4th Symposium on Advances in Approximate Bayesian Inference, 2021 1–8

Sliced Wasserstein Variational Inference

Mingxuan Yi MINGXUAN.YI@BRISTOL.AC.UK
Song Liu SONG.LIU@BRISTOL.AC.UK
University of Bristol, UK

Abstract

Variational Inference approximates an unnormalized distribution via the minimization of *Kullback-Leibler* (KL) divergence. Although this divergence is efficient for computation and has been widely used in applications, it suffers from some unreasonable properties. For example, it is not a proper metric, i.e., it is non-symmetric and does not preserve the triangle inequality. On the other hand, optimal transport distances recently have shown some advantages over KL divergence. To make use of these advantages, we propose a new variational inference method by minimizing sliced Wasserstein distance. This sliced Wasserstein distance can be approximated simply by running very few MCMC steps without solving any optimization problem. Our approximation also does not require a tractable density function of variational distributions so that approximating families can be amortized by generators like neural networks. Experiments on synthetic and real data are illustrated to show the performance of the proposed method.

1. Introduction

Variational inference (VI) is a method that recasts Bayesian inference as an optimization problem where it utilizes *Kullback-Leibler* (KL) divergence as a measurement to capture the discrepancy of two probability distributions. Unlike inference methods that utilizes Monte Carlo Markov Chains which needs to sample from the target probability space, VI is fast and lightweight in terms of computation. Therefore, it is preferred in many modern machine learning tasks.

Optimal Transport (OT) (Villani, 2009) has recently gained significant attention in the machine learning community. Compared to KL divergence, OT gives a valid metric that is symmetric and preserves triangular inequality. It is reported to show good performances in some downstream applications (Arjovsky et al., 2017; Gulrajani et al., 2017). While OT provides us with a new horizon to some old machine learning scenarios, the original OT problem requires a computationally demanding optimization procedure which impedes the popularity of applying the original optimal transport methods. To address this difficulty, sliced Wasserstein distance (Bonneel et al., 2015) reduces the computational inefficiency of OT by projecting high dimensional probability distributions into univariate slices where OT has a closed form solution. Sliced Wasserstein distance is successfully used in many practical tasks (Deshpande et al., 2018; Kolouri et al., 2018b,a) but it has not yet been applied to variational inference tasks.

In this paper, we extend sliced Wasserstein distance to variational inference tasks. Our methods utilize a small number of steps of MCMC to obtain an estimation of the distance.

The advantage is that by leveraging the sliced Wasserstein distance, our method does not rely on simultaneous adversarial training (Mescheder et al., 2017; Li et al., 2017; Zhang et al., 2020) to estimate the discrepancy but can still perform amortized inference (Gershman and Goodman, 2014), i.e., use a parameterized function as a sampler to capture target distributions. Similar work is contrastive divergence Hinton (2002), in which the gradient approximation is obtained with MCMC. However, our method uses MCMC to estimate the discrepancy between two distributions.

2. Background

2.1. Variational Inference

Given a joint probability distribution $p(x, z)$, where z is latent variables and x refers to observations, variational inference is finding a distribution $q_\phi(z)$ which approximates the posterior $p(z|x)$ as close as possible. Such approximations can be obtained via minimizing *Kullback-Leibler* (KL) divergence

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

Note that $D_{KL}(q_\phi(z)||p(z|x)) = 0$ if and only if $q_\phi(z) = p(z|x)$. However, Eq(1) is intractable because the posterior $p(z|x)$ is unnormalized. Instead, we can equivalently maximize *Evidence Lower Bound* (ELBO) $\mathcal{L}(\phi)$

$$\log p(x) \geq \mathcal{L}(\phi) = \mathbb{E}_{q_\phi(z)}\big[\log p(x, z) - \log q_\phi(z)\big] \tag{2}$$

Since we observe that the model evidence $\log p(x)$ is a constant w.r.t. ϕ and the above inequality becomes tight if $D_{KL}\big(q_\phi(z)||p(z|x)\big) = 0$. Optimization of ELBO requires the differentiation of the r.h.s. expectation. Gradient descent is a standard approach that allows for such optimization. To obtain a valid estimation of the gradient, a solution is to apply the score function method (Paisley et al., 2012) (Ranganath et al., 2014). An alternative solution to obtain the gradient of ELBO is the reparameterization trick (Kingma and Welling, 2014) (Rezende et al., 2014). Vanilla VI leverages KL divergence but this can be substituted by any other f-divergence and importance sampling (Jerfel et al., 2021; Wan et al., 2020) can be used to obtain gradient estimation for general f-divergences.

2.2. Wasserstein Distance

Wasserstern distance arises in optimal transport (Villani, 2009) in which a distribution is transformed to another by moving probability mass. Wasserstern distance measures the cost of such a transformation. Given two marginal distributions $p(x)$ and $q(y)$ over domains $\mathcal{X}$ and $\mathcal{Y}$, let $\Pi(p, q)$ be a set of any coupled joint distributions $\gamma(x, y)$ where $\int \gamma(x, y)dx = q(y)$ and $\int \gamma(x, y)dy = p(x)$. The $c-$Wasserstein distance is defined as

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

where $\|x - y\|$ is a cost function of moving a point from $\mathcal{X}$ to $\mathcal{Y}$. Intuitively, the $c-$Wasserstein distance aims to find an optimal joint distribution $\gamma(x, y)$ where the expected cost specified

by Eq(3) achieves its minimum. Solving this optimization problem is generally difficult but we can rewrite $p-$Wasserstein distance in a univariate case as

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

where $F(\cdot)$ is a cumulative distribution function and $F^{-1}(\cdot)$ is a quantile function of a probability distribution and the composition $F_v^{-1}F_u(\cdot)$ defines a transportation map that moves mass from $p(x)$ to $q(y)$. Given two empirical distributions, we can simply utilize Eq(4) to estimate $c-$Wasserstein distance by sorting samples.

2.3. Sliced Wasserstein Distance

Motivated by the computational efficiency of estimating Wasserstein distance with univariate distributions. We give a brief review of sliced Wasserstein distance (Bonneel et al., 2015). We first introduce *Radon* transformation (Beylkin, 1984).

Let $h(\cdot)$ be a function $h : \mathbb{R}^d \longrightarrow \mathbb{R}$. *Radon* transform is defined as

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

Eq(5) 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 $\mathbb{R}^d$. For any pair of vectors θ and h, we obtain a sliced function $h_\theta^R(\cdot)$. We note that marginalization of a high dimensional joint probability distribution can be regarded as a special case of the *Radon* transform with $\theta = e_i$, where e_i is an all-zero vector with only 1 at the i-th position. Note that the sliced function yielded by Eq(5) is univariate. Leveraging this property, we define sliced Wasserstein distance for probability distributions $p(x)$ and $q(y)$ as the averaged distance of these slices.

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

Given an empirical distribution described by $\hat{p} = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$, it is trivial to write down its *Radon* transformation defined in Eq(5) as $\hat{p}_\theta^R = \frac{1}{n}\sum_{i=1}^n \delta_{\langle x_i,\theta\rangle}$. We summarize the procedure of calculating sliced wasserstein distance via empirical samples in **Algorithm 1**.

3. Proposed Method

This method is named sliced Wasserstein variational inference (SWVI). We use the following notation: $q_\phi(z)$ as the variational distribution parameterized by ϕ, and $p(z|x)$ as the target distribution. The question we are interested in is finding an optimal parameter ϕ^* that minimizes sliced Wasserstein distance between the variational distribution and the target distribution.

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

Motivated by the intractability of the target distribution $p(z|x)$, the main idea of SWVI is to use MCMC methods to estimate the distance between the variational distribution and the target distribution.

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 $k = 0, 1 \cdots m$

1. Sample θ_k from $\mathbb{S}^{d-1}$ uniformly,
2. Obtain slices and sort $\{\langle x_i, \theta_k\rangle\} \longrightarrow \{\langle x_j, \theta_k\rangle\}$ and $\{\langle y_i, \theta_k\rangle\} \longrightarrow \{\langle y_j, \theta_k\rangle\}$

return $\mathcal{SW}_c(\hat{p}, \hat{q}) = \left(\frac{1}{mn}\sum_{k=1}^{m}\sum_{j=1}^{n}\left|\langle x_j, \theta_k\rangle - \langle y_j, \theta_k\rangle\right|^c\right)^{\frac{1}{c}}$

3.1. Estimating Distances with MCMC

Let $\mathcal{K}(\cdot|\cdot)$ be a transition kernel of a MCMC with the stationary distribution $p(z|x)$, and $q_\phi(z)$ be the initial distribution of the corresponding MCMC. We denote by $q^t(z)$ the marginal distribution of the Markov chain after applying t times transitions.

$$q^t(z) = \int q^{t-1}(z')\mathcal{K}(z|z')dz' \text{ where } q^0(z) = q_\phi(z) \tag{8}$$

Given a sufficiently long run, $q^t(z)$ converges to $p(z|x)$ because of the stationary property of Markov chain. At the current stage, one can directly evaluate sliced Wasserstein distance $\mathcal{SW}_c(p, q_\phi)$ via

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

Unfortunately, running a long enough MCMC chain is time consuming and it might be difficult to diagnose the burn-in period. To solve this difficulty, we instead evaluate a local distance $\mathcal{SW}_c(q^t, q_\phi)$ with a few number of steps t of iterating MCMC algorithms. Next we update parameters ϕ via gradient descent with the learning rate α

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

Since $q^t(z)$ is an improvement of q_ϕ, minimizing the sliced Wasserstein distance between them guides the variational distribution $q_\phi(z)$ towards the target distribution $p(z|x)$.

3.2. Stochastic Optimization

MCMC methods approximate an unnormalized probability distribution in a non-parametric manner, i.e., by particle approximations. We can thus use Monte Carlo methods to estimate $\mathcal{SW}_c(q^t, q_\phi)$, which is described by **Algorithm 1**. The approximation can be done by drawing samples respectively from $q_\phi(z)$ and $q^t(z)$. Suppose that $\{z_i^0\}_{i=1,2\cdots n} \sim q_\phi(z)$ and $\{z_i^t\}_{i=1,2\cdots n} \sim q^t(z)$. Sliced Wasserstein is then approximated by

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

Here we rewrite the approximate distance as a function $\mathcal{L}(\cdot, \cdot)$ of two sets of samples.

In order to optimize of the parameter of the variational distribution $q_\phi(z)$, we still need to reparameterize samples $\{z_i^0\}_{i=1,2\cdots n}$. This can be done by an amortized sampler that can

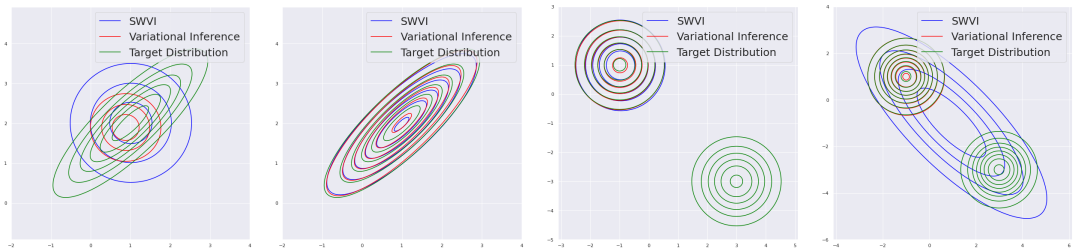

Figure 1: Approximating Diagonal 2D Gaussian Distributions with VI and SWVI

be either a parametric probability distribution or a flexible neural network. The amortized sampler is written as $z(\phi) = g_\phi(\epsilon),\ \epsilon \sim r(\epsilon)$, where $r(\epsilon)$ is a noise distribution and g_ϕ is a parametric model. We can use chain rule to obtain the gradient estimation of Eq(11),

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

This can be implemented easily via back-propagation. The overall procedure of SWVI is summarized in **Algorithm 2**

Algorithm 2: Sliced Wasserstein Variational Inference (SWVI)

Require: An unnormalized probability distribution $p(z|x)$ and learning rate α.
Initialize sampler $q_{\phi_0}(z)$.
for $m = 0, 1, 2 \cdots s-1$

1. Sample $\{z_i^0\}_{i=1,2\cdots n}$ from $q_{\phi_m}(z)$
2. Run MCMC towards $p(z|x)$ with particles initialized at $\{z_i^0\}_{i=1,2\cdots n}$ to get $\{z_i^t\}_{i=1,2\cdots n}$
3. $\phi_{m+1} = \phi_m - \alpha \nabla_\phi \mathcal{L}(\{z_i^0\}, \{z_i^t\})$

return $q_{\phi_s}(z)$

4. Experiments

For all experiments, we use sliced 1-Wasserstein distance.

4.1. 2D Experiment

In this experiment, we set target distributions as a bivariate Gaussian distribution and a bimodal Gaussian mixture. The sampler is a Gaussian distribution. We fit the sampler model to the target distribution with vanilla variational inference under KL divergence and the proposed method SWVI. In our method, we adopt the random walk Metropolis-Hastings algorithm as our MCMC instance and we run the Markov chain with 5 steps. The proposal distribution for Metropolis-Hastings algorithm is a standard Gaussian distribution. In first figure of Figure 1, we use a diagonal Gaussian distribution as the sampler model and it shows that SWVI results in an approximation with a larger variance compared to standard

VI which uses the KL divergence that is mode seeking. In the second figure, we use a regular Gaussian distribution with full trainable covariance matrix, both VI and SWVI can approximate the target distribution well. For the 3rd and 4th figures, the target distribution is set as a Gaussian mixture model, VI always fits one mode but SWVI can have different behaviours if we tune the step size of random walk in MCMC (step sizes are 1.0 and 2.5).

4.2. Neural Samplers

In this subsection, we show an experiment where we have amortized SWVI with a neural sampler, where the density function is no longer tractable. Note that the proposed method in **Algorithm 2** does not require a closed form of density function of $q_\phi(z)$. Hence, we can easily adapt a neural sampler that can generate samples from a more flexible distribution. For comparison, we also implemented amortized Stein variational gradient (SVGD) method (Liu and Wang, 2016; Feng et al., 2017). The target distribution is a mixture of two Gaussians but one with a larger variance and another with a smaller variance. We fit a neural sampler to this distribution with SWVI and amortized SVGD. This experiment shows that amortized SVGD fails to capture the other mode. The reason is that the kernel function (RBF) cannot adjust the bandwidth to the two modes with different ranges of variance. However, with the asymptotic guarantees of MCMC, the neural sampler trained with the proposed method can efficiently capture two different modes and outputs considerably better samples.

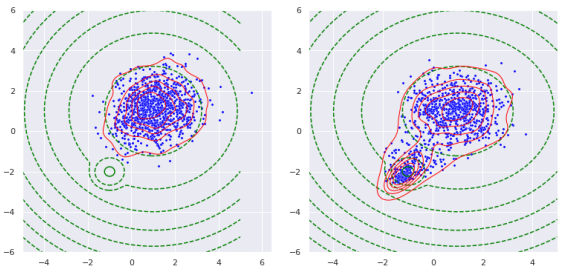

Figure 2: Fitting a mixture distribution with SVGD (L) and SWVI (R)

4.3. Bayesian Logistic Regression

We apply SWVI to binary classification tasks in the UCI repository (Asuncion and Newman, 2007) with Bayesian logistic regression models. We use a constant prior distribution for the Bayesian logistic model and 20 steps Langevin Dynamics (with learning rate 0.0001) (Neal, 2011; Welling and Teh, 2011) as the MCMC instance. The fitted distributions are both diagonal Gaussians for SWVI and VI. We present results in Table 1 and it can be seen that the performance of SWVI is on par with the vanilla VI.

Test Accuracy		
Dataset	SWVI	VI
Heart	0.852±0.019	**0.855**±0.030
Wine	0.716±0.025	**0.731**±0.012
Ionosphere	**0.771**±0.071	0.767±0.062

Table 1: Test accuracy for Bayesian logistic regression (32 posterior samples are used).

References

Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *ICML*, 2017.

Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

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

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

Ishan Deshpande, Ziyu Zhang, and Alexander G Schwing. Generative modeling using the sliced wasserstein distance. In *CVPR*, 2018.

Yihao Feng, Dilin Wang, and Qiang Liu. Learning to draw samples with amortized stein variational gradient descent. In *UAI*, 2017.

Samuel Gershman and Noah Goodman. Amortized inference in probabilistic reasoning. In *Proceedings of the annual meeting of the cognitive science society*, volume 36, 2014.

Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron Courville. Improved training of wasserstein gans. In *NeurIPS*, 2017.

Geoffrey E Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

Ghassen Jerfel, Serena Wang, Clara Fannjiang, Katherine A Heller, Yian Ma, and Michael I Jordan. Variational refinement for importance sampling using the forward kullback-leibler divergence. *arXiv preprint arXiv:2106.15980*, 2021.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

Soheil Kolouri, Phillip E Pope, Charles E Martin, and Gustavo K Rohde. Sliced wasserstein auto-encoders. In *ICLR*, 2018a.

Soheil Kolouri, Gustavo K Rohde, and Heiko Hoffmann. Sliced wasserstein distance for learning gaussian mixture models. In *CVPR*, pages 3427–3436, 2018b.

Yingzhen Li, Richard E Turner, and Qiang Liu. Approximate inference with amortised mcmc. *arXiv preprint arXiv:1702.08343*, 2017.

Qiang Liu and Dilin Wang. Stein variational gradient descent: A general purpose bayesian inference algorithm. In *NeurIPS*, 2016.

Lars Mescheder, Sebastian Nowozin, and Andreas Geiger. Adversarial variational bayes: Unifying variational autoencoders and generative adversarial networks. In *ICML*, 2017.

Radford Neal. *MCMC using Hamiltonian dynamics*. CRC press, 2011.

John Paisley, David Blei, and Michael Jordan. Variational bayesian inference with stochastic search. In *ICML*, 2012.

Rajesh Ranganath, Sean Gerrish, and David Blei. Black box variational inference. In *AIstats*, 2014.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

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

Neng Wan, Dapeng Li, and Naira Hovakimyan. f-divergence variational inference. In *NeurIPS*, 2020.

Max Welling and Yee W Teh. Bayesian learning via stochastic gradient langevin dynamics. In *ICML*, 2011.

Quan Zhang, Huangjie Zheng, and Mingyuan Zhou. Mcmc-interactive variational inference. *arXiv preprint arXiv:2010.02029*, 2020.