Provable Nonparametric Bayesian Inference

Bo Dai, Niao He, Hanjun Dai, Le Song

Georgia Institute of Technology

Motivation

Goal: scale up Bayesian inference with provable guarantee.

- Simplicity. Applicable to many probabilistic models, even with non-conjugate priors. Only require loglikelihood rather than its derivative.
- Flexibility. Approximate the posterior by kernel density estimation which could capture the multimodality.
- Stochasticity. Use a subset of the data each iteration.
- Theoretical guarantee. Converges to the true posterior in terms of KL-divergence in rate $O(\frac{1}{\sqrt{m}})$ with m particles.

Optimization View of Bayesian Inference

Given model $p(x|\theta)$ and prior $p(\theta)$, with the dataset $X = \{x_n\}_{n=1}^N$, the posterior of $\theta \in \mathbb{R}^d$ computed by **Bayes' rule**

$$p(\theta|X) = \frac{p(\theta)\prod_{n=1}^N p(x_n|\theta)}{\int \prod_{n=1}^N p(x_n|\theta)p(\theta)d\theta}$$

[Zellner(1988)] proposed that the posterior could be viewed as the solution of

$$\min_{q(\theta)\in\mathcal{P}} L(q) := KL(q(\theta)\,||\,p(\theta)) - \sum_{n=1}^N \left[\int q(\theta)\log p(x_n|\theta)\,d\theta\right],$$

which is 1-strongly convex w.r.t. KL-divergence.

Stochastic Mirror Descent in Density Space

The functional gradient $\nabla L(q)$ is defined as

$$L(q+\epsilon h) = L(q) + \epsilon\left\langle \nabla L(q), h\right\rangle_2 + O(\epsilon^2).$$

Randomly sample x_t from X, the stochastic functional gradient of $L(q)$ in L_2 is

$$g_t(\theta) = \log(q(\theta)) - \log(p(\theta)) - N\log p(x_t|\theta).$$

The stochastic mirror descent algorithm iteratively solves prox-mapping [Nemirovski et al.(2009)]

$$q_{t+1}(\theta) = \mathbf{P}_{q_t}(\gamma_t g_t) := \mathrm{argmin}_{\widetilde{q}(\theta)\in\mathcal{P}}\left\{\left\langle \widetilde{q}(\theta), \gamma_t g_t(\theta)\right\rangle_{L_2} + KL(\widetilde{q}(\theta)||q_t(\theta))\right\}$$

which leads to update

$$q_{t+1}(\theta) = q_t(\theta)\exp(-\gamma_t g_t(\theta))/Z = q_t(\theta)^{1-\gamma_t}p(\theta), \qquad (1)$$

where $Z := \int q_t(\theta)\exp(-\gamma_t g_t(\theta))\,d\theta$ is generally intractable.

Error Tolerant Stochastic Mirror Descent

Given $\epsilon \geqslant 0$ and $g \in L_2$, we define the ϵ-prox-mapping of q as the set

$$\mathbf{P}_q^{\epsilon}(g) := \{\widehat{q} \in \mathcal{P} : KL(\widehat{q}||q) + \langle g, \widehat{q}\rangle_{L_2} \leqslant \min_{\widetilde{q}\in\mathcal{P}}\{KL(\widetilde{q}||q) + \langle g, \widetilde{q}\rangle_{L_2}\} + \epsilon\}$$

Instead of solving prox-mapping exactly, we apply the updates

$$\widetilde{q}_{t+1}(\theta) \in \mathbf{P}_{\widetilde{q}_t}^{\epsilon_t}(\gamma_t g_t),\ t = 1, 2, \ldots.$$

Recall the objective function is 1-strongly convex, we have the recurrence, $\forall t \leqslant T$,

$$\mathbb{E}[KL(q^*||\widetilde{q}_{t+1})] \leqslant \epsilon_t + (1-\gamma_t)\mathbb{E}[KL(q^*||\widetilde{q}_t)] + \frac{\gamma_t^2\mathbb{E}\|g_t\|_\infty^2}{2}$$

Approximation using Weighted Particles

When the prior has the same support as posterior, based on the exact solution (1), we approximate $q_{t+1}(\theta)$ as a set of weighted particles

$$\tilde{q}_{t+1}(\theta) = \sum_{i=1}^m \alpha_i^{t+1}\delta(\theta_i),$$

$$\alpha_i^{t+1} := \frac{\alpha_i^t\exp(-\gamma_t g_t(\theta_i))}{\sum_{i=1}^m \alpha_i^t\exp(-\gamma_t g_t(\theta_i))}, \qquad \{\theta_i\}_{i=1}^m \overset{i.i.d.}{\sim} p(\theta).$$

One can simply update the set of working variables $\{\alpha_i\}_{i=1}^m$ in the algorithm.

Approximation using Weighted Kernel Density Estimator

In general, we may not have a good guess for the support of posterior. We propose weighted kernel density estimator as the approximation to $q(\theta)$. In t-step, we have $\widetilde{q}_t$ from last iteration, we derive the update rule from (1)

$$\widetilde{q}_{t+1}(\theta) = \sum_{i=1}^m \alpha_i K_h(\theta - \theta_i),$$

$$\alpha_i := \frac{\exp(-\gamma_t g_t(\theta_i))}{\sum_{i=1}^m \exp(-\gamma_t g_t(\theta_i))}, \quad \{\theta_i\}_{i=1}^m \overset{i.i.d.}{\sim} \widetilde{q}_t(\theta),$$

where $h > 0$ is the bandwidth parameter and $K_h(\theta) := \frac{1}{h^d}K(\theta/h)$ is a smoothing kernel.

Remark: 1) The update serves as an ϵ-prox-mapping. 2) The sampling procedure adjusts the support of intermediate estimation. 3) The computation of α_i does not need to evaluate $Z := \int q_t(\theta)\exp(-\gamma_t g_t(\theta))d\theta$.

Particle Mirror Descent Algorithm

Particle Mirror Descent

Input: Data set $X = \{x_n\}_{n=1}^N$, prior $p(\theta)$

Output: posterior density estimator $\tilde{q}_T(\theta)$

Initialize $\widetilde{q}_1(\theta) = p(\theta)$

for $t = 1, 2, \ldots, T-1$ **do**

  Sample $x_t \overset{unif.}{\sim} X$

  if Good $p(\theta)$ is provided **then**

    $\{\theta_i\}_{i=1}^{m_t} \overset{i.i.d.}{\sim} \pi(\theta)$ when $t = 1$

    $\alpha_i \leftarrow \alpha_i^{1-\gamma_t}p(x_t|\theta_i)^{N\gamma_t}, \forall i$

    $\alpha_i \leftarrow \frac{\alpha_i}{\sum_{i=1}^{m_t}\alpha_i}, \forall i$

    $\widetilde{q}_{t+1}(\theta) = \sum_{i=1}^{m_t}\alpha_i\,\delta(\theta_i)$

  else

    $\{\theta_i\}_{i=1}^{m_t} \overset{i.i.d.}{\sim} \widetilde{q}_t(\theta)$

    $\alpha_i \leftarrow \tilde{q}_t(\theta_i)^{-\gamma_t}p(\theta_i)^{\gamma_t}p(x_t|\theta_i)^{N\gamma_t}, \forall i$

    $\alpha_i \leftarrow \frac{\alpha_i}{\sum_{i=1}^{m_t}\alpha_i}, \forall i$

    $\widetilde{q}_{t+1}(\theta) = \sum_{i=1}^{m_t}\alpha_i K_{h_t}(\theta - \theta_i)$

  end if

end for

Theoretical Guarantees

Theorem 1 *Assume $p(\theta)$ has the same support as the true posterior $q^*(\theta)$, and the model $\|p(x|\theta)^N\|_\infty$ is bounded for any x. Then $\forall f(\theta)$ bounded and integrable, with stepsize $\gamma_t = \frac{\eta}{t}$, after m iteration, the PMD returns m weighted particles such that*

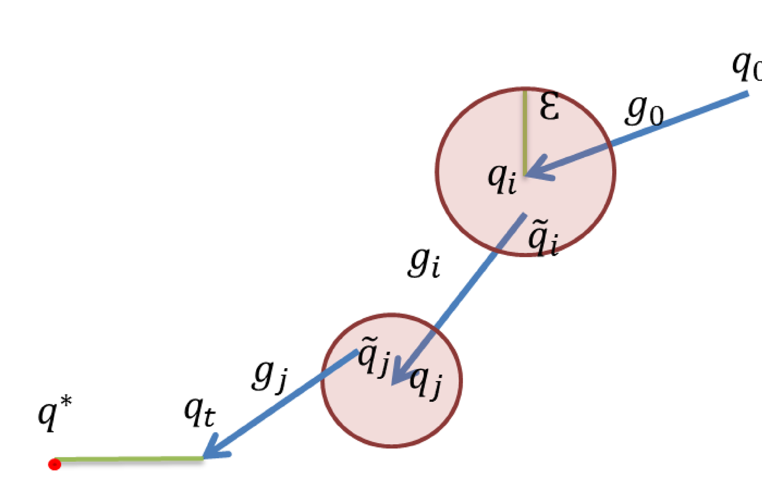
$$\mathbb{E}\left[|\langle\widetilde{q} - q^*, f\rangle|\right] \sim O(\frac{1}{\sqrt{m}}).$$

Theorem 2 *With some mild assumptions about kernel function, and the $p(\theta)$ and $p(x|\theta)$ are smooth enough, with stepsize $\gamma_t \sim O(1/t)$, after $\sqrt{m}$ iteration, the PMD returns weighted KDE such that*

$$\mathbb{E}[KL(q^*||\widetilde{q})] \sim O(\frac{1}{\sqrt{m}})$$

Proof idea Denote $\varrho_m(\theta) = \frac{1}{m}\sum_{i=1}^m \omega(\theta_i)K_h(\theta, \theta_i)$, we have $\mathbb{E}[\varrho_m(\theta)] = \mathbb{E}_{\theta_i}[\omega(\theta_i)K_h(\theta, \theta_i)] = q \star K_h$. The error can be decomposed as follows.

$$\begin{aligned}\epsilon &:= \mathbb{E}\,\|\widetilde{q}(\theta) - q(\theta)\|_1 \\ &\leqslant \underbrace{\mathbb{E}\,\|\widetilde{q}(\theta) - \varrho_m(\theta)\|_1}_{\text{normalization error}} \\ &+ \underbrace{\mathbb{E}\,\|\varrho_m(\theta) - \mathbb{E}\,\varrho_m(\theta)\|_1}_{\text{sampling error (variance)}} \\ &+ \underbrace{\|\mathbb{E}\,\varrho_m(\theta) - q(\theta)\|_1}_{\text{approximation error (bias)}}\end{aligned}$$

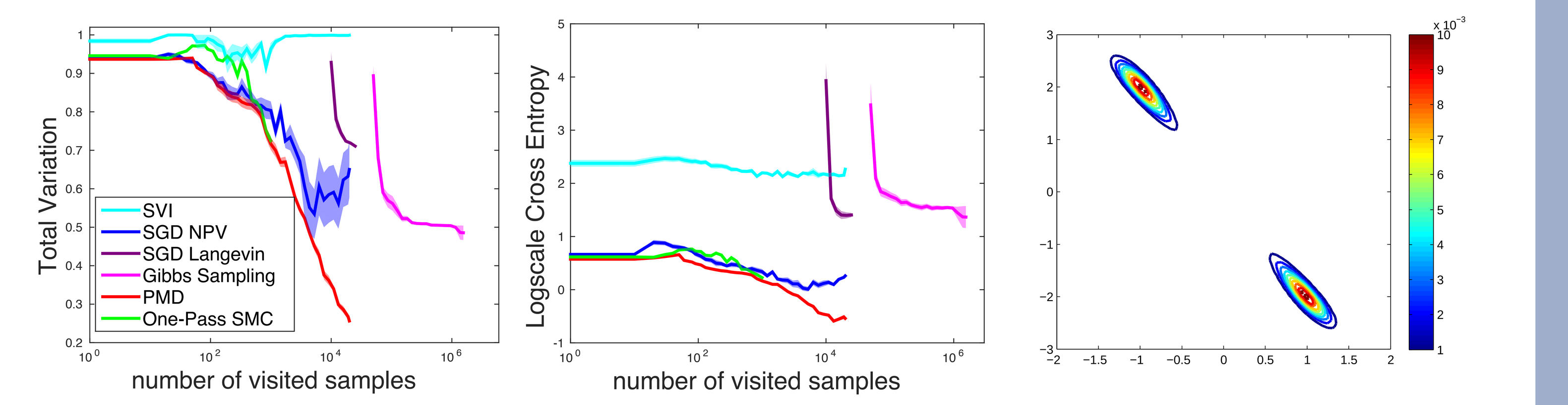


Experiments

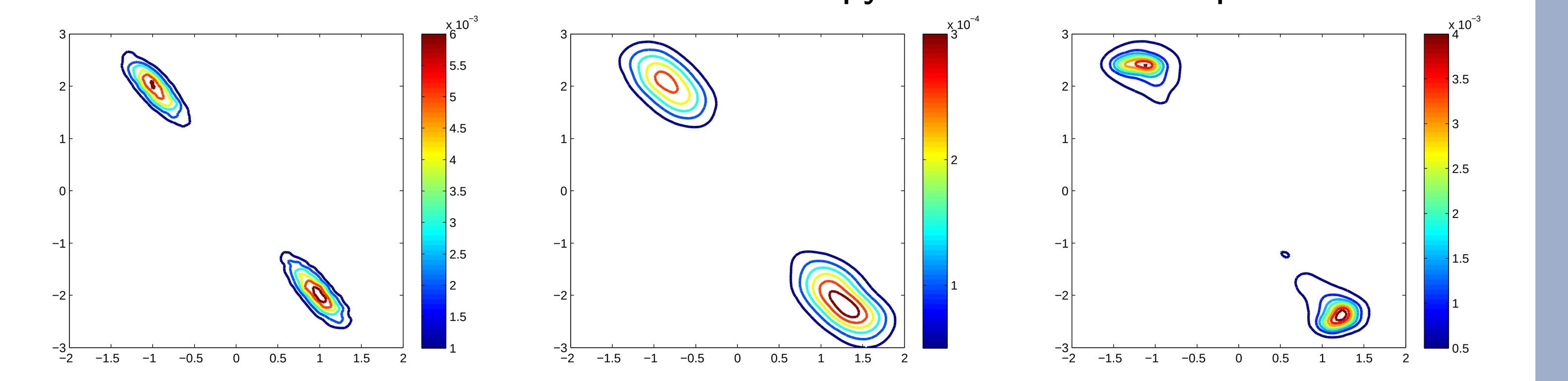
- Verification on multimodal model. We compare the alternatives on the mixture model

$$\theta_1 \sim \mathcal{N}(0, \sigma_1^2), \quad \theta_2 \sim \mathcal{N}(0, \sigma_2^2)$$
$$x_i \sim p\mathcal{N}(\theta_1, \sigma_x^2) + (1-p)\mathcal{N}(\theta_1+\theta_2, \sigma_x^2)$$

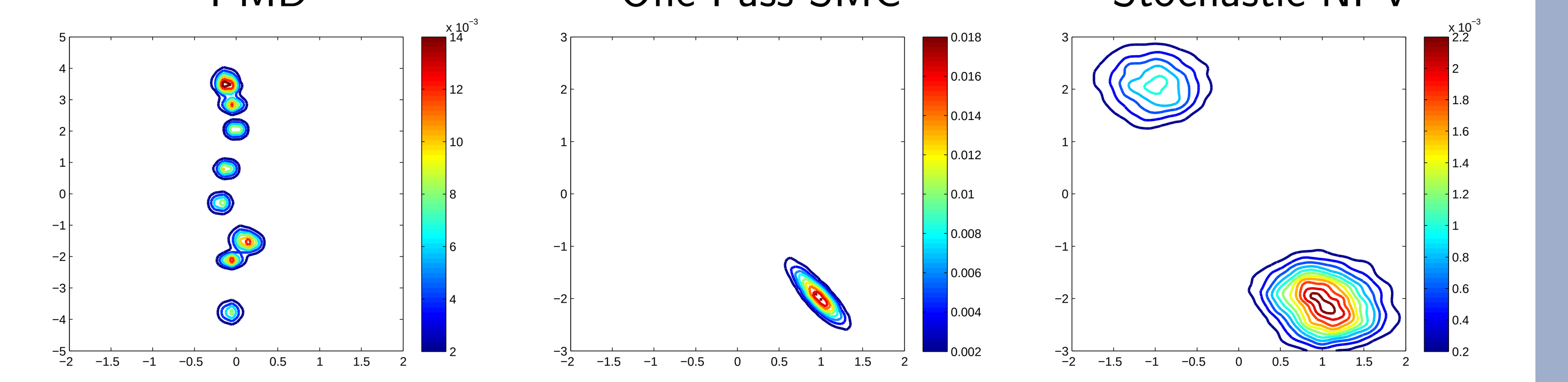
where $\sigma_1 = 1$, $\sigma_2 = 1$, $\sigma_x = 2.5$ and $p = 0.5$. The size of dataset is 1000.



Total Variation — Cross Entropy — True posterior

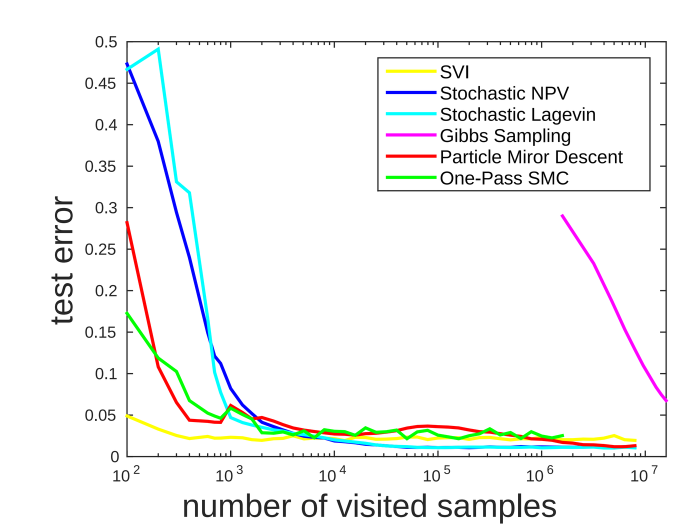


PMD — One-Pass SMC — Stochastic NPV

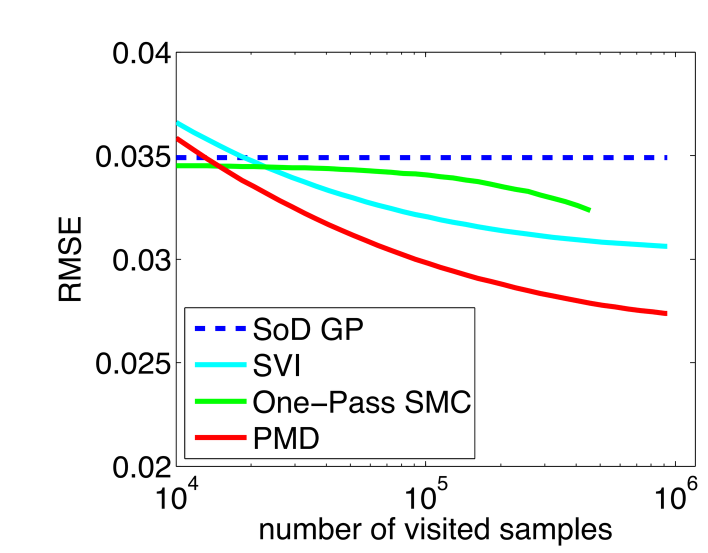


SVI — Gibbs Sampling — SGD Langevin

- Verification on non-conjugate model. We conduct comparison with logistic regression model on dataset MNIST8M 8 vs. 6 which contains about 1.6M data points.
- Verification on real-world application. We conduct comparison with sparse Gaussian Processes model on predicting the year of songs. The dataset contains 0.5M songs.



(1) Logistic Regression



Sparse Gaussian Processes

Reference

Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A.
Robust stochastic approximation approach to stochastic programming.
SIAM J. on Optimization, 19(4):1574–1609, January 2009.

Zellner, Arnold.
Optimal Information Processing and Bayes's Theorem.
The American Statistician, 42(4), November 1988.

Georgia Institute of Technology

Mailto: {bodai, nhe6, hdai8}@gatech.edu, lsong@cc.gatech.edu