Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm

Qiang Liu Dilin Wang
Dartmouth College

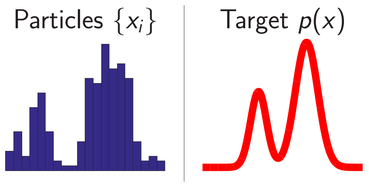
Introduction

Challenges of scalable Bayesian inference

- MCMC: often slow; difficult to access the convergence
- Variational Inference: critically depends on the set of distributions in which the approximation is defined

Stein Variational Gradient Descent (SVGD)

- Directly minimizes $\mathrm{KL}(\{x_i\} \,||\, p)$.
  1. no need to define variational approximation family
  2. leverages the gradient information



Main Idea

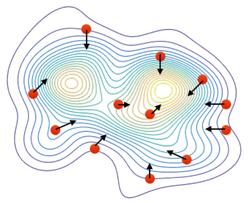
- Idea: Iteratively move $\{x_i\}_{i=1}^n$ towards the target p by updates of form

$$x_i' \leftarrow x_i + \epsilon \phi(x_i), \qquad (1)$$

where ϕ is a perturbation direction chosen to maximumly decrease the KL divergence with p, that is,

$$\phi = \arg\max_{\phi \in \mathcal{F}} \left\{ -\frac{\partial}{\partial \epsilon} \mathrm{KL}(q_{[\epsilon\phi]} \,||\, p)\big|_{\epsilon=0} \right\}, \qquad (2)$$

where $q_{[\epsilon\phi]}$ is the density of $x' = x + \epsilon\phi(x)$ and $\mathcal{F}$ is a set of perturbation directions that we optimize over.



- How to find the optimal ϕ ?

Stein Variational Gradient Descent (SVGD)

- It turns out the objective in (2) is a simple linear functional of ϕ,

$$-\frac{\partial}{\partial \epsilon} \mathrm{KL}(q_{[\epsilon\phi]} \,||\, p)\big|_{\epsilon=0} = \mathbb{E}_{x\sim q}[\mathcal{A}_p \phi(x)]$$

$$\text{with} \quad \mathcal{A}_p \phi(x) \overset{\text{def}}{=} \nabla_x \log p(x)^\top \phi(x) + \nabla_x \cdot \phi(x)$$

- Therefore, the optimization in (2) reduces to

$$\mathcal{D}(q||p) \overset{\text{def}}{=} \max_{\phi \in \mathcal{F}} \left\{ \mathbb{E}_{x\sim q}[\mathcal{A}_p \phi(x)] \right\} \qquad (3)$$

- Stein's Identity: $\mathbb{E}_{x\sim q}[\mathcal{A}_p \phi(x)] = 0$ iff $q = p$

Stein Variational Gradient Descent (SVGD) (Cont.)

- Take $\mathcal{F}$ to be the unit ball of a vector-valued reproducing kernel Hibert space (RKHS) $\mathcal{H}$. [Liu et al., 16] showed that the optimal solution of (3) has a simple closed form:

$$\begin{aligned}\phi^*(x') &\propto \mathbb{E}_{x\sim q}[\mathcal{A}_p k(x, x')] \\ &= \mathbb{E}_{x\sim q}[\nabla_x \log p(x) k(x, x') + \nabla_x k(x, x')]\end{aligned}$$

- Approximating $\mathbb{E}_{x\sim q}$ by using empirical average of the current particles $\{x_i\}_{i=1}^n$, (1) reduces to,

$$x_i \leftarrow x_i + \epsilon \hat{\mathbb{E}}_{x\sim\{x_i\}_{i=1}^n}[\nabla_x \log p(x) k(x, x_i) + \nabla_x k(x, x_i)] \qquad (4)$$

Algorithm

Algorithm 1 Bayesian Inference via Variational Gradient Descent

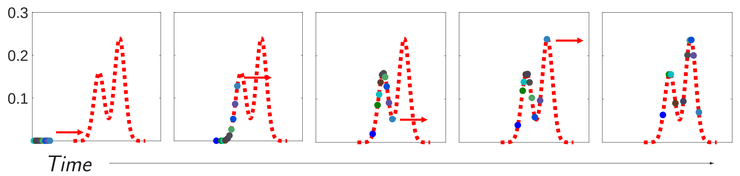
Input: A target distribution with density function $p(x)$ and a set of initial particles $\{x_i^0\}_{i=1}^n$.
Output: A set of particles $\{x_i\}_{i=1}^n$ that approximates the target distribution.

Repeat

$$x_i \leftarrow x_i + \epsilon \hat{\mathbb{E}}_{x\sim\{x_i\}_{i=1}^n}[\underbrace{\nabla_x \log p(x)}_{\text{gradient}} k(x, x_i) + \underbrace{\nabla_x k(x, x_i)}_{\text{repulsive force}}], \quad \forall i = 1 \cdots n.$$

where ϵ is the step size.

- $\nabla_x \log p(x)$: moves the particles $\{x_i\}$ towards high probability regions of $p(x)$.
- $\nabla_x k(x, x')$: enforce diversity in $\{x_i\}$ (otherwise all x_i collapse to modes of $p(x)$).
- Algorithm 1 reduces to a single chain of typical gradient ascent for MAP when the number of particles $n = 1$.



Complexity and Efficient Implementation

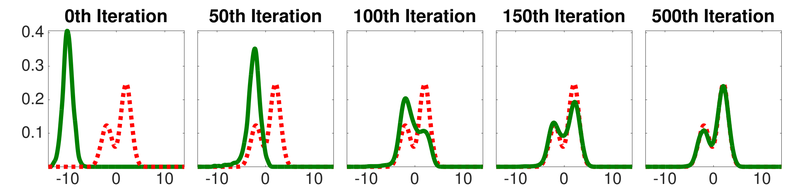
- In big data settings, $p(x) \propto p_0(x) \prod_{k=1}^N p(D_k \mid x)$ with a very large N
- Approximate $\nabla_x \log p(x)$ with subsampled mini-batches

$$\nabla_x \log p(x) \approx \nabla_x \log p_0(x) + \frac{N}{|\Omega|} \sum_{k\in\Omega} \nabla_x \log p(D_k \mid x)$$

Empirical Results

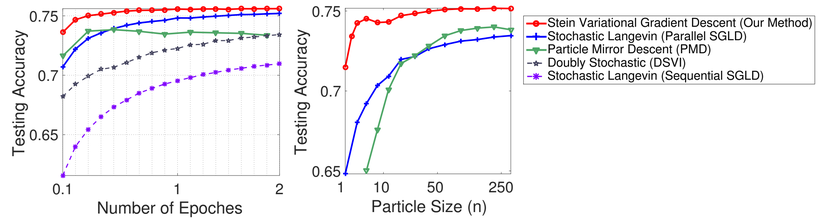
Toy Example on 1D Gaussian Mixture

- Target distribution, $p(x) = 1/3\mathcal{N}(x; -2, 1) + 2/3\mathcal{N}(x; 2, 1)$
- Initialization: $\mathcal{N}(x; -10, 1)$; 100 particles



Bayesian Logistic Regression

- Test on Convertype dataset with 581,012 data points
- Compared with Stochastic Langevin [Welling et al., 11], Particle Mirror Descent [Dai et al., 16] and Doubly Stochastic [Lázaro-Gredilla, 14]



(a) Particle size $n = 100$ (b) Results at 3000 iteration (≈ 0.32 epoches)

Bayesian Neural Network

- Test Bayesian neural nets on UCI datasets (with 20 particles)
- Compared with probabilistic back-propagation (PBP) [Hernández-Lobato and Adams, 15]

	Avg. Test RMSE		Avg. Test LL		Avg. Time (Secs)	
Dataset	**PBP**	**Our Method**	**PBP**	**Our Method**	**PBP**	**Ours**
Boston	2.977 ± 0.093	$\mathbf{2.957 \pm 0.099}$	-2.579 ± 0.052	$\mathbf{-2.504 \pm 0.029}$	18	**16**
Concrete	5.506 ± 0.103	$\mathbf{5.324 \pm 0.104}$	-3.137 ± 0.021	$\mathbf{-3.082 \pm 0.018}$	33	**24**
Energy	1.734 ± 0.051	$\mathbf{1.374 \pm 0.045}$	-1.981 ± 0.028	$\mathbf{-1.767 \pm 0.024}$	25	**21**
Kin8nm	0.098 ± 0.001	$\mathbf{0.090 \pm 0.001}$	0.901 ± 0.010	$\mathbf{0.984 \pm 0.008}$	118	**41**
Naval	0.006 ± 0.000	$\mathbf{0.004 \pm 0.000}$	3.735 ± 0.004	$\mathbf{4.089 \pm 0.012}$	173	**49**
Combined	4.052 ± 0.031	$\mathbf{4.033 \pm 0.033}$	-2.819 ± 0.008	$\mathbf{-2.815 \pm 0.008}$	136	**51**
Protein	4.623 ± 0.009	$\mathbf{4.606 \pm 0.013}$	-2.950 ± 0.002	$\mathbf{-2.947 \pm 0.003}$	682	**68**
Wine	0.614 ± 0.008	$\mathbf{0.609 \pm 0.010}$	-0.931 ± 0.014	$\mathbf{-0.925 \pm 0.014}$	26	**22**
Yacht	$\mathbf{0.778 \pm 0.042}$	0.864 ± 0.052	$\mathbf{-1.211 \pm 0.044}$	-1.225 ± 0.042	25	25
Year	$8.733 \pm \mathrm{NA}$	$\mathbf{8.684 \pm \mathrm{NA}}$	$-3.586 \pm \mathrm{NA}$	$\mathbf{-3.580 \pm \mathrm{NA}}$	7777	**684**

Our code is available at
`https://github.com/DartML/Stein-Variational-Gradient-Descent`

References

- Qiang Liu, Jason D. Lee, Michael I. Jordan. A Kernelized Stein Discrepancy for Goodness-of-fit Tests and Model Evaluation. *arXiv preprint arXiv:1602.03253*, 2016
- Welling, Max and Teh, Yee W. Bayesian learning via stochastic gradient Langevin dynamics. *IMCL*, 2011
- B. Dai, N. He, H. Dai, and L. Song. Provable Bayesian inference via particle mirror descent. *AISTATS*, 2016
- Lázaro-Gredilla, Miguel, Doubly stochastic variational Bayes for non-conjugate inference, *ICML*, 2014
- Hernández-Lobato, José Miguel and Adams, Ryan P. Probabilistic backpropagation for scalable learning of bayesian neural networks. *ICML*, 2015

Acknowledgment

This work is supported in part by NSF CRII 1565796.

mail: qiang.liu@dartmouth.edu www: http://www.cs.dartmouth.edu/~qliu/