Optimally Weighted Herding is Bayesian Quadrature

(Huszár and Duvenaud, 2012)

Dino S.

August 23, 2012

Computing Expectations of functions

$$\text{given } (\mathcal{X}, \mathcal{B}, P),\ f : \mathcal{X} \to \mathbb{R}, \qquad \text{compute:} \quad Z_{f,P} \;=\; \int f(x) dP(x)$$

- marginal distributions, posterior moments, Bayes risk

Computing Expectations of functions

$$\text{given } (\mathcal{X}, \mathcal{B}, P), \; f : \mathcal{X} \to \mathbb{R}, \qquad \text{compute:} \quad Z_{f,P} \;=\; \int f(x) dP(x)$$

- marginal distributions, posterior moments, Bayes risk
- **Monte Carlo**: $(\mathrm{x}_j)_{j=1}^n \overset{i.i.d.}{\sim} P \longrightarrow \frac{1}{n} \sum_{j=1}^n f(\mathrm{x}_j)$.
- Law of large numbers: $\left| \frac{1}{n} \sum_{j=1}^n f(\mathrm{x}_j) - Z_{f,P} \right| = \mathcal{O}_P(\frac{1}{\sqrt{n}})$.
  - Often: exact sampling impossible or impractical
  - Sample inexactly: convergence and rates may or may not suffer

Computing Expectations of Functions

$$\text{given } (\mathcal{X}, \mathcal{B}, P),\ f : \mathcal{X} \to \mathbb{R}, \qquad \text{compute:} \quad Z_{f,P} \ = \ \int f(x) dP(x)$$

- marginal distributions, posterior moments, Bayes risk

Computing Expectations of Functions

$$\text{given } (\mathcal{X}, \mathcal{B}, P),\ f : \mathcal{X} \to \mathbb{R}, \qquad \text{compute:} \quad Z_{f,P} \;=\; \int f(x) dP(x)$$

- marginal distributions, posterior moments, Bayes risk
- **Quasi Monte Carlo**: Deterministic sequence $\mathbf{x} = (x_j)_{j=1}^n$, and weights $\mathbf{w} = (w_j)_{j=1}^n$ which are an (approximate) solution of:

$$\arg\min_{(\mathbf{x},\mathbf{w})} \mathrm{d}(P, \sum_{j=1}^{n} w_j \delta_{x_j})$$

for some discrepancy d, suited for the assumption $f \in \mathcal{F} \subset \mathcal{X}^{\mathbb{R}}$.

- potentially results in better rates

Herding

- Let k be a kernel on $\mathcal{X}$, with an RKHS $\mathcal{H}_k$. Kernel herding is a Quasi Monte Carlo with $w_j = 1/n$, and discrepancy d given by the MMD:

$$\gamma_k(P, \frac{1}{n}\sum_{j=1}^{n}\delta_{x_j}) = \left\| \mu_P - \frac{1}{n}\sum_{j=1}^{n} k(\cdot, x_j) \right\|_{\mathcal{H}_k},$$

  where μ_P is the kernel mean embedding of P, and $\mu_{\hat{P}} = \frac{1}{n}\sum_{j=1}^{n} k(\cdot, x_j)$ is the kernel mean embedding of the empirical measure $\hat{P} = \frac{1}{n}\sum_{j=1}^{n}\delta_{x_j}$,

$$\mu_P = \mathbb{E}_{\mathrm{x}\sim P} k(\cdot, \mathrm{x}).$$

- Recall that for $f \in \mathcal{H}_k$, $\langle f, \mu_P \rangle_{\mathcal{H}_k} = \int f(x) dP(x)$.

Herding (2)

- Herding's objective: $\arg\min_{\mathbf{x}} \left\| \mu_P - \frac{1}{n} \sum_{j=1}^{n} k(\cdot, x_j) \right\|_{\mathcal{H}_k}$

Herding (2)

- Herding's objective: $\arg\min_{\mathbf{x}} \left\| \mu_P - \frac{1}{n} \sum_{j=1}^{n} k(\cdot, x_j) \right\|_{\mathcal{H}_k}$
- **Herding is greedy**: already chosen $(x_j)_{j=1}^{r-1}$

$$\begin{aligned} x_r &\leftarrow \arg\min_{x_r} \left\| \mu_P - \frac{1}{r} \sum_{j=1}^{r} k(\cdot, x_j) \right\|_{\mathcal{H}_k}^2 \\ &= \arg\min_{x_r} \left[\|\mu_P\|_{\mathcal{H}_k}^2 + \frac{1}{r^2} \sum_{i=1}^{r} \sum_{j=1}^{r} k(x_i, x_j) - \frac{2}{r} \sum_{j=1}^{r} \mu_P(x_j) \right] \\ &= \arg\max_{x_r} \left[\mu_P(x_r) - \underbrace{\frac{1}{r} \sum_{j=1}^{r} k(x_r, x_j)}_{\mu_{\hat{P}}(x_r)} \right]. \end{aligned}$$

Herding (2)

- Herding's objective: $\arg\min_{\mathbf{x}} \left\| \mu_P - \frac{1}{n} \sum_{j=1}^{n} k(\cdot, x_j) \right\|_{\mathcal{H}_k}$
- **Herding is greedy**: already chosen $(x_j)_{j=1}^{r-1}$

$$
\begin{aligned}
x_r \quad \leftarrow \quad & \arg\min_{x_r} \left\| \mu_P - \frac{1}{r} \sum_{j=1}^{r} k(\cdot, x_j) \right\|_{\mathcal{H}_k}^2 \\
= \quad & \arg\min_{x_r} \left[\|\mu_P\|_{\mathcal{H}_k}^2 + \frac{1}{r^2} \sum_{i=1}^{r} \sum_{j=1}^{r} k(x_i, x_j) - \frac{2}{r} \sum_{j=1}^{r} \mu_P(x_j) \right] \\
= \quad & \arg\max_{x_r} \left[\mu_P(x_r) - \underbrace{\frac{1}{r} \sum_{j=1}^{r} k(x_r, x_j)}_{\mu_{\hat{P}}(x_r)} \right].
\end{aligned}
$$

- mode-seeking behaviour

Bayesian Quadrature

- Nevermind that f is a given function. Bayesians *frequently* put priors on such things[1]
- f becomes a random function, Z_f becomes a random variable induced by f. Denote $f(\mathbf{x}) = (f(x_j))_{j=1}^n$

[1] things like RKHS functions, and the priors like Gaussian process priors

Bayesian Quadrature

- Nevermind that f is a given function. Bayesians *frequently* put priors on such things[1]
- f becomes a random function, Z_f becomes a random variable induced by f. Denote $f(\mathbf{x}) = (f(x_j))_{j=1}^n$
- Posterior mean of $Z_f | f(\mathbf{x})$:

$$\begin{aligned}
\mathbb{E}\left[Z_f | f(\mathbf{x})\right] &= \int \left[\int f(x) dP(x)\right] p(f | f(\mathbf{x})) df \\
&= \int \left[\int f(x) p(f | f(\mathbf{x})) df\right] dP(x) \\
&= Z_{\mathbb{E}[f | f(\mathbf{x})]} \\
&= \mu_P(\mathbf{x})^\top K^{-1} f(\mathbf{x}) \\
&= \sum_{j=1}^n w_j(\mathbf{x}) f(x_j).
\end{aligned}$$

[1] things like RKHS functions, and the priors like Gaussian process priors

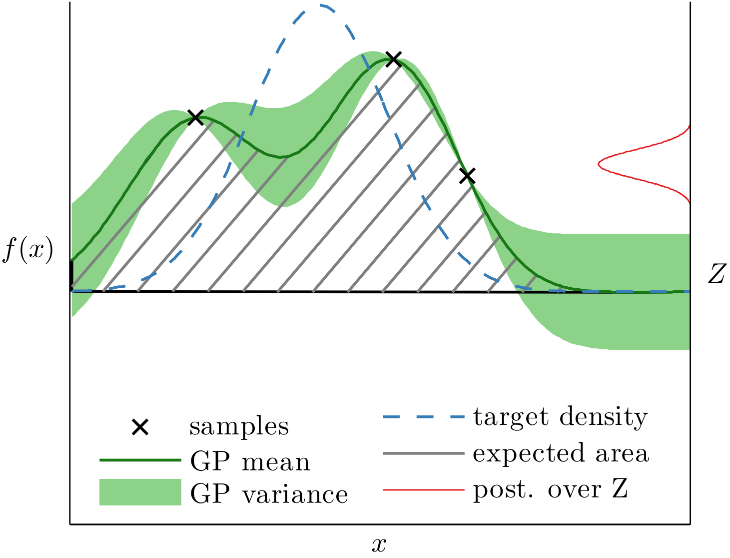
Bayesian Quadrature (2)

- Integration of a *known* function f evaluated at a finite number of points $\rightarrow$ estimation of a smooth, easy to integrate function $\hat{f}$, consistent with values at a finite number of points + integration of $\hat{f}$

$$\int f(x)dP(x) \approx \int \mathbb{E}[f|f(\mathbf{x})](x)dP(x) \quad = \quad \sum_{j=1}^{n} w_j(\mathbf{x})f(x_j)$$

- "Smoothness" assumption implied through the Gaussian process framework
- $\mathbf{w}$ is a deterministic function of $\mathbf{x}$ (does not depend on f!)
- No need for weights to sum up to a probability distribution **nor** to be positive...

Bayesian Quadrature (3)



Connection to MMD

- Full posterior of $Z_f | f(\mathbf{x})$, not just the mean,

$$
\begin{aligned}
Var\left[Z_f | f(\mathbf{x})\right] &= \mathbb{E}_f \left(Z_f - \sum_{j=1}^{n} w_j(\mathbf{x}) f(x_j) \right)^2 \\
&= \mathbb{E}_f \left\langle f, \mu_P - \sum_{j=1}^{n} w_j(\mathbf{x}) k(\cdot, x_j) \right\rangle^2 \\
&= \left\| \mu_P - \underbrace{\sum_{j=1}^{n} w_j(\mathbf{x}) k(\cdot, x_j)}_{\mu_{\hat{P}_{\mathbf{w}}}} \right\|_{\mathcal{H}_k}^2
\end{aligned}
$$

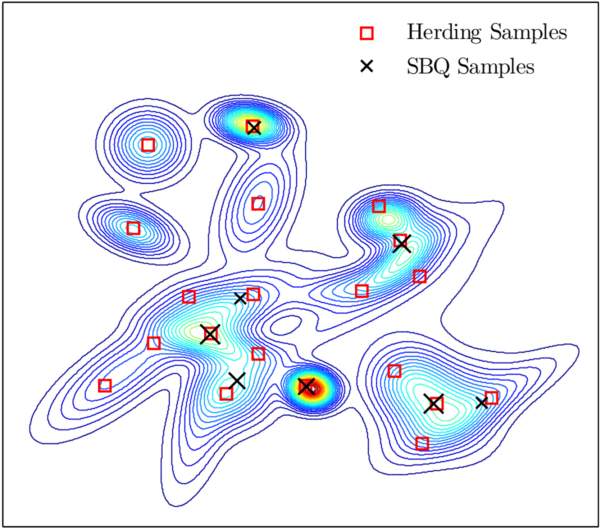
- Comes from $\langle f, g \rangle \sim \mathcal{N}(0, \|g\|_{\mathcal{H}_k}^2)$
- No dependence on $f(\mathbf{x})$, just on $\mathbf{x}$

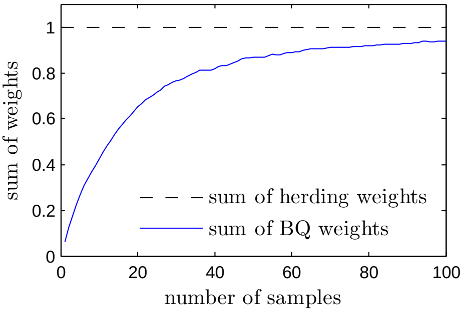
Sequential Bayesian Quadrature (SBQ)

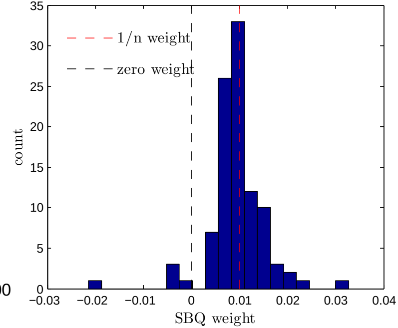
- Now we have a framework to find "optimal" weights given $\mathbf{x}$.
- **Greedy BQ optimization**: already chosen $(x_j)_{j=1}^{r-1}$

$$
\begin{aligned}
x_r &\leftarrow \arg\min_{x_r} Var\left[Z_f | f(\mathbf{x})\right] \\
&= \arg\max_{x_r} \left[\mu_P(x_r) - \underbrace{\sum_{j=1}^{r} w_j k(x_r, x_j)}_{\mu_{\hat{P}_{\mathbf{w}}}(x_r)} \right] \\
&= \arg\max_{x_r} \mu_P(\mathbf{x})^\top K^{-1} \mu_P(\mathbf{x}).
\end{aligned}
$$

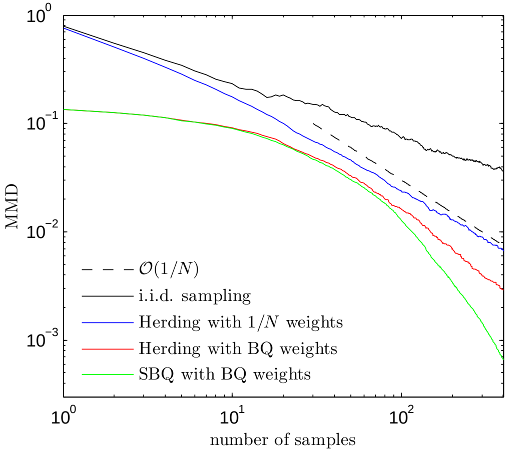
Experiments




1
0.8
0.6
0.4
0.2
0
0
20
40
60
80
100
sum of weights
number of samples
sum of herding weights
sum of BQ weights

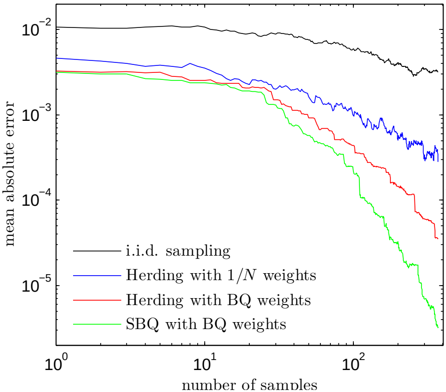

35
30
25
20
15
10
5
0
−0.03
−0.02
−0.01
0
0.01
0.02
0.03
0.04
count
SBQ weight
1/n weight
zero weight

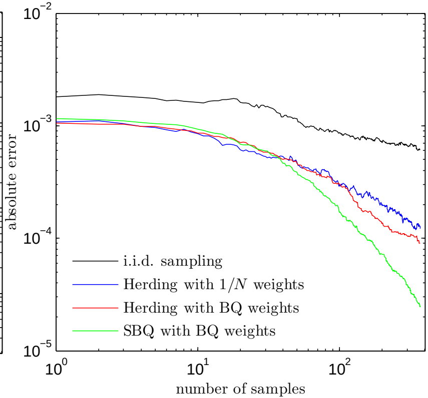
How does the MMD decrease?



SBQ: $\mathcal{O}(\frac{1}{n^{1+?}})$?

Performance for f in- and out-of-model





Rates

method	complexity	rate	guarantee
MCMC	$\mathcal{O}(N)$	variable	ergodic theorem
i.i.d. MC	$\mathcal{O}(N)$	$\frac{1}{\sqrt{N}}$	law of large numbers
herding	$\mathcal{O}(N^2)$	$\frac{1}{\sqrt{N}} \geq \cdot \geq \frac{1}{N}$	(Chen et al., 2010; Bach et al., 2012)
SBQ	$\mathcal{O}(N^3)$	unknown	approximate submodularity