

Variational Inference 04 Stochastic Gradient Variational Inference

Chen Gong

01 December 2019

在上一小节中，我们分析了 Mean Field Theory Variational Inference，通过平均假设来得到变分推断的理论，是一种 classical VI，我们可以将其看成 Coordinate Ascend。而另一种方法是 Stochastic Gradient Variational Inference (SGVI)。

对于隐变量参数 z 和数据集 x 。 $z \rightarrow x$ 是 Generative Model，也就是 $p(x|z)$ 和 $p(x, z)$ ，这个过程也被我们称为 Decoder。 $x \rightarrow z$ 是 Inference Model，这个过程被我们称为 Encoder，表达关系也就是 $p(z|x)$ 。

1 SGVI 参数规范

我们这节的主题就是 Stochastic Gradient Variational Inference (SGVI)，参数的更新方法为：

$$\theta^{(t+1)} = \theta^{(t)} + \lambda^{(t)} \nabla \mathcal{L}(q) \quad (1)$$

其中， $q(z|x)$ 被我们简化表示为 $q(z)$ ，我们令 $q(z)$ 是一个固定形式的概率分布， ϕ 为这个分布的参数，那么我们将把这个概率写成 $q_\phi(z)$ 。

那么，我们需要对原等式中的表达形式进行更新，

$$ELBO = \mathbb{E}_{q_\phi(z)} [\log p_\theta(x^{(i)}, z) - \log q_\phi(z)] = \mathcal{L}(\phi) \quad (2)$$

而，

$$\log p_\theta(x^{(i)}) = ELBO + KL(q||p) \geq \mathcal{L}(\phi) \quad (3)$$

而求解目标也转换成了：

$$\hat{p} = \arg \max_{\phi} \mathcal{L}(\phi) \quad (4)$$

2 SGVI 的梯度推导

$$\begin{aligned} \nabla_{\phi} \mathcal{L}(\phi) &= \nabla_{\phi} \mathbb{E}_{q_{\phi}} [\log p_{\theta}(x^{(i)}, z) - \log q_{\phi}] \\ &= \nabla_{\phi} \int q_{\phi} [\log p_{\theta}(x^{(i)}, z) - \log q_{\phi}] dz \\ &= \int \nabla_{\phi} q_{\phi} [\log p_{\theta}(x^{(i)}, z) - \log q_{\phi}] dz + \int q_{\phi} \nabla_{\phi} [\log p_{\theta}(x^{(i)}, z) - \log q_{\phi}] dz \end{aligned} \quad (5)$$

我们把这个等式拆成两个部分，其中：

$\int \nabla_{\phi} q_{\phi} [\log p_{\theta}(x^{(i)}, z) - \log q_{\phi}] dz$ 为第一个部分；

$\int q_{\phi} \nabla_{\phi} [\log p_{\theta}(x^{(i)}, z) - \log q_{\phi}] dz$ 为第二个部分。

2.1 关于第二部分的求解

第二部分比较好求，所以我们才首先求第二部分的，哈哈！因为 $\log p_{\theta}(x^{(i)}, z)$ 与 ϕ 无关。

$$\begin{aligned}
 2 &= \int q_{\phi} \nabla_{\phi} [\log p_{\theta}(x^{(i)}, z) - \log q_{\phi}] dz \\
 &= - \int q_{\phi} \nabla_{\phi} \log q_{\phi} dz \\
 &= - \int q_{\phi} \frac{1}{q_{\phi}} \nabla_{\phi} q_{\phi} dz \\
 &= - \int \nabla_{\phi} q_{\phi} dz \\
 &= - \nabla_{\phi} \int q_{\phi} dz \\
 &= - \nabla_{\phi} 1 \\
 &= 0
 \end{aligned} \tag{6}$$

2.2 关于第一部分的求解

在这里我们用到了一个小 trick，这个 trick 在公式 (6) 的推导中，我们使用过的。那就是 $\nabla_{\phi} q_{\phi} = q_{\phi} \nabla_{\phi} \log q_{\phi}$ 。所以，我们代入到第一项中可以得到：

$$\begin{aligned}
 1 &= \int \nabla_{\phi} q_{\phi} [\log p_{\theta}(x^{(i)}, z) - \log q_{\phi}] dz \\
 &= \int q_{\phi} \nabla_{\phi} \log q_{\phi} [\log p_{\theta}(x^{(i)}, z) - \log q_{\phi}] dz \\
 &= \mathbb{E}_{q_{\phi}} [\nabla_{\phi} \log q_{\phi} \log p_{\theta}(x^{(i)}, z) - \log q_{\phi}]
 \end{aligned} \tag{7}$$

那么，我们可以得到：

$$\nabla_{\phi} \mathcal{L}(\phi) = \mathbb{E}_{q_{\phi}} [\nabla_{\phi} \log q_{\phi} \log p_{\theta}(x^{(i)}, z) - \log q_{\phi}] \tag{8}$$

那么如何求这个期望呢？我们采用的是蒙特卡罗采样法，假设 $z^l \sim q_{\phi}(z)$ $l = 1, 2, \dots, L$ ，那么有：

$$\nabla_{\phi} \mathcal{L}(\phi) \approx \frac{1}{L} \sum_{l=1}^L \nabla_{\phi} \log q_{\phi}(z^{(l)}) [\log p_{\theta}(x^{(i)}, z) - \log q_{\phi}(z^{(l)})] \tag{9}$$

由于第二部分的结果为 0，所以第一部分的解就是最终的解。但是，这样的求法有什么样的问题呢？因为我们在采样的过程中，很有可能采到 $q_{\phi}(z) \rightarrow 0$ 的点，对于 \log 函数来说， $\lim_{x \rightarrow 0} \log x = -\infty$ ，那么梯度的变化会非常的剧烈，非常的不稳定。对于这样的 High Variance 的问题，根本没有办法求解。实际上，我们可以通过计算得到这个方差的解析解，它确实是一个很大的值。事实上，这里的梯度的方差这么的大，而 $\hat{\phi} \rightarrow q(z)$ 也有误差，误差叠加，直接爆炸，根本没有办法用。也就是不会 work，那么我们如何解决这个问题？

3 Variance Reduction

这里采用了一种比较常见的方差缩减方法，称为 Reparameterization Trick，也就是对 q_ϕ 做一些简化。

我们怎么可以较好的解决这个问题？如果我们可以得到一个确定的解 $p(\epsilon)$ ，就会变得比较简单。因为 z 来自于 $q_\phi(z|x)$ ，我们就想办法将 z 中的随机变量给解放出来。也就是使用一个转换 $z = g_\phi(\epsilon, x^{(i)})$ ，其中 $\epsilon \sim p(\epsilon)$ 。那么这样做，有什么好处呢？原来的 $\nabla_\phi \mathbb{E}_{q_\phi}[\cdot]$ 将转换为 $\mathbb{E}_{p(\epsilon)}[\nabla_\phi(\cdot)]$ ，那么不在是连续的关于 ϕ 的采样，这样可以有效的降低方差。并且， z 是一个关于 ϵ 的函数，我们将随机性转移到了 ϵ ，那么问题就可以简化为：

$$z \sim q_\phi(z|x^{(i)}) \longrightarrow \epsilon \sim p(\epsilon) \quad (10)$$

而且，这里还需要引入一个等式，那就是：

$$|q_\phi(z|x^{(i)})dz| = |p(\epsilon)d\epsilon| \quad (11)$$

为什么呢？我们直观性的理解一下， $\int q_\phi(z|x^{(i)})dz = \int p(\epsilon)d\epsilon = 1$ ，并且 $q_\phi(z|x^{(i)})$ 和 $p(\epsilon)$ 之间存在一个变换关系。

那么，我们将改写 $\nabla_\phi \mathcal{L}(\phi)$ ：

$$\begin{aligned} \nabla_\phi \mathcal{L}(\phi) &= \nabla_\phi \mathbb{E}_{q_\phi} [\log p_\theta(x^{(i)}, z) - \log q_\phi] \\ &= \nabla_\phi \int [\log p_\theta(x^{(i)}, z) - \log q_\phi] q_\phi dz \\ &= \nabla_\phi \int [\log p_\theta(x^{(i)}, z) - \log q_\phi] p(\epsilon) d\epsilon \\ &= \nabla_\phi \mathbb{E}_{p(\epsilon)} [\log p_\theta(x^{(i)}, z) - \log q_\phi] \\ &= \mathbb{E}_{p(\epsilon)} \nabla_\phi [(\log p_\theta(x^{(i)}, z) - \log q_\phi)] \\ &= \mathbb{E}_{p(\epsilon)} \nabla_z [(\log p_\theta(x^{(i)}, z) - \log q_\phi(z|x^{(i)})) \nabla_\phi z] \\ &= \mathbb{E}_{p(\epsilon)} \nabla_z [(\log p_\theta(x^{(i)}, z) - \log q_\phi(z|x^{(i)})) \nabla_\phi z] \\ &= \mathbb{E}_{p(\epsilon)} \nabla_z [(\log p_\theta(x^{(i)}, z) - \log q_\phi(z|x^{(i)})) \nabla_\phi g_\phi(\epsilon, x^{(i)})] \end{aligned} \quad (12)$$

那么我们的问题就这样愉快的解决了， $p(\epsilon)$ 的采样与 ϕ 无关，然后对先求关于 z 的梯度，然后再求关于 ϕ 的梯度，那么这三者之间就互相隔离开了。最后，我们再对结果进行采样， $\epsilon^{(l)} \sim p(\epsilon)$ ， $l = 1, 2, \dots, L$ ：

$$\nabla_\phi \mathcal{L}(\phi) \approx \frac{1}{L} \sum_{i=1}^L \nabla_z [(\log p_\theta(x^{(i)}, z) - \log q_\phi(z|x^{(i)})) \nabla_\phi g_\phi(\epsilon, x^{(i)})] \quad (13)$$

其中 $z \leftarrow g_\phi(\epsilon^{(i)}, x^{(i)})$ 。而 SGVI 为：

$$\phi^{(t+1)} \longrightarrow \phi^{(t)} + \lambda^{(t)} \nabla_\phi \mathcal{L}(\phi) \quad (14)$$

4 小结

那么 SGVI，可以简要的表述为：我们定义分布为 $q_\phi(Z|X)$ ， ϕ 为参数，参数的更新方法为：

$$\phi^{(t+1)} \longrightarrow \phi^{(t)} + \lambda^{(t)} \nabla_\phi \mathcal{L}(\phi) \quad (15)$$

$\nabla_{\phi}\mathcal{L}(\phi)$ 为:

$$\nabla_{\phi}\mathcal{L}(\phi) \approx \frac{1}{L} \sum_{i=1}^L \nabla_z [\log p_{\theta}(x^{(i)}, z) - \log q_{\phi}(z|x^{(i)})] \nabla_{\phi} g_{\phi}(\epsilon, x^{(i)}) \quad (16)$$