Unifying Probabilistic Inference

Dinghuai Zhang

May 2020

Abstract

This note provides a perspective to unify all three probabilistic inference approaches, namely MCMC, variational inference and particle-based optimization. The main part is not my contribution but from several drafts / workshop papers.

1 Notation

x is particle of some distribution of interest. p is target distribution, q_t is current distribution (consists of many x) at time t (when doing continuous time analysis). f is some invertible transformation applied to x.

2 Analysis

2.1 SVGD

Let's analyze SVGD [1] first from a continuous time view [2]. SVGD mechanism push the samples to go along the following gradient flow:

$$\frac{dx}{dt} = \mathbb{E}_{y\sim q_t}\left[k\left(x,y\right)\nabla_y \log p(y) + \nabla_y k\left(x,y\right)\right] \tag{1}$$

where q_t is the mean-field limit empirical distribution at time t. Invoking stein identity, this becomes

$$\frac{dx}{dt} = \mathbb{E}_{y\sim q_t}\left[k\left(x,y\right)\nabla_y (\log p(y) - \log q_t(y))\right] = \mathbb{E}_{y\sim q_t}\left[k\left(x,y\right)\nabla_y \left(\log \frac{p(y)}{q_t(y)}\right)\right]. \tag{2}$$

2.2 Variatinal Inference

Next let's dive into the gradient flow of variational inference. Denote the optimization is $\max_\omega L(\omega)$, where L is the ELBO, then

$$\frac{d\omega}{dt} = \nabla_\omega L(\omega). \tag{3}$$

Notice that for sampling, reparametrization trick is commonly used, we formulate this as

$$x \sim q_\omega(x) \Leftrightarrow \epsilon \sim p_0(\epsilon), x = f_\omega(\epsilon) \tag{4}$$

thus (recall the definition of ELBO)

$$\nabla_\omega L(\omega) = \mathbb{E}_\epsilon\left[\nabla_\omega f_\omega(\epsilon)\cdot\nabla_y\left(\log\frac{p(y)}{q_\omega(y)}\right)|_{y=f_\omega(\epsilon)}\right]. \tag{5}$$

Furthermore, define $\Theta_\omega(\varepsilon,\epsilon) := \left(\nabla_\omega f_\omega(\varepsilon)\right)^T \nabla_\omega f_\omega(\epsilon)$ and $k_\omega(x,y) := \Theta_\omega\left(f_\omega^{-1}(x), f_\omega^{-1}(y)\right)$, we have

$$
\begin{aligned}
\frac{dx}{dt} &= (\nabla_\omega f_\omega(\epsilon))^T \frac{d\omega}{dt} &(6)\\
&= \mathbb{E}_{\epsilon'}\left[\Theta(\epsilon,\epsilon')\cdot\nabla_y\left(\log\frac{p(y)}{q_\omega(y)}\right)|_{y=f_\omega(\epsilon')}\right] &(7)\\
&= \mathbb{E}_{y\sim q_\omega}\left[k_\omega(x,y)\cdot\nabla_y\left(\log\frac{p(y)}{q_\omega(y)}\right)\right] &(8)
\end{aligned}
$$

It's surprising that Eq 2 and Eq 8 share the same form, indicating that these two methods implicitly follow the same continuous time regime, where SVGD is guided by a human specified kernel and VI is guided by a neural tangent kernel [3].

2.3 MCMC

Then it's natural to apply the same analysis to MCMC [4]. From previous chapter we know the Langevin dynamics follows

$$
dX_t = \nabla\log q_t(X_t)dt + \sqrt{2}dW_t. \tag{9}
$$

From the JKO theorem [5] we know that this Langevin dynamics is the steepest one in the sense of

$$
q_{t+\eta}(\cdot) = \arg\min_q\left\{\frac{1}{2}\mathcal{W}_2^2(q,q_t(\cdot)) + \eta\mathbb{E}_q\left[\log\frac{q(x)}{p(x)}\right]\right\}. \tag{10}
$$

The optimal transport problem can be understood under the Monge formulation, i.e., the optimal transportation map f at time t:

$$
\begin{aligned}
f_t =& \arg\min_f \int_x q_t(x)\left\|x - f(x)\right\|^2 dx &(11)\\
&\text{s.t. } q_t(x) = q_{t+\eta}\left(f(x)\right)\left|\frac{\partial f}{\partial x}\right| &(12)
\end{aligned}
$$

The equality constrain comes from the law of changes of variables. Still from optimal transport literature [6] one can show that the optimal transportation function in the JKO formulation satisfies

$$
f_t(x) = x + \eta\nabla_x\left(\log\frac{p(x)}{q_t(x)}\right) = x + \eta\mathbb{E}_{y\sim q_t}\left[k_\delta(x,y)\cdot\nabla_y\left(\log\frac{p(y)}{q_t(y)}\right)\right]. \tag{13}
$$

where $k_\delta(x,y) = \mathbb{I}\{x=y\}$. As a result, if we use the transformation f_t to push current distribution q_t to $q_{t+\eta}$, then

$$
\begin{aligned}
\log q_{t+\eta}(x) =& \log q_t\left(f_t^{-1}(x)\right) - \log\left|\frac{\partial f_t}{\partial x}\right| &(14)\\
=& \log q_t\left(x - \eta\nabla_x\frac{\log p(x)}{\log q_t(x)} + O\left(\eta^2\right)\right) - \log\left|I + \eta\nabla_x^2\frac{\log p(x)}{\log q_t(x)}\right| &(15)\\
=& \log q_t(x) + \eta\nabla_x\log q_t(x)^\top\nabla_x\frac{\log q_t(x)}{\log p(x)} + \eta\,\mathrm{tr}\left(\nabla_x^2\frac{\log q_t(x)}{\log p(x)}\right) + O\left(\eta^2\right). &(16)
\end{aligned}
$$

It is also surprising that when the time step $\eta\to 0$, this is exactly the Fokker Planck equation of Langevin dynamics:

$$
\frac{\partial\log q_t(x)}{\partial t} = \nabla_x\log p(x)^\top\left(\nabla_x\log\frac{q_t(x)}{p(x)}\right) + \mathrm{tr}\left(\nabla_x^2\log\frac{q_t(x)}{p(x)}\right). \tag{17}
$$

2.4 In a word, ...

Again, in Eq 13 a $\nabla_y \left(\log \frac{p(y)}{q(y)}\right)$ term emerges. Actually, this is a functional derivative of KL. Suppose we want to find a transformation f where $y = f(x)$, $x \sim q_1(x)$ and $y \sim q_2^f(y)$, such that f minimizes $\mathrm{KL}(q_2^f \| p)$ given q_1 and p. (y and x has same number of dimensionality). Notice

$$F[f] := \mathrm{KL}(q_2^f \| p) = \mathbb{E}_{y \sim q_2^f}\left[\log \frac{q_2^f(y)}{p(y)}\right] = \underbrace{\mathbb{E}_{x\sim q_1}\left[\log q_2^f(f(x))\right]}_{F_1[f]} - \underbrace{\mathbb{E}_{x\sim q_1}\left[\log p(f(x))\right]}_{F_2[f]} \tag{18}$$

and

$$q_2^f(y) \cdot \nabla f(x) = q_1(x), \tag{19}$$

then

$$\begin{aligned}
\lim_{\epsilon\to 0} \frac{F_2[f+\epsilon g] - F_2[f]}{\epsilon} &= \int q_1(x) \log \frac{q_1(f(x)+\epsilon g(x))}{q_1(f(x))} dx & (20)\\
&= \frac{1}{\epsilon}\int q_1(x) \log \frac{q_1(f(x)) + \epsilon g(x)^T \cdot \nabla q_1(f(x)) + \mathcal{O}(\epsilon)}{q_1(f(x))} dx & (21)\\
&= \frac{1}{\epsilon}\int q_1(x) \log \left(1 + \epsilon \frac{g(x)^T \cdot \nabla q_1(f(x))}{q_1(f(x))} + \mathcal{O}(\epsilon)\right) dx & (22)\\
&= \int q_1(x) \left(\frac{g(x)^T \cdot \nabla q_1(f(x))}{q_1(f(x))}\right) + \mathcal{O}(1) dx & (23)\\
&= \mathbb{E}_{q_1}\left[g(x)^T \cdot \frac{\nabla q_1(f(x))}{q_1(f(x))}\right] & (24)
\end{aligned}$$

This tells us that
$$\frac{\delta F_2[f]}{\delta f} = \frac{\nabla q_1(f(x))}{q_1(f(x))} = \nabla_y \log q_1(y)|_{y=f(x)}.$$

Also,

$$\begin{aligned}
\lim_{\epsilon\to 0} \frac{F_1[f+\epsilon g] - F_1[f]}{\epsilon} &= \int q_1(x) \log \frac{|\nabla f(x)|}{|\nabla f(x) + t\nabla g(x)|} dx & (25)\\
&= -\int q_1(x) \mathrm{tr}\left((\nabla g(x))^{-1} \nabla g(x)\right) dx = -\int \mathrm{tr}\left(q_1 (\nabla f)^{-1} \cdot \nabla g(x)\right) dx & (26)\\
&= \int g(x)^T \cdot \left(\nabla^T \cdot \left(q_1 (\nabla f)^{-1}\right)\right) dx & (27)\\
&= \int q_1(x) g(x)^T \cdot \left(\frac{1}{q_1(x)} \nabla^T \cdot \left(q_1 (\nabla f)^{-1}\right)\right) dx & (28)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\frac{\delta F_1[f]}{\delta f} &= \frac{1}{q_1(x)} \nabla^T \cdot \left(q_1 (\nabla f)^{-1}\right) & (29)\\
&= \nabla_x \log q_1(x) \cdot \left((\nabla f)^{-1}\right) + \nabla^T \cdot \left((\nabla f)^{-1}\right) & (30)\\
&\overset{(*)}{=} \nabla_x \log q_1(x) \cdot (\nabla f)^{-1} + |\nabla f| \left(\nabla \left(\frac{1}{|\nabla f|}\right)^T\right) \cdot (\nabla f)^{-1} & (31)\\
&= \frac{\nabla_x \left(q_1(x) |\nabla_x f(x)|^{-1}\right)^T \cdot (\nabla_x f(x))^{-1}}{q_1(x) |\nabla_x f(x)|^{-1}} = \frac{\nabla_y q_2^f(y)}{q_2^f(y)}|_{y=f(x)} & (32)\\
&= \nabla_y \log q_2^f(y)|_{y=f(x)}. & (33)
\end{aligned}$$

This tells us that

$$\frac{\delta F[f]}{\delta f} = \nabla_y \log q_2^f(y)|_{y=f(x)} - \nabla_y \log q_1(y)|_{y=f(x)} = \nabla_y \log \left(\log \frac{q_2^f(y)}{q_1(y)} \right), \tag{34}$$

thus demonstrating $\nabla_y \left(\log \frac{p(y)}{q(y)} \right)$ is a functional derivative term of KL divergence. All in all, we show that all three probabilistic inference dynamics follow the same functional derivative term, using different kernel smoothing method (compare Eq 2, 8 and 13)[1].

References

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

[2] Casey Chu, Kentaro Minami, and Kenji Fukumizu. The equivalence between stein variational gradient descent and black-box variational inference, 2020.

[3] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks, 2018.

[4] Matt Hoffman and Yian Ma. Langevin dynamics as nonparametric variational inference. 2019.

[5] Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker-planck equation. *SIAM J. Math. Anal*, 29:1–17, 1999.

[6] C. Villani. *Topics in Optimal Transportation*. Graduate studies in mathematics. American Mathematical Society, 2003.

[1] To be honest, I don't check the correctness of (*) as I am not very familiar with matrix calculus. I believe it's right, at least it's indeed true for 1-dimensional case.