\[ \epsilon_{\text{herding}}(\{x_1, \ldots, x_N\}) = \text{MMD}_H \left( p, \frac{1}{N} \sum_{n=1}^{N} \delta_{x_n} \right) \]

\[
= \int \int k(x, y)p(x)p(y)dxdy - 2 \frac{1}{N} \sum_{n=1}^{N} k(x_n, x_m) + \frac{1}{N^2} \sum_{n,m=1}^{N} k(x_n, x_m)
\]

(6)

(7)

The herding procedure greedily minimizes its objective \( \epsilon_{\text{herding}}(\{x_1, \ldots, x_N\}) \), adding pseudosamples \( x_n \) one at a time. When selecting the \( n+1 \)-st pseudosample:

\[
x_{n+1} \leftarrow \arg\min_{x \in X} \epsilon_{\text{herding}}(\{x_1, \ldots, x_n, x\})
\]

\[
= \arg\max_{x \in X} 2E_{x' \sim p}[k(x, x')] - \frac{1}{n+1} \sum_{m=1}^{n} k(x, x_m),
\]

assuming \( k(x, x) = \text{const} \). The formula (8) admits an intuitive interpretation: the first term encourages sampling in areas with high mass under the target distribution \( p(x) \). The second term discourages sampling at points close to existing samples.

Evaluating (8) requires us to compute \( E_{x' \sim p}[k(x, x')] \), that is to integrate the kernel against the target distribution. Throughout the paper we will assume that these integrals can be computed in closed form. Whilst the integration can indeed be carried out analytically in several cases (Song et al., 2008; Chen et al., 2010), this requirement is the most pertinent limitation on applications of kernel herding, Bayesian quadrature and related algorithms.

### 2.2 Complexity and Convergence Rates

Criterion (8) can be evaluated in only \( \mathcal{O}(n) \) time. Adding these up for all subsequent samples, and assuming that optimisation in each step has \( \mathcal{O}(1) \) complexity, producing \( N \) pseudosamples via kernel herding costs \( \mathcal{O}(N^2) \) operations in total.

In finite dimensional Hilbert spaces, the herding algorithm has been shown to reduce MMD at a rate \( \mathcal{O}(\frac{1}{N}) \), which compares favourably with the \( \mathcal{O}(\frac{1}{\sqrt{N}}) \) rate obtained by non-deterministic Monte Carlo samplers. However, as pointed out by Bach et al. (2012), this fast convergence is not guaranteed in infinite dimensional Hilbert spaces, such as the RKHS corresponding to the Gaussian kernel.

3 \hspace{1em} \text{BAYESIAN QUADRATURE}

So far, we have only considered integration methods in which the integral (1) is approximated by the empirical mean of the function evaluated at some set of samples, or pseudo-samples. Equivalently, we can say that Monte Carlo and herding both assign an equal \( \frac{1}{N} \) weight to each of the samples.

In (Rasmussen & Ghahramani, 2003), an alternate method is proposed: Bayesian Monte Carlo, or Bayesian quadrature (BQ). BQ puts a prior distribution on \( f \), then estimates integral (1) by inferring a posterior distribution over the function \( f \), conditioned on the observations \( f(x_n) \) at some query points \( x_n \). The posterior distribution over \( f \) then implies a distribution over \( Z_{f,p} \). This method allows us to choose sample locations \( x_n \) in any desired manner. See Figure 2 for an illustration of Bayesian Quadrature.

3.1 \hspace{1em} BQ Estimator

Here we derive the BQ estimate of (1), after conditioning on function evaluations \( f(x_1) \ldots f(x_N) \), denoted as \( f(X) \). The Bayesian solution implies a distribution over \( Z_{f,p} \). The mean of this distribution, \( E[Z] \) is the optimal Bayesian estimator for a squared loss.

For simplicity, \( f \) is assigned a Gaussian process prior with kernel function \( k \) and mean 0. This assumption is very similar to the one made by kernel herding in Eqn. (7).