2nd Symposium on Advances in Approximate Bayesian Inference, 2019 1–7

Stein Variational Gradient Descent for Approximate Bayesian Computation

Chunlin Ji chunlin.ji@kuang-chi.org
Kuang-Chi Institute of Advanced Technology
Shenzhen, China

Jiangsheng Yi yijiangsheng@hotmail.com
Zhejiang University
Hangzhou, China

Wanchuang Zhu wanchuang.zhu@sydney.edu.au
Centre for Translational Data Science & School of Mathematics and Statistics, University of Sydney
Sydney, Australia

Abstract

Approximate Bayesian Computation (ABC) provides a generic framework of Bayesian inference for likelihood-free models, but sampling based posterior approximation is often time-consuming and has difficulty accessing the convergence. Stochastic variational inference forms the posterior inference to a optimization problem and enable the ABC scalable for large dataset. However, complex simulation models involved in ABC always lead to complex posteriors, which is not easy to approximate by simple parametric variational distributions. We draw upon recent advances in the implicit model of variational distribution and introduce the Stein variational gradient descent (SVGD) approach to approximate the posterior by nonparametric particles. We also find that the kernel in the SVGD algorithm helps in reducing the large variance of the gradient estimators of ABC likelihood. Moreover, energy distance is proposed as the statistics in the evaluation of ABC likelihood, which reduce the difficulty in selecting proper statistics. Simulation studies are provided to demonstrate the correctness and efficiency of our algorithm.

1. Introduction

ABC constitutes a class of computational methods that can be used to estimate the posterior distributions of various complex models, where the analytical likelihood is elusive or the likelihood is computationally very costly to evaluate.Various Monte Carlo sampling methods has been proposed for ABC computation. Sampling methods are widely used in ABC literature: rejection sampling (Tavaré et al., 1997), Markov Chain Monte Carlo (MCMC)(Paul Marjoram and Tavaré, 2003), and population-based sampling (Beaumont et al., 2009; Moral et al., 2006; Sisson et al., 2007). However, these methods tend to converge slowly and requires many calls to the simulator, making them ineffective for large-scale problems. Variational inference provides an alternative for Bayesian inference by forming the posterior inference to an optimization problem. In previous work (Moreno et al., 2016), they provide a general framework to incorporate variational inference (VI) (Jordan et al., 1999; Hoffman et al., 2012) with ABC and emphasis the variance reduction issue in using the ABC likelihood. Another difficulty of variational inference for ABC problem is that ABC problem always involve complex posteriors, such as multi-modal distribution or long tail distribution. Simple parametric variational distribution can not approximate such posteriors well. Recently several implicit models, such as stein variational gradient descent (Liu and Wang, 2016), normalizing flows (Rezende and Mohamed, 2015), are proposed for variational distribution to extend the ability to approximate complex posteriors. In this work, we study the SVGD in ABC. We discuss two advantage in SVGD for ABC, first, the SVGD is an nonparametric method, which can in theory approximate any complex posterior given sufficient particles; second, SVGD tends to smooth the gradient and reduce its variance due to the kernel term in the evaluation of the gradient, so it benefits the gradient of ABC. Moreover, the selection of statistics is a trick work, we provide a statistics-energy distance, which has shown being a good choice in most of case.

2. Stein Variational inference ABC

2.1. Variational inference for ABC

VI frames the posterior estimation to an optimization problem by introduce a surrogate loss, the evidence lower bound (ELBO), $\mathcal{L} = \int q(\theta;\lambda)[\log(p(\boldsymbol{y}|\theta)p(\theta)) - \log q(\theta;\lambda)]$ (Jordan et al., 1999), which is a lower bound of the model evidence $\log p(\boldsymbol{y})$, where $\boldsymbol{y}$ denotes the given dataset. When the likelihood $p(\boldsymbol{y}|\theta)$ is intractable or extremely expensive to compute, the ABC method introduce a ϵ-kernel to approximate $p(\boldsymbol{y}|\theta)$ (Moreno et al., 2016),

$$p_\epsilon(\boldsymbol{y}|\theta) = \int K_\epsilon[S(\boldsymbol{y}), S(\boldsymbol{x})]p(\boldsymbol{x}|\theta) \approx \frac{1}{M}\sum_{m=1}^{M} K_\epsilon[S(\boldsymbol{y}), S(\boldsymbol{x}^{(m)})] \tag{1}$$

where the simulator generates synthetic data x according to the parameters θ, $S(\boldsymbol{y})$ and $S(\boldsymbol{x})$ are summary statistics to represent the entire deta set, K_ϵ measures the discrepancy between $S(\boldsymbol{y})$ and $S(\boldsymbol{x})$ with a controlling parameter bandwidth ϵ. Replace the true likelihood $p(\boldsymbol{y}|\theta)$ with the ABC likelihood, we have

$$\mathcal{L} = \int q(\theta;\lambda)\log\int K_\epsilon[S(\boldsymbol{y}), S(\boldsymbol{x})]p(\boldsymbol{x}|\theta)d\boldsymbol{x}d\theta - \mathrm{KL}[q(\theta;\lambda)||p(\theta)] \tag{2}$$

then we replace expectations by samples to obtain the noisy estimation of $\mathcal{L}$ and apply the SGD algorithm to optimize the variational parameter λ iteratively. But the parametric form of variational distribution $q(\theta;\lambda)$ limits its ability to approximate complex $p(\boldsymbol{y}|\theta)$.

2.2. Stein variational gradient descent

Conventional variational inference approximates the target distribution $p(\theta|\boldsymbol{y})$ using a simple distribution $q(\theta;\lambda)$ found in a predefined set of distributions by minimizing the KL divergence $\mathrm{KL}[q(\theta;\lambda)||p(\theta|\boldsymbol{y})]$. The choice of the proposal distribution $q(\theta;\lambda)$ is critical, while simple parametric form prevents it to express complex posterior. Recently implicit models are proposed to utilize a set of distributions obtained by smooth transforms from a tractable reference distribution to approximate the posterior (Rezende and Mohamed, 2015). Stein variational inference belongs to the family of VI with implicit model. It takes a set of particles $\{\theta_i^0\}_{i=1}^n$ from a tractable reference distribution $q_0(\theta)$ with iteration of the form below:

$$\theta_i^{t+1} \leftarrow \theta_i^t + \epsilon_t \hat{\phi}^*(\theta_i^t), \quad \text{where} \quad \hat{\phi}^*(\theta) = \frac{1}{n}\sum_{j=1}^{n}[k(\theta_j^t, \theta)\nabla_{\theta_j^t} \log p(\theta_j^t|\boldsymbol{y})) + \nabla_{\theta_j^t} k(\theta_j^t, \theta)],$$

and ϵ_t is a step size. The form of $\hat{\phi}^*(\cdot)$ is from Stein identity(Liu et al., 2016), and $\hat{\phi}^*(\cdot)$ is optimal perturbation direction in reproducing kernel Hilbert space (RKHS) $\mathcal{H}^d$ as

$$\hat{\phi}^* = \underset{\phi \in \mathcal{H}^d}{\operatorname{argmax}} \{-\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{KL}[q_{\epsilon\phi}||p(|\boldsymbol{y})]\Big|_{\epsilon=0}\},$$

where $q_{\epsilon\phi}$ denotes the density of $T(\theta) = \theta + \epsilon\phi(\theta)$. This relation is the key founding of SVGD in (Liu and Wang, 2016). Furthermore, $\max_{\phi \in \mathcal{H}^d}\{-\frac{\mathrm{d}}{\mathrm{d}\epsilon}\mathrm{KL}[q_{\epsilon\phi}||p(|\boldsymbol{y})]\Big|_{\epsilon=0}\}$ is defined as kernelized Stein discrepancy (KSD), which is a powerful measurement of the difference between two distributions (Liu et al., 2016). So when we iteratively transform θ by $T(\theta)$, the KL-divergence $\mathrm{KL}(q_T||p)$ decrease at the steepest descent direction, which make q_T approximates p gradually.

SVGD holds MCMC's consistency and VI's efficiency, and it can deal with complex distribution particle-efficiently due to its diversity. (Liu, 2017; Liu et al., 2019; Lu et al., 2019) have studied SVGD's convergence in different perspectives. And SVGD also has many extensions and applications (Han and Liu, 2018; Gong et al., 2019; Wang and Liu, 2019).

2.3. Energy distance

In the ABC likelihood evaluation, we have to select proper statistics $S(\cdot)$ and ϵ-kernel. If two or more statistics are selected, we have to weighting these statistics, which introduce extra efforts. Here, we introduce energy distance as the ϵ-kernel. Energy distance (Rizzo, 2003) between d-dimensional independent random variable X and Y is defined as follows, $\mathcal{E}(X,Y) = 2\mathbb{E}||X-Y||_p - \mathbb{E}||X-X^{'}||_p - \mathbb{E}||Y-Y^{'}||_p$, where $\mathbb{E}||X||_p < \infty$, $\mathbb{E}||Y||_p < \infty$, $X^{'}, Y^{'}$ are iid copy of X, Y respectively. Energy distance measures the distance between two distributions. Therefore it can be used in various fields, including two-sample test (Székely and Rizzo, 2004), one sample goodness-of-fit test (Székely and Rizzo, 2005). It has been shown in (Székely and Rizzo, 2013) that, $\mathcal{E}(X,Y) \geq 0$ with equality to zero if and only if X and Y are identically distributed. Let $\boldsymbol{x} = \{x_1, \cdots, x_{n_1}\}$ and $\boldsymbol{y} = \{y_1, \cdots, y_{n_2}\}$ denote two independent random sample sets drawn from two distribution $F_{\boldsymbol{x}}$ and $F_{\boldsymbol{y}}$ respectively. Energy statistics is defined as follows,

$$\begin{aligned} \mathcal{E}(\boldsymbol{x}, \boldsymbol{y}) &= \frac{n_1 n_2}{n_1+n_2}(\frac{2}{n_1 n_2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_2}||x_i - y_j||_p && (3) \\ &- \frac{1}{n_1^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}||x_i - x_j||_p - \frac{1}{n_2^2}\sum_{i=1}^{n_2}\sum_{j=1}^{n_2}||y_i - y_j||_p), && (4) \end{aligned}$$

where $||\cdot||_p$ denotes the p-norm.

2.4. The algorithm

To implement the SVGD algorithm to approximate the posterior in ABC, we first draw a set of particles $\{\theta_i^0\}_{i=1}^n$ from a simple initial distribution q_0, and then iteratively update the particles with a empirical smooth transform $\hat{T}(\theta) = \theta + \epsilon\hat{\phi}^*(\theta)$. This procedure allows to deterministically transport the points $\{\theta_i\}_{i=1}^n$ to match the posterior distribution. The kernel $k(\theta, \theta')$ in the perturbation direction affects the diversity of these particles, which forces the particles to spread enough to cover complex posterior. Moreover, the SVGD tends to smooth the gradient due to the kernel $k(\theta, \theta')$ which weights the gradient at other particles positions. This variance reduction behavior benefits the convergence of the optimization procedure. In the implementation of the algorithm, we obtain the numerical gradient in $\phi^*(\theta)$ leveraging on the automatic differentiation tools, such as Pytorch (Paszke et al., 2017). Advanced SGD type algorithms, such as Adam (Kingma and Ba, 2014), help in accelerating the convergence. In addition, the bandwidth of the kernel in $\hat{\phi}^*$ can be selected by heat equation (HE) method from (Liu et al., 2019). The ϵ in $p_\epsilon(\boldsymbol{y}|\theta)$ matters bias and variance tradeoff like the way (Wilkinson, 2013), and specific form of energy statistics can be chosen as randomly projected energy statistics (RPES) (Huang and Huo, 2017), which can speed up the computation of energy statistics by fast approximations. The proposal algorithm is summarized in Algorithm 1.

Algorithm 1: Stein variational gradient descent ABC

```
• Input: Initialize $p$ for the $p$-norm of energy distance, temprature $\tau$ in $\epsilon$ -kernel, and a set of initial particles $\{\theta_i^0\}_{i=1}^n$.
• For $t = 0 : T$
    – For each particles $\theta_i^t$, run the simulator to produce out put $\boldsymbol{x}^{(1:M)}$
    – Evaluate the ABC likelihood $p_\epsilon(\boldsymbol{y}|\theta_i^t)$ using equation (1)
    – Update each particles according to $\theta_i^{t+1} = \theta_i^t + \epsilon_t\hat{\phi}^*(\theta_i^t)$, where $\hat{\phi}^*(\theta) = \frac{1}{n}\sum_{j=1}^n \Big[k(\theta_j^t, \theta)\nabla_{\theta_j^t}\log[p_\epsilon(\boldsymbol{y}|\theta_j^t)p(\theta_j^t)] + \nabla_{\theta_j^t}k(\theta_j^t, \theta)\Big]$, $\epsilon_t$ is the step size at the $t$-iteration
• Output: a set of particles $\{\theta_i\}_{i=1}^n$ that approximates the target distribution.
```

3. Simulation studies

Bimodal posterior: consider a linear regression model that contains a univariate response variable y_i with a p-dimensional predictor X_i for $i = 1, ..., n$. The regression coefficient is denoted by $\theta \in \mathbb{R}^p$; i.e., $y_i = X_i^T\theta + \epsilon_i$, where $\epsilon_i \sim N(0, \sigma^2)$. We consider the case of $p = 1$ and $n = 500$, where X_i is i.i.d. standard Gaussian. The data-generating process follows $y_i = \gamma_i X_i - (1 - \gamma_i)X_i + \epsilon_i$, where $\gamma_i \sim \mathit{Bernoulli}(1/2)$ and $\epsilon_i \sim N(0, 0.2^2)$. The scatter plot of the synthetic data set is illustrated in Figure 1 (left), which has the 'scissors' like shape. The SVGD algorithm takes 30 particles and a mini-batch of 50 data. We plot the

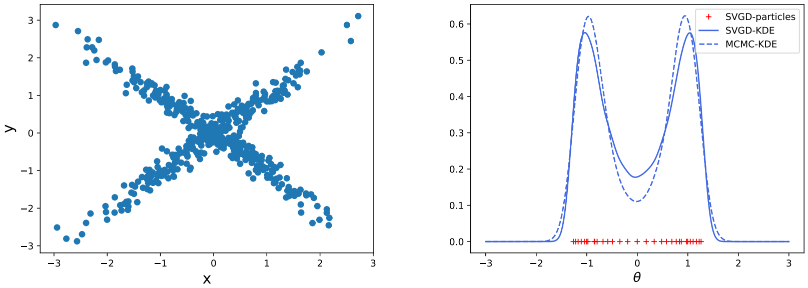

Figure 1: The scatter plot of a 'scissor' example (left); the estimated posterior distribution by SVGD, and MCMC.

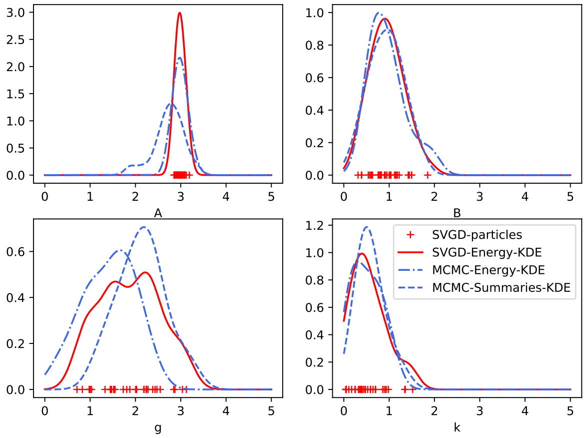

Figure 2: The estimated posterior distribution by SVGD with energy distance as statistics, MCMC with summary statistics and energy distance.

kernel density estimation based on optimized particles of SVGD and MCMC samples. The SVGD produces similar results with MCMC, and captures the bimodal.

g-and-k distribution: the g-and-k and related distributions have been analysed in the ABC setting by Allingham et al. (2009). Its density function is defined through a quantile function, $Q(q|A,B,g,k) = A + B\left[1 + 0.8 * \frac{1-\exp(-gz(q))}{1+\exp(-gz(q))}\right](1+z(q)^2)^k z(q)$, where $z(q) = \Phi^{-1}(q)$ is the q-th quantile of the standard normal distribution function. Given $\theta = (A, B, g, k)$, simulations $z(q) \sim N(0,1)$ can be transformed into samples from the g-and-k distribution. A simulated dataset of length $n = 1000$ generated from the g-and-k distribution with parameter vector $\theta = (3, 1, 2, 0.5)$. The SVGD algorithm takes 30 particles and a mini-batch of 200 data. For comparison, we use the MCMC method with traditional summary statistics and energy distance. The SVGD produces competitive results with MCMC.

References

David Allingham, R. A. R. King, and Kerrie L. Mengersen. Bayesian estimation of quantile distributions. *Statistics and Computing*, 19:189–201, 2009.

Mark A. Beaumont, Jean-Marie Cornuet, Jean-Michel Marin, and Christian P. Robert. Adaptive approximate bayesian computation. *Biometrika*, 96(4):983–990, 2009.

Chengyue Gong, Jian Peng, and Qiang Liu. Quantile stein variational gradient descent for batch Bayesian optimization. In *Proceedings of the International Conference on Machine Learning*, pages 2347–2356, 2019.

Jun Han and Qiang Liu. Stein variational gradient descent without gradient. In *Proceedings of the International Conference on Machine Learning*, pages 1900–1908, 2018.

Matthew D. Hoffman, David M. Blei, Chong Wang, and John W. Paisley. Stochastic variational inference. *J. Mach. Learn. Res.*, 14:1303–1347, 2012.

Cheng Huang and Xiaoming Huo. An efficient and distribution-free two-sample test based on energy statistics and random projections. arXiv preprint,arXiv:1707.04602, 2017.

Michael Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, 1999.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.

Chang Liu, Jingwei Zhuo, Pengyu Cheng, Ruiyi Zhang, and Jun Zhu. Understanding and accelerating particle-based variational inference. In *Proceedings of the 36th International Conference on Machine Learning*, pages 4082–4092, 2019.

Qiang Liu. Stein variational gradient descent as gradient flow. In *Proceedings of the Neural Information Processing Systems*, pages 3115–3123. Curran Associates, Inc., 2017.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *Proceedings of the Neural Information Processing Systems*, 2016.

Qiang Liu, Jason D. Lee, and Michael I. Jordan. A kernelized stein discrepancy for goodness-of-fit tests. In *Proceedings of the International Conference on Machine Learning*, 2016.

Jianfeng Lu, Yulong Lu, and James Nolen. Scaling limit of the stein variational gradient descent: The mean field regime. *Siam Journal on Mathematical Analysis*, 51(2):648–671, 2019.

Pierre Del Moral, Arnaud Doucet, and Ajay Jasra. Sequential monte carlo samplers. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(3):411–436, 2006.

Alexander Moreno, Tameem Adel, Edward Meeds, James M. Rehg, and M. Welling. Automatic variational ABC. *ArXiv*, abs/1606.08549, 2016.

Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

Vincent Plagnol Paul Marjoram, John Molitor and Simon Tavaré. Markov chain monte carlo without likelihoods. *Proc Natl Acad Sci U S A*, 100(26):15324–8, 2003.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. *ArXiv*, abs/1505.05770, 2015.

Maria L Rizzo. A test of homogeneity for two multivariate population. In *2002 Proceedings of the American Statistical Association, Physical and Engineering Science Section. American Statistical Association, Alexandria, VA*, 2003.

Scott Anthony Sisson, Yale Fan, and Mark M. Tanaka. Sequential monte carlo without likelihoods. *Proceedings of the National Academy of Sciences of the United States of America*, 104 6:1760–5, 2007.

Gábor J Székely and Maria L Rizzo. Testing for equal distributions in high dimension. *Interstat*, page 2004, 2004.

Gábor J. Székely and Maria L. Rizzo. A new test for multivariate normality. *Journal of Multivariate Analysis*, 93(1):58–80, 2005.

Gábor J. Székely and Maria L. Rizzo. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning & Inference*, 143(8):1249–1272, 2013.

Simon Tavaré, David J. Balding, R. C. Griffiths, and Peter Donnelly. Inferring coalescence times from dna sequence data. *Genetics*, 145 2:505–18, 1997.

Dilin Wang and Qiang Liu. Nonlinear stein variational gradient descent for learning diversified mixture models. In *Proceedings of the International Conference on Machine Learning*, pages 6576–6585, 2019.

Richard D Wilkinson. Approximate bayesian computation (abc) gives exact results under the assumption of model error. *Statistical Applications in Genetics and Molecular Biology*, 12(2):129–141, 2013.