selection as implemented by Huggins et al. [2016], since these algorithms were tested by Khanna et al. [2019] on the same problem. To keep the focus on our goal of distributing WKH/SBQ, we do not compare against other coreset algorithms, since coreset construction is not the central goal of this paper. We note that generally SBQ has better performance numbers than WKH for same k across different values of k. Note that `WebSpam` and `CovType` were too big to run on a single machine, and they are thus perfect examples to illustrate the impact and usefulness of the distributed algorithm. All the experiments were run on 12-core 16Gb RAM machines. For all the experiments we conducted, the variance over multiple runs of the distributed algorithm was very low (almost 0), and the trend of relative performance remained the same.

The results are presented in Figure 2. The algorithms we run are WKH, SBQ, s-SBQ and s-WKH, where s represents the number of machines used to select m samples for different values of m. The s-WKH-1 experiment is obtained by using only the output of a single split out of s of the dataset run on only one of the machines (as was done by Khanna et al. [2019] to scale up since they do not have the distributed algorithm to work with). For completeness, we also include "Random" which selects the data points uniformly at random, and "Coreset" which was proposed by Huggins et al. [2016].

7 CONCLUSION

We have analysed two existing algorithms — WKH and SBQ — as well as a new distributed algorithm for estimating expectations. Our results help to bridge the gap between theory and empirical performance by showing that these algorithms perform comparably to the theoretical best possible sampling method over MMD, and exhibit geometric convergence rates for finitely atomic target measures. Our realizability assumption is the key insight that allows us to improve upon previous results. However, we were unable to develop convergence rates for the distributed algorithm over arbitrary atomic target measures using the techniques presented. Developing new methods to address this is the subject of future work.

References

Jason Altschuler, Aditya Bhaskara, Gang Fu, Vahab S. Mirrokni, Afshin Rostamizadeh, and Morteza Zadimoghaddam. Greedy column subset selection: New bounds and distributed algorithms. In *ICML*, 2016.

Francis Bach, Simon Lacoste-Julien, and Guillaume Obozinski. On the equivalence between herding and conditional gradient algorithms. In *ICML*, pages 1355–1362, 2012.

François-Xavier Briol, Chris Oates, Mark Girolami, and Michael A Osborne. Frank-Wolfe Bayesian Quadrature: Probabilistic integration with theoretical guarantees. In *NIPS*, pages 1162–1170, 2015.

François-Xavier Briol, Chris J. Oates, Mark Girolami, Michael A. Osborne, and Dino Sejdinovic. Probabilistic integration: A role in statistical computation? *Statistical Science*, 34(1):1–22, Feb 2019. ISSN 0883-4237. doi: 10.1214/18-sts660.

Yutian Chen, Max Welling, and Alexander J. Smola. Supersamples from kernel herding. In *UAI*, 2010.

Josef Dick and Friedrich Pillichshammer. *Digital Nets and Sequences: Discrepancy Theory and Quasi-Monte Carlo Integration*. Cambridge University Press, New York, NY, USA, 2010.

Marguerite Frank and Philip Wolfe. An algorithm for quadratic programming. *Naval research logistics quarterly*, 1956.

Zoubin Ghahramani and Carl E. Rasmussen. Bayesian Monte Carlo. In *NIPS*, pages 505–512. 2003.

Arthur Gretton, Karsten Borgwardt, Malte Rasch, Bernhard Schölkopf, and Alex J. Smola. A kernel method for the two-sample-problem. In *NIPS*, pages 513–520. 2007.

Jonathan H. Huggins, Trevor Campbell, and Tamara Broderick. Coresets for scalable Bayesian logistic regression. In *NIPS*, pages 4080–4088, 2016.

Ferenc Huszar and David K. Duvenaud. Optimally-weighted herding is Bayesian quadrature. In *UAI*, 2012.

Tommi S. Jaakkola and David Haussler. Exploiting generative models in discriminative classifiers. In *NIPS*, pages 487–493, 1999.

Martin Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML*, pages 427–435, 2013.

Motonobu Kanagawa and Philipp Hennig. Convergence guarantees for adaptive bayesian quadrature methods. In *NeurIPS*, pages 6234–6245. 2019.

Motonobu Kanagawa, Bharath K. Sriperumbudur, and Kenji Fukumizu. Convergence Analysis of Deterministic Kernel-Based Quadrature Rules in Misspecified Settings. *Foundations of Computational Mathematics*, 20(1):155–194, February 2020.

Rajiv Khanna, Ethan Elenberg, Alexandros G. Dimakis, Joydeep Ghosh, and Sahand Negahban. On approximation guarantees for greedy low rank optimization. In *ICML*, 2017.

Rajiv Khanna, Been Kim, Joydeep Ghosh, and Oluwasanmi Koyejo. Interpreting black box predictions using Fisher kernels. In *AISTATS*. PMLR, 2019.

Simon Lacoste-Julien and Martin Jaggi. On the Global Linear Convergence of Frank-Wolfe Optimization Variants. In *NIPS 2015*, pages 496–504, 2015.

Francesco Locatello, Rajiv Khanna, Michael Tschannen, and Martin Jaggi. A unified optimization view on generalized matching pursuit and Frank-Wolfe. In *AISTATS*, 2017.

Anthony O'Hagan. Bayes-Hermite quadrature. *Journal of Statistical Planning and Inference*, 29(3):245 – 260, 11 1991.

Gabriele Santin and Bernard Haasdonk. Convergence rate of the data-independent p-greedy algorithm in kernel-based approximation. *arXiv 1612.02672*, 2016.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Bernhard Schölkopf, and Gert R.G. Lanckriet. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11:1517–1561, August 2010. ISSN 1532-4435.

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

Max Welling. Herding dynamic weights for partially observed random field models. In *UAI*, pages 599–606, 2009a.

Max Welling. Herding dynamical weights to learn. In *ICML*, pages 1121–1128, 2009b.

Max Welling and Yutian Chen. Statistical inference using weak chaos and infinite memory. *Journal of Physics: Conference Series*, 233:012005, jun 2010.