Progress Report

Alex Loomis

December 7, 2022

Recall the Stein variational descent algorithm:

1. Choose a target density f, and a collection of points $\{x_i^0\}_{i=1}^n$.
2. Let $\hat{\phi}_\ell^*(x) = \frac{1}{n}\sum_{j=1}^n \gamma(x_j^\ell, x)\nabla_{x_j^\ell} \log f(x_j^\ell) + \nabla_{x_j^\ell}\gamma(x_j^\ell, x)$.
3. Define recursively $x_i^{\ell+1} = x_i^\ell + \varepsilon_\ell \hat{\phi}_\ell^*(x_i^\ell)$.

We can think about this as a discrete time-step approximation of an interacting particle system, where $\phi^*(x_k)$ is the momentum of x_k. Passing to continuous time, we can define an interacting particle system by

$$\frac{d}{dt}x_\gamma(t) = \phi_\gamma(t),$$

where $\phi_k(t) = \phi^*(x_k(t))$.

Let L_k be a function such that given $X = (x_1, \dots x_n)$ quantiles of a distribution with PDF f, $L_k(X) \approx (\log f)'(x_k)$.

Consider a dynamical system defined by the Hamiltonian

$$H = \frac{1}{n} \sum \frac{1}{2} p_k^2 + (\log f)'(x_k)^2 + L_k^2.$$

Hamiltonian dynamical systems obey the equations of motion

$$\frac{dx_k}{dt} = \frac{\partial H}{\partial p_k} = p_k \quad \text{and} \quad \frac{dp_k}{dt} = -\frac{\partial H}{\partial x_k}.$$

We can then relate the particle systems. If they simultaneously govern the same system, the equality

$$-\frac{\partial H}{\partial x_k} = \frac{dp_k}{dt} = \frac{d\phi_k}{dt}$$

must be satisfied.

The derivative

$$\frac{d\phi_k}{dt} = \frac{1}{n}\sum \frac{d}{dt}\big[\gamma(x_j, x_k)\big](\log f)'(x_j)$$
$$+ \gamma(x_j, x_k)(\log f)''(x_j)\phi(x_j) + \frac{d}{dt}\partial_1\gamma(x_j, x_k)$$

Taking the partial derivative of H,

$$\frac{\partial H}{\partial x_k} = \frac{1}{n} 2(\log f)'(x_k)(\log f)''(x_k) + \frac{1}{n}\sum \frac{\partial}{\partial x_k} L_j^2.$$

Thus we wish to choose a function L and kernel γ satisfying

$$\begin{aligned}
-2(\log f)'(x_k)(\log f)''(x_k) - \sum \frac{\partial}{\partial x_k} L_j^2 \\
= \sum \frac{d}{dt}\big[\gamma(x_j, x_k)\big](\log f)'(x_j) \\
+ \gamma(x_j, x_k)(\log f)''(x_j)\phi(x_j) + \frac{d}{dt}\partial_1 \gamma(x_j, x_k),
\end{aligned}$$

for every k.

This will perhaps be more manageable if we pick a distribution whose PDF has a simple log derivative, and then attempt to generalize later. If we work with the normal distribution, then $(\log f)'(x) = -x$, and so the previous equation can be rewritten as

$$2x_k + \sum \frac{\partial}{\partial x_k} L_j^2 = \sum \frac{d}{dt}\left[\gamma(x_j, x_k)\right] x_j + \gamma(x_j, x_k)\phi(x_j) - \frac{d}{dt}\partial_1\gamma(x_j, x_k).$$

I have had no success in finding L, γ satisfying this.

We could choose instead $(\log f)'(x) = -1$, and so the previous equation would be rewritten as

$$\sum \frac{\partial}{\partial x_k} L_j^2 = \sum \frac{d}{dt} \left[\gamma(x_j, x_k) \right] - \frac{d}{dt} \partial_1 \gamma(x_j, x_k).$$

Restricting $n = 2$ and choosing $\gamma(x, y) = e^{-\frac{1}{2}(y-x)^2}$, we can search for an L_k that will satisfy this.

After a lot of arithmetic, we derive that

$$L_x^2 + L_y^2 = (y-x)(y-x-1)\gamma(x,y)^2 + \frac{\sqrt{\pi}}{2}\operatorname{erf}(y-x) + g(x)$$

$$L_x^2 + L_y^2 = (y-x)(y-x+1)\gamma(x,y)^2 - \frac{\sqrt{\pi}}{2}\operatorname{erf}(y-x) + h(y),$$

for some functions g, h. Subtracting one from the other,

$$h(y) - g(x) = \sqrt{\pi}\operatorname{erf}(y-x) - 2(y-x)\gamma(x,y)^2,$$

which may not be satisfiable.

However,

$$\sqrt{\pi}\,\mathrm{erf}(y-x) - 2(y-x)\gamma(x,y)^2 = O((y-x)^3)$$

near $x = y$, so we'll approximate $h(y) = g(x) = 0$ and see if that leads to anything useful. Adding the two equations for $L_x^2 + L_y^2$,

$$L_x^2 + L_y^2 = \frac{1}{2}(y-x)^2\gamma(y,x)^2.$$

From the physical interpretation of L, we should have that $L_x = -L_y$, so

$$L_y = \frac{1}{2}(y-x)\gamma(y,x).$$

Using the physical interpretation of L again, we will extend this to more than two particles by

$$L_k = \frac{1}{n}\sum_{j=1}^{n}(x_k - x_j)e^{-\frac{1}{2}(x_k - x_j)^2}.$$

To see whether this is a meaningful choice, we can try to verify that

- the function L_k satisfies the desired approximation property, $L_k \approx (\log f)'(x_k)$,
- the same choice can be justified for other target distributions,
- the stable states of the Hamiltonian systems behave like expected, and
- the SVGD algorithm with the given kernel converges to a low energy state of the Hamiltonian system.

Repeating the derivation with the uniform distribution as the target yeilds

$$L_y = \frac{i}{\sqrt{2}}(y - x)\gamma(y, x),$$

which hopefully indicates a dropped factor of -2. This demonstrates the biggest roadblock so far with this approach: there is a lot of involved arithmetic where it is very easy to make difficult to find errors.

Qiang Liu and Dilin Wang. “Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm”. In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016. URL: `https://proceedings.neurips.cc/paper/2016/file/b3ba8f1bee1238a2f37603d90b58898d-Paper.pdf`.