[Paper review 23]

Unbiased Implicit Variational Inference

(Michalis K. Titsias, Francisco J. R. Ruiz, 2019)

[Contents]

1. Key Idea

Gradient of ELBO : $\nabla_\theta \mathcal{L}(\theta)=\mathbb{E}_{q(\varepsilon)}\left[\nabla_\theta\left(\log p\left(x, f_\theta(\varepsilon)\right)-\log q_\theta\left(f_\theta(\varepsilon)\right)\right)\right]$

- (1) model term : (with MC approximation)
  $\mathbb{E}_{q(\varepsilon)}\left[\nabla_\theta \log p\left(x, f_\theta(\varepsilon)\right)\right] \approx \frac{1}{S} \sum_{s=1}^S \nabla_\theta \log p\left(x, f_\theta\left(\varepsilon^{(s)}\right)\right), \quad \varepsilon^{(s)} \sim q(\varepsilon)$
- (2) entropy term :
  $\nabla_\theta \log q_\theta\left(f_\theta(\varepsilon)\right)=\nabla_z \log q_\theta(z) \times \nabla_\theta f_\theta(\varepsilon)+\underbrace{\left.\nabla_\theta \log q_\theta(z)\right|_{z=f_\theta(\varepsilon)}}_{=0(\text { in expectation })}=\nabla_z \log q_\theta(z) \times \nabla_\theta f_\theta(\varepsilon)$

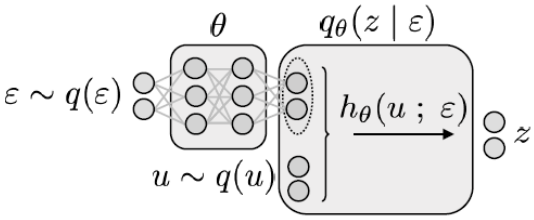
  but $\nabla_z \log q_\theta(z)$ is not available !

use "UNBIASED MC estimator" of $\nabla_z \log q_\theta(z)$

- using density ratio (X)
- using lower bound of ELBO (X)
- directly optimize ELBO (O)

Key idea: as a form of... $\nabla_z \log q_\theta(z)=\mathbb{E}_{\text{distrib }(\cdot)}[\text{ function }(z, \cdot)]$

2. UIVI



$$
\begin{aligned}
\nabla_z \log q_\theta(z) &= \mathbb{E}_{q_\theta(\varepsilon' \mid z)}\left[\nabla_z \log q_\theta\left(z \mid \varepsilon'\right)\right] \\
&\approx \nabla_z \log q_\theta\left(z \mid \varepsilon'\right), \quad \varepsilon' \sim q_\theta\left(\varepsilon' \mid z\right)
\end{aligned}
$$

(use MC estimation)

gradient of ELBO

$$
\begin{aligned}
\nabla_\theta \mathcal{L}(\theta) &= \mathbb{E}_{q(\varepsilon)}\left[\nabla_\theta\left(\log p\left(x, f_\theta(\varepsilon)\right) - \log q_\theta\left(f_\theta(\varepsilon)\right)\right)\right] \\
&= \mathbb{E}_{q(\varepsilon)}\left[\nabla_\theta \log p\left(x, f_\theta(\varepsilon)\right)\right] - \mathbb{E}_{q(\varepsilon)}\left[\nabla_\theta \log q_\theta\left(f_\theta(\varepsilon)\right)\right] \\
&= \mathbb{E}_{q(\varepsilon)}\left[\nabla_\theta \log p\left(x, f_\theta(\varepsilon)\right)\right] - \mathbb{E}_{q(\varepsilon)}\left[\nabla_z \log q_\theta(z) \times \nabla_\theta f_\theta(\varepsilon)\right] \\
&\approx \mathbb{E}_{q(\varepsilon)}\left[\nabla_\theta \log p\left(x, f_\theta(\varepsilon)\right)\right] - \mathbb{E}_{q(\varepsilon)}\left[\nabla_z \log q_\theta\left(z \mid \varepsilon'\right) \times \nabla_\theta f_\theta(\varepsilon)\right] \quad \varepsilon' \sim q_\theta\left(\varepsilon' \mid z\right)
\end{aligned}
$$

$$
\nabla_\theta \mathcal{L}(\theta) = \mathbb{E}_{q(\varepsilon) q(u)}\left[\nabla_z\left(\log p(x, z) - \log q_\theta(z)\right)\Big|_{z=h_\theta(u;\varepsilon)} \times \nabla_\theta h_\theta(u;\varepsilon)\right]
$$

2.1 Full Algorithm

Estimate the gradient based on samples :

- 1) sample$\epsilon \sim q(\epsilon)$, $u \sim q(u)$ (standard Gaussian)
- 2) set $z = h_\theta(\varepsilon; u) = \mu_\theta(\varepsilon) + \Sigma_\theta(\varepsilon)^{1/2} u$
- 3) evaluate $\nabla_z \log p(x, z)$ and $\nabla_\theta h_\theta(u; \varepsilon)$
- 4) sample $\varepsilon' \sim q_\theta\left(\varepsilon' \mid z\right)$
- 5) approximate $\nabla_z \log q_\theta(z) \approx \nabla_z \log q_\theta\left(z \mid \varepsilon'\right)$

(How to do step 4 & step 5?)

2.2 Reverse Conditional

in "step 4) sample $\varepsilon' \sim q_\theta\left(\varepsilon' \mid z\right)$"...

- conditional : $q_\theta(z \mid \epsilon)$
- reverse conditional : $q_\theta\left(z \mid \varepsilon'\right)$

sample from reverse conditional using HMC

- $q\left(\varepsilon' \mid z\right) \propto q\left(\varepsilon'\right) q_\theta\left(z \mid \varepsilon'\right)$ (unnormalized density)
- but HMC is slow

  Thus, start with a GOOD STARTING(INITIAL) POINT , which is ϵ

[proof]

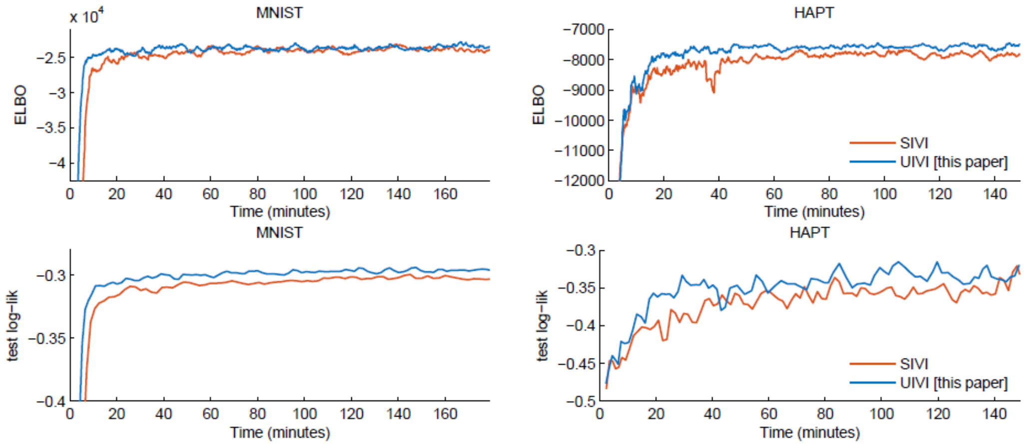
$(\varepsilon, z) \sim q_\theta(\varepsilon, z) = q(\varepsilon) q_\theta(z \mid \varepsilon) = q_\theta(z) q_\theta(\varepsilon \mid z)$

Thus, ε is a sample from $q_\theta(\varepsilon \mid z)$

To accelerate sampling $\varepsilon' \sim q\left(\varepsilon' \mid z\right)$, initialize HMC at ε

(after few iterations, the correlation between ϵ and $\epsilon^{'}$ will decrease!)

3. SIVI vs UIVI



4. VAE Experiments with UIVI

Model ：

- $p_\phi(x, z) = \prod_n p(z_n) p_\phi(x_n \mid z_n)$

Amortized variational distribution :

- $q_\theta(z_n \mid x_n) = \int q(\varepsilon_n) q_\theta(z_n \mid \varepsilon_n, x_n) d\varepsilon_n$

Goal:

- Find model parameters ϕ and variational parameters θ

method	average test log-likelihood MNIST	Fashion-MNIST
Explicit (standard VAE)	−98.29	−126.73
SIVI	−97.77	−121.53
UIVI	**−94.09**	**−110.72**

UIVI provides better predictive performance