The Impact of Kernel Parameters on Performance of the Stein Variational Gradient Descent Algorithm

Ji Won Park and Junmo Ryang

July 31, 2017

1 Introduction

In this paper, we investigate Stein Variational Gradient Descent (SVGD), an algorithm that allows a set of particles to evolve according to a combination of gradient descent and convolution with a Gaussian kernel, so that the particles' empirical density approximates a target probability distribution. Unlike Monte Carlo (MC) algorithms, which operate via randomness, SVGD deterministically updates the particle positions so that each iteration decreases the KL divergence between the particles and the target distribution.[1] For some distributions, SVGD may prove to be a superior to MC in terms of accuracy and computation time.[2] It often requires fewer numbers of particles than the counterpart, because it takes full advantage of the gradient information of the target distribution; for one particle, it degenerates to the gradient descent algorithm.[1] It may be difficult, on the other hand, to analyze the convergence of MC.[2]

We restrict our discussion to one-dimensional distributions. Given a target probability density function (PDF) $\rho(x)$, we seek to understand an algorithm that constructs particles $\{x_1, \ldots, x_N\}$ in $\mathbb{R}$ such that

$$\int_{\mathbb{R}} f(x)\rho(x)\,dx \approx \frac{1}{N}\sum_{k=1}^{N} f(x_k) \tag{1.1}$$

for some general function $f(x)$. We call this set of particles and their distribution the *particle approximation*. Since we are essentially replacing a weighted average across the entire real line with an even average across a finite number of points, the position of these particles determines the approximate distribution.

The algorithm is a *particle method* that spawns and moves a number of *particles* with positions $\{X_1, \ldots, X_N\}$ on the real line. The resultant position of the particles after long time is set of approximation points given by the algorithm. The dynamics of the particles and indeed the method itself can be characterized by a system of ordinary differential equations (ODEs) of the form:

$$\frac{d}{dt}X_i = \phi(X_i), \quad i = 1, \ldots, N \tag{1.2}$$

In all particle methods, the velocity field ϕ must somehow incorporate features of the target PDF, so that the dynamics of the particles will depend upon the target distribution. In particular, SVGD is an *interacting* particle method, meaning that $\phi(X_i)$ depends on the position of all particles $\{X_1, \ldots, X_N\}$ rather than just X_i itself.

The following sections present theory behind the derivation and expected behavior of a particular interacting particle method in addition to interesting results found from the implementation and testing of the algorithm.

2 Theory

2.1 Particle Approximations

Let X be a continuous random variable with PDF ρ. That is,

$$\mathbb{P}(X \in (a, b)) = \int_a^b \rho(x)\, dx \tag{2.1}$$

In addition, let F_X be the cumulative distribution function (CDF) of X. It

can be represented:

$$F_X(x) = \int_{-\infty}^{x} \rho(\tau)\, d\tau \tag{2.2}$$

In the coming sections, we will commonly refer to continuous random variables and their distribution functions with X, ρ, and F_X. The target distribution we are attempting to approximate comes in this form.

Let Y be a discrete random variable. Its distribution can be described by a set of weights $\{w_1, \ldots, w_N \mid \sum w_i = 1\}$ corresponding to points $\{x_1, \ldots, x_N\}$ such that

$$\mathbb{P}(Y = x_i) = w_i, \quad i = 1, \ldots, N \tag{2.3}$$

Note that Y does not have a traditional PDF, since it is a discrete random variable, but it does have a CDF, which we call F_Y.

$$F_Y(x) = \begin{cases} 0 & x \leq x_1 \\ \sum_{j=1}^{i} w_j & x_i \leq x \leq x_{i+1}, \quad i = 1, \ldots, (N-1) \\ 1 & x_N \leq x \end{cases} \tag{2.4}$$

Using the Dirac delta function, we can also define a generalized probability density function μ

$$\mu(x) = \sum_{i}^{N} w_i\, \delta(x - x_i) \tag{2.5}$$

that fulfills the role for the PDF of Y:

$$\begin{aligned} \mathbb{P}(a \leq Y \leq b) &= \int_a^b \mu(x)\, dx \\ &= \int_a^b \sum_{i}^{N} w_i\, \delta(x - x_i)\, dx \end{aligned} \tag{2.6}$$

which is the sum of w_i for all x_i between a and b.

Our particle approximation yields a specific case of this general form. We use equal weights $w_i = \frac{1}{N}$ for all of the points $\{X_1, \ldots, X_N\}$ produced by the

interacting particle method, resulting in the following cases of the equations above:

$$\mathbb{P}(Y = X_i) = \frac{1}{N}, \quad i = 1, \ldots, N \tag{2.7}$$

$$F_Y(x) = \begin{cases} 0 & x \leq X_1 \\ \frac{i}{N} & X_i \leq x \leq X_{i+1}, \quad i = 1, \ldots, (N-1) \\ 1 & X_N \leq x \end{cases} \tag{2.8}$$

$$\mu(x) = \frac{1}{N} \sum_i^N \delta(x - X_i) \tag{2.9}$$

$$\begin{aligned} \mathbb{P}(a \leq Y \leq b) &= \int_a^b \mu(x)\, dx \\ &= \int_a^b \frac{1}{N} \sum_i^N \delta(x - X_i)\, dx \end{aligned} \tag{2.10}$$

We will commonly refer to discrete random variables and their distribution functions with Y, μ, and F_Y. The particle approximation to the target distribution comes in this form.

2.2 Metric for Evaluating Particle Approximations

In order to evaluate the accuracy with which a particle approximation approximates a target distribution, we employ a distance metric that measures the maximum the difference between CDFs. Given random variables X and Y that have CDFs F_X and F_Y respectively, we define the metric d as follows:

$$d(X, Y) = \max_{x \in \mathbb{R}} |F_X(x) - F_Y(x)| \tag{2.11}$$

If we apply this metric to random variables X and Y with PDF ρ and generalized PDF μ, we quickly arrive at a lower bound for the distance metric.

Theorem 2.1. *Let X be a continuous random variable with PDF ρ and CDF F_X, and let Y be a discrete random variable with generalized PDF μ and CDF F_Y. Then,*

$$d(X, Y) \geq \frac{1}{2N}$$

Proof. Suppose $d(X,Y) < \frac{1}{2N}$. We will proceed by contradiction.

By Lemma A.1 (proven in the Appendix), if P is the set of the points $\{X_1, \ldots, X_N\}$ for which μ is non-zero, the following is true:

$$\begin{aligned} d(X,Y) &= \max_{x \in \mathbb{R}} |F_X(x) - F_Y(x)| \\ &= \max_{p \in P} \left(|F_X(p) - F_Y(p)|, \lim_{x \to p^-} |F_X(x) - F_Y(x)| \right) \end{aligned}$$

Combined with our initial supposition,

$$d(X,Y) = \max_{p \in P} \left(|F_X(p) - F_Y(p)|, \lim_{x \to p^-} |F_X(x) - F_Y(x)| \right) < \frac{1}{2N},$$

$$|F_X(p) - F_Y(p)| < \frac{1}{2N} \quad \textit{and} \quad \lim_{x \to p^-} |F_X(x) - F_Y(x)| < \frac{1}{2N} \qquad \forall\, p \in P.$$

Because F_X is everywhere differentiable and therefore everywhere continuous, $\lim_{x \to p^-} F_X(x) = F_X(p)$. Using this fact, we rewrite the two inequalities above:

$$-\frac{1}{2N} < F_X(p) - F_Y(p) < \frac{1}{2N} \quad \textit{and} \quad -\frac{1}{2N} < F_X(p) - \lim_{x \to p^-} F_Y(x) < \frac{1}{2N}$$

If we flip the parity of the left inequality and sum the two, we arrive at a new inequality:

$$-\frac{1}{2N} < F_Y(p) - F_X(p) < \frac{1}{2N} \quad \textit{and} \quad -\frac{1}{2N} < F_X(p) - \lim_{x \to p^-} F_Y(x) < \frac{1}{2N}$$

$$-\frac{1}{N} < F_Y(p) - F_X(p) + F_X(p) - \lim_{x \to p^-} F_Y(x) < \frac{1}{N}$$

$$-\frac{1}{N} < F_Y(p) - \lim_{x \to p^-} F_Y(x) < \frac{1}{N}$$

$$\left| F_Y(p) - \lim_{x \to p^-} F_Y(x) \right| < \frac{1}{N} \qquad \forall\, p \in P$$

Now note that by the construction of $F_Y(x)$:

$$F_Y(X_i) - \lim_{x \to X_i^-} F_Y(x) = w_i$$

Substituting into the inequality above:

$$|w_i| < \frac{1}{N} \qquad \forall\, p = X_i \in P = \{X_1, \ldots, X_N\}$$

Since the weights w_i are always positive, we can drop the absolute value. If we sum over all N values of i, we arrive at the problematic:

$$\sum_{i=1}^{N} w_i < \sum_{i=1}^{N} \frac{1}{N} = 1$$

This is a contradiction, since by definition, $\sum w_i = 1$. Our supposition that $d(X, Y) < \frac{1}{2N}$ must be incorrect. Therefore,

$$d(X, Y) \geq \frac{1}{2N}$$

□

Although it may be obscured in the proof above, the intuition behind this bound is quite simple. In Figure 1 below, we can see that lengths a and b sum to w_i.

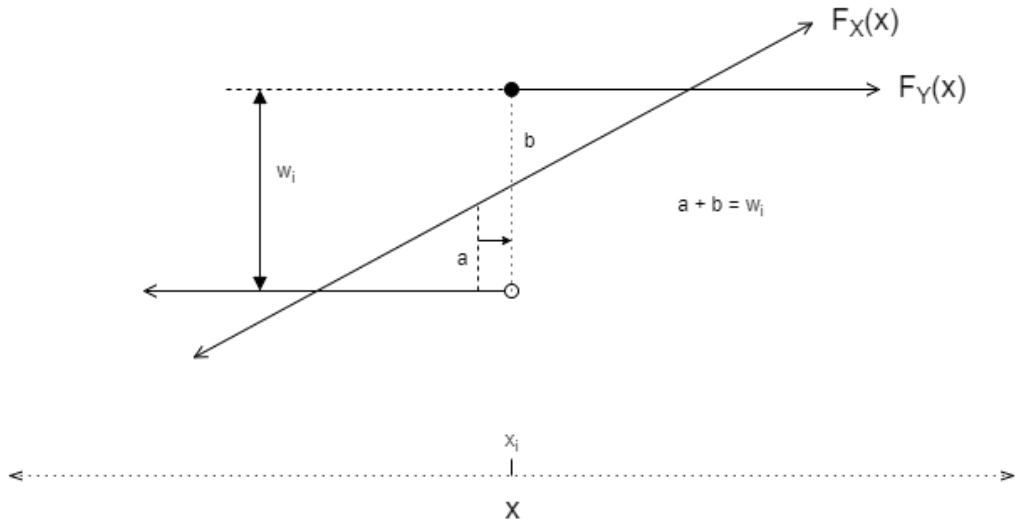

Figure 1: Intuition for Metric Lower Bound

Since both a and b are of the form $|F_X(x) - F_Y(x)|$, they each lower bound $d(X, Y)$. If we wish to maximize the lowest lower bound, they must be equal. In total, there are N points x_i each with their own a and b constructions that lower bound $d(X, Y)$. Since $\sum w_i = 1$, if we make all $2N$ construction lengths equal, each would be $\frac{1}{2N}$.

2.2.1 Note on the Choice of Metric

The distance metric featured above is not the only usable metric. For example, if we revisit the goal of producing points $\{x_1, \ldots, x_N\}$ in $\mathbb{R}$ such that

$$\int_{\mathbb{R}} f(x)\rho(x)\,dx \approx \frac{1}{N}\sum_{k=1}^{N} f(x_k) \tag{2.12}$$

we can derive another metric based on the difference between the left- and right-hand sides of the equation. If we let X be a continuous random variable with PDF ρ, and if we let Y be a discrete random variable with equal weights across the points $\{x_1, \ldots, x_N\}$, and generalized PDF μ, then we would want to minimize the following metric:

$$\begin{aligned} d(X, Y) &= \sup_{f\in\mathcal{H}} \left| \int_{\mathbb{R}} f(x)\rho(x)\,dx - \frac{1}{N}\sum_{k=1}^{N} f(x_k) \right| \\ &= \sup_{f\in\mathcal{H}} \left| \int_{\mathbb{R}} f(x)\rho(x)\,dx - \int_{\mathbb{R}} f(x)\mu(x)\,dx \right| \end{aligned} \tag{2.13}$$

for a certain class of functions $\mathcal{H}$. The problem then becomes how to choose this class of functions against which we will measure our particle approximations.

Ultimately, we chose the distance metric of the difference between CDFs for one main reasons. Its ease of computation allowed for faster testing of simulations. Lemma A.1 helped to decrease the number of computations, and Theorem 2.2 gave a benchmark against which to compare algorithm performance.

2.3 Derivation of ODE Dynamics

2.3.1 Deterministic ODE: SVGD

The SVGD algorithm can be expressed by the following ODE[2],

$$\frac{d}{dt}X_i = \phi(X_i) = -\frac{1}{N}\sum_{j=1}^{N}\left[K(X_i - X_j)E'(X_j) + \frac{1}{\beta}K'(X_i - X_j)\right] \quad (2.14)$$

where the input score function $E(x)$ is related to the target distribution by

$$\rho(x) = \frac{1}{Z}e^{-\beta E(x)} \quad (2.15)$$

,with Z a normalizing constant, and K is a positive-definite kernel. We choose a Gaussian kernel with variance σ^2, i.e.

$$K(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}. \quad (2.16)$$

The parameter β has the physical significance of inverse time, such that small betas allow the particles to move faster toward a steady state, if there is one. In the implementation, we use a uniform step size ϵ such that, at each iteration, $X_i \to x_i + \epsilon\phi(X_i)$. We assume all particles have the same weight, so at each time step t, the empirical distribution is

$$\mu(x,t) = \frac{1}{N}\sum_{i}^{N}\delta(x - X_i). \quad (2.17)$$

In Equation 2.14, the first term of $\phi(X_i)$ in square brackets is the kernalized gradient descent which takes particles to highest-probability regions and the second term spreads particles out in space. Note that, if $N = 1$ or if $K(x) = \delta(x)$, Equation 2.14 degenerates to gradient descent.

2.3.2 The Continuous Time Limit: Fokker-Planck Equation

It can be shown that, as $\epsilon \to 0$, the time evolution of the empirical density $\mu(x,t)$ in 2.3.1 reduces to a nonlinear partial differential equation,

$$\begin{aligned}\frac{\partial}{\partial t}\mu(x,t) &= -\frac{\partial}{\partial x}\left(\phi(x)\mu(x,t)\right) \\ &= \frac{\partial}{\partial x}\left(E'(x)\mu(x,t)\right) + \frac{1}{\beta}\frac{\partial^2}{\partial x^2}\left(\mu(x,t)\right) \quad (2.18)\end{aligned}$$

Equation 2.17 is called the Fokker-Planck equation. See B.3 in [1] for the derivation. In particular, if $N \to \infty$ as well, $\mu(x,t) \to \rho(x,t)$.

2.3.3 Stochastic ODE: Langevin Dynamics

It is worth comparing SVGD, a deterministic ODE system, to Langevin dynamics, a stochastic one. The equivalence of Langevin dynamics and the Fokker-Planck equation is given in [5]. Langevin dynamics can be expressed by the ODE[3],

$$\frac{d}{dt}X_i = \psi(X_i) = -E'(X_i) + \sqrt{\frac{2}{\beta}}\,\eta(t) \tag{2.19}$$

where $\eta(t) = \frac{dw(t)}{dt}$ and $w(\tau) = \int_0^\tau \eta(t)dt$ is a Wiener process, the sum of the steps of a random walk after time τ.[4] A property of random walks is that the $w(\tau)$ scales with $\sqrt{\tau}$. In implementation, η_t can be simulated by $\frac{\xi}{\sqrt{\epsilon}}$ where $\xi \sim N(0,1)$ and ϵ is a small time step. The physical interpretation of this white noise term is that solvent molecules randomly kick a given particle without contributing to the average velocity of the particle.[4]

Note that the first term of $\psi(X_i)$ in Equation 2.19 is identical to the first term of $\phi(X_i)$ in Equation 2.14 in the absence of a kernel, i.e. when the kernel is a delta function.

It will be instructive to compare the effect of the inverse-time parameter β on SVGD and Langevin algorithms. Because β changes the target distribution, an algorithm for which the convergence time is reasonable for a large range of βs will be one that can practically accommodate a large variety of distributions.

2.4 Simple Cases

We would like to explore the steady state of the Gaussian Kernel ODE system in order to predict the effectiveness of the algorithm. We can directly compute the distance metric for very simple cases.

2.4.1 Single Well Potential, N=2

Consider the case where

$$\rho(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2} \quad and \quad N = 2 \tag{2.20}$$

Since $N = 2$, we begin with a pair of ODE's:

$$\begin{aligned} \frac{d}{dt}X_1 &= -\frac{1}{N}\sum_{j=1}^{N}\left[K(X_1 - X_j)E'(X_j) + \frac{1}{\beta}K'(X_1 - X_j)\right] \\ \frac{d}{dt}X_2 &= -\frac{1}{N}\sum_{j=1}^{N}\left[K(X_2 - X_j)E'(X_j) + \frac{1}{\beta}K'(X_2 - X_j)\right] \end{aligned} \tag{2.21}$$

Since we are interested in the steady state, we can take advantage of the symmetry in the steady state for the $N = 2$ case.

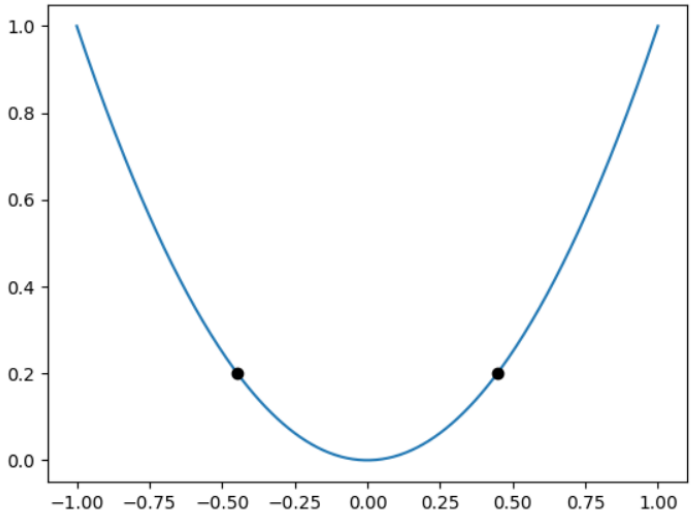

Figure 2: Steady State for N=2

Call the left particle X_{left} and the right particle X_{right}. We will assume that $X_{left} = -X_{right}$ and only solve for the steady state of X_{right}. This allows us

to reduce the system of equations to just one:

$$\frac{d}{dt}X_{right} = -\frac{1}{N}\sum_{j=1}^{N}\left[K(X_{right} - X_j)E'(X_j) + \frac{1}{\beta}K'(X_{right} - X_j)\right] = 0 \tag{2.22}$$

Substituting proper values/expressions for N, E', β, K, and K' and solving for X_{right}, we arrive at the following result for the position of the particle on the right as a function of the kernel parameter sigma:

$$X_{right} = \sigma\sqrt{\log\sqrt{\frac{1+\sigma^2}{\sigma^2}}} \tag{2.23}$$

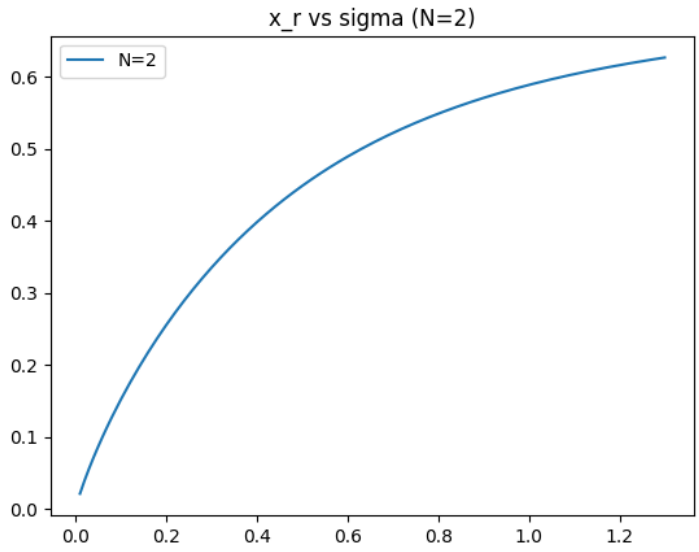

Figure 3: Position of Right Particle vs Sigma for N=2

We can also find the distance between the particle approximation and the target distribution as a function of sigma. We begin with the statement of Lemma A.1:

$$d(X,Y) = \max_{p\in P}\left(\left|F_X(p) - F_Y(p)\right|, \lim_{x\to p^-}\left|F_X(x) - F_Y(x)\right|\right) \tag{2.24}$$

Due to symmetry, the values of the absolute value expressions for $p = X_{left}$ and $p = X_{right}$ are the same. Computing for just the $p = X_{right}$ case:

$$\begin{aligned} d(X,Y) &= \max\left(\left|F_X(X_{right}) - F_Y(X_{right})\right|, \lim_{x \to X_{right}^-} \left|F_X(x) - F_Y(x)\right| \right) \\ &= \max\left(\left|F_X(X_{right}) - 1\right|, \left|F_X(X_{right}) - 0.5\right| \right) \\ d(\sigma) &= \max\left(\left|F_X(X_{right}(\sigma)) - 1\right|, \left|F_X(X_{right}(\sigma) - 0.5\right| \right) \end{aligned} \tag{2.25}$$

The equation derived above is shown graphically below in Figure 4. The two 'branches' of the graph correspond to the two absolute value expressions in the equation. The first branch persists as long as the value of $F_X(X_{right}(\sigma))$ is farther away from 1 than from 0.5, that is, less than 0.75. Then, the second branch takes over once $F_X(X_{right}(\sigma))$ is farther from 0.5 than from 1, that is, greater than 0.75. The minimum distance occurs between the two branches. The optimal sigma value that produces the minimum distance is about 0.57.

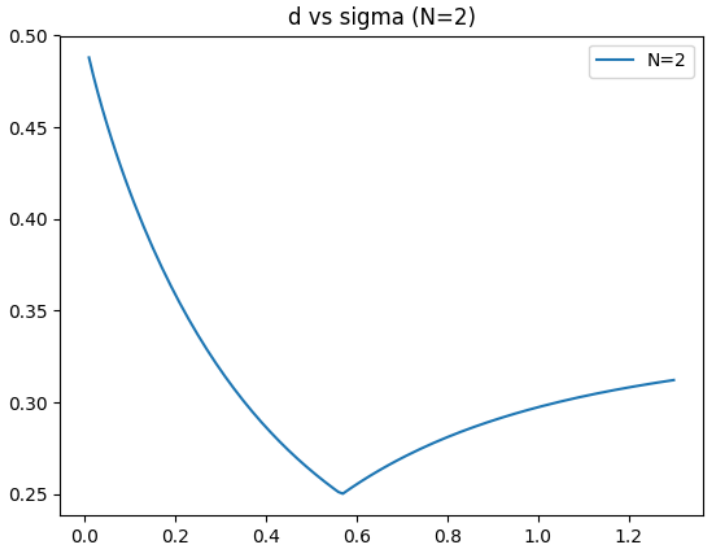

Figure 4: Distance Metric vs Sigma for N=2

2.4.2 Single Well, N=3

Consider the case where

$$\rho(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2} \quad and \quad N = 3 \tag{2.26}$$

Since $N = 3$, we begin with a trio of ODE's:

$$\begin{aligned}
\frac{d}{dt}X_1 &= -\frac{1}{N}\sum_{j=1}^{N}\left[K(X_1 - X_j)E'(X_j) + \frac{1}{\beta}K'(X_1 - X_j)\right] \\
\frac{d}{dt}X_2 &= -\frac{1}{N}\sum_{j=1}^{N}\left[K(X_2 - X_j)E'(X_j) + \frac{1}{\beta}K'(X_2 - X_j)\right] \\
\frac{d}{dt}X_3 &= -\frac{1}{N}\sum_{j=1}^{N}\left[K(X_3 - X_j)E'(X_j) + \frac{1}{\beta}K'(X_3 - X_j)\right]
\end{aligned} \tag{2.27}$$

Again, since we are interested in the steady state, we can take advantage of the symmetry in the steady state for the $N = 3$ case.

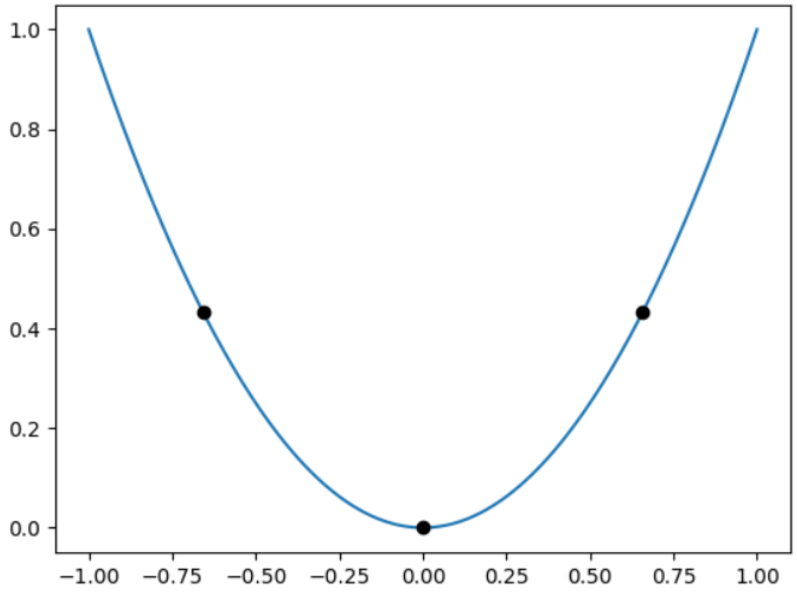

Figure 5: Steady State for N=3

Call the left particle X_{left}, the center particle X_{center}, and the right particle X_{right}. We will assume that $X_{left} = -X_{right}$, $X_{center} = 0$, and only solve for the steady state of X_{right}. This allows us to reduce the system of equations to just one:

$$\frac{d}{dt}X_{right} = -\frac{1}{N}\sum_{j=1}^{N}\left[K(X_{right} - X_j)E'(X_j) + \frac{1}{\beta}K'(X_{right} - X_j)\right] = 0 \tag{2.28}$$

After substitution, the equation is not easily solvable by analytical methods. Although a numerical solution gives the graphical results below, overlaid with the graphs for the $N = 2$ case:

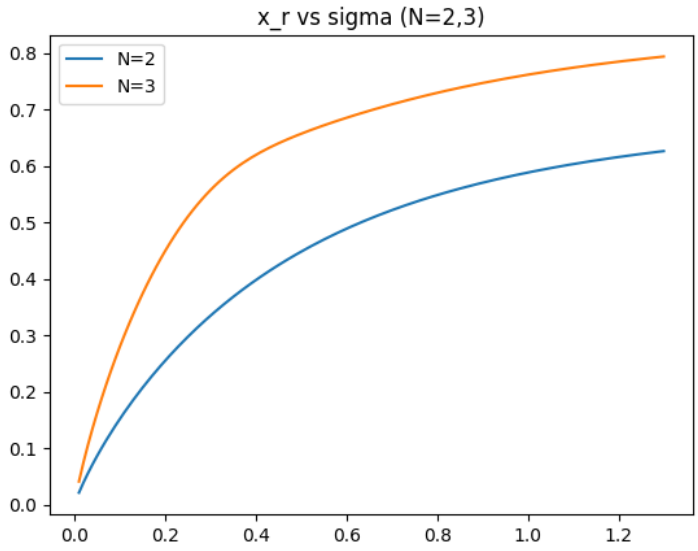

Figure 6: Position of Right Particle vs Sigma for N=3

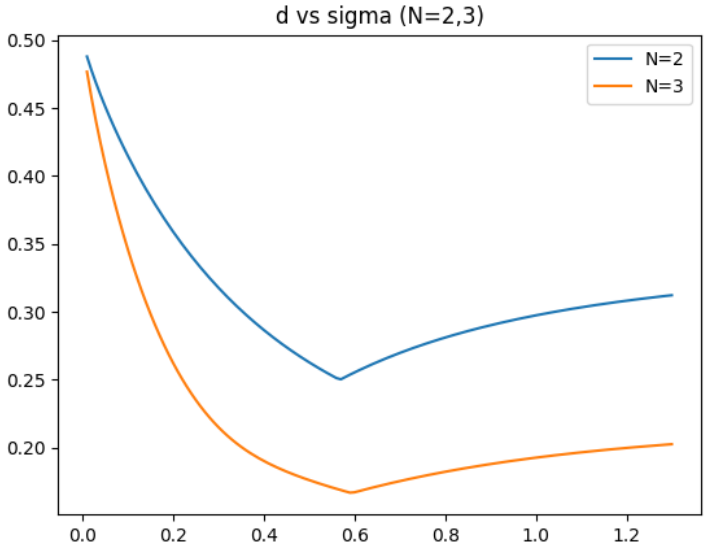

Figure 7: Distance Metric vs Sigma for N=3

2.5 Central Assumption

We have yet to prove that symmetric steady states exist and/or are attracting for all N and σ in the single well case, making the basis of the analysis not entirely valid. However, we make the unsubstantiated assumption that the SVGD algorithm has one steady state for any choice of N and σ, and proceed to explore the distance as a function of σ for larger N.

2.6 N-Scalability of Simulations

For N particles, the SVGD system has N equations. Ideally we would like to extend the analysis in 2.4 for increasing N, but symmetry can only halve the number of equations, so the system of equations become increasingly difficult to solve analytically. Thus we use numerical simulations to guide our exploration of the system's behavior.

A brute-force implementation of the SVGD algorithm is $O(N^2)$. Due to computational time considerations, we use a Barnes-Hut tree which recursively branches the entire number line into nodes representing half of the

range, so that particles which are sufficiently far away from a given particle are treated as one effective particle at the center of the range represented by their node.[6] The simulation with Barnes-Hut approximation is $O(N \log N)$.

3 Results and Analysis

The plot of steady state distance against the kernel sigma for varying N is shown in Figure 8.

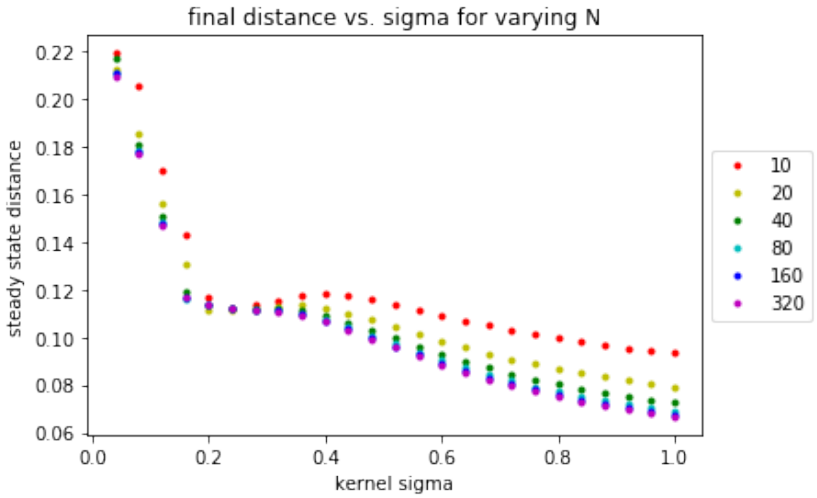

Figure 8: Steady state distance vs. kernel sigma

As expected, larger N results in a smaller final distance at all of the evaluated sigma values. A notable feature in the plot presented in Figure 8 is the local minimum found for all N values at $\sigma = 0.24$. As this dip is independent of N, we suspect that the value of the dip may also be obtained analytically from the kernalized PDE, which assumes infinite N. It is not clear, however, whether the dip persists for N $>$ 320 because, as N increases, the plot flattens. The flattening is as expected, because we know that, for the kernalized PDE (N$= \infty$), steady state distance should be 0 at all sigma values.

It is interesting that, following the dip, the distance decreases with sigma up until $\sigma = 1$. To test whether the decrease is monotonic, we will need data for larger sigma values.

The steady state width is plotted against kernel sigma for varying N in Figure 9. The width is operationally defined as as the difference between the maximum and minimum position values of particles at steady state. As mentioned, bigger width is desirable because it represents the ability of the algorithm to prevent the particles from lingering in the highest-probability regions. Since we have defined distance using the CDF metric, which accounts for the spread of particles across the real line, it makes sense that plots in Figures 8 and 9 display a strictly inverse relationship, i.e. there is no case where a smaller width resulted in a smaller distance.

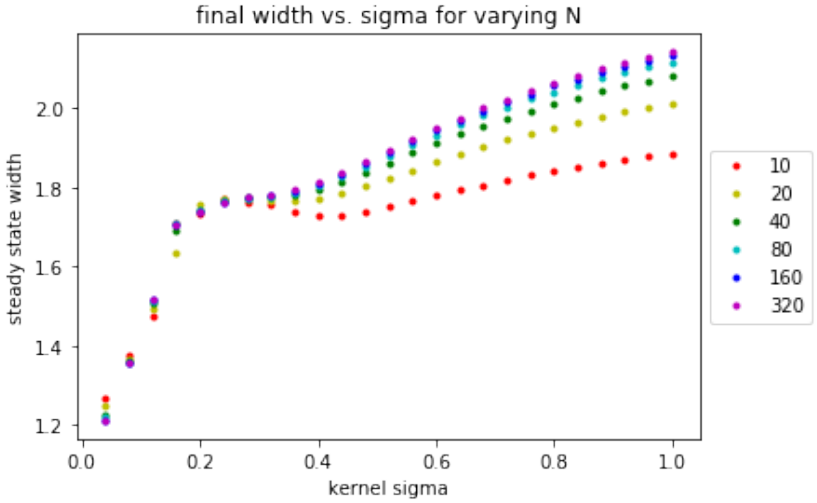

Figure 9: Steady state width vs. kernel sigma

Figure 10 presents a plot of distance over time for varying N. It is difficult to observe differences across varying N, but we see that for the first 3000 steps, the distance changes minimally for all N. The plot in Figure 11 zooms into the steps near convergence (the 20000th step onward). For convergence, we have used the operational definition that the distance between the empirical density and the target distribution does not vary by a threshold difference of $1.e-5$ for 500 consecutive time steps. Our code stops collecting data as soon as the system converges. This definition of convergence is reasonable so long as there is one attracting steady state. According to Figure 11, however, the code stopped because there was an upkick, which adds weight to the argument that the steady state is repelling. It would be instructive to run

the system for longer; perhaps there is more than one steady state.

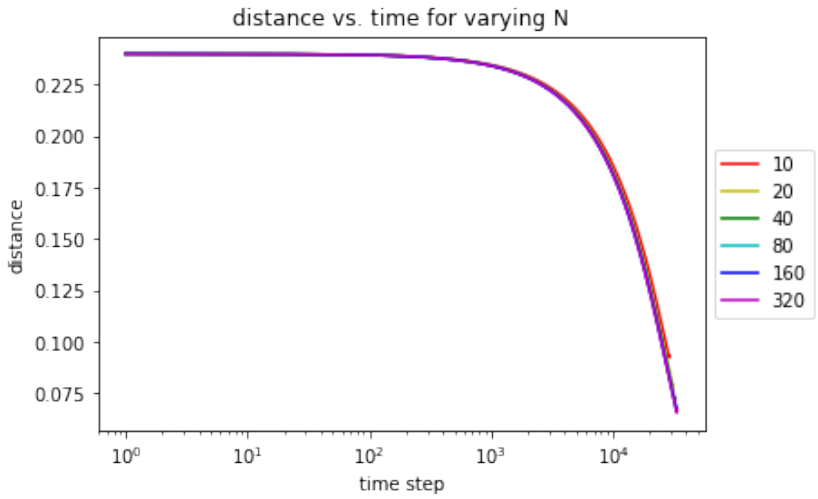

Figure 10: Distance vs. time step

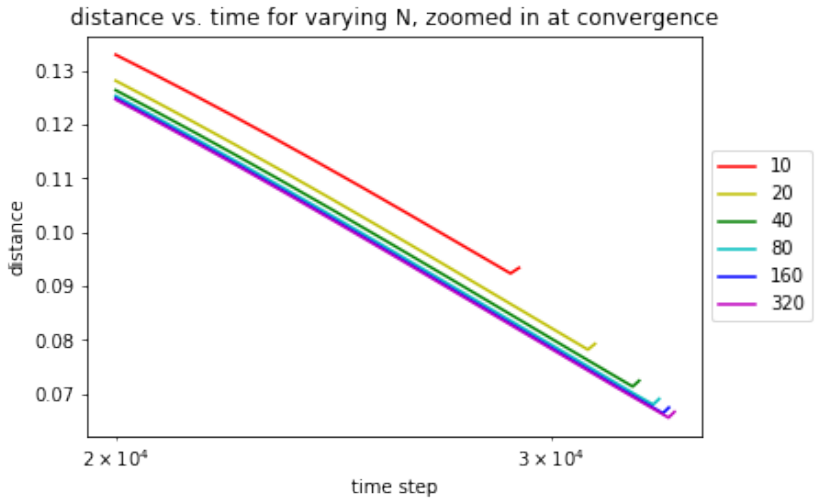

Figure 11: Distance vs. time step zoomed in at convergence

Lastly, Figure 12 overlays the distance vs. time for SVGD and Langevin, at $\sigma = 1$ (which was the optimal sigma from the range of sigmas investigated) and N= 320.

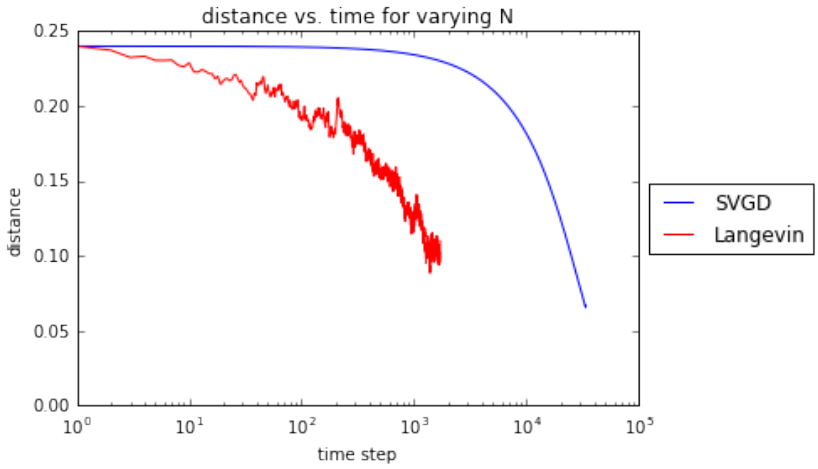

Figure 12: Distance vs. time step for Langevin and SVGD

As mentioned, the challenge of using the Langevin algorithm is that it is stochastic, so convergence is difficult to analyze. The code for Langevin dynamics stopped before that for SVGD, because the variation happened to be within $1.e-5$ for 500 consecutive steps. But the distance at which the system met our criteria for convergence is higher than with SVGD.

The Langevin algorithm, however, decreases distance at a higher rate than does SVGD. It also does not tend to remain at one distance for many initial steps.

4 Conclusion and Future Work

In our study, we have assumed that SVGD converges to a steady state, and one steady state only even as we vary parameters such as σ, β and N. It remains to be proven whether steady states exist and, if so, what the properties of the steady states are. The same questions remain for the Langevin system.

Appendix

A

Lemma A.1. *Let X be a continuous random variable with PDF ρ and CDF F_X, and let Y be a discrete random variable with generalized PDF μ and CDF F_Y. Also, let P be the set of the points $\{X_1, \ldots, X_N\}$ for which μ is non-zero. Then,*

$$\begin{aligned} d(X,Y) &= \max_{x\in\mathbb{R}} |F_X(x) - F_Y(x)| \\ &= \max_{p\in P} \left(|F_X(p) - F_Y(p)|\,,\ \lim_{x\to p^-} |F_X(x) - F_Y(x)| \right) \end{aligned}$$

Proof. Consider the intervals $(-\infty,\, X_1)$, $[X_1,\, X_2)$, $\ldots$, $[X_{N-1},\, X_N)$, $[X_N,\, \infty)$. Over each of these intervals, both F_X and F_Y are monotonically increasing, or entirely non-decreasing, because they are CDFs. More strictly, F_Y is actually constant over each of these intervals. As a result, the quantity $F_X(x) - F_Y(x)$ over these intervals is also monotonically increasing. We now explore the quantity $|F_X(x) - F_Y(x)|$.

In the case that $F_X(x) - F_Y(x)$ is strictly non-negative over an interval, $|F_X(x) - F_Y(x)| = F_X(x) - F_Y(x)$, so $|F_X(x) - F_Y(x)|$ is also monotonically increasing. This would mean that the maximum value occurs at the right endpoint of the interval, or in the limit as we approached the right endpoint from the left if the interval is open on the right.

In the case that $F_X(x) - F_Y(x)$ is strictly non-positive over an interval, $|F_X(x) - F_Y(x)| = -(F_X(x) - F_Y(x))$, so $|F_X(x) - F_Y(x)|$ is monotonically decreasing. This would mean that the maximum value occurs at the left endpoint of the interval, or in the limit as we approached the left endpoint from the right if the interval is open on the left.

In the case that $F_X(x) - F_Y(x)$ is neither non-negative nor non-positive over an interval, $F_X(x') - F_Y(x') = 0$ for some x' on the interval. $|F_X(x) - F_Y(x)|$ would be monotonically decreasing from the left endpoint to x' and monotonically increasing from x' to the right endpoint. The maximum value would occur at the left or right endpoint, or in the limit as we approached either endpoint if the interval is open on the respective side.

On the interval $(-\infty, X_1)$, $F_X(x) - F_Y(x)$ is strictly non-negative, since $F_X(x) > 0$ and $F_Y(x) = 0$ over the entire interval, so the maximum value of $|F_X(x) - F_Y(x)|$ occurs in the limit as we approach X_1 from the left.

On the interval $[X_N, \infty)$, $F_X(x) - F_Y(x)$ is strictly non-positive, since $F_X(x) < 1$ and $F_Y(x) = 1$ over the entire interval, so the maximum value of $|F_X(x) - F_Y(x)|$ occurs at X_N.

On the rest of the intervals, $F_X(x) - F_Y(x)$ is indeterminate in terms of these three cases, but all cases result in the the maximum value of $|F_X(x) - F_Y(x)|$ occurring either at X_i or in the limit as we approach X_i from the left, for some i.

All of the intervals together comprise the entire real line, and so the maximum value of $|F_X(x) - F_Y(x)|$ over $\mathbb{R}$ is the maximum value of the maximum values over each of the intervals. This value therefore occurs somewhere in the set of points that produce maximum values for each of the intervals. The following expression details the maximum value over all of this set of points:

$$\max_{p \in P} \left(|F_X(p) - F_Y(p)|, \lim_{x \to p^-} |F_X(x) - F_Y(x)| \right)$$

□

References

[1] Liu, Qiang. "Stein Variational Gradient Descent as Gradient Flow." arXiv preprint arXiv:1704.07520 (2017).

[2] Liu, Qiang, and Dilin Wang. "Stein variational gradient descent: A general purpose bayesian inference algorithm." Advances In Neural Information Processing Systems. 2016.

[3] Beck, C., and G. Roepstorff. "From dynamical systems to the Langevin equation." Physica A: Statistical Mechanics and its Applications 145.1-2 (1987): 1-14.

[4] Carlon, E. Laleman, M. and Nomidis, S. "Computational Physics: Molecular Dynamics Simulations." (2015): 19.

[5] Denisov, S. I., Werner Horsthemke, and Peter Hänggi. "Generalized Fokker-Planck equation: Derivation and exact solutions." The European Physical Journal B-Condensed Matter and Complex Systems 68.4 (2009): 567-575.

[6] Barnes, Josh, and Piet Hut. "A hierarchical O (N log N) force-calculation algorithm." nature 324.6096 (1986): 446-449.