Stein Variational Gradient Descent: Theory and Applications

Qiang Liu
Department of Computer Science
Dartmouth College
Hanover, NH 03755
qiang.liu@dartmouth.edu

Abstract

Although optimization can be done very efficiently using gradient-based optimization these days, Bayesian inference or probabilistic sampling has been considered to be much more difficult. Stein variational gradient descent (SVGD) is a new particle-based inference method derived using a functional gradient descent for minimizing KL divergence without explicit parametric assumptions. SVGD can be viewed a natural counterpart of gradient descent for optimization, and in fact exactly reduces to the typical gradient ascent for MAP using only a single particle. This short paper gives a brief introduction to SVGD, and discusses its theoretical foundation and applications.

1 Introduction

Bayesian inference provides a unified, powerful framework for reasoning about complex phenomena under uncertainty, and has been widely adopted as a powerful tool for statistical data analysis in many scientific areas, as well as state-of-the-art machine learning and AI techniques. Unfortunately, the current Bayesian computation techniques tend to scale poorly to big data and big models. Markov chain Monte Carlo (MCMC) has been routinely used for Bayesian computation, but is widely criticized for its convergence issues and has particular difficulty scaling up to big data settings. Variational inference has been widely used in machine learning, especially for deep learning and probabilistic graphical models, but does not guarantee asymptotic exactness due to deterministic approximation errors, and requires users to make careful, often case-by-case choices of parametric approximation families. These difficulties form a major barrier to developing highly scalable, fully automatic Bayesian inference tools easily assessable to practitioners.

This situation is in sharp contrast with the optimization techniques for point estimation, or maximum a posteriori (MAP) estimation, for which simple gradient-based methods provide efficient, generic and easy-to-use tools, scalable to big data settings via stochastic gradient, and has been known to be surprisingly efficient in deep neural models. It is highly desirable to develop new Bayesian inference methods that incorporate the key advantages of gradient-based optimization to enable scalable and automatic inference.

Stein variational gradient descent (SVGD) [1] is a new Bayesian inference algorithm that seeks a set of points (or particles) to approximate the target distribution using iterative gradient based updates. It has a simple form that closely mimics the typical gradient descent for optimization, and in fact reduces to the typical gradient descent for optimization when using only one particle. This makes SVGD highly flexible and scalable, and can be easily combined with various state-of-the-art techniques that have been responsible for the success of gradient optimization, including stochastic gradient, adaptive learning rates (such as adagrad), and momentum.

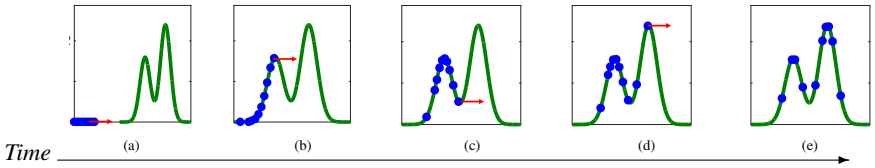

Figure 1: Illustrating how the interactive particle dynamics of SVGD in (4) can escape local modes and obtain diverse points to approximate the full distribution. In (a), 10 particles are initialized far away on the left, but move quickly towards the high probability area of the target distribution $p(x)$ driven by the gradient information. In (b), (c), (d), the leading particle has arrived stationary points (with zero gradient itself), but is pushed further to the right side by the repulsive force and shared gradient from the other particles. In (e), the particles have reached equilibrium and form a close approximation for the target distribution. Note that typical non-gradient based particle methods such as sequential Monte Carlo may experience weight degeneration in this example due to the poor initialization we set up in (a).

{fig:1dgmmparticle}

This short paper gives a brief overview of the key idea of SVGD, and outline several directions for future work, including a new theoretical framework that interprets SVGD as a natural gradient descent of the KL divergence functional on a Riemannian-like metric structure on the space of distributions, and extensions of SVGD that allow us to train neural networks to draw approximate samples from given distributions, and develop new adaptive importance sampling methods without assuming parametric forms on the proposals.

2 Stein Variational Gradient Descent

To give a quick overview of the main idea of SVGD, let $p(x)$ be the positive density function on $\mathcal{X} \subseteq \mathbb{R}^d$ which we want to approximate with a set of particles $\{x_i\}_{i=1}^n$. We initialize the particles with some simple distribution q_0, and iteratively update them via

$$x_i \leftarrow x_i + \epsilon \boldsymbol{\phi}(x_i), \quad \forall i = 1, \ldots, n,$$

where ϵ is a small step size, and $\boldsymbol{\phi}(x)$ is a perturbation direction, or velocity field, chosen to maximumly decrease the KL divergence between the distribution of the updated particles and the target distribution, in the sense that

{equ:ff00}
$$\boldsymbol{\phi} = \arg\max_{\boldsymbol{\phi} \in \mathcal{F}} \Big\{ -\frac{d}{d\epsilon} \mathrm{KL}(q_{[\epsilon\boldsymbol{\phi}]} \,||\, p)\big|_{\epsilon=0} \Big\}, \tag{1}$$

where $q_{[\epsilon\boldsymbol{\phi}]}$ denotes the density of the updated particle $x' = x + \epsilon\boldsymbol{\phi}(x)$ when the density of the original particle x is q, and $\mathcal{F}$ is a set of perturbation directions that we optimize over. We take $\mathcal{F}$ to be the unit ball of a vector-valued reproducing kernel Hilbert space (RKHS) $\mathcal{H} = \mathcal{H}_0 \times \cdots \mathcal{H}_0$, where $\mathcal{H}_0$ is a scalar-valued RKHS associated with a scalar-valued positive definite kernel $k(x, x')$, which is a dense subset of the space of continuous vector-valued functions with typical universal kernels such as the RBF kernel. Extension to matrix-valued positive definite kernels is also straightforward.

A key observation is that the objective in (1) is a simple linear functional of $\boldsymbol{\phi}$. In fact, we have

{equ:klstein00}
$$-\frac{d}{d\epsilon} \mathrm{KL}(q_{[\epsilon\boldsymbol{\phi}]} \,||\, p)\big|_{\epsilon=0} = \mathbb{E}_{x\sim q}[\mathcal{T}_p \boldsymbol{\phi}(x)], \tag{2}$$

$$\text{with} \quad \mathcal{T}_p \boldsymbol{\phi}(x) \stackrel{def}{=} \nabla_x \log p(x)^\top \boldsymbol{\phi}(x) + \nabla_x \cdot \boldsymbol{\phi}(x),$$

where $\mathcal{T}_p$ is considered as a linear operator acting on function $\boldsymbol{\phi}$ and is called the Stein operator in connection with Stein's identity, which shows that the RHS of (2) equals zero if $p = q$:

$$\mathbb{E}_p[\mathcal{T}_p \boldsymbol{\phi}] = \mathbb{E}_p[\nabla_x \log p^\top \boldsymbol{\phi} + \nabla_x \cdot \boldsymbol{\phi}] = 0.$$

This is a result of integration by parts assuming the value of $p(x)\boldsymbol{\phi}(x)$ vanishes on the boundary of the integration domain $\mathcal{X}$.

Therefore, the optimization in (2) reduces to

{equ:ksd}
$$\mathbb{D}(q \,||\, p) \stackrel{def}{=} \max_{\boldsymbol{\phi} \in \mathcal{F}} \big\{ \mathbb{E}_{x\sim q}[\mathcal{T}_p \boldsymbol{\phi}(x)] \big\}, \tag{3}$$

where $\mathbb{D}(q \,||\, p)$ is defined as the kernelized Stein discrepancy (KSD) between p and q [2–4]. It has been shown that the optimal solution of (2) has a simple closed form:

$$\boldsymbol{\phi}^*(x') \propto \mathbb{E}_{x\sim q}[\mathcal{T}_p k(x, x')] = \mathbb{E}_{x\sim q}[\nabla_x \log p(x) k(x, x') + \nabla_x k(x, x')].$$

By approximating the expectation under q with the empirical averaging of the current particles $\{x_i\}_{i=1}^n$, our algorithm admits a simple form of updates:

{equ:update11}
$$x_i \;\leftarrow\; x_i \;+\; \epsilon\, \hat{\mathbb{E}}_{x\sim\{x_i\}_{i=1}^n}[\nabla_x \log p(x) k(x, x_i) + \nabla_x k(x, x_i)], \quad \forall i = 1, \ldots, n, \tag{4}$$

where $\hat{\mathbb{E}}_{x\sim\{x_i\}_{i=1}^n}$ denotes empirical averaging over $\{x_i\}$, that is, $\hat{\mathbb{E}}_{x\sim\{x_i\}_{i=1}^n}[f(x)] = \sum_i f(x_i)/n$.

Intuitively, this update pushes the particles towards the high probability regions of the target probability via the gradient term $\nabla_x \log p$, while maintaining a degree of diversity via the second term $\nabla_x k(x, x_i)$, which can be shown to serve as a repulsive force between the particles. Overall, this particle update produces an interesting "momentum" effect in which the particles move collaboratively to escape the local optima to converge to diverse points to approximate the target distribution (see Figure 1).

It is easy to see that (4) reduces to the typical gradient descent if we use only a single particle ($n = 1$) and $\nabla_x k(x, x') = 0$ when $x = x'$; with more particles, (4) allows a full Bayesian sampling to cover more local modes and provide uncertainty measure.

3 Gradient Flow, Optimal Transport, Nonparametric Information Geometry

As the number of the particles becomes large ($n \to \infty$), our process can be interpreted as a particle approximation of a functional gradient descent of the KL divergence functional on a new Remamian-like metric structure defined on the space of probability distributions.

Here we briefly introduce this perspective and we will elaborate the details in our incoming work. Let q_ℓ be the limit distribution of the particles at the ℓ-th iteration as we take $n \to \infty$, and $F(\log q) = \mathrm{KL}(q \,||\, p)$ the KL divergence as a functional of $\log q$. As the step-size ϵ approaches zero ($\epsilon \to 0$), we can define a continuous time $t = \epsilon\ell$, and we can show that the evolution of density q_t is governed by a gradient flow:

{equ:gradklqt}
$$\frac{d \log q_t}{dt} = -\tilde{\nabla} F(\log q_t), \tag{5}$$

where $\tilde{\nabla} F(\log q)$ represents a type of functional gradient of $F(\log q)$ w.r.t. $\log q$ induced by a Riemannian metric structure on the space of log-density functions defined by the RKHS-related cost of transforming one distribution to another.

We now define this Riemannian metric. For each log-density $\log q$ for which Stein's identity holds for q, we define the tangent space $\mathcal{H}_q$ around $\log p$ to be the set of functions formed by the outputs of Stein operator:

$$\mathcal{H}_q \stackrel{def}{=} \{f = \mathcal{T}_q \boldsymbol{\phi},\ \ \forall \boldsymbol{\phi} \in \mathcal{H}\},$$

where $\mathcal{H}$ is any $d \times 1$ vector-valued RKHS. By stein's identity, all the functions in $\mathcal{H}_p$ have zero-expectation under q, that is, $\mathbb{E}_q[f] = 0, \forall f \in \mathcal{H}_q$.

In addition, It has been shown that $\mathcal{H}_q$ forms a RKHS whose kernel is denoted by $\kappa_q(x, x')$ [3]. This allows us to define a Riemannian metric with inner product $\langle f, g\rangle_{\mathcal{H}_q}$ on the elements $f, g \in \mathcal{H}_q$ in the tangent space. It is then possible to define a gradient of functional $F(\log q)$ to be any function $\tilde{\nabla} F(\log q)$ that satisfies

$$F(\log q + df) = F(\log q) + \langle \tilde{\nabla} F(\log q),\ \ df\rangle_{\mathcal{H}_q},$$

for any infinite element df in the tangent space $\mathcal{H}_q$.

This Remannian inner product structure then induces a geodesic distance between two points q and p in the space, which can be interpreted as an optimal transport metric induced by RKHS norm. To be specific, let $\Phi = \{\boldsymbol{\phi}_t(x) \colon t \in [0, T]\}$ be a collection of velocity field index by continuous time t and

x_t the solution of ODE $dx = \boldsymbol{\phi}_t(x)dt$. Then $\Phi = \{\boldsymbol{\phi}_t(x) \colon t \in [0,1]\}$ is called to transport q to p if when $x_0 \sim q$ and have $x_1 \sim p$. This allows us define a metric between q and p via

{equ:roptd}
$$d(q,p) = \min_{\Phi}\{\int ||\boldsymbol{\phi}_t||_{\mathcal{H}} dt \colon \quad \Phi = \{\boldsymbol{\phi}_t \colon t \in [0,1]\} \text{ transports } q \text{ to } p\}. \tag{6}$$

It is also possible to define the gradient $\tilde{\nabla} F(\log q)$ as the steepest descent direction for $F(\log q)$ in the neighborhood defined by metric $d(q,p)$.

Overall, SVGD can be treated as an **Optimization-Then-Approximate** approach for solving the infinite dimensional optimization problem of $\min_q \mathrm{KL}(q \,||\, p)$: it first derives an infinite dimensional gradient descent (5) (effectively a partial differential equation (PDE)) that attempts to solve the infinite dimensional optimization in that it converges to the exact p as $t \to \infty$, and then develop a finite dimensional particle approximation for the gradient flow (or effectively a numerical solution of the corresponding PDE). This is in contrast with the traditional variational inference method that can be treated as **Approximate-Then-Optimization** approaches, in which we first approximate the infinite dimensional optimization $\min_q \mathrm{KL}(q \,||\, p)$ with a finite dimensional one by restricting q in a parametric family, and then solve the corresponding parametric optimization. We think it is possible to derive a host of new approaches based on the new Optimization-Then-Approximate framework, which allows us to leverage the large literature of PDE and their numerical solutions, and has the advantage of requiring no explicit parametric assumptions, amending asymptotic consistency analysis.

This framework has deep theoretical implication that deserves in-depth study to establish a solid mathematical foundation and connection with existing theories. In particular, the RKHS-based optimal transport metric (6) seems to yield an extension to the classical theory of optimal transport and gradient flow in optimal transport (Weierstrass) metric space in which the transport cost is often defined by L^p, rather than RKHS cost [5–7], and extensive studies are need to understand its basic properties in parallel to the L^p-based Weierstrass metric. In addition, the gradient flow (5) can be shown to be an instance of Vlasov equation [8] known in physics, with rich connection to stochastic systems with mean field interactions [9]. Concepts related to displacement convexity [10] and logarithmic Sobolev inequalities [11] can play critical roles in establishing exponential convergence rate of the nonlinear evolutionary PDF in (5). The Riemannian metric structure here induced by Stein operator seems to open the possibility of establishing a new theory of nonparametric information geometry (in contrast with the typical information geometry which only considers distributions that are indexed by infinite dimensional parameters), connections to existing theories of nonparametric information geometry is needed [e.g., 12, 13].

4 Applications and Extensions

In addition to providing an efficient general-purpose particle inference algorithm, SVGD and its basic idea provides a foundation for developing new tools for solving difficult problems. Here we present two examples: one for training stochastic neural networks to "learn to sample" from given distributions, another for developing efficient adaptive importance sampling based on the optimal variable transforms given by SVGD.

Amortized SVGD SVGD and other particle-based methods become inefficient when we need to repeatedly infer a large number of different target distributions for multiple tasks, because they can not improve based on the experience from the past tasks, and may require a large memory to restore a large number of particles; this happens, for example, in MLE training of intractable distributions when fast inference is needed as the inner loop. One possible solution for this is to "amortize SVGD", training a stochastic neural network so that its output mimics the SVGD dynamics, and hence closely approximates the target distribution when it converges. This essentially allows us to train neural samplers to *learn to draw samples* for given target distributions, and hence yield wide application.

In Wang and Liu [14], this idea is used to train neural samplers to approximate MLE for intractable energy-based models, yielding an algorithm that can be interpreted as an adversarial game [15] between the deep energy model and the neural sampler, which allows us to generate high quality realistic-looking images. We refer the readers to Wang and Liu [14] for more details.

Figure 2: Samples drawn by our neural samplers trained by amortized SVGD for approximating MLE in a deep energy model on the CelebA dataset Upper: images generated by DCGAN [16] and our neural sampler. Lower: images generated by our method when the random seed of the neural sampler changes gradually; we can see that a man with glasses and black hair gradually changes to a woman with blonde hair. See Wang and Liu [14] for more examples.

{fig:face}

Stein Variational Importance Sampling Because SVGD reduces to MAP when using one single particle, it is "particle-efficient" in that good practical results can be achieved using a small number of particles. This is contrast with the typical Monte Carlo methods which often require to average a large number of particles to obtain good results. On the other hand, the theoretical properties of typical Monte Carlo is much more understood, and can be more attractive when emphasis is on getting well-calibrated confidence intervals or unbiased estimates. In fact, it is easy to turn SVGD into an efficient adaptive importance sampling procedure, in which the SVGD update (4) serves to iteratively improve the proposal distribution using optimal variable transforms.

To give a brief overview, note that our method can be treated as constructing a path of distributions $\{q_\ell\}$ that connects the initial distribution q_0 with the target distribution p, in which q_ℓ gets closer to the target p along a steepest descent direction of KL divergence. Therefore, we can leverage q_ℓ as excellent proposal distributions for importance sampling estimates of p. In order to draw i.i.d. sample from q_ℓ needed for importance sampling, we introduce an additional set of particles $\{y_i\}$ in additional to the $\{x_i\}$ updated by (4), where $\{x_i\}$ is responsible for constructing q_ℓ, while $\{y_i\}$ simply follows the updates constructed by $\{x_i\}$ (without influence the trajectory of $\{x_i\}$); as a result $\{y_i\}$ can be viewed as i.i.d. samples from q_ℓ, whose importance weights $w_i = p(x_i)/q_\ell(x_i)$ can be also calculated efficiently with an iterative update.

This method can be viewed as a special adaptive importance sampling, where at each iteration the proposal distribution q^ℓ is improved by applying a variable transform defined by (4) that promises to design its KL divergence with the target distribution. This distinguishes us with the traditional adaptive importance sampling in which the proposals are optimized in predefined parametric families (usually mixture families).

This method can also be used to estimate the partition function for unnormalized distributions. It is interesting to compare it with the path integration ideas (e.g., Gelman and Meng [17], Neal [18]) which are also based on a path of distributions that connects the target distribution with a simple reference distribution. Typically, these methods construct the annealing path using simple geometric mean of the probabilities, while our path moves along the steepest descent direction of KL divergence and can potentially better with less intermediate distributions.

5 Conclusion

In this short paper, we introduced the key idea of Stein variational gradient descent (SVGD), and discussed its theoretical properties and several extensions.

References

[1] Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. *arXiv preprint arXiv:1608.04471*, 2016.

[2] Qiang Liu, Jason D Lee, and Michael I Jordan. A kernelized Stein discrepancy for goodness-of-fit tests and model evaluation. In *ICML*, 2016.

[3] Chris J Oates, Mark Girolami, and Nicolas Chopin. Control functionals for Monte Carlo integration. *Journal of the Royal Statistical Society, Series B*, 2017.

[4] Kacper Chwialkowski, Heiko Strathmann, and Arthur Gretton. A kernel test of goodness-of-fit. *arXiv preprint arXiv:1602.02964*, 2016.

[5] Cédric Villani. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

[6] Felix Otto. The geometry of dissipative evolution equations: the porous medium equation. 2001.

[7] Cédric Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.

[8] Anatoly Alexandrovich Vlasov. On vibration properties of electron gas. *J. Exp. Theor. Phys*, 8(3):291, 1938.

[9] Pierre Del Moral. *Mean field simulation for Monte Carlo integration*. CRC Press, 2013.

[10] Robert J McCann. A convexity principle for interacting gases. *Advances in mathematics*, 128(1):153–179, 1997.

[11] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

[12] Giovanni Pistone. Nonparametric information geometry. In *Geometric Science of Information*, pages 5–36. Springer, 2013.

[13] Kenji Fukumizu. Exponential manifold by reproducing kernel hilbert spaces. In *Algebraic and Geometric methods in statistics*, pages 291–306. 2009.

[14] Dilin Wang and Qiang Liu. Learning to draw samples: With application to amortized mle for generative adversarial learning. *https://arxiv.org/pdf/1611.01722*, 2016.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 2672–2680, 2014.

[16] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

[17] Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

[18] Radford M Neal. Annealed importance sampling. *Statistics and Computing*, 11(2):125–139, 2001.