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

$$x_{t+1} = \arg\max_{x \in \mathcal{X}} \langle w_t, \phi(x) \rangle \tag{1}$$

$$w_{t+1} = w_t + \mathbb{E}_{x \sim p}[\phi(x)] - \phi(x_{t+1})$$

$$x_t \in \mathbb{R}^n, \phi : \mathbb{R}^n \to \mathbb{R}^m$$

Under some conditions, it is a greedy minimization of

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

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 $\phi(x)$ with $k(x, x') = \langle \phi(x), \phi(x') \rangle$:

1. Define $w_0 = \mu := \mathbb{E}_{x \sim p}[\phi(x)]$
2.

$$x_{t+1} = \arg\max_{x \in \mathcal{X}} \langle w_t, \phi(x) \rangle$$

$$= \arg\max_{x \in \mathcal{X}} \langle w_0 + T\, \mathbb{E}_{x' \sim p}[\phi(x')] - \sum_{t=1}^{T} \phi(x_t), \phi(x) \rangle$$

$$= \arg\max_{x \in \mathcal{X}} (T+1)\, \mathbb{E}_{x' \sim p}\, k(x, x') - \sum_{t=1}^{T} k(x, x_t)$$

Interpretation

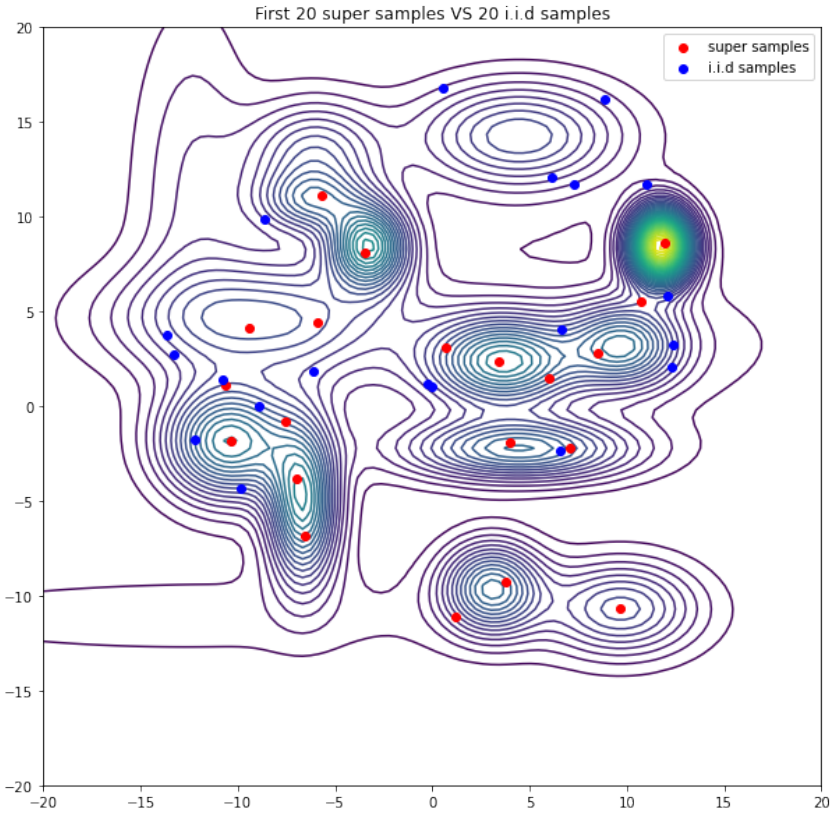
$$\varepsilon_{\mathrm{T}}^2 = \left\| \mu_{\mathrm{p}} - \frac{1}{\mathrm{T}} \sum_{\mathrm{t}=1}^{\mathrm{T}} \phi(\mathrm{x}_{\mathrm{t}}) \right\|_{\mathcal{H}}^2$$

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

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

Expiriment 0. Toy dataset

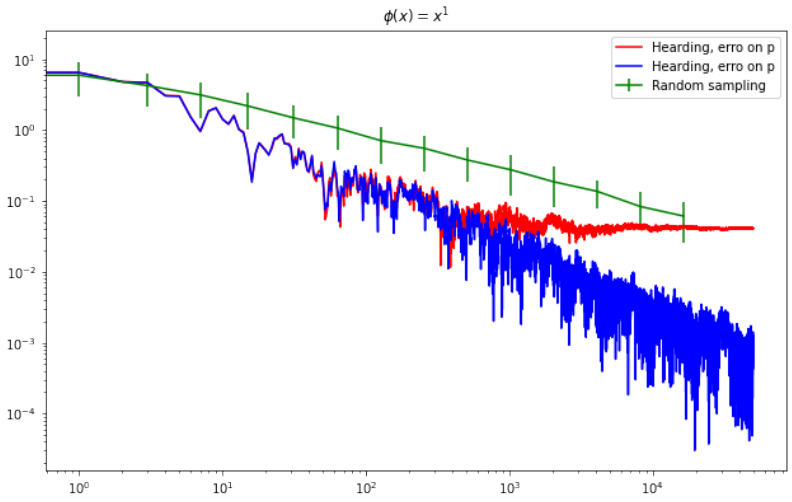
Mixture of 20 2D Gaussians:

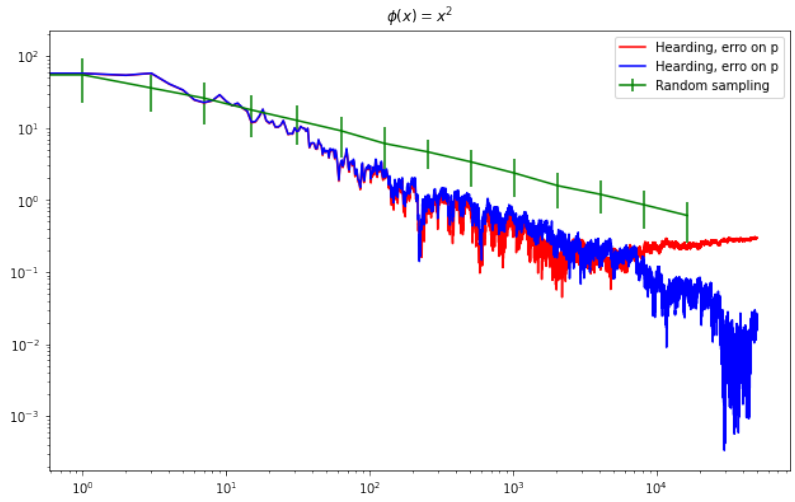


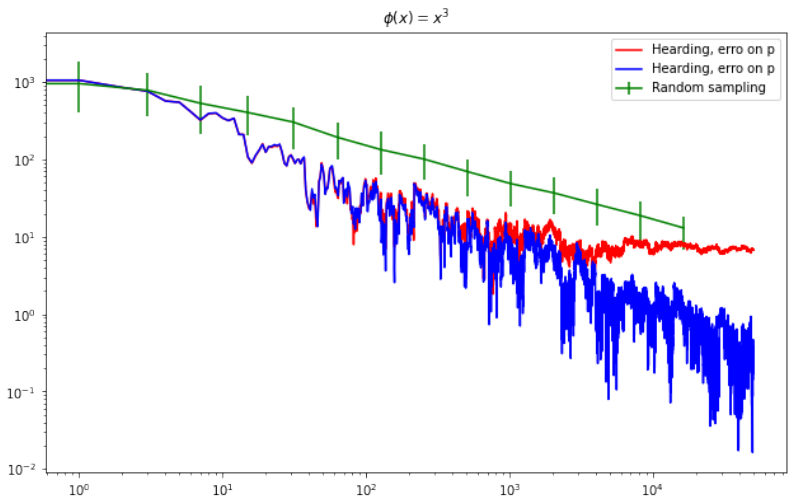
Expiriment 1. Empirical Matching

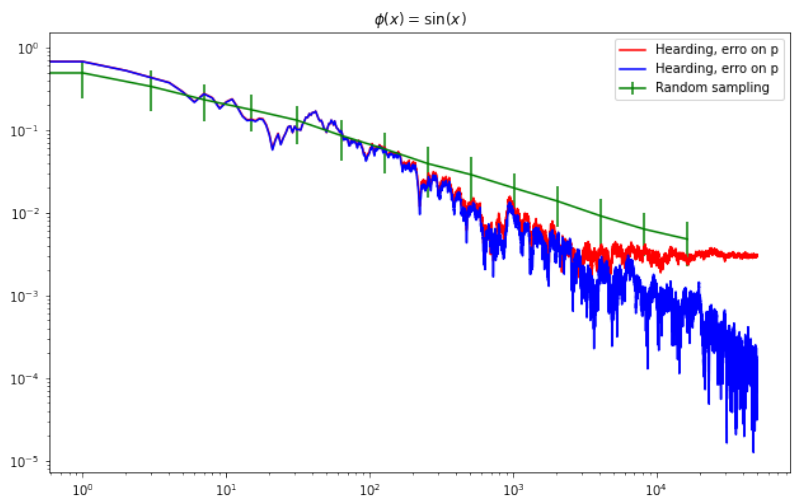
1. $\mathrm{p(x)}$ = mixture of 10 5D Gaussians
2. $\mathcal{D} = 10^5$ i.i.d samples
3. Gaussian kernel k (with $\sigma = 10$)
4. Herding vs random samples on 4 functions of interest:
   - 4.1 $\phi(\mathrm{x}) = \mathrm{x}^{\mathrm{i}}, \mathrm{i} \in 1, 2, 3$
   - 4.2 $\phi(\mathrm{x}) = \sin(\mathrm{x})$

Expiriment 1. Empirical Matching. Results







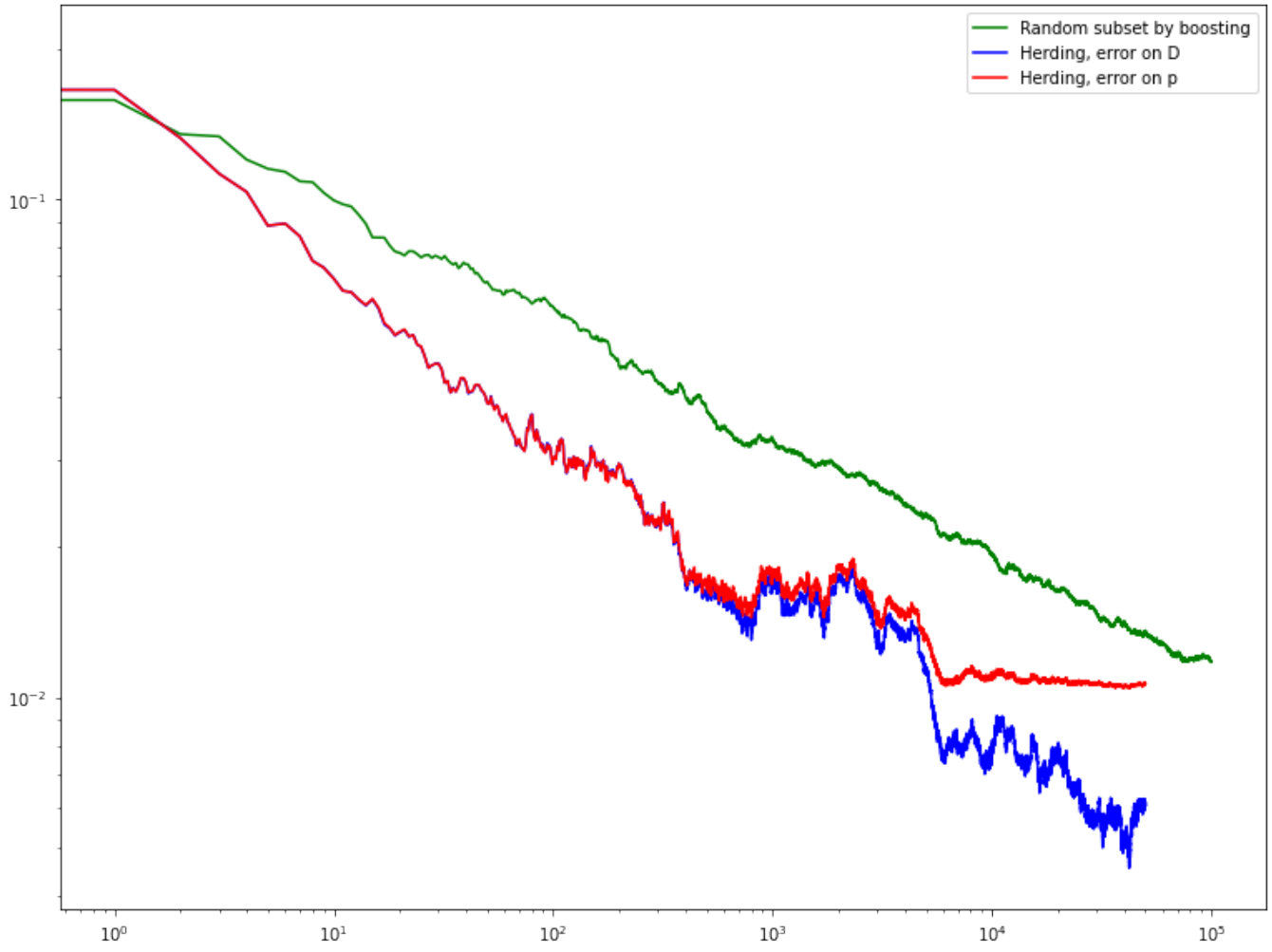


Expiriment 2. Approximating the Bayesian Posterior

- 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

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

Expiriment 2. Approximating the Bayesian Posterior. Result



Problems

Dependence on the parameters of Gaussian kernel is not clear.