Bayesian Methods of Machine Learning. Project Presentation. Super-Samples from Kernel Herding

Georgii Novikov

October 2020

Skoltech
Skolkovo Institute of Science and Technology

Super-Samples from Kernel Herding

Weakly chaotic, non-linear dynamical system

$$
\begin{aligned}
x_{t+1} &= \operatorname*{arg\,max}_{x \in \mathcal{X}} \langle w_t, \phi(x) \rangle \qquad (1) \\
w_{t+1} &= w_t + \mathbb{E}_{x \sim p}[\phi(x)] - \phi(x_{t+1}) \\
x_t &\in \mathbb{R}^n, \phi : \mathbb{R}^n \to \mathbb{R}^m
\end{aligned}
$$

Under some conditions, it is a greedy minimization of

$$
\varepsilon_T^2 = \left\| \mu_p - \frac{1}{T} \sum_{t=1}^{T} \phi(x_t) \right\|^2, \text{where } \mu_p = \mathbb{E}_{x \sim p}\, \phi(x) \qquad (2)
$$

Skoltech
Skolkovo Institute of Science and Technology

Theoretical Guarantees

Theorem

x_t in (1) is optimal on each step $\Rightarrow$ error in (2) decreases at a rate $\mathcal{O}(T^{-1})$.

Advantage

1. I.i.d samples have rate $\mathcal{O}(T^{-\frac{1}{2}})$
2. MCMC converges even slower then $\mathcal{O}(T^{-\frac{1}{2}})$ (due to positive correlations)

Kernel Trick

We want to replace $\phi(\mathrm{x})$ with $\mathrm{k}(\mathrm{x}, \mathrm{x}') = \langle \phi(\mathrm{x}), \phi(\mathrm{x}') \rangle$:

1. Define $\mathrm{w}_0 = \mu := \mathbb{E}_{\mathrm{x} \sim \mathrm{p}}[\phi(\mathrm{x})]$
2.

$$\begin{aligned}
\mathrm{x}_{\mathrm{t}+1} &= \operatorname*{arg\,max}_{\mathrm{x} \in \mathcal{X}} \langle \mathrm{w}_\mathrm{t}, \phi(\mathrm{x}) \rangle \\
&= \operatorname*{arg\,max}_{\mathrm{x} \in \mathcal{X}} \langle \mathrm{w}_0 + \mathrm{T}\, \mathbb{E}_{\mathrm{x}' \sim \mathrm{p}}[\phi(\mathrm{x}')] - \sum_{\mathrm{t}=1}^{\mathrm{T}} \phi(\mathrm{x}_\mathrm{t}), \phi(\mathrm{x}) \rangle \\
&= \operatorname*{arg\,max}_{\mathrm{x} \in \mathcal{X}} (\mathrm{T}+1)\, \mathbb{E}_{\mathrm{x}' \sim \mathrm{p}}\, \mathrm{k}(\mathrm{x}, \mathrm{x}') - \sum_{\mathrm{t}=1}^{\mathrm{T}} \mathrm{k}(\mathrm{x}, \mathrm{x}_\mathrm{t})
\end{aligned}$$

Interpretation

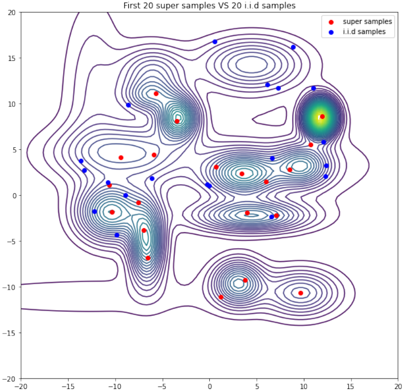
$$\varepsilon_T^2 = \left\| \mu_p - \frac{1}{T}\sum_{t=1}^{T} \phi(x_t) \right\|_{\mathcal{H}}^2$$

$$= \mathbb{E}_{x,x' \sim p}\, k(x,x') - \frac{2}{T}\sum_{i=1}^{T} \mathbb{E}_{x\sim p}\, k(x,x_t) + \frac{1}{T^2}\sum_{t,t'=1}^{T} k(x,x')$$

This is a measure between distribution p and empirical distribution $\hat{p}_T(x) = \sum_{t=1}^{T} \delta(x, x_t)$. And we still have a rate of convergence $\mathcal{O}(T^{-1})$!

Expiriment 0. Toy dataset

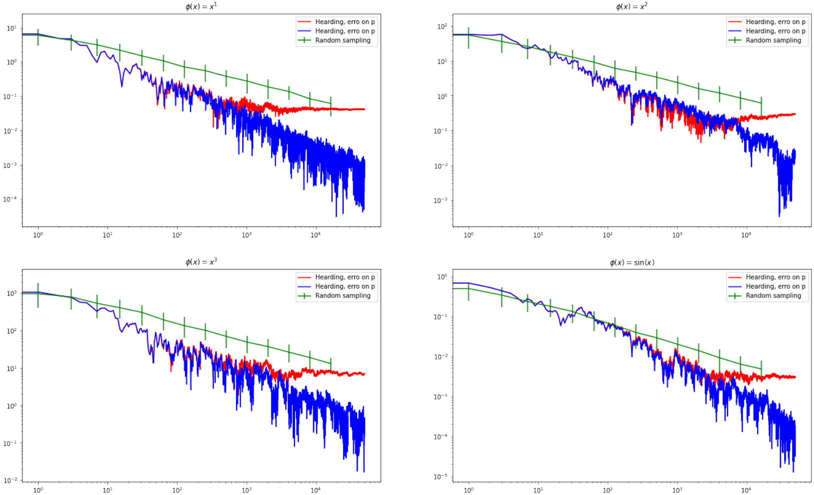
Mixture of 20 2D Gaussians:



Expiriment 1. Empirical Matching

1. p(x) = mixture of 10 5D Gaussians
2. $\mathcal{D} = 10^5$ i.i.d samples
3. Gaussian kernel k (with $\sigma = 10$)
4. Herding vs random samples on 4 functions of interest:
   4.1 $\phi(x) = x^i, i \in 1, 2, 3$
   4.2 $\phi(x) = \sin(x)$

Expiriment 1. Empirical Matching. Results



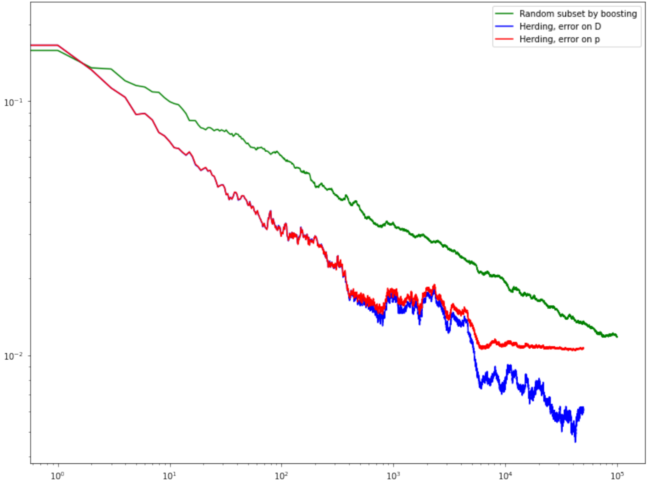
Expiriment 2. Approximating the Bayesian Posterior

- UCI spambase: 57 features. 4601 samples (3000 for train and 1601 for test).
- Whiten with PCA.
- Sample 10^7 logistic regression parameters from the poster distribution with gaussian prior with Metropolis-Hasting and subsample to 10^5 to reduce the autocorrelation.
- Whiten parameters with PCA.
- Herd super samples.
- Compare with metric

$$\mathrm{RMSE}(S_T, D) = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{1}{T}\sum_{t=1}^{T} p(y_n|x_n,\theta_t) - \frac{1}{|D|}\sum_{d=1}^{|D|} p(y_n|x_n,\theta_d\right]$$

Skoltech
Skolkovo Institute of Science and Technology

Expiriment 2. Approximating the Bayesian Posterior. Result



Problems

Dependence on the parameters of Gaussian kernel is not clear.