Depth First Learning: SVGD

Week 5: Stein Variational Gradient Descent as Gradient Flow

June 13, 2019

SVGD as Gradient Flow is one of the first papers that analyzes the dynamics and theoretical properties of SVGD. In this week, we structure the notes a bit differently, as the paper is (relatively) more dense than weeks previous. After a brief review of SVGD and notation, we will follow the structure of the paper; each section will contain a background (relevant to that subsection of the main paper), and a discussion on the section. Each main header in this week is relatively self-contained, allowing you to pick and choose what you would like to take a look at.

The goal of this week is to work through SVGD as Gradient Flow [7], specifically understanding the following part of the Introduction: "characterize the SVGD dynamics using an evolutionary process of the empirical measures of the particles that is known as Vlasov process in physics, and establish that empirical measures of the particles weakly converge to the given target distribution. We develop a geometric interpretation of SVGD that views SVGD as a gradient flow of KL divergence, defined on a new Riemannian-like metric structure imposed on the space of density functions."

Throughout this note, we will redefine terms that have been seen in previous references, as some play a more central role in this week's discussion than they have in previously. In certain cases, the background will use different notation that described in the paper; we will ensure that we connect the notation (wherever possible) back to the paper, often after the background is fully covered.

1 Table of Contents

    (a) Brownian Motion
    (b) Fokker-Planck Equations

5. SVGD as Gradient Flow
    (a) Optimal Transport and Wasserstein Distances
    (b) Tangent Spaces
    (c) Riemmanian Metrics
    (d) Geodesics
    (e) The Geometry of SVGD

2 SVGD Review

We are dealing with a probability measure ν_p with a positive, (weakly) differentiable density $p(x)$ on some open set $\mathcal{X} \subseteq \mathbf{R}^d$. Because we do not have access to the measure of interest, we want to approximate ν_p with some set of N particles $\{x_i\}$.

Weak Convergence A measure μ_n *converges weakly* to μ_0 in some metric space $(\mathcal{X}, \rho)$ if for every bounded, continuous function f on $\mathcal{X}$, we have $\int f d\mu_n \to \int f d\mu_0$ as $n \to \infty$. If some random elements $\{\xi_i\}$ for $i = 1, 2...$ taking values in $\mathcal{X}$ s.t the distribution of $\xi_n = \mu_n$, we write $\xi_n \to \xi_0$ and say ξ_n converges *in distribution* to ξ_0 if μ_n converges weakly to μ_0 [2].

Convergence in Distribution Let $X_1, X_2, ...$ be random variables defined on some probability space $(\Omega, \mathcal{F}, \rho)$. $X_1, X_2, ...$ converge to random variable X if for all f, $\mathbf{E}f(X_n) = \mathbf{E}f(X)$ as $n \to \infty$ [1].

Our set of particles $\{x_i\}$ has an empirical measure $\hat{\mu}_n(dx) = \sum^n \delta(x - x_i)/ndx$ which weakly converges to measure ν_p as $n \to \infty$. For some test function h (bounded and continuous)[1], this means $\mathbf{E}_{\hat{\mu}_n} h = \mathbf{E}_{\nu_p} h$.

SVGD [8] does this by initializing the particles randomly, updating them with the following map:

$$\mathbf{T}(x) = x + \epsilon\phi(x) \tag{1}$$

where ϵ is a step size and $\phi(x)$ is a perturbation direction (referred to as the **velocity field** with regards to later analysis). SVGD, as we have seen, chooses $\phi(x)$ to maximally decrease the KL divergence of the current particle distribution with the target distribution. They solve the following optimization:

[1] h from the paper plays the role of f in the definition

$$\max_{\phi \in \mathcal{H}} = \left\{ -\frac{d}{d\epsilon} KL(\mathbf{T}\mu||\nu_p)|_{\epsilon=0} \quad s.t \quad ||\phi||_{\mathcal{H}} \leq 1 \right\} \tag{2}$$

which, through connections through Stein's Lemma and Identity, turns the Equation 2 into one that gives the *Stein Discrepancy*:

$$\mathbb{D}(\mu||\nu_p) := \max_{\phi \in \mathcal{H}} \mathbf{E}_\mu[\mathcal{S}_p \phi] \quad s.t. \quad ||\phi||_{\mathcal{H}} \leq 1 \tag{3}$$

where $\mathcal{S}_p$ is the Stein Operator. This discrepancy, given that $\mathcal{H}$ is sufficiently large, provides a "distance" between the measures ($\mathbb{D}(\mu||\nu_p) = 0$ iff $\mu = \nu_p$). The authors then use an RKHS as the function class $\mathcal{H}$, leading to a closed-form solution and tractable optimization procedure. In the rest of this work, we *assume*, as the authors do, that $\mathcal{H}$ *is* expressive enough for the previous iff statement holds [2]. Namely, as shown in [5], the work assumes that the following holds true: Given a sequence of probability measures $\{\mu_l\}_{l=1}^{\infty}$,

$$\mathbb{D}(\mu_l||\nu_p) \to 0 \iff \mu_l \to \nu_p \quad \text{as} \quad l \to \infty \tag{4}$$

3 Large Sample Asymptotic of SVGD

Of course, there exists an optimal transform to the optimization problems in the previous sections, denoted $\mathbf{T}_{\mu,p}$ with velocity field that we denote as $\phi^*_{\mu,p}$. This fully characterizes SVGD dynamics, meaning that the empirical measure at any iteration can be found by recursively applying $\mathbf{T}_{\mu,p}$ to the initial empirical measure. This map is defined in the paper as $\mathbf{\Phi}_p$, which is nonlinear due to its dependence on the current empirical measure.

Dirac Delta Functions Dirac Delta functions were introduced to model densities of an idealized point mass. A Dirac delta function is one that is equal to 0 everywhere except 0, and whose integral over the real number line is 1.

"Equal in the Sense of Distributions" Distributions p and q are said to be equal in distribution if $p[\phi] = q[\phi]$ $\forall \phi \in D(U)$, where $D(U)$ is the set of test functions on open set U. If F and G are generalized functions, then F is equal to G if $\int_U F(x)\phi(x)dx = \int_U G(x)\phi(x)dx$ $\forall \phi \in D(U)$ [4]. [3]

If our current measure μ has density q and step size ϵ is small enough, the optimal map, $\mathbf{T}$, is invertible, giving us the density q' of $\mu' = \mathbf{T}\mu$ via change of variables:

$$q'(z) = q(\mathbf{T}^{-1}_{\mu,p}(z)) \cdot |\det(\nabla \mathbf{T}^{-1}_{\mu,p}(z))| \tag{5}$$

[2] If $\mathcal{H}$ *wasn't* expressive / large enough, the discrepancy could be 0 even if the measures are not equal

[3] These definitions will help you understand the extra statement in the paper: "When μ is an empirical measure and q is a Dirac delta function, this equation still holds formally in the sense of distribution (generalized functions)"

If we assume that the initial empirical measure $\hat{\mu}_0^n$ converges to some measure μ_0^∞ as $n \to \infty$[4], we can assume that, at any finite iteration, this same idea applies *if* the map $\mathbf{\Phi}_p$ satisfies some Lipchitz condition.

Lipschitz Constant Intuitively, the Lipschitz condition bounds how fast a function can change in value: $|f(x_1) - f(x_2)| \leq K|x_1 - x_2|$, where K is known as the Lipschitz constant.

Bounded Lipschitz Metric Define $BL(X, d)$ given some metric space X and distance function d to be $BL(X, d) := \{f : \mathcal{X} \to \mathbf{R}; f \text{ bounded and Lipschitz.}\}$. Then, for $f \in BL(X, d)$, define $||f||_{BL} = ||f||_\infty + Lip(f)$, where $||f||_\infty$ is the uniform norm and $Lip(f) = ||f||_{Lip} = \sup_{x,y \in X, x \neq y} \frac{|f(x)-f(y)|}{d(x,y)}$. The BL metric, given two measures, is defined to be the difference of the two means of the measures. The BL Metric metricizes weak convergence, which means that $BL(\mu_n, \nu_p) \to 0$ iff $\mu_n \to \nu_p$.

Since BL implies weak convergence, and at each iteration, the empirical measure converges to the true measure as $n \to \infty$, but only with an extremely-quickly decaying step size (Equation 11, Lemma 3.1, and Theorem 3.2 in [7]). In addition, the authors note that Lemma 3.1 can only be used when $\mathcal{X}$ is compact (which is generally not the case in these applications), and to my knowledge, no further result has improved upon this one. However, assuming these conditions hold, Theorem 3.2 explains that we only need to ensure that the initial measure has a finite KL-Divergence with ν_p; after, a rapidly-enough decaying step size and the map $\mathbf{\Phi}_p$ take care of the rest.

4 Large Time Asymptotic of SVGD and Continuous Time Limits

We skip Section 3.2 (the theorems and results are laid out in a relatively self-explanatory way), which further implies how we should set how we should set our step size. Instead, we look at the continuous time limit (set $t = \epsilon l$ and infinitesimally small step size ϵ), which generates a Partial Differential Equation. Before we describe the density dynamics, we will talk about a few extra topics that will clarify Section 3.3 in the original paper.

Brownian Motion In fluid dynamics, particle interactions are so chaotic that we model the resulting system assuming that particles move randomly and independently of their past motion. Brownian Motion can also be though of as the limit of a random walk as time and space increments approach zero. Formally, we describe a stochastic process $B(t)$ as a Brownian motion if it satisfies four characteristics [3].

1. *Grounding in Space*: $B(0) = x$

[4] Can be achieved by MCMC.

2. *Continual Randomness and Independent Increments*: for all $t_i \leq t_{i+1}$, $B(t_{i+1}) - B(t)$ are independent random variables. This says that each particle at all times is getting (randomly) affected by fluid molecules.

3. *Normality*: for all $t \geq 0, h > 0$, $B(t+h) - B(t)$ is normally distributed according to $\mathcal{N}(0, h)$. The expected displacement of any particle should be proportional to the time it has been travelling, and should be symmetrically distributed about its starting point.

4. *Continuity* As this is a physical system, almost surely, $B(t)$ is continuous.

Fokker-Planck For an Ito Process driven by a standard Brownian variable W_t, we can use the following SDE to describe it:

$$dX_t = \mu(X_t, t)dt + \sigma(X_t, t)dW_t \tag{6}$$

with *drift* $\mu(X_t, t)$ and *diffusion coefficient* $D(X_t, t) = \frac{1}{2}\sigma^2(X_t, t)$, we can get the Fokker-Planck equation [6]:

$$\frac{\partial}{\partial t}p(x,t) = \frac{\partial}{\partial x}[\mu(x,t)p(x,t)] + \frac{\partial^2}{\partial x^2}[D(x,t)p(x,t)] \tag{7}$$

The Fokker-Planck equation describes the evolution of a the probability distribution of a field, say a particle's velocity under random forces (Brownian motion).

In SVGD, we can describe the evolution of the density (Equation 5) with a non-linear, deterministic Fokker-Planck equation:

$$\frac{\partial}{\partial t}q_t(x) = -\nabla(\phi^*_{q_t,p} q_t(x)) \tag{8}$$

The deterministic forces in our setting zero out the diffusion term, but the velocity field's dependence on the *current* particle density makes the drift term nonlinear ($\phi^*_{q_t,p} = \mathbf{E}_{x' \sim q_t}[\mathcal{S}_p^{x'} \otimes k(x, x')$). While we didn't discuss it, Theorem 3.3(2) in the original paper describes the convergence of the empirical measure to the true measure with sufficiently-decaying step size; the same type of result applies for the continuous time limit (Theorem 3.4), restated below:

Theorem 3.4 *Assuming $\{\mu_t\}$ are probability measures whose densities satisfy Equation 8, and μ_0 has finite KL with the target measure, then*

$$\frac{d}{dt}KL(\mu_t || \nu_p) = -\mathbb{D}(\mu_t || \nu_p)^2$$

However, the original FP equation works only with differentiable densities, so the last part of the section (Equations 13 and 15) describes how the FP can be translated to measure-value PDEs, allowing us to use empirical measures. The resulting PDE is a special form of a Vlaslov Process, which, along with the original Fokker-Planck equation in Equation 8, generates a geometry that we explore in the next section.

5 SVGD as Gradient Flow

The resulting Vlaslov process in the previous section has a geometric interpretation as the "gradient flow for minimizing the KL divergence functional, defined on a new type of optimal transport metric on the space of density functions induced by Stein operator," but before we understand what this sentence means, let's cover some prerequisites.

Optimal Transport An easy way to understand Wasserstein Distances is to motivate the most common problem application (atleast w.r.t machine learning) it gets used in: Optimal Transport. Consider the job of a landscaper. Given some dirt (with a total mass of 1, which will become evident on "why" in a few paragraphs) μ, she is tasked with transforming that dirt into some other configuration ν. As in the physical world, there will be some cost associated with transporting the mass from some point x to some point y, which is governed by some non-negative cost function $c(x, y)$. In order to generate the new configuration, she can generate a *transport plan* $\gamma(x, y)$, which describes the amount of mass to move from x to y. Of course, this plan won't be unique, so we define the optimal transport plan as the one with minimum cost.

Now, back in mathematics, it turns out $\mu(x)$ and $\nu(x)$ are probability distributions on some space $\mathcal{X}$. An additional constraint in the optimal transport problem is that the plan γ must be a joint probability distribution with marginals μ and ν.

Wasserstein Distances Wasserstein Distances characterize and formalize distances between probability distributions, which as we will see soon, connect back to our cost function $c(x, y)$ in the optimal transport problem. Given some metric space (M, d), for $p \geq 1$, define $P_p(M)$ to be all of of the probability measures μ on M with a finite p^{th} moment. The Wasserstein distance between two measures W is then defined to be:

$$W_p(\mu, \nu) := (\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{M \times M} d(x, y)^p d\gamma(x, y)^{(1/p)}) \tag{9}$$

where Γ is the set of all couplings (in the OT terminology, plans) of μ and ν. In machine learning, we often use the W_1 distance; in the OT setting, if we take the cost function to simply be $d(x, y)$, we recover the W_1 Wasserstein distance.

SVGD's geometric analysis focuses on the relationship between two densities: q and q', which is obtained by applying $\mathbf{T}(x) = x + \phi(x)dt$ with infinitesimal dt on some particle x sampled from q with some ϕ that lives in Hilbert space $\mathcal{H}$. This allows us to define the relationships between the log-density and density of q' to q:

$$\log q'(x) = \log q(x) - \mathcal{S}_q \phi(x) dt, \tag{10a}$$

$$q'(x) = q(x) - q(x)\mathcal{S}_q \phi(x) dt \tag{10b}$$

Using the fact that $\mathcal{S}_q \phi = \frac{\nabla \cdot (\phi q)}{q}$[5], the authors define an operator $q\mathcal{S}_q$ by $q\mathcal{S}_q \phi(x) = q(x)\mathcal{S}_q = \nabla \cdot (\phi(x)q(x))$. This, along with Equations 10, tells us that the Stein operator $\mathcal{S}_q$[6] translates an $\phi-$perturbation on x to a perturbation on the log-density.

If we define $\mathcal{H}_q$ to be the space of functions of the form $\mathcal{S}_q \phi$ with $\phi \in \mathcal{H}$, and define $q\mathcal{H}_q$ to be the space of functions of form qf where $f \in \mathcal{H}_q$, then we can look at the inverse of the Stein operator for functions in $\mathcal{H}_q$. For each $f \in \mathcal{H}_q$, there is some unique function ψ that has minimum $|| \cdot ||_{\mathcal{H}}$ norm in the set of functions that satisfy $\mathcal{S}_q \psi = f$ (which is the Stein equation). Due to RKHS, this means that $\mathcal{H}_q$ inherits an inner product from $\mathcal{H}$ (shown in Equation 19 of the paper), given by the fact that $\mathcal{H}_{\text{II}}$ is itself an RKHS.

Taking q and some infinitesimally-perturbed density $q' = q + qf\text{d}t$, the $\psi_{q,f}$ can be seen as the optimal transform that has minimum $||\cdot||_{\mathcal{H}}$ norm, and $\psi_{q,f}$ takes the place of the optimal pertubation direction in Equation 1. $\psi_{q,f}$ therefore defines a distance between q and q', which generates a metric structure in the distribution space:

$$W_{\mathcal{H}}(q, q') \coloneqq ||\psi_{q,f}||_{\mathcal{H}} \text{d}t = ||q - q'||_{q\mathcal{H}_q} \tag{11}$$

Before we continue onto the implications of this Wasserstein distance, we will need to cover two more topics related to geometry. We will not need a ton of background on this, and the few concepts we will cover are relatively basic. A great overview of differential geometry can be found on Roger Grosse's Metacademy Roadmap for Differential Geometry.

Tangent Spaces The intutition behind tangent spaces can be described in a single picture, seen in Figure 1:

Given some manifold M and point x, the tangent space $T_x M$ consists of all possible directions which you can pass tangentially through x. If $\gamma(t) \in M$ defines a positional curve, the tangent vector is the velocity, whereas the tangent space could be all possible velocities you can have at that point. A more formal definition regarding charts and equivalence relations can be found, but for our understanding of this last section, is not strictly necessary. One point that is important to note is that tangent spaces are *chart-invariant*, meaning they do not depend on the type of chart (with lots of simplifications, coordinate frame) being used.

[5] Unfortunately, I'm not sure if this is a past result, or one derived from Equation (2) in the original paper.

[6] For the rest of this note, I will forgo including the "respectively" results, which correspond to Equation 10b

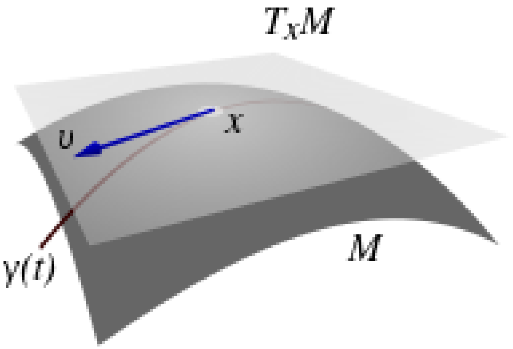

Figure 1: The tangent space T_xM of point x lying on manifold M. Source: Wikipedia

Riemmanian Metric Taken directly from Roger Grosse's notes, a Riemmanian Metric "assigns an inner product to the tangent space at each point of a differentiable manifold. It gives a local notion of distance, and allows one to define (and generalize) notions such as orthogonal vectors, the norm of a vector, the length of a path, and the distance between two points."

Back to the paper, we can now try to understand the implication of the distance in Equation 11. In the infinitesimal neighborhood $\{q' : W_{\mathcal{H}}(q, q') \leq \mathrm{d}t$, we have densities of the form:

$$q' = q + g\mathrm{d}t, \quad \forall g \in q\mathcal{H}_q, \quad ||g||_{q\mathcal{H}q} \leq 1 \tag{12}$$

This means $q\mathcal{H}_q$ can be seen as the tangent space around density q (where q takes the place of x in Figure 1). Then, the norms inherited from the virtue of RKHS both define a metric structure related to the distance in Equation 11.

This gives us an optimal-transport based metric, dependent on $\mathcal{H}$, between two distributions, allowing us to understand the implications of Equation 20 in the original paper. Its connections are highlighted with its relation to Langevian dynamics (Section 3.5), which we have decided not to cover in the interest of time.

Understanding the main result of the paper, Theorem 3.5, requires the basic understanding of one more topic: *covariant derivatives*.

Covariant Derivative A covariant derivative is a generalization of directional derivatives, specified using the tangent vectors of a manifold. Much like a standard, directional derivative, the covariant derivative $\mathrm{grad}_u v$ takes in a vector u at point p and a vector field v defined in the neighborhood of p. A covariant derivative must be independent of the manner in which it is expressed in a coordinate system (which makes the tangent space a good fit to define it).

Given some functional $F(q)$ given our density q, the covariant gradient $\mathrm{grad}_{\mathcal{H}} F(q)$ of $F(q)$ is a map from q to an element of the tangent space of q ($q\mathcal{H}_q$)) that satisfies the following condition:

$$F(q + f dt) = F(q) + \langle \text{grad}_{\mathcal{H}} F(q), f dt \rangle_{q\mathcal{H}_q} \tag{13}$$

for any f in tangent space $q\mathcal{H}_q)$ where the $\langle \cdot, \cdot \rangle_{q\mathcal{H}_q}$ is the induced norm from the RKHS.

Theorem 3.5 then relates the gradient of this functional, when taken to be the KL-divergence between q and p, to the original gradient flow of KL-Divergence under the metric $\mathbb{W}_{\mathcal{H}}(\cdot, \cdot)$.

References

[1] Encyclopedia of math: Convergence in distribution.

[2] Encyclopedia of math: Weak convergence of probability measure.

[3] A. Dahl. A rigorous introduction to brownian motion.

[4] P. DuChateau. Introduction to the theory of distributions.

[5] J. Gorham and L. Mackey. Measuring sample quality with kernels, 2017.

[6] S. Hottovy. The fokker-planck equation.

[7] Q. Liu. Stein variational gradient descent as gradient flow, 2017.

[8] Q. Liu and D. Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm, 2016.