Particle VI methods

Tony

February 2022

1 Introduction

This document contains information and and overview of papers that introduce Particle VI methods.

2 Variational Inference

Given a joint prior distribution $p(x, z)$ with latent variables z and observations x we find a variational distribution $q_\phi(z)$ that approximates the target posterior $p(z|x)$. This is Variational Inference. The most popular and common way to do that is by minimizing KL divergence:

$$D_{KL}(q_\phi(z)||p(z|x)) = \int q_\phi(z) \log \frac{q_\phi(z)}{p(z|x)} dz \tag{1}$$

We have that $D_{KL}(q_\phi(z)||p(z|x)) = 0$ if and only if $q_\phi(z) = p(z|x)$. If we continue Eq. (1) we have the following:

$$
\begin{aligned}
&D_{KL}(q_\phi(z)||p(z|x)) = \\
&= \int q_\phi(z) \log \frac{q_\phi(z)}{p(z|x)} dz \\
&= -\int q_\phi(z) \log \frac{p(z|x)}{q_\phi(z)} dz \\
&= -\int q_\phi(z) \log \frac{p(z,x)}{p(x) q_\phi(z)} dz \\
&= -\int q_\phi(z) \log \frac{p(z,x)}{q_\phi(z)} dz + \int q_\phi(z) \log p(x) dz \\
&= \underbrace{-\int q_\phi(z) \log p(z,x) dz + \int q_\phi(z) \log q_\phi(z) dz}_{-L = -ELBO = -\mathbb{E}_{q_\phi(z)}[\log p(z,x)] + H[q_\phi(z)]} + \underbrace{\int q_\phi(z) \log p(x) dz}_{\log p(x)} \\
&\Rightarrow D_{KL}(q_\phi(z)||p(z|x)) = -L + \log p(x) \\
&\Rightarrow L = \log p(x) - D_{KL}(q_\phi(z)||p(z|x)) \\
&\Rightarrow \log p(x) = L + D_{KL}(q_\phi(z)||p(z|x)) \\
&\Rightarrow \log p(x) \geq L \\
&= \mathbb{E}_{q_\phi(z)}[\log p(z,x) - q_\phi(z)]
\end{aligned}
$$

Eq. (1) is in general intractable since the target posterior distribution $p(z|x)$ is unormalized and we need to integrate over all configurations of the hidden variables in order to compute the denominator in the target posterior distribution:

$$p(z|x) = \frac{p(x|z)p(z)}{p(x)} = \frac{p(x|z)p(z)}{\int_z p(x|z)} \tag{2}$$

3 Sliced Wasserstein Variational Inference

This VI method does not utilize KL divergence, but minimizes Sliced Wasserstein distance (Bonneel et al., 2015), which reduces computational inefficiency by projecting high dimensional probability distributions into univariate slices. Sliced Wasserstein distance can be easily approximated by a few MCMC steps. This method called Sliced Wasserstein Variational Interence(SWVl) (Yi and Liu, 2022).

3.1 Wasserstein distance

Wasserstein distance arises in the context of optimal transport problem (Villani, 2009), and measures the cost of moving probability mass to transform a probability distribution to another. We define marginal distribution $p(x)$ in $\mathcal{X}$, $q(y)$ in $\mathcal{Y}$, $\prod(p,q)$ a set of any coupled joint distributions $\gamma(x,y)$ where $\int \gamma(x,y)dx =$

$q(y)$ and $\int \gamma(x,y)dy = p(x)$. These two last properties, $\int \gamma(x,y)dx = q(y)$ and $\int \gamma(x,y)dy = p(x)$, must be satisfied and they basically mean that the two qualities $p(x)$ and $q(y)$ need to be of equivalent size. We note that $\int \gamma(x,y)$ denotes a transportation plan. Then the c-Wasserstein distance is defined as :

$$\mathcal{W}_c(p,q) = \left\{ \inf_{\gamma \in \prod(p,q)} \int_{\mathcal{X}\times\mathcal{Y}} ||x-y||^c d\gamma(x,y) \right\}^{\frac{1}{c}} \tag{3}$$

where $||x-y||$ is the cost function of moving a point from $\mathcal{X}$ to $\mathcal{Y}$. The inf symbol in Eq. (3) means the greatest lower bound of the set. Also in Eq. (3) $c \geq 1$, and if $c = 1$ then $\mathcal{W}_c$ is called Earth Mover distance. The choice of c affects the Wasserstein distance for instance in the case of outliers. Let us compare the choices of $c = 1$ and $c = 2$. We define a probability density that is 95% in the range of $[0, 1]$ but there are some outliers, 5% in the range [5,6]. The goal is to move this probability density. In this case the $\mathcal{W}_2$ distance will be much higher than the $\mathcal{W}_1$. This happens because $\mathcal{W}_1$ penalizes less the outliers and therefore is more robust than $\mathcal{W}_2$. So the best transportation plan is dependent on the outliers and moving the outliers' mass effectively is important [1].

Intuitively, the c-Wasserstein distance finds eventually an optimal joint distribution $\gamma(x,y)$ that minimizes the expected cost function $\mathcal{W}_c$ in Eq. (3). Minimizing Eq. (3) is generally difficult and computationally expensive since we have to find the infimum of all sets. We can rewrite the c-Wasserstein distance in a univariate case as an analytical solution:

$$\mathcal{W}_c(p,q) = \left\{ \int_0^1 |F_p^{-1}(t) - F_q^{-1}(t)|^c dt \right\}^{\frac{1}{c}} = \left\{ \int_{\mathcal{X}} |x - F_q^{-1}(F_p(x))|^c dx \right\}^{\frac{1}{c}} \tag{4}$$

where $F(\cdot)$ is a cumulative distribution function (CDF) and $F^{-1}(\cdot)$ is a quantile function of a probability distribution (or inverse cumulative distribution function), and the $F_q^{-1}(F_p(\cdot))$ is the transportation map that moves probability density mass from $p(x)$ to $q(y)$. We can use Eq. (4) to estimate c-Wasserstein distance by sorting samples. We note that optimal transport preserves the order of probability mass elements so mass at quantile t of p moves to quantile t of q. Fig. 1 depicts an intuitive example of how optimal transport, and therefore how c-Wasserstein distance works.

3.2 Sliced Wasserstein distance

Drawing motivation from the univariate case for c-Wasserstein distance we briefly present Sliced Wasserstein distance. We start by introducing Radon transformation (Beylkin, 1984). For a density f the Radon transform represents the projection data obtained as the output of a tomographic scan. So the inverse can reconstruct the density. Let $h(\cdot)$ be a function $h(\cdot) : \mathbb{R}^d \to \mathbb{R}$. The Radon transform is the following:

[1] A more detailed explanation can be found in `https://stats.stackexchange.com/questions/490069/what-is-the-intuitive-difference-between-wasserstein-1-distance-and-wasserstein`

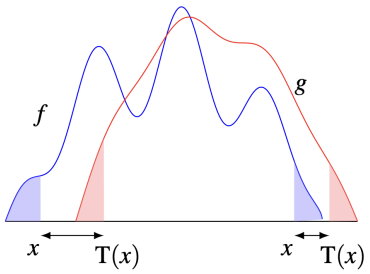

Figure 1: "Horizontal" distances where the transport T is calculated in the picture on the right as in the 1D case by imposing equality between the blue and red areas of functions f and g respectively by using CFD functions. In this specific case and according to Eq. (4) where $T(x) = F_q^{-1}(F_p(x))$. Figure taken from Santambrogio (2015).

$$h_\theta^R(l) = \int_{S:l=\langle x,\theta\rangle} h(x)dS \tag{5}$$

Radon transform defines a surface integral on a hyper-plane $S : l = \langle x, \theta\rangle$ where $l \in \mathbb{R}$ and $\theta \in \mathbb{S}^{d-1}$, where $\mathbb{S}^{d-1}$ is a unit ball embedded in $\mathbb{R}^d$. So for any pair of vectors θ and h we obtain a sliced function $h_\theta^R(\cdot)$, and the sliced function in Eq. (5) is univariate since $\mathbb{S}^{d-1}$ is a unit ball. Basically Radon transform projects a high dimensional distribution into a univariate distribution. Radon transform of a density can be defined as a series of line integrals through that density at different offsets from the origin. The value of the density at a particular line is equal to the line integral of the density over that line. This is depicted in Fig. 2.

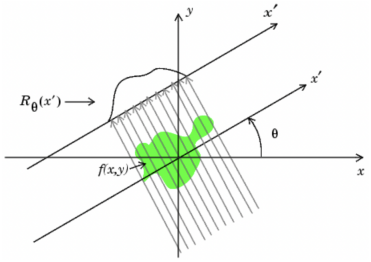

Figure 2: Radon transform of a density can be defined as a series of line integrals through that density at different offsets from the origin. Figure taken from this link.

Leveraging the fact that Eq. (5) is univariate we define the Sliced Wasserstein distance for distributions $p(x)$ and $q(y)$ as the average distance of these slices as follows:

$$\mathcal{SW}_c(p,q) = \left(\int_{\theta \in \mathbb{S}^{d-1}} \mathcal{W}_c^c(p_\theta^R, q_\theta^R) d\theta \right)^{\frac{1}{c}} \tag{6}$$

So given an empirical distribution $\hat{p} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, its Radon transformation is $\hat{p}_\theta^R = \frac{1}{n}\sum_{i=1}^n \delta_{\langle x_i, \theta \rangle}$. We calculate Sliced Wasserstein distance in Eq. (6) via estimating samples as shown in Algorithm 1. We sort the samples since we want to calculate the closest distances between the slices of the two distributions p and q.

Algorithm 1 Estimation of Sliced Wasserstein Distance with Samples

Require: $\hat{p} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$ and $\hat{q} = \frac{1}{n}\sum_{i=1}^n \delta_{y_i}$
  for $k = 0, 1, \ldots, m$ **do**
    Sample θ_k from $\mathbb{S}^{d-1}$ uniformly ▷ we sample θ from the unit ball.
    Obtain slices and sort $\{\langle x_i, \theta_k \rangle\} \to \{\langle x_j, \theta_k \rangle\}$ and $\{\langle y_i, \theta_k \rangle\} \to \{\langle y_j, \theta_k \rangle\}$
  ▷ we take slices and sort them.
  return $\mathcal{SW}_c(\hat{p}, \hat{q}) = \left(\frac{1}{mn} \sum_{k=1}^m \sum_{j=1}^n | \langle x_j, \theta_k \rangle - \langle y_j, \theta_k \rangle |^c \right)^{\frac{1}{c}}$
  end for

3.3 Proposed Method

Following from the previous subsections we finally end up with the proposal, Sliced Wasserstein Variational Inference (SWVI). Let $q_\phi(z)$ be the variational distribution parameterized by ϕ, and $p(z|x)$ the target posterior distribution. We need to find optimal parameter $\phi^\star$ that minimizes Sliced Wasserstein distance between $q_\phi(z)$ and $p(z|x)$. So we have the same problem to solve:

$$\phi^\star = \arg\min_\phi \mathcal{SW}_c(p, q_\phi) \tag{7}$$

Due to the intractability of $p(z|x)$ as we mentioned in Eq. (1), we use MCMC to estimate the distance between $q_\phi(z)$ and $p(z|x)$. Let $\mathcal{K}(\cdot)$ be a transition kernel of an MCMC with the stationary distribution $p(z|x)$, and $q_\phi(z)$ to be the initial distribution that we start sampling from with MCMC. Let $q^t(z)$ be the marginal distribution of MCMC after t transitions. Then we would have:

$$q^t(z) = \int q^{t-1}(z') \mathcal{K}(z|z') dz' \tag{8}$$

where $q^0(z) = q_\phi(z)$. For $t \to \infty$ $q^t(z)$ converges to $p(z|x)$ because of the stationary property of MCMC. So we could evaluate $\mathcal{SW}_c(p,q)$ directly via:

$$\mathcal{SW}_c(p, q_\phi) = \mathcal{SW}_c(q^t, q_\phi) \text{ as } t \to \infty \tag{9}$$

This is of course time consuming and there are also other problems to consider such as the burn-in period. Instead we evaluate a local distance $\mathcal{SW}_c(q^t, q_\phi)$

with just a few t steps of MCMC. Then we optimize use this local distance $\mathcal{SW}_c(q^t, q_\phi)$ to update parameters ϕ with Gradient Descent as follows:

$$\phi' \leftarrow \phi - a\nabla_\phi \mathcal{SW}_c(q^t, q_\phi) \tag{10}$$

$q^t(z)$ is an improvement of q_ϕ and minimizing the Sliced Wasserstein distance between those two guides the variational distribution $q_\phi(z)$ towards distribution $p(z|x)$. What we can do in this case is to use Monte Carlo methods to estimate $\mathcal{SW}_c(q^t, q_\phi)$ as described in Algorithm 1. The difference in Algorithm 1 is that we sample from $q_\phi(z)$ and $q^t(z)$ instead of $q_\phi(z)$ and $p(z|x)$. Let $\left\{z_i^0\right\}_{i=1,2,\ldots,n} \sim q_\phi(z)$ and $\{z_i^t\}_{i=1,2,\ldots,n} \sim q^t(z)$. Then we approximate Sliced Wasserstein distance with:

$$\mathcal{SW}_c(q^t, q_\phi) \approx \mathcal{L}(\left\{z_i^0\right\}, \left\{z_i^t\right\}) \tag{11}$$

Basically we rewrite Sliced Wasserstein distance as a function of two sets of samples. In order to optimize the parameters of variational distribution $q_\phi(z)$ we need to reparameterize the samples of its set $\left\{z_i^0\right\}_{i=1,2,\ldots,n}$, since the samples are not differentiable yet. For that we use an amortized sampler (a parametric probability distribution or a flexible neural network) as $z(\phi) = g_\phi(\epsilon), \epsilon \sim r(\epsilon)$, where $r(\epsilon)$ is a noise distribution and $g_\phi(\epsilon)$ is a parametric model. Using the chain rule on Eq. (11) we have:

$$\nabla_\phi \mathcal{L}(\left\{z_i^0\right\}, \left\{z_i^t\right\}) = \sum_{i=1}^{n} \nabla_{z_i} \mathcal{L}(\left\{z_i^0\right\}, \left\{z_i^t\right\}) \nabla_\phi z_i^0 \tag{12}$$

Algorithm 2 Sliced Wasserstein Variational Inference (SWVI)

Require: An unnormalized probability distribution $p(z|x)$ and learning rate a.
sampler $q_{\phi 0}(z)$
for $m = 0, 1, \ldots, s-1$ **do**
Sample $\left\{z_i^0\right\}_{i=1,2,\ldots,n}$ from $q_{\phi_m}(z)$
Run MCMC towards $p(z|x)$ with particles (samples) initialized at $\left\{z_i^0\right\}_{i=1,2,\ldots,n}$ to get $\{z_i^t\}_{i=1,2,\ldots,n}$
$\phi_{m+1} = \phi_m - a\nabla_\phi \mathcal{L}(\left\{z_i^0\right\}, \{z_i^t\}$ **return** $q_{\phi_s}(z)$
end for

References

Gregory Beylkin. 1984. The inversion problem and applications of the generalized radon transform. *Communications on pure and applied mathematics* 37(5):579–599.

Nicolas Bonneel, Julien Rabin, Gabriel Peyré, and Hanspeter Pfister. 2015. Sliced and radon wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision* 51(1):22–45.

Filippo Santambrogio. 2015. Optimal transport for applied mathematicians. *Birkäuser, NY* 55(58-63):94.

Cédric Villani. 2009. *Optimal transport: old and new*, volume 338. Springer.

Mingxuan Yi and Song Liu. 2022. Sliced wasserstein variational inference. In *Proceedings of the Advances in Approximate Bayesian Inference*.