ANIRUDH SETH
aniset@kth.se

Tutorial 10
Combining Variational Inference and Sequential MC methods using python

2020-03-03

# Exercise 1

$$\begin{aligned}\log p(x) &= \log \int_z p_\theta(x,z)dz \\ &= \log \int_z p_\theta(x,z)\frac{q_\phi(z|x)}{q_\phi(z|x)}dz \\ &= \log\left(\mathrm{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right]\right) \\ &\geq \mathrm{E}_{q_\phi(z|x)}\left(\log\frac{p_\theta(x,z)}{q_\phi(z|x)}\right) \text{ due to concavity of log and Jensen's inequality[2]}\end{aligned} \tag{1}$$

We can also derive the result from the fact that KL divergence is non negative[3]. The KL divergence between the true posterior and the proposal can be written as

$$\begin{aligned}KL\left[q_\phi(z|x)\|p_\theta(z|x)\right] &= \int_Z q_\phi(z|x)\log\frac{q_\phi(z|x)}{p_\theta(z|x)} \\ &= -\int_Z q_\phi(z|x)\log\frac{p_\theta(z|x)}{q_\phi(z|x)} \\ &= -\left(\int_Z q_\phi(z|x)\log\frac{p_\theta(x,z)}{q_\phi(z|x)} - \int_Z q_\phi(z|x)\log p(x)\right)\text{, by applying bayes rule} \\ &= -\int_Z q_\phi(z|x)\log\frac{p_\theta(x,z)}{q_\phi(z|x)} + \log p(x)\int_Z q_\phi(z|x) \\ &= -\mathrm{E}_{q_\phi(z|x)}\left(\log\frac{p_\theta(x,z)}{q_\phi(z|x)}\right) + \log p(x) \\ \mathrm{E}_{q_\phi(z|x)}\left(\log\frac{p_\theta(x,z)}{q_\phi(z|x)}\right) &= \log p(x) - KL\left[q_\phi(z|x)\|p_\theta(z|x)\right]\end{aligned} \tag{2}$$

KL divergence can be shown to be $\geq 0$ in [3]. Hence

$$\mathcal{L}(\boldsymbol{x},\boldsymbol{\theta},\boldsymbol{\phi}) = \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\left[\log\frac{p_{\boldsymbol{\theta}}(\boldsymbol{x},\boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}\right] \leq \log p(\boldsymbol{x}) \tag{3}$$

# Exercise 2

Since $\hat{p}_N(x)$ is an unbiased positive estimator of $p(x)$ , meaning $\mathbb{E}\left[\hat{p}_N(x)\right] = p(x),$ . As mentioned in [1] , from Jensen's inequality the log of any unbiased positive Monte Carlo estimator of the marginal likelihood results in a lower bound that can be optimized for MLE. This is shown below

Let the Monte Carlo objective $\mathcal{L}_{\mathrm{N}}(x,p)$ over $p \in \mathcal{P}$ defined by $\hat{p}_N(x)$ is

$$\mathcal{L}_{\mathrm{N}}(x,p) = \mathbb{E}\left[\log\hat{p}_N(x)\right]$$

By the concavity of log and Jensen's inequality,

$$\mathcal{L}_N(x,p) = \mathbb{E}\left[\log\hat{p}_N(x)\right] \leq \log\mathbb{E}\left[\hat{p}_N(x)\right] = \log p(x)$$

# PRACTICAL

The results of the training for ELBO, IWAE and FIVO bounds can be seen in Figure 1 for 2k steps and Figure 2 for 5k steps.

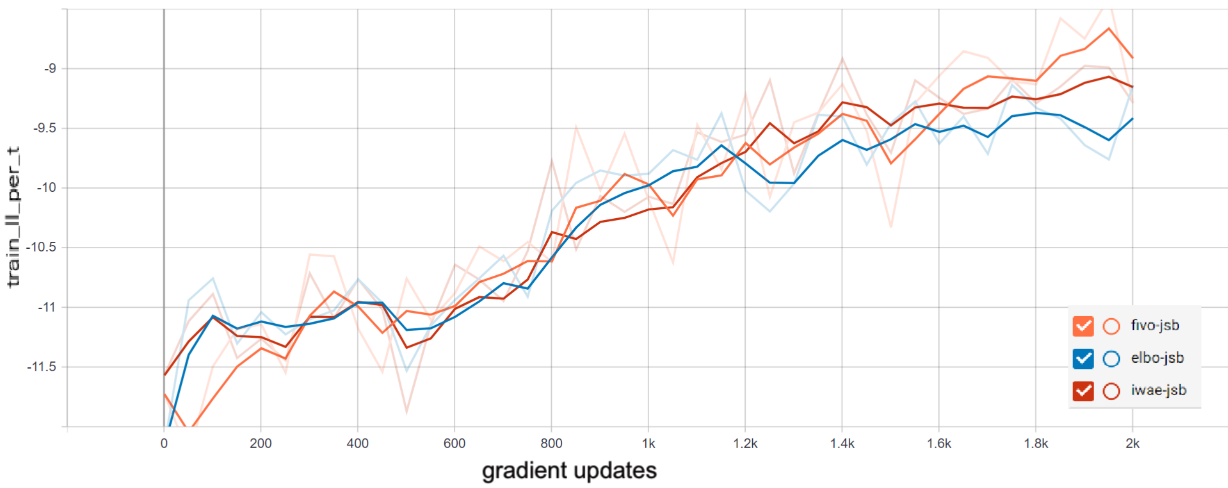


Figure 1: train log likelihood (ll/t) vs the number of training steps(2k steps)

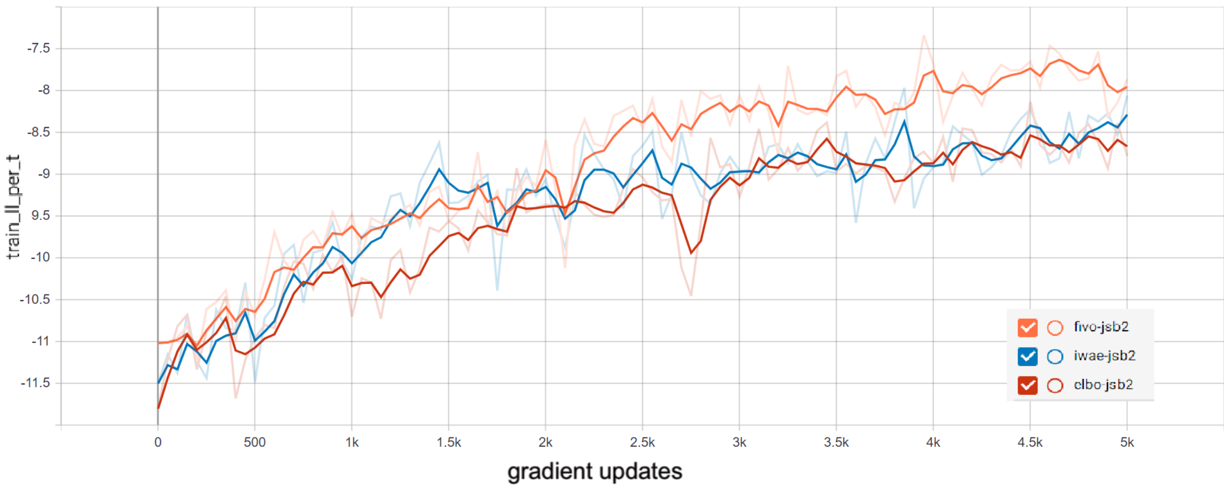


Figure 2: train log likelihood (ll/t) vs the number of training steps(5k steps)

Evaluation was run, the test log likelihood after 5K steps is summarized in Table 1.

| | **FIVO** | **IWAE** | **ELBO** |
|---|---|---|---|
| **ll/t** | -7.965333 | -8.480670 | -8.760990 |
| **ll/seq** | -488.781799 | -520.404730 | -537.606202 |

Table 1: Evaluation results(5k steps)

# Exercise 5

Figure 3 shows a comparison between the relative entropy and ESS re-sampling criterion as implemented in [1] by the authors. A small improvements can be observed in bound during training. The evaluation results however , don't vary significantly.

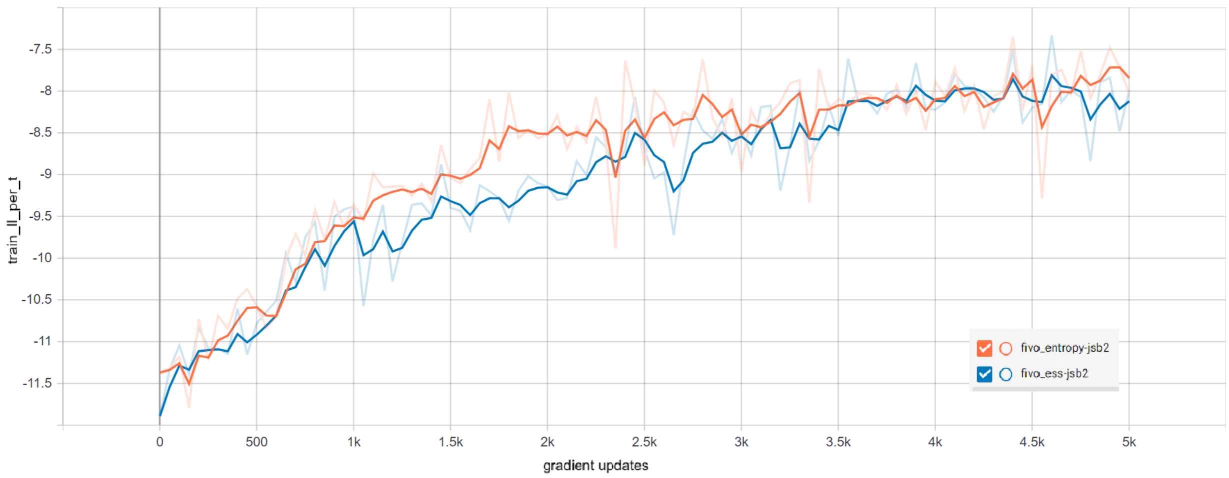


Figure 3: train log likelihood (ll/t) vs the number of training steps for relative entropy and ess re-sampling criterion

| | **FIVO Entropy** | **FIVO ESS** |
|---|---|---|
| **ll/t** | -8.070633 | -8.078573 |
| **ll/seq** | -495.243389 | -495.730597 |

Table 2: Evaluation results(5k steps)

# References


[1] Chris J Maddison, John Lawson, George Tucker, Nicolas Heess, Mohammad Norouzi, Andriy Mnih, Arnaud Doucet, and Yee Teh. Filtering variational objectives. In Advances in Neural Information Processing Systems, pages 65766586, 2017.

[2] https://en.wikipedia.org/wiki/Jensen%27s_inequality

[3] Elements of Information Theory by Thomas M. Cover and Joy A. Thomas