Notes on Flexible and Efficient Inference with Particles for the Variational Gaussian Approximation

Yuanxing Cheng

April 2, 2022

1 before starting

Transform between variable flow and particle flow for Gaussian variational inference.

2 Abstruct

Variational Inference. A flexible and efficient algorithm based on a linear flow leading to a particle based approximation. With sufficient number of particles, algorithm converges linearly to the exact solution for Gaussian targets. On a set of synthetic and high-dimension problems, algorithm outperforms.

3 Introduction

Introducing Gaussian particle flow (GPF), that approximate a Gaussian variational distribution with particles. A stochastic version, Gaussian Flow (GF). Prove the decreasing of empirical version of free energy. Comparison with other VGA algorithm.

4 Related work

Bayesian Inference is to find posterior distribution of latent variable $\mathbf{x} \in \mathbb{R}^D$ given observations y. To use Bayes theorem that $p(\mathbf{x} \mid y) = \frac{p(y \mid x)p(x)}{p(y)}$ we need to compute $p(y)$ which is hard. Variational inference (VI) turns this into an optimization problem. The measure of closeness of densities is Kullback-Leibler (KL) divergence

$$\mathrm{KL}\left[q(x) \parallel p(x)\right] = \mathbb{E}_q\left[\log q(x) - \log p(x)\right] = \int q(x) \log \frac{q(x)}{p(x)}\, \mathrm{d}x$$

Denote by $\mathcal{Q}$ a family of distributions, we look for

$$\operatorname*{arg\,min}_{q \in \mathcal{Q}} \mathrm{KL}\left[q(x) \parallel p(x \mid y)\right] = \operatorname*{arg\,min}_{q \in \mathcal{Q}} \int q(x) \log \frac{q(x)p(y)}{p(y \mid x)p(x)}\, \mathrm{d}x$$

Equivalently, we minimize the upper bound, the variational free energy $\mathcal{F}$

$$\mathrm{KL}\left[q(x) \parallel p(x \mid y)\right] \leq \mathcal{F}[q] = \int q(x) \log q(x) - q(x) \log\left(p(y \mid x)p(x)\right)\, \mathrm{d}x = -\mathbb{E}_q\left[\log\left(p(y \mid x)p(x)\right)\right] - \mathbb{H}_q \quad (4.1)$$

where $\mathbb{H}_q = -\mathbb{E}_q\left[\log q(x)\right]$ is the entropy of q.

Following are developed approaches in the literature.

4.1 The Variational Gaussian Approximation

We restrict the distribution family $\mathcal{Q}$ to be multivariate Gaussian distribution: $q(x) = \mathcal{N}(m, C)$ where $m \in \mathbb{R}^D$ is the mean and $C \in \left\{A \in \mathbb{R}^{D\times D} \,|\, x^\top A x \geq 0, \forall x \in \mathbb{R}^D\right\}$ is the covariance matrix. As in the definition, C is positive semi-definite. Then we use the result that the entropy of multivariate normal is $\frac{1}{2}\log\left(\det\left(2\pi e C\right)\right)$, so we can rewrite the energy as follows, ignoring the constant.

$$\mathcal{F}[q] = -\mathbb{E}_q\left[\log\left(p(y\,|\,x)p(x)\right)\right] - \mathbb{H}_q = -\frac{1}{2}\log|C| + \mathbb{E}_q\left[\phi(x)\right] \tag{4.2}$$

where $\phi(x) = -\log(p(y\,|\,x)p(x))$.

Issues with this method: hard to compute gradient wrt C, non-sparse matrix from gradient of entropy, and positive-definiteness of covariance leads to non-trivial constraints on parameter updates and thus the instabilities in the algorithm.

To solve above issues, first focus on factorizable models. For problems with likelihoods that can be rewritten as $p(y\,|\,x) = \prod_{d=1}^{D} p(y\,|\,x_d)$, the number of independent variational parameters is reduced to $2D$. Then the Gaussian expectation of free energy split into a sum of 1-d integrals.

To extend to the general case, gradients of the gree energy are estimated by a stochastic sampling approach. And this relies on the *reparametrization trick*, where the expectation over the parameter dependent variational density q_θ is replaced by an expectation over a fixed density q^0, and thus $\nabla_\theta q_\theta$ is avoided.

For the Gaussian case, this reparametrization is a linear transformation of an arbitrary D dimensional Gaussian random varaible $x \sim q_\theta(x)$ in terms of a D dimensional Gaussian rv $x^0 \sim q^0 = \mathcal{N}(m^0, C^0)$

$$x = \Gamma(x^0 - m^0) + m \tag{4.3}$$

where $\Gamma \in \mathbb{R}^{D\times D}$ and $m \in \mathbb{R}^d$ are the variational parameters. Assuming C^0 non-degenerate and for simplicity, we set it as identity matrix I. Then we can write the gradient of the expectation given q over a function f given mean m: $\nabla_m \mathbb{E}_q\left[f(x)\right] = \mathbb{E}_{q^0}\left[\nabla_m f\left(\Gamma\left(x^0 - m^0\right) + m\right)\right]$

5 Gaussian (Particle) Flow

Below denote $\frac{\mathrm{d}(\cdot)}{\mathrm{d}t}$ indicates the total derivative given time, and $\frac{\partial(\cdot)}{\partial t}$

6 Gaussian Variable Flows

Based on idea of variable flows, define $x^{n+1} = x^n + \epsilon f^n(x^n)$ where $f^n : \mathbb{R}^D \to \mathbb{R}^D$. Using reparametrization trick, we choose a linear map f and write

$$\frac{\mathrm{d}x^t}{\mathrm{d}t} = f^t(x^t) = A^t(x^t - m^t) + b^t \tag{6.1}$$

where A^t is a matrix and $m^t := \mathbb{E}_{q^t}\left[x\right]$. And when initial x^0 is Gaussian, x^t are also Gaussian for any t. Then we construct a flow that decreases the free energy.

$$\frac{\mathrm{d}\mathcal{F}[q^t]}{\mathrm{d}t} = \frac{\mathrm{d}}{\mathrm{d}t}\int q^t(\log q^t(x) + \phi(x))\,\mathrm{d}x \tag{6.2}$$

$$= \int \frac{\partial q^t(x)}{\partial t}(\log q^t(x) + \phi(x))\,\mathrm{d}x + \int q^t(x)\left(\frac{\partial q^t(x)}{\partial t}\frac{1}{q^t(x)} + \frac{\partial \phi(x)}{\partial t}\right)\mathrm{d}x \tag{6.3}$$

$$= \int \frac{\partial q^t(x)}{\partial t}(\log q^t(x) + \phi(x))\,\mathrm{d}x \tag{6.4}$$

Then use continuity equation for the density

$$\frac{\partial q^t(x)}{\partial t} = -\nabla_x \cdot (q^t(x) f^t(x))$$

$$
\begin{aligned}
\frac{\mathrm{d}\mathcal{F}[q^t]}{\mathrm{d}t} &= \int -\nabla_x \cdot (q^t(x) f^t(x)) \big(\log q^t(x) + \phi(x)\big) \, \mathrm{d}x \\
&= \int (q^t(x) f^t(x)) \cdot \nabla_x \big(\log q^t(x) + \phi(x)\big) \, \mathrm{d}x \\
&= \int - \big(\nabla_x \cdot (q^t(x) f^t(x)) + q^t(x) f^t(x) \cdot \nabla_x \phi(x)\big) \, \mathrm{d}x \\
&= \int - \big(\nabla_x q^t(x) \cdot f^t(x) + q^t(x) f^t(x) \cdot \nabla_x \phi(x)\big) \, \mathrm{d}x \\
&= -\mathbb{E}_{q^t} \big[\nabla_x \cdot f^t(x)\big]
\end{aligned}
$$