Variational Inference 01 Background

Chen Gong

30 November 2019

这一小节的主要目的是清楚我们为什么要使用 Variational Inference，表达一下 Inference 到底有什么用。机器学习，我们可以从频率角度和贝叶斯角度两个角度来看，其中频率角度可以被解释为优化问题，贝叶斯角度可以被解释为积分问题。

1 优化问题

为什么说频率派角度的分析是一个优化问题呢？我们从回归和 SVM 两个例子上进行分析。我们将数据集描述为：$D = \{(x_i, y_i)\}_{i=1}^N, x_i \in \mathbb{R}^p, y_i \in \mathbb{R}$。

1.1 回归

回归模型可以被我们定义为：$f(w) = w^T x$，其中 loss function 被定义为：$L(w) = \sum_{i=1}^N ||w^T x_i - y_i||^2$，优化可以表达为 $\hat{w} = argmin\ L(w)$。这是个无约束优化问题。

求解的方法可以分成两种，数值解和解析解。解析解的解法为：

$$\frac{\partial L(w)}{\partial w} = 0 \Rightarrow w^* = (X^T X)^{-1} X^T Y \tag{1}$$

其中，X 是一个 $n \times p$ 的矩阵。而数值解中，我们常用的是 GD 算法，也就是 Gradient Descent，或者 Stochastic Gradient descent (SGD)。

1.2 SVM (Classification)

SVM 的模型可以被我们表述为：$f(w) = sign(w^T + b)$。loss function 被我们定义为：

$$\begin{cases} \min\ \frac{1}{2} w^T w \\ s.t.\quad y_i(w^T x_i + b) \geq 1 \end{cases} \tag{2}$$

很显然这是一个有约束的 Convex 优化问题。常用的解决条件为，QP 方法和 Lagrange 对偶。

1.3 EM 算法

我们的优化目标为：

$$\hat{\theta} = \arg\max_{\theta}\ \log p(X|\theta) \tag{3}$$

优化的迭代算法为：

$$\theta^{(t+1)} = \arg\max_{\theta} \int_z \log p(X, Z|\theta) \cdot p(Z|X, \theta^{(t)}) dz \tag{4}$$

2 积分问题

从贝叶斯的角度来说，这就是一个积分问题，为什么呢？我们看看 Bayes 公式的表达：

$$p(\theta|x) = \frac{p(x|\theta)p(\theta)}{p(x)} \tag{5}$$

其中，$p(\theta|x)$ 称为后验公式，$p(x|\theta)$ 称为似然函数，$p(\theta)$ 称为先验分布，并且 $p(x) = \int_\theta p(x|\theta)p(\theta)d\theta$。什么是推断呢？通俗的说就是求解后验分布 $p(\theta|x)$。而 $p(\theta|x)$ 的计算在高维空间的时候非常的复杂，我们通常不能直接精确的求得，这是就需要采用方法来求一个近似的解。而贝叶斯的方法往往需要我们解决一个贝叶斯决策的问题，也就是根据数据集 X(N 个样本)。我们用数学的语言来表述也就是，$\widetilde{X}$ 为新的样本，求 $p(\widetilde{X}|X)$：

$$\begin{aligned} p(\widetilde{X}|X) &= \int_\theta p(\widetilde{X}, \theta|X)d\theta \\ &= \int_\theta p(\widetilde{X}|\theta) \cdot p(\theta|X)d\theta \\ &= \mathbb{E}_{\theta|X}[p(\widetilde{X}|\theta)] \end{aligned} \tag{6}$$

其中 $p(\theta|X)$ 为一个后验分布，那么我们关注的重点问题就是求这个积分。

3 Inference

Inference 的方法可以被我们分为精确推断和近似推断，近似推断可以被我们分为确定性推断和随机近似。确定性推断包括 Variational Inference (VI)；随机近似包括 MCMC，MH，Gibbs Sampling 等

Variational Inference 02 Algorithm

Chen Gong

30 November 2019

我们将 X：Observed data；Z：Latent Variable + Parameters。那么 (X, Z) 为 complete data。根据我们的贝叶斯分布公式：

$$p(X) = \frac{p(X, Z)}{p(Z|X)} \tag{1}$$

在两边同时取对数并且引入函数 $q(Z)$ 我们可以得到：

$$\begin{aligned} \log p(X) &= \log \frac{p(X, Z)}{p(Z|X)} \\ &= \log p(X, Z) - \log p(Z|X) \\ &= \log \frac{p(X, Z)}{q(Z)} - \log \frac{p(Z|X)}{q(Z)} \end{aligned} \tag{2}$$

1 公式化简

左边 $= p(X) = \int_Z log\ p(X)q(Z)dZ$。

右边 $=$

$$\int_Z q(Z) \log \frac{p(X, Z)}{q(Z)} dZ - \int_Z q(Z) \log \frac{p(Z|X)}{q(Z)} dZ \tag{3}$$

其中，$\int_Z q(Z) \log \frac{p(X,Z)}{q(Z)} dZ$ 被称为 Evidence Lower Bound (ELBO)，被我们记为 $\mathcal{L}(q)$，也就是变分。

$-\int_Z q(Z) \log \frac{p(Z|X)}{q(Z)} dZ$ 被称为 $KL(q||p)$。这里的 $KL(q||p) \geq 0$。

由于我们求不出 $p(Z|X)$,我们的目的是寻找一个 $q(Z)$,使得 $p(Z|X)$ 近似于 $q(Z)$,也就是 $KL(q||p)$ 越小越好。并且，$p(X)$ 是个定值，那么我们的目标变成了 $\arg\max_{q(z)} \mathcal{L}(q)$。那么，我们理一下思路，我们想要求得一个 $\widetilde{q}(Z) \approx p(Z|X)$。也就是

$$\widetilde{q}(Z) = \arg\max_{q(z)} \mathcal{L}(q) \Rightarrow \widetilde{q}(Z) \approx p(Z|X) \tag{4}$$

2 模型求解

那么我们如何来求解这个问题呢?我们使用到统计物理中的一种方法,就是平均场理论 (mean field theory)。也就是假设变分后验分式是一种完全可分解的分布：

$$q(z) = \prod_{i=1}^{M} q_i(z_i) \tag{5}$$

在这种分解的思想中，我们每次只考虑第 j 个分布，那么令 $q_i(1,2,\cdots,j-1,j+1,\cdots,M)$ 个分布 fixed。

那么很显然：

$$\mathcal{L}(q) = \int_Z q(Z)\log p(X,Z)dz - \int_Z q(Z)\log q(Z)dZ \tag{6}$$

我们先来分析第一项 $\int_Z q(Z)\log p(X,Z)dZ$。

$$\begin{aligned}\int_Z q(Z)\log p(X,Z)dZ &= \int_Z \prod_{i=1}^M q_i(z_i)\log p(X,Z)dZ \\ &= \int_{z_j} q_j(z_j)\left[\int_{z_1}\int_{z_2}\cdots\int_{z_M}\prod_{i=1}^M q_i(z_i)\log p(X,Z)dz_1dz_2\cdots dz_M\right]dz_j \\ &= \int_{z_j} q_j(z_j)\mathbb{E}_{\prod_{i\neq j}^M q_i(x_i)}\left[\log p(X,Z)\right]dz_j\end{aligned} \tag{7}$$

然后我们来分析第二项 $\int_Z q(Z)\log q(Z)dZ$，

$$\begin{aligned}\int_Z q(Z)\log q(Z)dZ &= \int_Z \prod_{i=1}^M q_i(z_i)\sum_{i=1}^M \log q_i(z_i)dZ \\ &= \int_Z \prod_{i=1}^M q_i(z_i)\left[\log q_1(z_1) + q_2(z_2) + \cdots + q_M(z_M)\right]dZ\end{aligned} \tag{8}$$

这个公式的计算如何进行呢？我们抽出一项来看，就会变得非常的清晰：

$$\begin{aligned}\int_Z \prod_{i=1}^M q_i(z_i)\log q_1(z_1)dZ &= \int_{z_1z_2\cdots z_M} q_1q_2\cdots q_M\log q_1dz_1dz_2\cdots z_M \\ &= \int_{z_1} q_1\log q_1dz_1\cdot\int_{z_2}q_2dz_2\cdot\int_{z_3}q_3dz_3\cdots\int_{z_M}q_Mdz_M \\ &= \int_{z_1} q_1\log q_1dz_1\end{aligned} \tag{9}$$

因为，$\int_{z_2} q_2dz_2$ 每一项的值都是 1。所以第二项可以写为：

$$\sum_{i=1}^M\int_{z_i} q_i(z_i)\log q_i(z_i)dz_i = \int_{z_j} q_j(z_j)\log q_i(z_i)dz_j + C \tag{10}$$

因为我们仅仅只关注第 j 项，其他的项都不关注。为了进一步表达计算，我们将：

$$\mathbb{E}_{\prod_{i\neq j}^M q_i(z_i)}\left[\log p(X,Z)\right] = \log\hat{p}(X,z_j) \tag{11}$$

那么 (8) 式可以写作：

$$\int_{z_j} q_j(z_j)\log\hat{p}(X,z_j)dz_j \tag{12}$$

这里的 $\hat{p}(X,z_j)$ 表示为一个相关的函数形式，假设具体参数未知。那么 (7) 式将等于 (13) 式减 (11) 式：

$$\int_{z_j} q_j(z_j)\log q_i(z_i)dz_j - \int_{z_j} q_j(z_j)\log\hat{p}(X,z_j)dz_j + C = -KL(q_j||\hat{p}(x,z_j)) + C \tag{13}$$

$\arg\max_{q_j(z_j)} -KL(q_j||\hat{p}(x,z_j))$ 等价于 $\arg\min_{q_j(z_j)} KL(q_j||\hat{p}(x,z_j))$。那么这个 $KL(q_j||\hat{p}(x,z_j))$ 要如何进行优化呢？我们下一节将回归 EM 算法，并给出求解的过程。

Variational Inference 03 Algorithm Solution

Chen Gong

01 December 2019

在上一小节中，我们介绍了 Mean Field Theory Variational Inference 的方法。在这里我需要进一步做一些说明，z_i 表示的不是一个数，而是一个数据维度的集合，它表示的不是一个维度，而是一个类似的最大团，也就是多个维度凑在一起。在上一节中，我们得出：

$$\log q_j(z_j) = \mathbb{E}_{\prod_{i\neq j} q_i(z_i)} \left[\log p(X, Z|\theta)\right] + C \tag{1}$$

并且，我们令数据集为 $X = \{x^{(i)}\}_{i=1}^N$，$Y = \{y^{(i)}\}_{i=1}^N$。variation 的核心思想是在于用一个分布 q 来近似得到 $p(z|x)$。其中优化目标为，$\hat{q} = \arg\min\ KL(q||p)$。其中：

$$\log p(X|\theta) = ELBO(\mathcal{L}(q)) + KL(q||p) \geq \mathcal{L}(q) \tag{2}$$

在这个求解中，我们主要想求的是 $q(x)$，那么我们需要弱化 θ 的作用。所以，我们计算的目标函数为：

$$\hat{q} = \arg\min_q KL(q||p) = \arg\max_q \mathcal{L}(q) \tag{3}$$

在上一小节中，这是我们的便于观察的表达方法，但是我们需要严格的使用我们的数学符号。

1 数学符号规范化

在这里我们弱化了相关参数 θ，也就是求解过程中，不太考虑 θ 起到的作用。我们展示一下似然函数，

$$\log p_\theta(X) = \log \prod_{i=1}^N p_\theta(x^{(i)}) = \sum_{i=1}^N \log p_\theta(x^{(i)}) \tag{4}$$

我们的目标是使每一个 $x^{(i)}$ 最大，所以将对 ELBO 和 $KL(p||q)$ 进行规范化表达：

ELBO：

$$\mathbb{E}_{q(z)}\left[\log \frac{p_\theta(x^{(i)}, z)}{q(z)}\right] = \mathbb{E}_{q(z)}\left[\log p_\theta(x^{(i)}, z)\right] + H(q(z)) \tag{5}$$

KL:

$$KL(q||p) = \int q(z) \cdot \log \frac{q(z)}{p_\theta(z|x^{(i)})} dz \tag{6}$$

而，

$$\begin{aligned}\log q_j(z_j) =& \mathbb{E}_{\prod_{i\neq j} q_i(z_i)}\left[\log p_\theta(x^{(i)}, z)\right] + C \\ =& \int_{q_1}\int_{q_2}\cdots\int_{q_{j-1}}\int_{q_{j+1}}\cdots\int_{q_M} q_1 q_2 \cdots q_{j-1} q_{j+1} \cdots q_M dq_1 dq_2 \cdots dq_{j-1} dq_{j+1} \cdots dq_M\end{aligned} \tag{7}$$

2 迭代算法求解

在上一步中，我们已经将所有的符号从数据点和划分维度上进行了规范化的表达。在这一步中，我们将使用迭代算法来进行求解：

$$\hat{q}_1(z_1) = \int_{q_2} \cdots \int_{q_M} q_2 \cdots q_M \left[\log p_\theta(x^{(i)}, z)\right] dq_2 \cdots dq_M \tag{8}$$

$$\hat{q}_2(z_2) = \int_{\hat{q}_1(z_1)} \int_{q_3} \cdots \int_{q_M} \hat{q}_1 q_3 \cdots q_M \left[\log p_\theta(x^{(i)}, z)\right] \hat{q}_1 dq_2 \cdots dq_M \tag{9}$$

$$\vdots$$

$$\hat{q}_M(z_M) = \int_{\hat{q}_1} \cdots \int_{\hat{q}_{M-1}} \hat{q}_1 \cdots \hat{q}_{M-1} \left[\log p_\theta(x^{(i)}, z)\right] d\hat{q}_1 \cdots d\hat{q}_{M-1} \tag{10}$$

如果，我们将 $q_1, q_2, \cdots, q_M$ 看成一个个的坐标点，那么我们知道的坐标点越来越多，这实际上就是一种坐标上升的方法 (Coordinate Ascend)。

这是一种迭代算法，那我们怎么考虑迭代的停止条件呢？我们设置当 $\mathcal{L}^{(t+1)} \leq \mathcal{L}^{(t)}$ 时停止迭代。

3 Mean Field Theory 的存在问题

1. 首先假设上就有问题，这个假设太强了。在假设中，我们提到，假设变分后验分式是一种完全可分解的分布。实际上，这样的适用条件挺少的。大部分时候都并不会适用。

2. Intractable。本来就是因为后验分布 $p(Z|X)$ 的计算非常的复杂，所以我们才使用变分推断来进行计算，但是有个很不幸的消息。这个迭代的方法也非常的难以计算，并且

$$\log q_j(z_j) = \mathbb{E}_{\prod_{i \neq j} q_i(z_i)} \left[\log p(X, Z|\theta)\right] + C \tag{11}$$

的计算也非常的复杂。所以，我们需要寻找一种更加优秀的方法，比如 Stein Disparency 等等。Stein 变分是个非常 Fashion 的东西，机器学习理论中非常强大的算法，我们以后会详细的分析。

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。$z \longrightarrow x$ 是 Generative Model，也就是 $p(x|z)$ 和 $p(x,z)$，这个过程也被我们称为 Decoder。$x \longrightarrow z$ 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：

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

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，

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

而，

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

而求解目标也转换成了：

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

2 SGVI 的梯度推导

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

我们把这个等式拆成两个部分，其中：
$\int \nabla_\phi q_\phi \left[\log p_\theta(x^{(i)}, z) - \log q_\phi\right] dz$ 为第一个部分；
$\int q_\phi \nabla_\phi \left[\log p_\theta(x^{(i)}, z) - \log q_\phi\right] dz$ 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，哈哈哈！因为 $\log p_\theta(x^{(i)}, z)$ 与 ϕ 无关。

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

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：

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

那么，我们可以得到：

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

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 $z^l \sim q_\phi(z)\ l = 1, 2, \cdots, L$，那么有：

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

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 $q_\phi(z) \longrightarrow 0$ 的点，对于 log 函数来说，$\lim_{x \longrightarrow 0} \log x = \infty$，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 $\hat{\phi} \longrightarrow q(z)$ 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:

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

而且，这里还需要引入一个等式，那就是:

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

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 $\nabla_\phi \mathcal{L}(\phi)$:

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

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，$\epsilon^{(l)} \sim p(\epsilon)$，$l = 1, 2, \cdots, L$:

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

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

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

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:

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

$\nabla_\phi \mathcal{L}(\phi)$ 为：

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