In this paper we highlight and explore the connections between two deterministic sampling and integration methods: Bayesian quadrature (bq) (O’Hagan, 1991; Rasmussen & Ghahramani, 2003) (also known as Bayesian Monte Carlo) and kernel herding (Chen et al., 2010). Bayesian quadrature estimates integral (1) by inferring a posterior distribution over f conditioned on the observed evaluations $f_{x_n}$, and then computing the posterior expectation of $Z_{f,p}$. The points where the function should be evaluated can be found via Bayesian experimental design, providing a deterministic procedure for selecting sample locations.

Herding, proposed recently by Chen et al. (2010), produces pseudosamples by minimising the discrepancy of moments between the sample set $\{x_1, \ldots, x_N\}$ and the target distribution $p$.

**Summary of contributions** In this paper, we make two main contributions. First, we show that the Maximum Mean Discrepancy (MMD) criterion used to choose samples in kernel herding is identical to the expected error in the estimate of the integral $Z_{f,p}$ under a Bayesian procedure prior for $f$. This expected error is the criterion being minimized when choosing samples for Bayesian quadrature. Because Bayesian quadrature assigns different weights to each of the observed function values $f(x)$, we can view Bayesian quadrature as a weighted version of kernel herding. We show that these weights are optimal in a minimax sense over all functions in the Hilbert space defined by our kernel. This implies that Bayesian quadrature dominates uniformly-weighted kernel herding and other non-optimally weighted herding in rate of convergence.

Second, we show that minimizing the MMD, when using BQ weights is closely related to the sparse dictionary selection problem studied in (Krause & Cevher, 2010), and therefore is approximately submodular with respect to the samples chosen. This allows us to reason about the performance of greedy forward selection algorithms for Bayesian Quadrature. We call this greedy method Sequential Bayesian Quadrature (SBQ).

We then demonstrate empirically the relative performance of herding, i.i.d random sampling, and SBQ, and demonstrate that SBQ attains a rate of convergence faster than $O(1/N)$.

## 2 HERDING

Herding was introduced by Welling (2009) as a method for generating pseudo-samples from a distribution in such a way that certain nonlinear moments of the sample set closely match those of the target distribution. The empirical mean $\frac{1}{N} \sum_{n=1}^{N} f_{x_n}$ over these pseudosamples is then used to estimate integral (1).

### 2.1 Maximum Mean Discrepancy

For selecting pseudosamples, herding relies on an objective based on the maximum mean discrepancy (MMD; Sriperumbudur et al., 2010). MMD measures the divergence between two distributions, $p$ and $q$ with respect to a class of integrand functions $\mathcal{F}$ as follows:

$$
\text{MMD}_\mathcal{F}(p, q) = \sup_{f \in \mathcal{F}} \left| \int f_x p(x) dx - \int f_x q(x) dx \right| 
$$

Intuitively, if two distributions are close in the MMD sense, then no matter which function $f$ we choose from $\mathcal{F}$, the difference in its integral over $p$ or $q$ should be small. A particularly interesting case is when the function class $\mathcal{F}$ is functions of unit norm from a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$. In this case, the MMD between two distributions can be conveniently expressed using expectations of the associated kernel $k(x, x')$ only (Sriperumbudur et al., 2010):

$$
\text{MMD}_\mathcal{H}^2(p, q) = \sup_{\|f\|_{\mathcal{H}} = 1} \left| \int f_x p(x) dx - \int f_x q(x) dx \right|^2 
$$

$$
= \|\mu_p - \mu_q\|^2_{\mathcal{H}} 
$$

$$
= \int \int k(x, y) p(x) p(y) dx dy 
$$

$$
- 2 \int \int k(x, y) p(x) q(y) dx dy 
$$

$$
+ \int \int k(x, y) q(x) q(y) dx dy, 
$$

where in the above formula $\mu_p = \int \phi(x) p(x) dx \in \mathcal{H}$ denotes the mean element associated with the distribution $p$. For characteristic kernels, such as the Gaussian kernel, the mapping between a distribution and its mean element is bijective. As a consequence $\text{MMD}_\mathcal{H}(p, q) = 0$ if and only if $p = q$, making it a powerful measure of divergence.

Herding uses maximum mean discrepancy to evaluate of how well the sample set $\{x_1, \ldots, x_N\}$ represents the target distribution $p$. 

378