The equivalence between Stein variational gradient descent and black-box variational inference

Casey Chu [1] **Kentaro Minami** [2] **Kenji Fukumizu** [2,3]

Abstract

We formalize an equivalence between two popular methods for Bayesian inference: Stein variational gradient descent (SVGD) and black-box variational inference (BBVI). In particular, we show that BBVI corresponds precisely to SVGD when the kernel is the neural tangent kernel. Furthermore, we interpret SVGD and BBVI as *kernel gradient flows*; we do this by leveraging the recent perspective that views SVGD as a gradient flow in the space of probability distributions and showing that BBVI naturally motivates a Riemannian structure on that space. We observe that kernel gradient flow also describes dynamics found in the training of generative adversarial networks (GANs). This work thereby unifies several existing techniques in variational inference and generative modeling and identifies the kernel as a fundamental object governing the behavior of these algorithms, motivating deeper analysis of its properties.

1. Introduction

The goal of Bayesian inference is to compute the *posterior* $P(x|z)$ over a variable of interest x. In principle, this posterior may be computed from the *prior* $P(x)$ and the *likelihood* $P(z|x)$ of observing data z, using the equation

$$p(x) := P(x|z) = \frac{P(z|x)P(x)}{\int P(z|x)P(x)\,dx}. \qquad (1)$$

We denote the posterior as $p(x)$ for convenience of notation. Unfortunately, the integral in the denominator is usually intractable, which motivates *variational inference* techniques, which approximate the true posterior $p(x)$ with an approximate posterior $q(x)$, often by minimizing the KL divergence $\mathrm{KL}(q(x) \,\|\, p(x))$. In this paper, we consider two

[1]Stanford University [2]Preferred Networks [3]The Institute of Statistical Mathematics. Correspondence to: Casey Chu <caseychu@stanford.edu>.

popular variational inference techniques, black-box variational inference (Ranganath et al., 2014) and Stein variational gradient descent (Liu & Wang, 2016), and show that they are equivalent when viewed as instances of *kernel gradient flow*.

2. Stein variational gradient descent

Stein variational gradient descent (Liu & Wang, 2016), or SVGD, is a technique for Bayesian inference that approximates the true posterior $p(x)$ with a set of particles $x_1, \ldots, x_n$.

In the continuous-time limit of small step size, each particle undergoes the update rule

$$\frac{dx_i}{dt} = \mathbb{E}_{y \sim q_t}[k(x_i, y)\nabla_y \log p(y) + \nabla_y k(x_i, y)], \quad (2)$$

where q_t denotes the empirical distribution of particles at time t:

$$q_t = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i(t)}, \qquad (3)$$

and $k(x, y)$ is a user-specified kernel function, such as the RBF kernel $k(x, y) = e^{-||x-y||^2}$.

In the mean-field limit as $n \to \infty$ (Lu et al., 2019), an equivalent form of the dynamics (2) is obtained by an application of Stein's identity (integration by parts on the second term):

$$\frac{dx}{dt} = \mathbb{E}_{y \sim q_t}[k(x, y)\nabla_y(\log p(y) - \log q_t(y))]. \qquad (4)$$

3. Black-box variational inference

Black-box variational inference (Ranganath et al., 2014), or BBVI, is another technique for Bayesian inference that approximates the true posterior $p(x)$ with an approximate posterior $q_\phi(x)$, where q_ϕ is a family of distributions parameterized by ϕ. In BBVI, we maximize the evidence lower bound, or ELBO, objective

$$L(\phi) := \mathbb{E}_{x \sim q_\phi}\left[\log \frac{P(z|x)P(x)}{q_\phi(x)}\right] \qquad (5)$$

by gradient ascent on ϕ. This procedure effectively minimizes the KL divergence between $q_\phi(x)$ and the true posterior $p(x) = P(x|z)$, since the KL divergence and the ELBO

objective differ by only the *evidence* $P(z)$, which is constant w.r.t. ϕ:

$$\mathrm{KL}(q_\phi(x) \,\|\, p(x)) = P(z) - L(\phi). \qquad (6)$$

Our claim is:

Claim 1. *The sequence of approximate posteriors generated by BBVI, when the reparameterization trick of Kingma & Welling (2014) is used, is governed by the SVGD dynamics* (4)*, where the kernel k is the neural tangent kernel of Jacot et al. (2018).*

To see this, we observe that the evolution of the parameters ϕ under gradient ascent is governed by

$$\frac{d\phi}{dt} = \nabla_\phi L(\phi). \qquad (7)$$

Next, we specialize to the case where the family of approximate posteriors is parameterized via the reparameterization trick (Kingma & Welling, 2014). That is, suppose that there exists a fixed distribution ω and a parameterized function f_ϕ such that the following two sampling methods result in the same distribution over x:

$$x \sim q_\phi \iff \varepsilon \sim \omega \text{ and } x = f_\phi(\varepsilon). \qquad (8)$$

As an example, the family of normal distributions $\mathcal{N}(\mu, \sigma)$ may be reparameterized as

$$x \sim \mathcal{N}(\mu, \sigma) \iff \varepsilon \sim \mathcal{N}(0, 1) \text{ and } x = \mu + \sigma\varepsilon. \qquad (9)$$

In this setting, Roeder et al. (2017) and Feng et al. (2017) noted that

$$\nabla_\phi L(\phi)$$
$$= \mathbb{E}_{w\sim\omega}[\nabla_\phi f_\phi(w) \cdot \nabla_y(\log p(y) - \log q_\phi(y))\big|_{y=f_\phi(w)}]. \qquad (10)$$

Now, we consider the dynamics of a sample $x = f_\phi(\varepsilon)$ under the parameter dynamics (7). By the chain rule, we have that

$$\frac{dx}{dt} = (\nabla_\phi f_\phi(\varepsilon))^T \frac{d\phi}{dt}. \qquad (11)$$

Let us introduce the *neural tangent kernel* of Jacot et al. (2018)

$$\Theta_\phi(\varepsilon, w) := (\nabla_\phi f_\phi(\varepsilon))^T \nabla_\phi f_\phi(w), \qquad (12)$$

and define

$$k_\phi(x, y) := \Theta_\phi(f_\phi^{-1}(x), f_\phi^{-1}(y)), \qquad (13)$$

making the additional assumption that $\varepsilon \mapsto f_\phi(\varepsilon)$ is injective. Note that if $x \in \mathbb{R}^n$, then $\Theta_\phi(\varepsilon, w)$ and $k_\phi(x, y)$ are

both n-by-n matrices that depend on ϕ. Then, substituting (7) and (10) into (11), we find that the samples satisfy

$$\frac{dx}{dt} = (\nabla_\phi f_\phi(\varepsilon))^T \frac{d\phi}{dt} \qquad (14)$$
$$= \mathbb{E}_{w\sim\omega}[\Theta_\phi(\varepsilon, w) \nabla_y(\log p(y) - \log q_\phi(y))\big|_{y=f_\phi(w)}] \qquad (15)$$
$$= \mathbb{E}_{y\sim q_\phi}[k_\phi(x, y) \nabla_y(\log p(y) - \log q_\phi(y))]. \qquad (16)$$

Comparing (16) with the SVGD dynamics (4), we find an exact correspondence between SVGD and BBVI, where in BBVI, the kernel is given by (13) and defined by the neural tangent kernel.

3.1. Example: a Gaussian variational family

As an example, consider the family of multivariate normal distributions $\mathcal{N}(\mu, \Sigma)$, parameterized by an invertible matrix A and a vector μ, with the relation $\Sigma = AA^T$. This variational family is reparameterizable with

$$x \sim \mathcal{N}(\mu, \Sigma) \iff \varepsilon \sim \mathcal{N}(0, I) \text{ and } x = \mu + A\varepsilon. \qquad (17)$$

In this setting, the kernel (13) becomes

$$k(x, y) = (1 + (x - \mu)^T \Sigma^{-1}(y - \mu))I, \qquad (18)$$

where I is the identity matrix. In the continuous-time and many-particle limit, BBVI with the parameterization (17) produces the same sequence of approximate posteriors as SVGD with the kernel (18). Figure 1 compares the sequence of approximate posteriors generated by BBVI and SVGD with the theoretically equivalent kernel (18) in fitting a bimodal 2D distribution; we see that the agreement is quite close.

It is instructive to perform the computation of (18) explicitly. We use index notation with Einstein summation notation, where indices that appear twice are implicitly summed over. We have that $f_i(\varepsilon) = \mu_i + A_{ik}\varepsilon_k$ and

$$\frac{\partial f_i(\varepsilon)}{\partial \mu_\ell} = \delta_{i\ell}, \quad \frac{\partial f_i(\varepsilon)}{\partial A_{\ell m}} = \delta_{i\ell}\delta_{km}\varepsilon_k, \qquad (19)$$

so that the neural tangent kernel is

$$\Theta_{ij}(\varepsilon, w) = \frac{\partial f_i(\varepsilon)}{\partial \mu_\ell}\frac{\partial f_j(w)}{\partial \mu_\ell} + \frac{\partial f_i(\varepsilon)}{\partial A_{\ell m}}\frac{\partial f_j(w)}{\partial A_{\ell m}} \qquad (20)$$
$$= \delta_{i\ell}\delta_{j\ell} + \delta_{i\ell}\delta_{km}\varepsilon_k\delta_{j\ell}\delta_{om}w_o \qquad (21)$$
$$= \delta_{ij} + \delta_{ij}\varepsilon_m w_m, \qquad (22)$$

or $\Theta(\varepsilon, w) = (1 + \varepsilon \cdot w)I$ in vector notation. Then, using the definition (13) and substituting $f^{-1}(x) = A^{-1}(x - \mu)$ and $\Sigma = AA^T$, we arrive at (18).
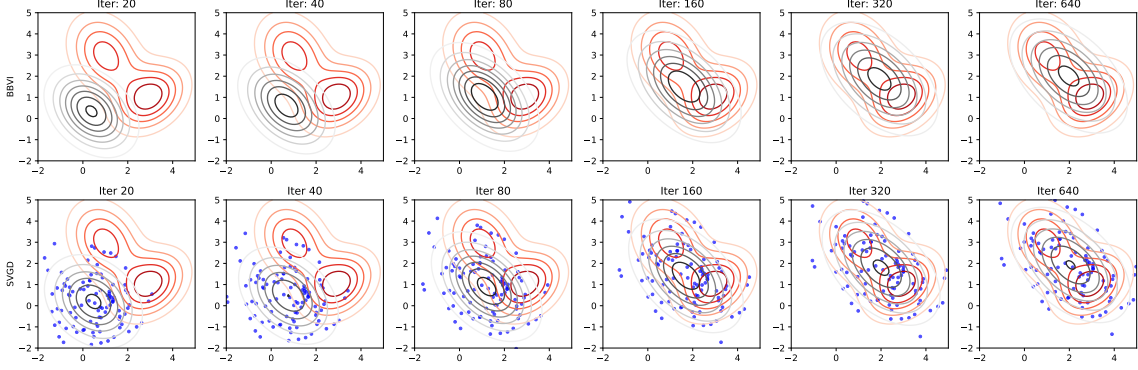
Figure 1. The sequence of approximate posteriors obtained by BBVI and SVGD with the theoretically equivalent kernel.

4. Motivating a Riemannian structure

In the previous section, we found that SVGD and BBVI both correspond to particle dynamics of the form

$$\frac{dx}{dt} = \mathbb{E}_{y \sim q_\phi}[k_\phi(x, y) \, \nabla_y(\log p(y) - \log q_\phi(y))]. \quad (23)$$

One peculiar feature of the BBVI dynamics is that the kernel k_ϕ depends on the current parameter ϕ, rather than being constant as the approximate posterior q_ϕ changes, as in the SVGD case.

In fact, we argue that this feature of BBVI is quite natural:

Claim 2. *The requirement of BBVI that the kernel depends on the current distribution naturally motivates a Riemannian structure on the space of probability distributions.*

To make this claim, let us first review Euclidean and Riemannian gradient flows. In Euclidean space, following the negative gradient of a function $J : \mathbb{R}^n \to \mathbb{R}$ according to

$$\frac{dx}{dt} = -\nabla J(x) \quad (24)$$

can lead to a minimizer of J. Analogously, on a *Riemannian manifold* M, following the negative *Riemannian gradient* of a function $J : M \to \mathbb{R}$ according to

$$\frac{dx}{dt} = -G(x)^{-1} \nabla J(x), \quad (25)$$

can lead to a minimizer of J. Here, G is a positive-definite matrix-valued function called the *Riemannian metric*, which defines the local geometry at x and perturbs the Euclidean gradient ∇J pointwise. Note that in the case that $G(x)$ is the identity matrix for all x, Riemannian gradient flow reduces to the Euclidean gradient flow.

Next, we review Wasserstein gradient flows, which generalize gradient flows to the space of probability distributions (Ambrosio et al., 2008). Here, we consider the set of all probability distributions over a particular space formally as an "infinite-dimensional" manifold $\mathcal{P}$, and we consider a function $J : \mathcal{P} \to \mathbb{R}$. In variational inference, the most relevant such function is the KL divergence $J(q) := \mathrm{KL}(q \, \| \, p)$, where we are interested in finding an approximate posterior that minimizes J. Analogous to before, a minimizer of J may be obtained by following the analogue of a gradient; the trajectory of the distribution q turns out to take the form of the PDE

$$\frac{\partial q}{\partial t} = \nabla \cdot (q \nabla \Psi_q). \quad (26)$$

Here, $\nabla \Psi_q$ serves as the correct analogue of the gradient of J evaluated at q, and it turns out that $\Psi_q(x) = \log q(x) - \log p(x)$ for the variational inference case $J(q) := \mathrm{KL}(q \, \| \, p)$. This function Ψ_q is known variously as the functional derivative, first variation, or von Mises influence function.

Now, we review the recent perspective that SVGD can be interpreted as a generalized Wasserstein gradient flow under the *Stein geometry* (Liu, 2017; Liu et al., 2019; Duncan et al., 2019). We follow the presentation of Duncan et al. (2019) and refer to it for a rigorous treatment. To set the stage, we take a non-parametric view of the SVGD update (2), in which the dependence on ϕ is interpreted as dependence on the distribution q itself:

$$\frac{dx}{dt} = \mathbb{E}_{y \sim q}[k(x, y) \, \nabla_y(\log p(y) - \log q(y))]. \quad (27)$$

Substituting $\Psi_q(y) = \log q(y) - \log p(y)$ and the linear operator T_q defined by

$$(T_q \varphi)(x) := \mathbb{E}_{y \sim q}[k(x, y)\varphi(y)], \quad (28)$$

we have

$$\frac{dx}{dt} = -(T_q \nabla \Psi_q)(x). \quad (29)$$

Under these dynamics, the probability distribution q evolves according to the PDE

$$\frac{\partial q}{\partial t} = \nabla \cdot (q T_q \nabla \Psi_q). \quad (30)$$

Comparing (30) with (26), we see that (30) defines a modified gradient flow in which the gradient $\nabla\Psi_q$ is perturbed by the operator T_q.

We now advocate for generalizing Wasserstein gradient flow in the same way that Riemannian gradient flow generalizes Euclidean gradient flow. The operator T_q perturbs the gradient in a way analogous to how the Riemannian metric perturbs the Euclidean gradient in (25), so the operator T_q thereby defines an analogue of a Riemannian metric on $\mathcal{P}$. However, there is no fundamental reason that T_q must have the restrictive form prescribed by (28). Indeed, because T_q is analogous to the Riemannian metric $G(x)$, it is natural to let the kernel, whose action defines the operator T_q, depend on the current value of q. It is also natural to allow the kernel to output a matrix rather than a scalar so that T_q may mix all components of φ. Duncan et al. (2019) in fact speculate on these possibilities (Remarks 17 and 1).

With these considerations in mind, we propose replacing (28) with

$$(T_q\varphi)(x) := \mathbb{E}_{y\sim q}[k_q(x,y)\varphi(y)], \qquad (31)$$

where the kernel k_q now depends on q and outputs a matrix. This defines a gradient flow by (30) that we will refer to as *kernel gradient flow*.[1]

Once T_q has the form (31), BBVI may naturally be regarded as an instance of kernel gradient flow, in which the kernel k_q is the neural tangent kernel which depends on the current distribution q. More abstractly, we see that the neural tangent kernel defines a Riemannian metric on the space of probability distributions. We summarize the perspective that this framework gives on variational inference:

Claim 3. *SVGD updates generate a kernel gradient flow of the loss function $J(q) := \mathrm{KL}(q\,\|\,p)$, with a Riemannian metric determined by the user-specified kernel.*

Claim 4. *BBVI updates generate a kernel gradient flow of the loss function $J(q) := \mathrm{KL}(q\,\|\,p)$, with a Riemannian metric determined by the neural tangent kernel of f_ϕ.*

5. Beyond variational inference: GANs as kernel gradient flow

We now argue that the kernel gradient flow perspective we have developed describes not only SVGD and BBVI, but also describes the training dynamics of *generative adversarial networks* (Goodfellow et al., 2014).

---

[1]To further the analogy between Euclidean and Riemannian gradient flow and Wasserstein and kernel gradient flow, note that just as setting the Riemannian metric to identity matrix for all x reduces Riemannian to Euclidean gradient flow, setting T_q to the identity operator for all q reduces kernel to Wasserstein gradient flow. The special "Euclidean" Riemannian metric obtained this way is the central object of the *Otto calculus* (Otto, 2001; Ambrosio et al., 2008).

Generative adversarial networks, or GANs, are a technique for learning a generator distribution q_ϕ that mimics an empirical data distribution p_{data}. The generator distribution q_ϕ is defined implicitly as the distribution obtained by sampling from a fixed distribution ω, often a standard normal, and running the sample through a neural network f_ϕ called the *generator*. The learning process is facilitated by another neural network D_θ called the *discriminator* that takes a sample and outputs a real number, and is trained to distinguish between a real sample from p_{data} and a fake sample from q_ϕ. The generator and discriminator are trained simultaneously until the discriminator is unable to distinguish between real and fake samples, at which point the generator distribution q_ϕ hopefully mimics the data distribution p_{data}.

For many GAN variants, the rule to update the generator parameters ϕ can be expressed in the continuous-time limit as

$$\frac{d\phi}{dt} = \nabla_\phi \mathbb{E}_{w\sim\omega}[D_\theta(f_\phi(w))], \qquad (32)$$

or by the chain rule,

$$\frac{d\phi}{dt} = \mathbb{E}_{w\sim\omega}[\nabla_\phi f_\phi(w)\cdot\nabla_y D_\theta(y)\big|_{y=f_\phi(w)}]. \qquad (33)$$

The discriminator parameters θ are updated simultaneously to minimize a separate discriminator loss $L_{\mathrm{disc}}(\theta,\phi)$, but it is common for theoretical purposes to assume that the discriminator achieves optimality at every training step. Denoting this optimal discriminator as $-\Psi_\phi$ (i.e. setting $\Psi_\phi := -D_{\theta^*}$ for $\theta^* := \arg\min_\theta L_{\mathrm{disc}}(\theta,\phi)$), we have

$$\frac{d\phi}{dt} = -\mathbb{E}_{w\sim\omega}[\nabla_\phi f_\phi(w)\cdot\nabla_y\Psi_\phi(y)\big|_{y=f_\phi(w)}]. \qquad (34)$$

This matches the BBVI update rule (10) with $\log q_\phi(y) - \log p(y)$ replaced by the discriminator $\Psi_\phi(y)$. Hence, analogous to (16), the generated points x satisfy

$$\frac{dx}{dt} = -\mathbb{E}_{y\sim q_\phi}[k_\phi(x,y)\nabla_y\Psi_\phi(y)], \qquad (35)$$

where here k_ϕ is defined as in (13) by the neural tangent kernel of the generator f_ϕ. Finally, it was observed that the optimal discriminator Ψ_q of the minimax GAN equals the functional derivative of the Jensen–Shannon divergence (Chu et al., 2019; 2020); hence we conclude:

Claim 5. *Minimax GAN updates generate a kernel gradient flow of the Jensen–Shannon divergence $J(q) := D_{\mathrm{JS}}(p_{data}, q)$, with a Riemannian metric determined by the neural tangent kernel of the generator f_ϕ.*

Similarly, non-saturating and Wasserstein GAN updates generate kernel gradient flows on the directed divergence $J(q) := \mathrm{KL}(\frac{1}{2}p_{\mathrm{data}} + \frac{1}{2}q\,\|\,p_{\mathrm{data}})$ and Wasserstein-1 distance $J(q) := W_1(p_{\mathrm{data}}, q)$ respectively.

6. Conclusion

We have cast SVGD and BBVI, as well as the dynamics of GANs, into the same theoretical framework of kernel gradient flow, thus identifying an area ripe for further study.

References

Ambrosio, L., Gigli, N., and Savaré, G. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Chu, C., Blanchet, J., and Glynn, P. Probability functional descent: A unifying perspective on GANs, variational inference, and reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1213–1222, 2019.

Chu, C., Minami, K., and Fukumizu, K. Smoothness and stability in GANs. In *International Conference on Learning Representations*, 2020.

Duncan, A., Nüsken, N., and Szpruch, L. On the geometry of Stein variational gradient descent. *arXiv preprint arXiv:1912.00894*, 2019.

Feng, Y., Wang, D., and Liu, Q. Learning to draw samples with amortized Stein variational gradient descent. In *UAI*, 2017.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pp. 8571–8580, 2018.

Kingma, D. P. and Welling, M. Auto-encoding variational Bayes. 2014.

Liu, C., Zhuo, J., Cheng, P., Zhang, R., and Zhu, J. Understanding and accelerating particle-based variational inference. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 4082–4092, 2019.

Liu, Q. Stein variational gradient descent as gradient flow. In *Advances in Neural Information Processing Systems*, pp. 3115–3123, 2017.

Liu, Q. and Wang, D. Stein variational gradient descent: A general purpose Bayesian inference algorithm. In *Advances in neural information processing systems*, pp. 2378–2386, 2016.

Lu, J., Lu, Y., and Nolen, J. Scaling limit of the Stein variational gradient descent: The mean field regime. *SIAM Journal on Mathematical Analysis*, 51(2):648–671, 2019.

Otto, F. The geometry of dissipative evolution equations: the porous medium equation. 2001.

Ranganath, R., Gerrish, S., and Blei, D. Black box variational inference. In *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, pp. 814–822, 2014.

Roeder, G., Wu, Y., and Duvenaud, D. K. Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, pp. 6925–6934, 2017.