Transport Score Climbing: Variational Inference using forward KL and Adaptive Neural Transport

Liyi Zhang, Christian A. Naesseth and David M. Blei

Preprint: 7$^{\text{th}}$ February, 2022

Presented by: Achint Kumar

Duke University

February 21, 2022

Desiderata

Bayesian Inference- Problem Statement

Goal

Given a probabilistic model $p(\mathbf{x}, \mathbf{z})$ for latent variables $\mathbf{z}$ and data $\mathbf{x}$ compute posterior distribution, $p(\mathbf{z}|\mathbf{x})$

Since, computing posterior is intractable one approach is to use Variational Inference (VI).
In VI, we consider an approximating distribution $q_\theta(z)$ parameterized by θ. We optimize θ such that $q_\theta(z) \approx p(z|x)$.

Question

Which distance measure $KL(q_\theta||p)$ or $KL(p||q_\theta)$ should we use for optimizing θ?

$$KL(q_\theta||p) = \int q_\theta \log\left(\frac{q_\theta}{p}\right) dz, \;\; KL(p||q_\theta) = \int p \log\left(\frac{p}{q_\theta}\right) dz$$

Bayesian Inference- Problem Statement

Goal

Given a probabilistic model $p(\mathbf{x}, \mathbf{z})$ for latent variables $\mathbf{z}$ and data $\mathbf{x}$ compute posterior distribution, $p(\mathbf{z}|\mathbf{x})$

Since, computing posterior is intractable one approach is to use Variational Inference (VI).
In VI, we consider an approximating distribution $q_\theta(z)$ parameterized by θ. We optimize θ such that $q_\theta(z) \approx p(z|x)$.

Question

Which distance measure $KL(q_\theta||p)$ or $KL(p||q_\theta)$ should we use for optimizing θ?

$$KL(q_\theta||p) = \int q_\theta \log\left(\frac{q_\theta}{p}\right) dz, \;\; KL(p||q_\theta) = \int p \log\left(\frac{p}{q_\theta}\right) dz$$

KL divergence is asymmetric

We have a distribution $p(x)$ and wish to approximate it with another distribution $q_\theta(x)$. There are two ways to do it:

Forward $KL(p||q_\theta)$

- To find optimal θ we require normalization wrt p (computationally expensive)
- Mean-seeking, inclusive of full distribution
- Convex in θ, for all distributions p

Reverse $KL(q_\theta||p)$

- To find optimal θ we don't require normalization wrt p (computationally cheap)
- Mode-seeking, exclusive to single mode of distribution
- Not convex in θ, for multimodal p

KL divergence is asymmetric

We have a distribution $p(x)$ and wish to approximate it with another distribution $q_\theta(x)$. There are two ways to do it:

Forward $KL(p||q_\theta)$

- To find optimal θ we require normalization wrt p (computationally expensive)
- Mean-seeking, inclusive of full distribution
- Convex in θ, for all distributions p

Reverse $KL(q_\theta||p)$

- To find optimal θ we don't require normalization wrt p (computationally cheap)
- Mode-seeking, exclusive to single mode of distribution
- Not convex in θ, for multimodal p

KL divergence is asymmetric

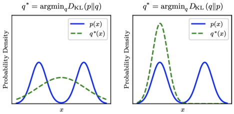
We have a distribution $p(x)$ and wish to approximate it with another distribution $q_\theta(x)$. There are two ways to do it:

Forward $KL(p||q_\theta)$

- To find optimal θ we require normalization wrt p (computationally expensive)
- Mean-seeking, inclusive of full distribution
- Convex in θ, for all distributions p

Reverse $KL(q_\theta||p)$

- To find optimal θ we don't require normalization wrt p (computationally cheap)
- Mode-seeking, exclusive to single mode of distribution
- Not convex in θ, for multimodal p



Credit: *Deep Learning by Ian Goodfellow, et. al*

KL divergence is asymmetric

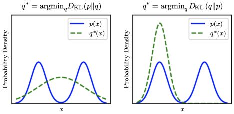
We have a distribution $p(x)$ and wish to approximate it with another distribution $q_\theta(x)$. There are two ways to do it:

Forward $KL(p||q_\theta)$

- To find optimal θ we require normalization wrt p (computationally expensive)
- Mean-seeking, inclusive of full distribution
- Convex in θ, for all distributions p

Reverse $KL(q_\theta||p)$

- To find optimal θ we don't require normalization wrt p (computationally cheap)
- Mode-seeking, exclusive to single mode of distribution
- Not convex in θ, for multimodal p



Credit: *Deep Learning by Ian Goodfellow, et. al*

Minimizing $KL(p||q_\theta)$

$$KL(p(z|x)||q_\theta(z)) := \mathbb{E}_{p(z|x)}\left[\log p(z|x) - \log q_\theta(z)\right]$$

The gradient with respect to variational parameter θ is given by,

$$g(\theta) = \nabla_\theta KL = -\mathbb{E}_{p(z|x)}\left[\nabla_\theta \log q_\theta(z)\right] = -\mathbb{E}_{p(z|x)}\left[s_\theta(z)\right]$$

$s_\theta(z)$ is called score function.
We now look at some proposed methods to minimize $g(\theta)$

Method 1: Stochastic Gradient Descent with Importance Sampling

SGD updates are given by,

$$\theta_k = \theta_{k-1} - \epsilon_k g(\theta_{k-1})$$

For gradient $g(\theta)$ is estimated by Importance Sampling. So,

$$g(\theta) = -\mathbb{E}_{p(z|x)}\left[\nabla_\theta \log q_\theta(z)\right] = -\mathbb{E}_{q_\theta(z)}\left[\frac{p(z|x)}{q_\theta(z)}\nabla_\theta \log q_\theta(z)\right]$$

If the proposal distribution $q_\theta(z)$ is not well matched with true distribution $p(z|x)$ then samples have low effective sample size which leads to samples having large variance

Method 2: Stochastic Gradient Descent with MCMC

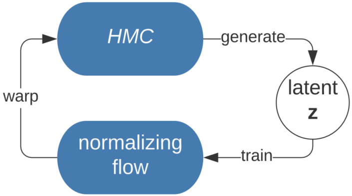
This idea is described in Markov Score Climbing paper by Naesseth, 2021. Current paper is built on this work.
The steps involved are:

- Create a Markov chain with $p(z|x)$ as the stationary distribution using MCMC algorithm. This gives the associated Markov kernel $M(z'|z;\theta)$.
- Samples are not generated independently. New sample $z_k \sim M(.|z_{k-1};\theta)$
- Compute score, $s(z_k;\theta) = \nabla_\theta \log q_\theta(z_k)$
- Update θ, $\theta_k = \theta_{k-1} + \epsilon_k s(z_k;\theta)$

Slow to converge

Transport Score Climbing: Motivation

- Transport Score Climbing(TSC) replaces MCMC of Markov Score Climbing with Hamiltonian Monte Carlo(HMC) on a transported space.
- HMC in transported space involves sampling from isotropic Gaussian(easy!) giving samples z_0
- Normalizing flow learns the map to take the samples from transported space(z_0) to the real space(z).
- The samples are used to update the parameters θ of the posterior $q_\theta(z)$



Hamiltonian Monte Carlo

- Sampling technique which combines Hamiltonian dynamics and MCMC.
- Faster than MCMC and works in high dimensions

$$F = Ma \iff \begin{aligned} \frac{dz}{dt} &= \frac{\partial \mathcal{H}}{\partial m} \\ \frac{dm}{dt} &= -\frac{\partial \mathcal{H}}{\partial q} \end{aligned} \tag{1}$$

Here, $\mathcal{H}(z, m) = \frac{m^2}{2M} + U(z)$. But, $U(z) = -\log[p(z|x)]$. The algorithm has the following steps:

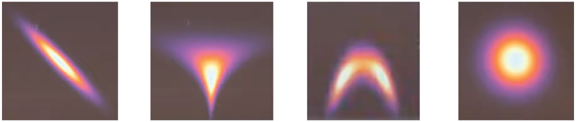
- Initialize z_0 and $m_0 \sim \mathcal{N}(0, M)$.
- Evolve (z_0, m_0) according to Hamiltonian dynamics to (z, m).
- Accept the new state z with probability given by $\min\{1, \exp(\delta\mathcal{H})\}$ where $\delta\mathcal{H}(z, m) = \mathcal{H}(z, m) - \mathcal{H}(z_0, m_0)$

Transport Map: HMC on Warped Space

- HMC takes $\sigma_{max}/\sigma_{min}$ iterations to get acceptable samples
- HMC is slow if the target distribution has mix of low curvature and high curvature directions

Non-isotopic Gaussian, Neal's funnel distribution and Banana distribution are difficult to efficiently sample from while Isotopic Gaussian is easy. How about we transport the difficult distributions to Isotopic Gaussians?

Non-isotopic Gaussian | Neal's funnel distribution | Banana distribution | Isotopic Gaussian

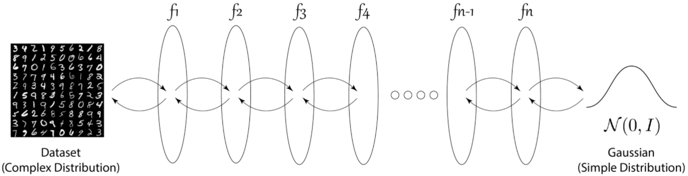


Normalizing Flows

- It is a generative model constructed out of sequence of invertible transformations $f(z_i)$ based on the change of variable formula,

$$p(z_i) = p(z_{i-1}) \left| \det \frac{df(z_{i-1})}{dz_{i-1}} \right|$$

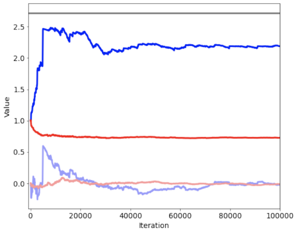
- Unlike VAE, Normalizing Flows learn the exact data distribution $p(x)$

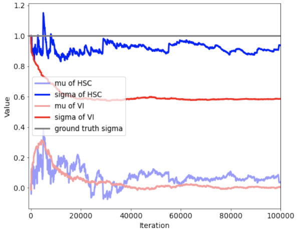


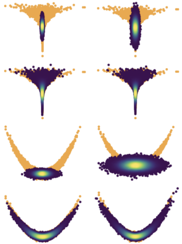
Credit: *Lillian Weng*

Synthetic Data Experiment

- TSC was trained to learn the funnel and banana distributions. Both distributions are known to be difficult to sample from HMC.
- Left: VI underestimates uncertainty in sigma
- Right: Row 1,3 is Gaussian fit. Row 2,4 is VI fit(left) and TSC fit(right)







Theorem

Theorem

The parameter θ of variational distribution converges to a local optima of the forward KL.

Let $\theta(t)$ satisfy the following differential equation,

$$\frac{d\theta(t)}{dt} = -\mathbb{E}_{p(z|x)}\log[s_\theta(z)],\ \theta(0) = \theta_0$$

We need to show that $\theta(t)$ has a basin of attraction and converges to the fixed point in it.
The proof of MSC and TSC is taken from Gu and Kong, 1998 with close to no modification