Notes on Natural-Gradient Variational Inference

Yuanxing Cheng

March 23, 2022

1 About natural gradient VI

1.1 Exponential familiy

$$q(\theta \,|\, \eta) = q_\eta(\theta) = h(\theta) \exp\left\{\langle \eta, \phi(\theta)\rangle - A(\eta)\right\} \tag{1.1}$$

Here q is an exponential familiy over parameters θ with natural parameters η.

- $\phi(\theta)$ is the vector of sufficient statistics.
- $A(\eta) = \log \int \exp\left(\phi(\theta)^\top \eta\right) \mathrm{d}\theta$ is the log-partition function.
- $h(\theta)$ is a scaling constant.

As in our case, this $q(\theta \,|\, \eta)$ is our $\rho(x)$. This paper update q through updating η.

Also assume a minimal exponential familiy, then $\phi(\theta)$ are lin-idp. And this leads to the result that there's a 1-to-1 mapping between η and mean parameters m.

$$m = \mathrm{E}\left[\phi(\theta)\right] = \nabla_\eta A(\eta)$$

Above equation is obtained by considering it as the first order dirivative of first momentum.

Next, the objective function is ELBO defined as following:

$$\mathcal{L}(\eta) = \mathbb{E}_{q_\eta(\theta)}\left[\log p(\mathcal{D} \,|\, \theta)\right] + \mathbb{E}_{q_\eta(\theta)}\left[\log \frac{p_0(\theta)}{q_\eta(\theta)}\right] \tag{1.2}$$

where $\mathcal{D}$ are data. Above it's a expecation of likelihood plus the KL divergence.

1.2 Updating strategy

$$\eta_{t+1} = \eta_t + \beta_t \mathbf{F}^{-1}(\eta_t) \nabla_\eta \mathcal{L}(\eta_t) \tag{1.3}$$

with $\mathbf{F}(\eta_t) = \mathbb{E}_{q_\eta(\theta)}\left[\nabla_\eta \log q_\eta(\theta) \nabla_\eta \log q_\eta(\theta)^\top\right]$ the Fisher Information matrix. i.e., $I = \int \rho \left|\nabla \log \rho\right|^2 \mathrm{d}x$ in our case. And β_t is the learning rate.

A simplification with result: $\mathbf{F}(\eta) = \nabla^2_{\eta\eta} A(\eta)$, then by consider $\mathcal{L}$ as a function of m instead of η (denote as $\mathcal{L}_*$), we have

$$\nabla_\eta \mathcal{L}(\eta_t) = \nabla_\eta m_t \nabla_m \mathcal{L}_*(m_t) = \nabla^2_{\eta\eta} A(\eta) \nabla_m \mathcal{L}_*(m_t) = \mathbf{F}(\eta) \nabla_m \mathcal{L}_*(m_t) \tag{1.4}$$

And thus the updating strategy is reduced to

$$\eta_{t+1} = \eta_t + \beta_t \nabla_m \mathcal{L}_*(m_t) \tag{1.5}$$

Then we plug in $\mathcal{L}_*$, first we notice the gradient of KL term is easily obtained.

$$\nabla_m \mathrm{KL} = \nabla_m \mathbb{E}_{q_\eta \theta}\left[\phi(\theta)^\top (\eta_0 - \eta) + A(\eta) + \mathrm{const}\right]$$

$$
\begin{aligned}
&= \nabla_m (m^\top (\eta_0 - \eta)) + \nabla_m A(\eta) \\
&= (\eta_0 - \eta - \nabla_m \eta^\top) + \nabla_m A(\eta) \\
&= \eta_0 - \eta - \mathbf{F}^{-1}(\eta) m + \mathbf{F}^{-1}(\eta) m = \eta_0 - \eta
\end{aligned}
$$

Then the update further reduced through

$$
\eta_{t+1} = \eta_t + \beta_t (\nabla_m \text{likelihood} + (\eta_0 - \eta_t)) \tag{1.6}
$$