A Connection Between DPM Object Function and Cross Entropy

郑镇鑫

blair.star@163.com

对任意概率分布的随机变量 $X \sim q(x)$，都可按照如下的变换，逐渐把概率分布转变成标准的正态分布。

$$Z = \sqrt{\alpha}X + \sqrt{1-\alpha}\epsilon \qquad where \quad \alpha < 1, \quad \epsilon \sim \mathcal{N}(0, I) \tag{1}$$

上述的变换分为两个部分，第一部分是对随机变量 X 执行一个线性变换 $(\sqrt{\alpha}X)$，线性变换使 X 的概率分布“变窄变高”，具体可参考图1的例子，左图为一维随机变量的概率分布，右图是经过线性变换后的概率分布，可以看出，相对于左图，右图的曲线“变窄变高”了；第二部分是“加上随机噪声”$(\sqrt{1-\alpha}\epsilon)$，“加上随机噪声”相当于对已有的概率分布执行“高斯模糊”，具体可参考图2，可以看出，相对于左图，右图的棱角变光滑了。关于此结论的进一步解释可参考文献 [1]。

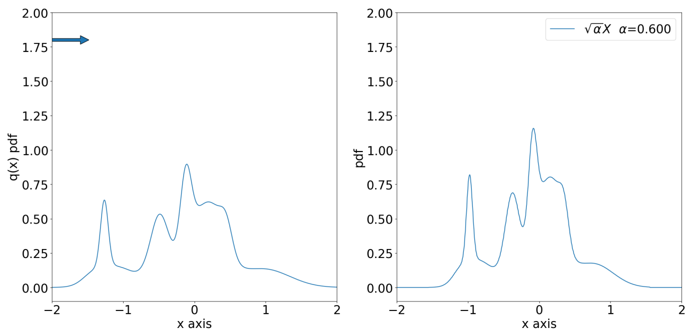

图 1: linear transform

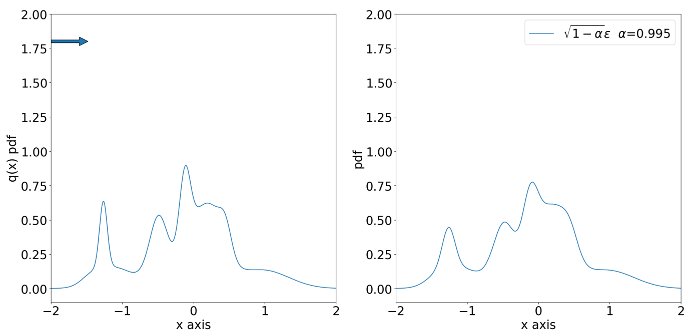

图 2: add noise

连续应用上述的变换，最终输出的概率分布将变得越来越接近于标准高斯分布。

$$
\begin{aligned}
Z_1 &= \sqrt{\alpha_1} X + \sqrt{1-\alpha_1}\epsilon_1 \\
Z_2 &= \sqrt{\alpha_2} Z_1 + \sqrt{1-\alpha_2}\epsilon_2 \\
&\dots \\
Z_t &= \sqrt{\alpha_t} Z_{t-1} + \sqrt{1-\alpha_t}\epsilon_t \\
&\dots \\
Z_T &= \sqrt{\alpha_T} Z_{T-1} + \sqrt{1-\alpha_T}\epsilon_T \\
&where \quad \alpha_t < 1 \qquad t \in 1, 2, \dots, T
\end{aligned} \tag{2}
$$

可参考图3的例子，最左侧子图是一维随机变量的概率分布，逐步加了五次噪声，最终的概率分布如最右侧子图所示，与标准高斯分布 (绿色曲线) 非常相似。

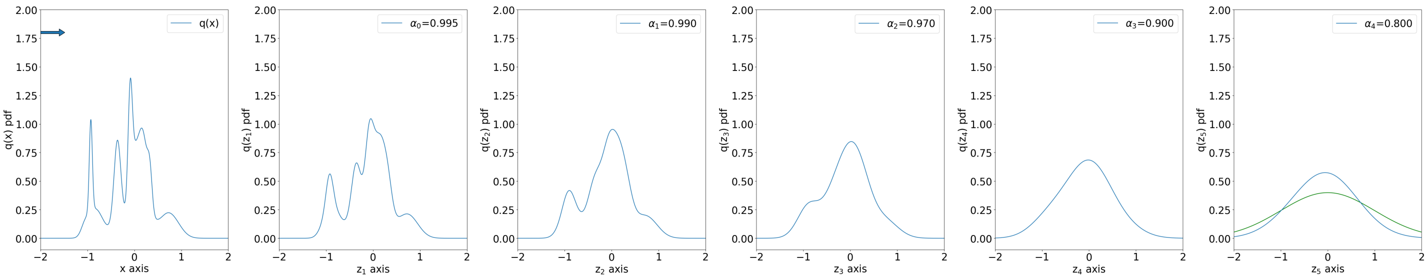

图 3: gradually add noises

如果知道了概率分布 $q(z_t)$ 及后验概率分布 $q(z_{t-1}|z_t)$，则可通过贝叶斯公司和全概率公式计算得到 $q(z_{t-1})$，如公式(5)。另外，当加了足够多的噪声后，$q(z_T)$ 的概率分布接近于标准高斯分布，所以 $q(z_T)$ 是已知的。于是，如果知道了各个后验概率分布 $\{q(z_{t-1}|z_t)\}_{t=1}^{T}$，则可从末端反向计算各个随机变量的概率分布，包括初始变量的 X 的概率分布，如式(3)至(8)。同时，也可通过祖先采样 (Ancestor Sampling) 的方法，从末端反向采样得到联合概率 $q(x, z_{1:T})$ 的样本；舍弃 $z_{1:T}$，从而进一步得到 $q(x)$ 的样本。

$$q(z_{T-1}) = \int q(z_{T-1}, z_T) dz_T = \int q(z_{T-1}|z_T) q(z_T) dz_T \tag{3}$$

$$\dots \tag{4}$$

$$q(z_{t-1}) = \int q(z_{t-1}, z_t) dz_t = \int q(z_{t-1}|z_t) q(z_t) dz_t \tag{5}$$

$$\dots \tag{6}$$

$$q(z_1) = \int q(z_1, z_2) dz_1 = \int q(z_1|z_2) q(z_2) dz_2 \tag{7}$$

$$q(x) = \int q(x, z_1) dz_1 = \int q(x|z_1) q(z_1) dz_1 \tag{8}$$

那如何学习各个后验概率分布 $q(z_{t-1}|z_t)$ 呢？

参数化一个新的概率分布函数 $p(z_{t-1}|z_t)$，依赖于 Z_t，比如条件高斯概率分布，然后通过优化交叉熵损失(cross entropy loss)，估计 $p(z_{t-1}|z_t)$ 函数的参数，使 $p(z_{t-1}|z_t)$ 接近于 $q(z_{t-1}, z_t)$。由于后验概率是条件概率分布，所以需综合考虑各个条件，并以各个条件发生的概率进行加权平均。最终的损失函数形式如下

$$loss = -\int q(z_t) \underbrace{\int q(z_{t-1}|z_t) \log p(z_{t-1}|z_t) dz_{t-1}}_{cross\ entropy} dz_t \tag{9}$$

$$= -\iint q(z_{t-1}, z_t) \log p(z_{t-1}|z_t) dz_{t-1} dz_t \tag{10}$$

对上式应用蒙特卡罗 (Monte Carlo) 积分近似，可得

$$\begin{aligned} loss &= -\iint q(z_{t-1}, z_t) \log p(z_{t-1}|z_t) dz_{t-1} dz_t \qquad (11) \\ &\approx -\sum_{i=0}^{N} \log p(Z_{t-1}^i | Z_t^i) \qquad where \quad (Z_{t-1}^i, Z_t^i) \sim q(z_{t-1}, z_t) \qquad (12) \end{aligned}$$

样本 $\{(Z_{t-1}^i, Z_t^i)\}_{i=1}^N$ 可通过式(2)变换的方式采样得到。上述形式也是一种优化方式。

对上述损失函数的形式进行转化，可得到 DPM 优化目标中一致项 (Consistent Term)。

$$\begin{aligned} loss &= -\iint q(z_{t-1}, z_t) \log p(z_{t-1}|z_t) dz_{t-1} dz_t \\ &= -\iint \int q(x) q(z_{t-1}, z_t|x) dx \log p(z_{t-1}|z_t) dz_{t-1} dz_t \\ &= \underbrace{\iint \int q(x) q(z_{t-1}, z_t|x) \log q(z_{t-1}|z_t, x) dx dz_{t-1} dz_t}_{This\ Term\ Is\ Constant,\ Represented\ By\ C_1} - \iint \int q(x) q(z_{t-1}, z_t|x) \log p(z_{t-1}|z_t) dx dz_{t-1} dz_t - C_1 \\ &= \iint \int q(x) q(z_{t-1}, z_t|x) \log \frac{q(z_{t-1}|z_t, x)}{p(z_{t-1}|z_t)} dx dz_{t-1} dz_t - C_1 \\ &= \iint q(x) q(z_t|x) \int q(z_{t-1}|z_t, x) \log \frac{q(z_{t-1}|z_t, x)}{p(z_{t-1}|z_t)} dz_{t-1}\ dz_x dz_t - C_1 \\ &= \iint \ q(x) q(z_t|x) KL[q(z_{t-1}|z_t, x) \| p(z_{t-1}|z_t)] dx dz_t - C_1 \\ &\propto \iint \ q(x) q(z_t|x) KL[q(z_{t-1}|z_t, x) \| p(z_{t-1}|z_t)] dx dz_t \end{aligned} \qquad (13)$$

上式中的 C_1 项是一个固定值，不包含待优化的参数。其中，$q(x)$ 是固定的概率分布，不知其具体的形式，但知道服从此概率分布的一批样本；$q(z_{t-1}|z_t)$ 也是固定概率分布，具体形式由 $q(x)$ 及系数 α 确定。

于是，得到了 DPM 优化目标中的一致项 (Consistent Term)。

对于重构项 (Reconstruction Term)，可通过类似的方式得到

$$\begin{aligned} loss &= -\int q(z_1) \underbrace{\int q(x|z_1) \log p(x|z_1) dx}_{Cross\ Entropy} dz_1 \\ &= -\iint q(z_1, x) \log p(x|z_1) dx dz_1 \\ &= -\int q(x) \int q(z_1|x) \log p(x|z_1) dz_1\ dx \end{aligned} \qquad (14)$$

上式即是 DPM 目标函数中的重构项 (Reconstruction Term)。

更好相关的细节可进一步阅读文献 [1]。

参考文献

[1] The Art of DPM