[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:

$$abla_{ heta} \log q_{ heta}\left(f_{ heta}(arepsilon)
ight) =
abla_{z} \log q_{ heta}(z) imes
abla_{ heta} f_{ heta}(arepsilon) + \underbrace{
abla_{ heta} \log q_{ heta}(z)|_{z=f_{ heta}(arepsilon)}}_{=0 ext{(in expectation)}} =
abla_{z} \log q_{ heta}(z) imes
abla_{ heta} f_{ heta}(arepsilon)$$

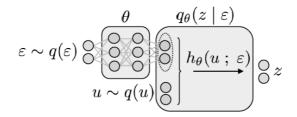
but $\nabla_z \log q_{\theta}(z)$ is not available!

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



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

gradient of ELBO

$$egin{aligned}
abla_{ heta} \mathcal{L}(heta) &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} (\log p \left(x, f_{ heta}(arepsilon)
ight) - \log q_{ heta} \left(f_{ heta}(arepsilon)
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \log p \left(x, f_{ heta}(arepsilon)
ight) - \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \log q_{ heta} \left(z
ight)
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \log p \left(x, f_{ heta}(arepsilon)
ight)
ight] - \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \log q_{ heta} \left(z
ight)
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \log p \left(x, f_{ heta}(arepsilon)
ight)
ight] - \log q_{ heta} \left(z
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \left(\log p \left(x, z
ight) - \log q_{ heta} \left(z
ight)
ight)
ight|_{z = h_{ heta}(u; arepsilon)}
ight.
abla_{ heta} \left(u; arepsilon
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \left(\log p \left(x, z
ight) - \log q_{ heta} \left(z
ight)
ight)
ight|_{z = h_{ heta}(u; arepsilon)}
ight.
abla_{ heta} \left(u; arepsilon
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \left(\log p \left(x, z
ight) - \log q_{ heta} \left(z
ight)
ight)
ight|_{z = h_{ heta}(u; arepsilon)}
ight.
abla_{ heta} \left(u; arepsilon
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \left(\log p \left(x, z
ight) - \log q_{ heta} \left(z
ight)
ight)
ight|_{z = h_{ heta}(u; arepsilon)}
ight.
abla_{ heta} \left(u; arepsilon
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \left(u; arepsilon
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \left(u; arepsilon
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \left(u; arepsilon
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \left(u; arepsilon
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \left(u; arepsilon
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \left(u; arepsilon
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \left(u; arepsilon
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \left(u; arepsilon
ight)
ight] \ &= \mathbb{E}_{q(arepsilon)} \left[
abla_{ heta} \left(u; arepsilon
ight)
ight] \$$

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 $\varepsilon' \sim q_{ heta}\left(\varepsilon' \mid z \right)$
- 5) approximate $\nabla_z \log q_{\theta}(z) pprox \nabla_z \log q_{\theta}\left(z \mid arepsilon'
 ight)$

(How to do step 4 & step 5?)

2.2 Reverse Conditional

in "step 4) sample $arepsilon' \sim q_{ heta} \, (arepsilon' \mid z)$ "...

- conditional : $q_{\theta}(z \mid \epsilon)$
- reverse conditional : $q_{\theta} (z \mid \varepsilon')$

sample from reverse conditional using HMC

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

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

[proof]

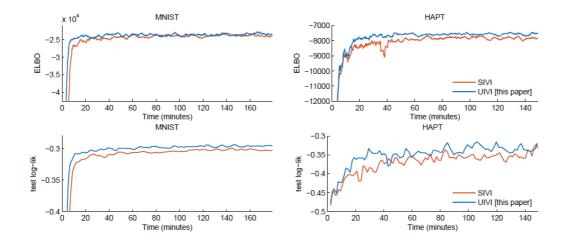
$$(arepsilon,z) \sim q_{ heta}(arepsilon,z) = q(arepsilon)q_{ heta}(z\midarepsilon) = q_{ heta}(z)q_{ heta}(arepsilon\mid z)$$

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

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

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

3. SIVI vs UIVI



4. VAE Experiments with UIVI

Model:

$$ullet \ p_{\phi}(x,z) = \prod_n p\left(z_n
ight) p_{\phi}\left(x_n \mid z_n
ight)$$

Amortized variational distribution:

•
$$q_{ heta}\left(z_{n}\mid x_{n}
ight)=\int q\left(arepsilon_{n}
ight)q_{ heta}\left(z_{n}\midarepsilon_{n},x_{n}
ight)darepsilon_{n}$$

Goal:

• Find model parameters ϕ and variational parameters θ

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

UIVI provides better predictive performance