

Understanding Diffusion Probability Model Interactively

English Chinese

Collapse Expand

0. Introduction

扩散模型[1][2]是当前图像生成和视频生成使用的主要方式，但由于其晦涩的理论，很多工程师并不能很好地理解。本文将提供一种非常直观易懂的方式，方便读者理解把握扩散模型的原理。特别地，将以互动的形式，以一维随机变量的扩散模型进行举例，直观解释扩散模型的多个有趣的性质。

扩散模型是一个概率模型。概率模型主要提供两方面的功能：计算给定样本出现的概率；采样生成新样本。扩散模型侧重于第二方面，方便采样新样本，从而实现“生成”的任务。

扩散模型与一般的概率模型(如GMM)不同，直接建模随机变量的概率分布。扩散模型采用一种间接方式，利用“随机变量变换”的方式(如图1a)，逐步将待建模的概率分布(数据分布)转变成“标准正态分布”，同时，建模学习各个变换对应的后验概率分布(图1b-c)。有了最终的标准正态分布和各个后验概率分布，则可通过祖先采样(Ancestral Sampling)的方式，从反向逐步采样得到各个随机变量 $Z_T \dots Z_2, Z_1, X$ 的样本。同时也可通过贝叶斯公式和全概率公式确定初始的数据分布 $q(x)$ 。

可能会有这样的疑问：间接的方式需要建模学习 T 个后验概率分布，直接方式只需要建模学习一个概率分布，为什么要选择间接的方式呢？是这样子的：初始的数据分布可能很复杂，很难用一个概率模型直接表示；而对于间接的方式，各个后验概率分布的复杂度会简单许多，可以用简单的概率模型进行拟合。下面将会看到，当满足一些条件时，后验概率分布将非常接近高斯分布，所以可以使用简单的条件高斯模型进行建模。

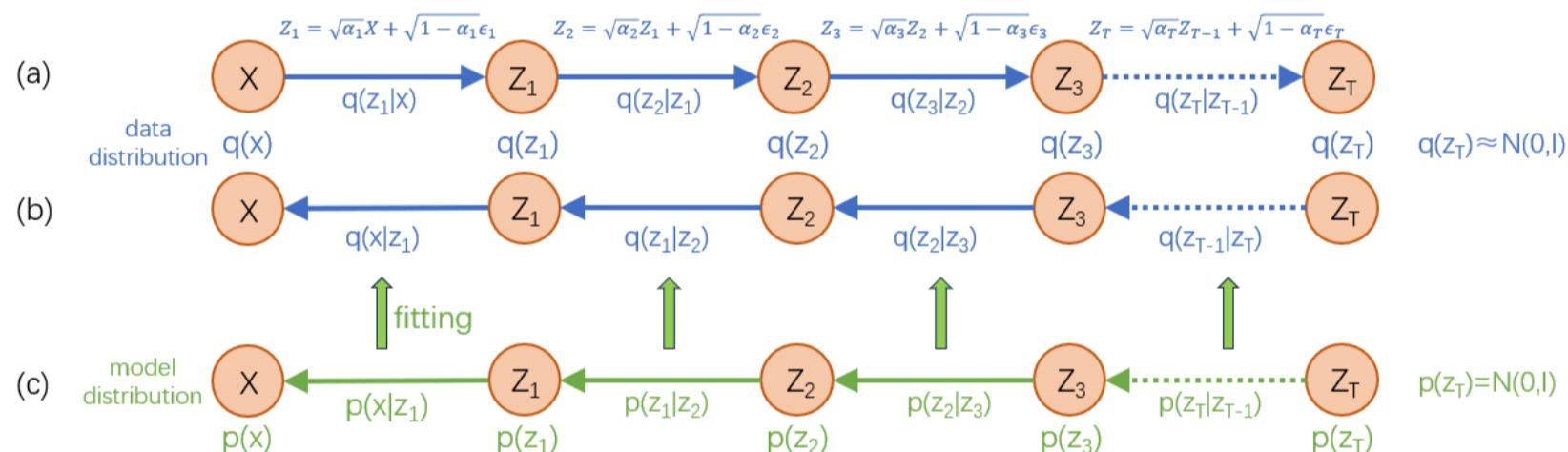


Figure 1: Diffusion model schematic

1. How To Transform

为了将初始的数据分布转换为简单的标准正态分布，扩散模型采用如下的变换方式

$$Z = \sqrt{\alpha}X + \sqrt{1 - \alpha}\epsilon \quad \text{where } \alpha < 1, \epsilon \sim \mathcal{N}(0, I) \quad (1.1)$$

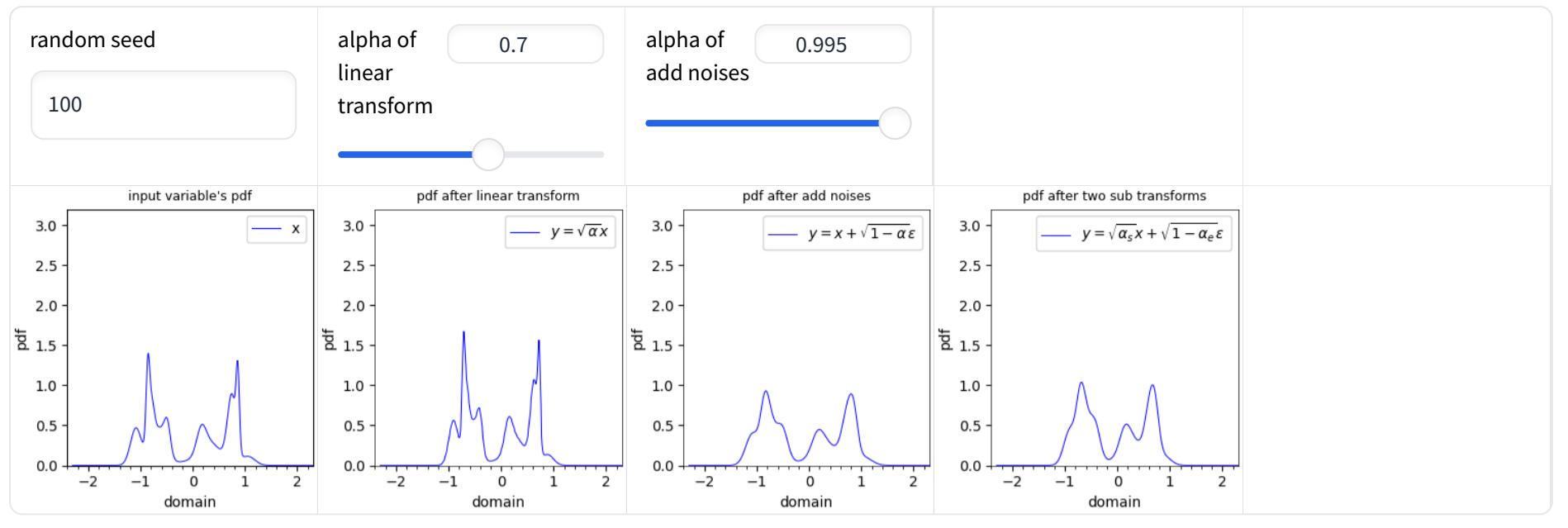
其中 $X \sim q(x)$ 是任意的随机变量， $Z \sim q(Z)$ 是变换后的随机变量。

此变换可分为两个子变换。

第一个子变换是对随机变量 X 执行一个线性变换($\sqrt{\alpha}X$)，根据文献[3]的结论，线性变换使 X 的概率分布“变窄变高”，并且“变窄变高”的程度与 α 的值成正比。具体可看[Demo 1](#)，左1图为随机生成的一维的概率分布，左2图是经过线性变换后的概率分布，可以看出，与左1图相比，左2图的曲线“变窄变高”了。读者可亲自测试不同的 α 值，获得更直观的理解。

第二个子变换是“加上独立的随机噪声”($\sqrt{1 - \alpha}\epsilon$)，根据文献[4]的结论，“加上独立的随机变量”等效于对两个概率分布执行卷积，由于随机噪声的概率分布为高斯形状，所以相当于执行“高斯模糊”的操作。经过模糊后，原来的概率分布将变得更加平滑，与标准正态分布将更加相似。模糊的程度与噪声大小($1 - \alpha$)正相关。具体可看[Demo 1](#)，左1图是随机生成的一维概率分布，左3图是经过变换后的结果，可以看出，变换后的曲线变光滑了，棱角变少了。读者可测试不同的 α 值，感受噪声大小对概率分布曲线形状的影响。左4图是综合两个子变换后的结果。

Demo 1 - Random Variable Transform In DPM



2. Likelihood of The Transform

由变换的方式(式1.1)可以看出，前向条件概率 $q(z|x)$ 的概率分布为高斯分布，且只与 α 的值有关，与 $q(x)$ 的概率分布无关。

$$q(z|x) = \mathcal{N}(\sqrt{\alpha}x, 1 - \alpha) \quad (2.1)$$

具体可看[Demo 2](#)，左3图展示了 $q(z|x)$ 的形状，从图中可以看到一条均匀的斜线，这意味着 $q(z|x)$ 的均值与 x 线性相关，方差固定不变。 α 值的大小将决定斜线宽度和倾斜程度。

3. Posterior of The Transform

后验概率分布没有闭合的形式，但可以通过一些方法，推断其大概的形状，并分析影响其形状的因素。

根据Bayes公式，有

$$q(x|z) = \frac{q(z|x)q(x)}{q(z)} \quad (3.1)$$

当 z 是取固定值时， $q(z)$ 是常数，所以 $q(x|z)$ 是关于 x 的概率密度函数，并且其形状只与 $q(z|x)q(x)$ 有关。

$$q(x|z) \propto q(z|x)q(x) \quad \text{where } z \text{ is fixed} \quad (3.2)$$

实际上， $q(z) = \int q(z|x)q(x)dx$ ，也就是说， $q(z)$ 是对函数 $q(z|x)q(x)$ 遍历 x 求和，所以， $q(z|x)q(x)$ 除以 $q(z)$ 相当于对 $q(z|x)q(x)$ 执行归一化。

$$q(x|z) = \text{Normalize}(q(z|x)q(x)) \quad (3.3)$$

由式2.1可知， $q(z|x)$ 为高斯分布，于是有

$$\begin{aligned} q(x|z) &\propto \frac{1}{\sqrt{2\pi(1-\alpha)}} \exp \frac{-(z - \sqrt{\alpha}x)^2}{2(1-\alpha)} q(x) \quad \text{where } z \text{ is fixed} \\ &= \frac{1}{\sqrt{\alpha}} \frac{1}{\sqrt{2\pi \frac{1-\alpha}{\alpha}}} \exp \frac{-\left(\frac{z}{\sqrt{\alpha}} - x\right)^2}{2\frac{1-\alpha}{\alpha}} q(x) \\ &= \underbrace{\frac{1}{\sqrt{\alpha}} \frac{1}{\sqrt{2\pi\sigma^2}} \exp \frac{-(x - \mu)^2}{2\sigma^2}}_{\text{GaussFun}} q(x) \quad \text{where } \mu = \frac{z}{\sqrt{\alpha}} \quad \sigma = \sqrt{\frac{1-\alpha}{\alpha}} \end{aligned} \quad (3.4)$$

可以看出，GaussFun部分是关于 x 的高斯函数，均值为 $\frac{z}{\sqrt{\alpha}}$ ，标准差为 $\sqrt{\frac{1-\alpha}{\alpha}}$ ，所以 $q(x|z)$ 的形状由“GaussFun与 $q(x)$ 相乘”决定。

根据“乘法”的特点，可以总结 $q(x|z)$ 函数形状具有的特点。

- $q(x|z)$ 的支撑集应该包含于GaussFun的支撑集，GaussFun的支撑集是一个超球体，中心位于均值 μ ，半径约为3倍标准差 σ 。
- 当高斯函数的方差较小(较小噪声)，或者 $q(x)$ 线性变化时， $q(x|z)$ 的形状将近似于高斯函数，函数形式较简单，方便建模学习。
- 当高斯函数的方差较大(较大噪声)，或者 $q(x)$ 剧烈变化时， $q(x|z)$ 的形状将较复杂，与高斯函数有较大的差别，难以建模学习。

[Appendix B](#)给出了较严谨的分析，当 σ 满足一些条件时， $q(x|z)$ 的近似于高斯分布。

具体可看[Demo 2](#), 左4图给出后验概率分布 $q(x|z)$ 的形态, 可以看出, 其形状较不规则, 像一条弯曲且不均匀的曲线。当 α 较大时(噪声较小), 曲线将趋于均匀且笔直。读者可调整不同的 α 值, 观察后验概率分布与噪声大小的关系; 左5图, 蓝色虚线给出 $q(x)$, 绿色虚线给出式3.4中的GaussFun, 黄色实线给出两者相乘并归一化的结果, 即固定 z 条件下后验概率 $q(x|z = \text{fixed})$ 。读者可调整不同 z 值, 观察 $q(x)$ 的波动变化对后验概率 $q(x|z)$ 形态的影响。

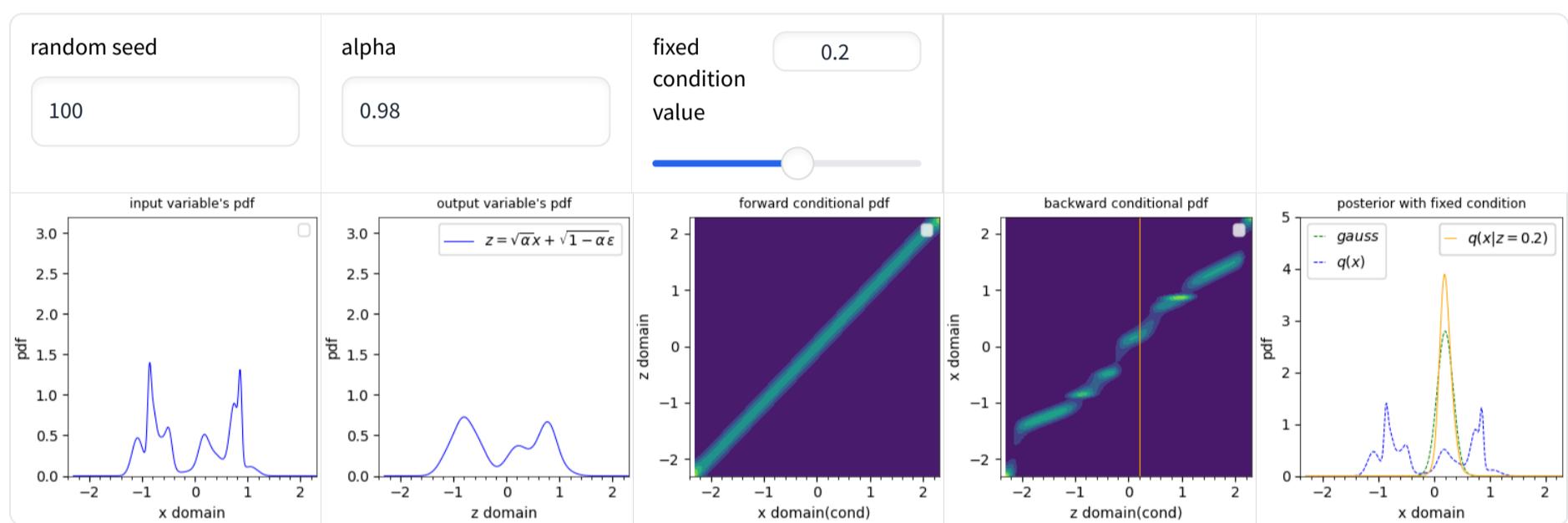
两个特殊状态下的后验概率分布 $q(x|z)$ 值得考虑一下。

- 当 $\alpha \rightarrow 0$ 时, GaussFun的标准差趋向于**无穷大**, GaussFun变成一个很大支撑集的近似的均匀分布, $q(x)$ 与均匀分布**相乘**结果仍为 $q(x)$, 所以, 不同 z 值对应的 $q(x|z)$ 几乎变成一致, 并与 $q(x)$ 几乎相同。读者可在[Demo 2](#)中, 将 α 设置为0.001, 观察具体的结果。
- 当 $\alpha \rightarrow 1$ 时, GaussFun的标准差趋向于**无穷小**, 不同 z 值的 $q(x|z)$ 收缩成一系列不同偏移量的Dirac delta函数, 偏移量等于 z 。但有一些例外, 当 $q(x)$ 存在为零的区域时, 其对应的 $q(x|z)$ 将不再为Dirac delta函数, 而是零函数。可在[Demo 2](#)中, 将 α 设置为0.999, 观察具体的结果。

有一点需要注意一下, 当 $\alpha \rightarrow 0$ 时, 较大 z 值对应的GaussFun的均值($\mu = \frac{z}{\sqrt{\alpha}}$)也急剧变大, 也就是说, GaussFun位于离原点较远的地方, 此时, $q(x)$ 的支撑集对应的GaussFun部分的“均匀程度”会略微有所下降, 从而会略微降低 $q(x|z)$ 与 $q(x)$ 的相似度, 但这种影响会随着 α 减小而进一步降低。读者可在[Demo 2](#)中观察此影响, 将 α 设置为0.001, $q(x|z = -2)$ 与 $q(x)$ 会略微有一点差别, 但 $q(x|z = 0)$ 与 $q(x)$ 却看不出区别。

关于高斯函数的“均匀程度”, 有如下两个特点: 标准差越大, 均匀程度越大; 离均值越远, 均匀程度越小。

Demo 2 - Likelihood and Posterior of Transform



4. Transform Data Distribution To Normal Distribution

对于任意的数据分布 $q(x)$, 均可连续应用上述的变换(如式4.1~4.4), 随着变换的次数的增多, 输出的概率分布将变得越来越接近于标准正态分布。对于较复杂的数据分布, 需要较多的次数或者较大的噪声。

具体可看[Demo 3.1](#), 第一子图是随机生成的一维概率分布, 经过7次的变换后, 最终的概率分布与标准正态分布非常相似。相似的程度与迭代的次数和噪声大小正相关。对于相同的相似程度, 如果每次所加的噪声较大(较小的 α 值), 那所需变换的次数将较少。读者可尝试不同的 α 值和次数, 观测最终概率分布的相似程度。

起始概率分布的复杂度会比较高, 随着变换的次数增多, 概率分布 $q(z_t)$ 的复杂度将会下降。根据第3节结论, 更复杂的概率分布对应更复杂的后验概率分布, 所以, 为了保证后验概率分布与高斯函数较相似(较容易学习), 在起始阶段, 需使用较大的 α (较小的噪声), 后期阶段可适当使用较小的 α (较大的噪声), 加快向标准正态分布转变。

在[Demo 3.1](#)的例子可以看到, 随着变换次数增多, $q(z_t)$ 的棱角变得越来越少, 同时, 后验概率分布 $q(z_{t-1}|z_t)$ 图中的斜线变得越来越笔直匀称, 越来越像条件高斯分布。

$$Z_1 = \sqrt{\alpha_1}X + \sqrt{1-\alpha_1}\epsilon_1 \quad (4.1)$$

$$Z_2 = \sqrt{\alpha_2}Z_1 + \sqrt{1-\alpha_2}\epsilon_2 \quad (4.2)$$

...

$$Z_t = \sqrt{\alpha_t}Z_{t-1} + \sqrt{1-\alpha_t}\epsilon_t \quad (4.3)$$

...

$$Z_T = \sqrt{\alpha_T}Z_{T-1} + \sqrt{1-\alpha_T}\epsilon_T \quad (4.4)$$

where $\alpha_t < 1 \quad t \in 1, 2, \dots, T$

把式4.1代入式4.2, 同时利用高斯分布的性质, 可得出 $q(z_2|x)$ 的概率分布的形式

$$z_2 = \sqrt{\alpha_2}(\sqrt{\alpha_1}x + \sqrt{1-\alpha_1}\epsilon_1) + \sqrt{1-\alpha_2}\epsilon_2 \quad (4.5)$$

$$= \sqrt{\alpha_2\alpha_1}x + \sqrt{\alpha_2 - \alpha_2\alpha_1}\epsilon_1 + \sqrt{1-\alpha_2}\epsilon_2 \quad (4.6)$$

$$= \mathcal{N}(\sqrt{\alpha_1\alpha_2}x, 1 - \alpha_1\alpha_2) \quad (4.7)$$

同理，可递推得出

$$q(z_t|x) = \mathcal{N}(\sqrt{\alpha_1\alpha_2\cdots\alpha_t}x, 1 - \alpha_1\alpha_2\cdots\alpha_t) = \mathcal{N}(\sqrt{\bar{\alpha}_t}x, 1 - \bar{\alpha}_t) \quad \text{where } \bar{\alpha}_t \triangleq \prod_{j=1}^t \alpha_j \quad (4.8)$$

比较式4.8和式2.1的形式，可发现，两者的形式是完全一致的。

如果只关注首尾两个变量之间的关系，那么连续t次的小变换可用一次大变换替代，大变换的 α 是各个小变换的 α 累积，因为两种变换对应的联合概率分布相同。

读者可在[Demo 3.1](#)中做一个实验，对同样的输入分布 $q(x)$ ，使用两种不同的变换方式：1) 使用三个变换， α 均为0.95；2) 使用一个变换， α 设置为0.857375。分别执行变换，然后比较变换后的两个分布，将会看到，两个分布是完全相同的。

在DDPM[2]论文中，作者使用了1000步($T=1000$)，将数据分布 $q(x)$ 转换至 $q(z_T)$ ， $q(z_T|x)$ 的概率分布如下：

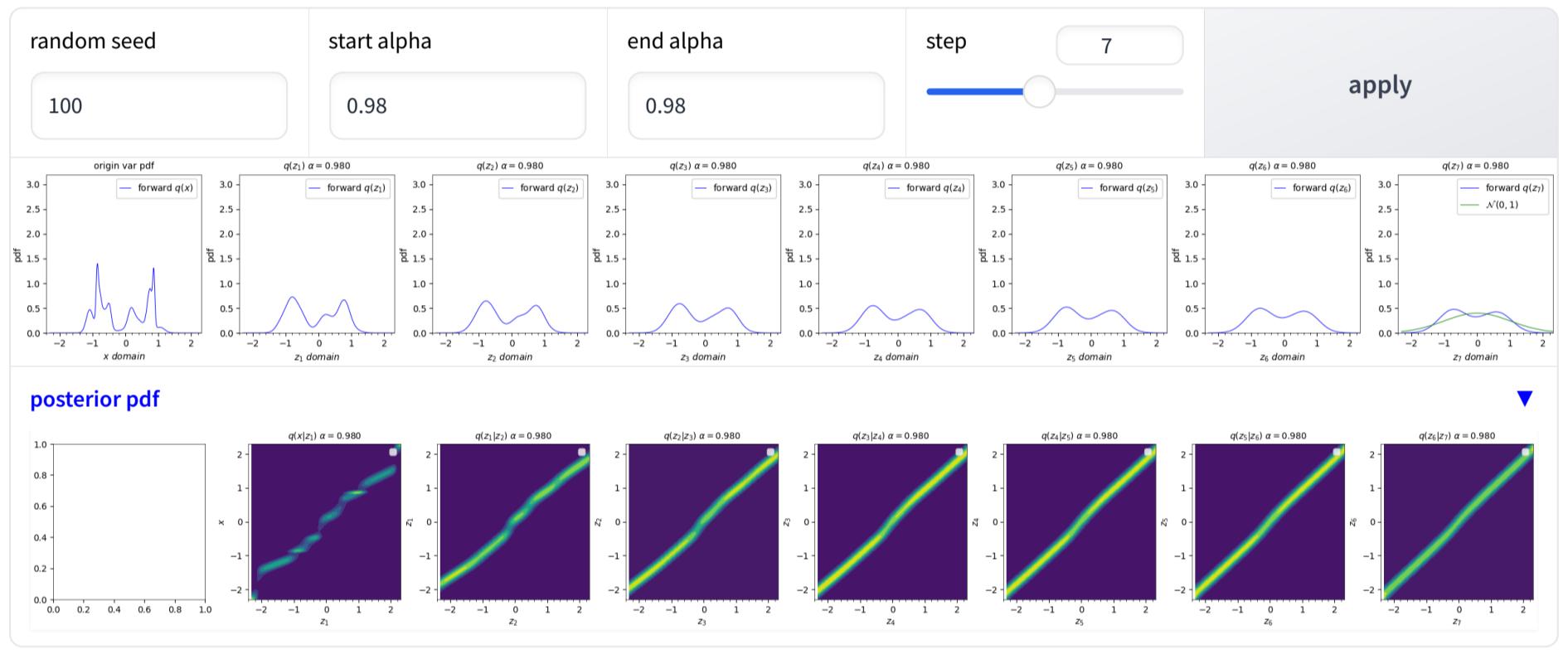
$$q(z_T|x) = \mathcal{N}(0.00635x, 0.99998) \quad (4.9)$$

如果只考虑 X, Z_T 的联合分布 $q(x, z_T)$ ，也可使用一次变换代替，变换如下：

$$Z_T = \sqrt{0.0000403}X + \sqrt{1 - 0.0000403}\epsilon = 0.00635X + 0.99998\epsilon \quad (4.10)$$

可以看出，应用两种变换后，变换后的分布 $q(z_T|x)$ 相同，因此， $q(x, z_T)$ 也相同。

Demo 3.1 - Transform To Normal Distribution Iteratively



5. Restore Data Distribution From Normal Distribution

如果知道了最终的概率分布 $q(z_T)$ 及各个转换过程的后验概率 $q(x|z), q(z_{t-1}|z_t)$ ，则可通过“贝叶斯公式”和“全概率公式”恢复数据分布 $q(x)$ ，见式5.1~5.4。当最终的概率分布 $q(z_T)$ 与标准正态分布很相似时，可用标准正态分布代替。

具体可看[Demo 3.2](#)。示例中 $q(z_T)$ 使用 $\mathcal{N}(0, 1)$ 代替，同时通过JS Div给出了误差大小。恢复的概率分布 $q(z_t)$ 及 $q(x)$ 使用绿色曲线标识，原始的概率分布使用蓝色曲线标识。可以看出，数据分布 $q(x)$ 能够被很好地恢复回来，并且误差(JS Divergence)会小于标准正态分布替换 $q(z_T)$ 引起的误差。

$$q(z_{T-1}) = \int q(z_{T-1}, z_T) dz_T = \int q(z_{T-1}|z_T) q(z_T) dz_T \quad (5.1)$$

$$\dots$$

$$q(z_{t-1}) = \int q(z_{t-1}, z_t) dz_t = \int q(z_{t-1}|z_t) q(z_t) dz_t \quad (5.2)$$

$$\dots$$

$$q(z_1) = \int q(z_1, z_2) dz_1 = \int q(z_1|z_2) q(z_2) dz_2 \quad (5.3)$$

$$q(x) = \int q(x, z_1) dz_1 = \int q(x|z_1) q(z_1) dz_1 \quad (5.4)$$

在本文中，将上述恢复过程(式5.1~5.4)所使用的变换称之为“后验概率变换”。例如，在式5.4中，变换的输入为概率分布函数 $q(z_1)$ ，输出为概率分布函数 $q(x)$ ，整个变换由后验概率分布 $q(x|z_1)$ 决定。此变换也可看作为一组基函数的线性加权和，基函数为不同条件下的 $q(x|z_1)$ ，各个基函数的权重为 $q(z_1)$ 。在[第7节](#)，将会进一步介绍此变换的一些有趣性质。

在[第3节](#)中，我们考虑了两个特殊的后验概率分布。接下来，分析其对应的“后验概率变换”。

- 当 $\alpha \rightarrow 0$ 时，不同 z 值的 $q(x|z)$ 均与 $q(x)$ 几乎相同，也就是说，线性加权和的基函数几乎相同。此状态下，**不管输入如何变化，变换的输出总为 $q(x)$** 。
- 当 $\alpha \rightarrow 1$ 时，不同 z 值的 $q(x|z)$ 收缩成一系列不同偏移量的Dirac delta函数及零函数。此状态下，只要输入分布的支撑集(support set)包含于 $q(x)$ 的支撑集，变换的输出与输入将保持一致。

在第4节中提到，DDPM[2]论文所使用的1000次变换可使用一次变换表示：

$$Z_T = \sqrt{0.0000403} X + \sqrt{1 - 0.0000403} \epsilon = 0.00635 X + 0.99998 \epsilon \quad (5.5)$$

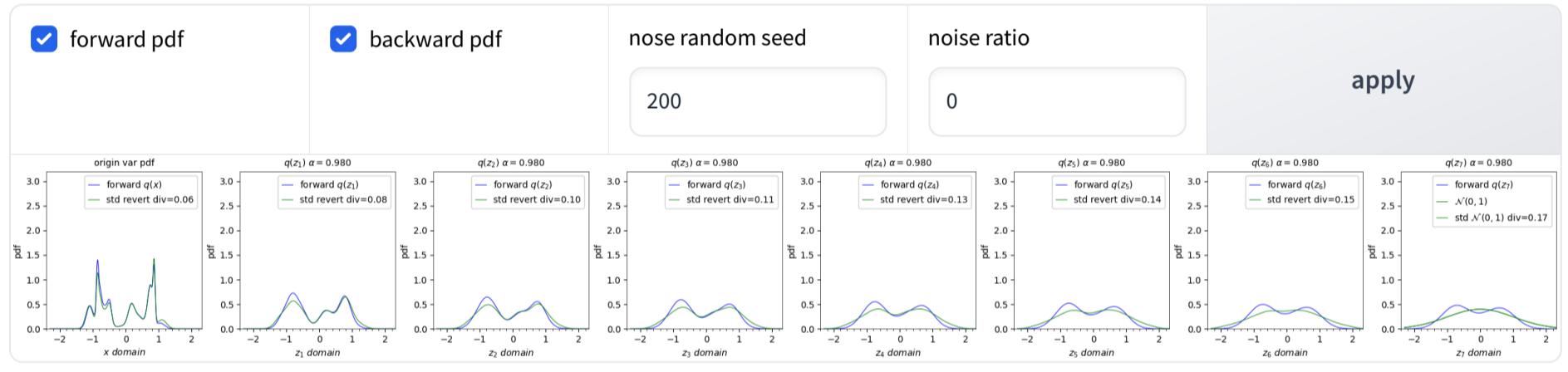
由于 $\alpha = 0.0000403$ 非常小，其对应的GaussFun(式3.4)的标准差达到157.52。如果把 $q(x)$ 的支撑集限制在单位超球范围内($\|x\|_2 < 1$)，那当 $z_T \in [-2, +2]$ 时，对应的各个 $q(x|z_T)$ 均与 $q(x)$ 非常相似。在这种状态下，对于 $q(x|z_T)$ 相应的后验概率变换，不管输入分布的形状如何，只要支撑集在 $[-2, +2]$ 范围内，其输出分布都将为 $q(x)$ 。

所以，可以总结，在DPM模型中，如果 $q(x)$ 的支撑集是有限的，并且最终变量 Z_T 的信噪比足够大，那恢复 $q(x)$ 的过程可以使用任意的分布，不必一定需要使用标准正态分布。

读者可亲自做一个类似的实验。在Demo 3.1中，将start_alpha设置0.25，end_alpha也设置为0.25，step设置为7，此时 $q(z_7) = \sqrt{0.000061}X + \sqrt{1 - 0.000061}\epsilon$ ，与DDPM的 $q(z_T)$ 基本相似。点击apply执行前向变换(蓝色曲线)，为接下来的反向恢复做准备。在Demo 3.2中，noise_ratio设置为1，为末端分布 $q(z_7)$ 引入100%的噪声，切换noise_random_seed的值可改变噪声的分布，取消选择backward_pdf，减少画面的干扰。点击apply将通过后验概率变换恢复 $q(x)$ ，将会看到，不管输入的 $q(z_7)$ 的形状如何，恢复的 $q(x)$ 均与原始的 $q(x)$ 完全相同，JS Divergence为0，恢复的过程使用红色曲线画出。

另外有一点值得注意一下，在深度学习任务中，常将输入样本的各个维度缩放在 $[-1, 1]$ 范围内，也是说在一个超立方体内(hypercube)。超立方体内任意两点的最大欧氏距离会随着维度的增多而变大，比如，对于一维，最大距离为2，对于二维，最大距离为 $2\sqrt{2}$ ，对于三维，最大距离为 $2\sqrt{3}$ ，对于n维，最大距离为 $2\sqrt{n}$ 。所以，对于维度较高的数据，需要 Z_T 变量有更高的信噪比，才能让恢复过程的起始分布接受任意的分布。

Demo 3.2 - Recover From Normal Distribution Iteratively



6. Fitting Posterior With Conditional Gaussian Model

由第3节前半部分可知，各个后验概率分布是未知的，并且与 $q(x)$ 有关。所以，为了恢复数据分布或者从数据分布中采样，需要对各个后验概率分布进行学习估计。

由第3节后半部分可知，当满足一定条件时，各个后验概率分布 $q(x|z)$ 、 $q(z_{t-1}|z_t)$ 近似于高斯概率分布，所以可通过构建一批条件高斯概率模型 $p(x|z)$, $p(z_{t-1}|z_t)$ ，学习拟合对应的 $q(x|z)$, $q(z_{t-1}|z_t)$ 。

由于模型表示能力和学习能力的局限性，拟合过程会存在一定的误差，进一步会影响恢复 $q(x)$ 的准确性。拟合误差大小与后验概率分布的形状有关。由第3节可知，当 $q(x)$ 较复杂或者所加噪声较大时，后验概率分布会较复杂，与高斯分布差别较大，从而导致拟合误差，进一步影响恢复 $q(x)$ 。

具体可看Demo 3.3，读者可测试不同复杂程度的 $q(x)$ 和 α ，观看后验概率分布 $q(z_{t-1}|z_t)$ 的拟合程度，以及恢复 $q(x)$ 的准确度。恢复的概率分布使用**橙色**标识，同时也通过JS divergence给出误差。

关于拟合的目标函数，与其它概率模型类似，可**优化交叉熵损失**，使 $p(z_{t-1}|z_t)$ 逼近于 $q(z_{t-1}|z_t)$ 。由于 $(z_{t-1}|z_t)$ 是条件概率，所以需要综合考虑各个条件，以**各个条件发生的概率** $q(z_t)$ 加权平均**各个条件对应的交叉熵**。最终的损失函数形式如下：

$$loss = - \int q(z_t) \overbrace{\int q(z_{t-1}|z_t) \log p(z_{t-1}|z_t) dz_{t-1}}^{\text{Cross Entropy}} dz_t \quad (6.1)$$

$$= - \iint q(z_{t-1}, z_t) \log p(z_{t-1}|z_t) dz_{t-1} dz_t \quad (6.2)$$

也可以KL散度作为目标函数进行优化，KL散度与交叉熵是等价的[10]。

$$loss = \int q(z_t) KL(q(z_{t-1}|z_t) \| p(z_{t-1}|z_t)) dz_t \quad (6.3)$$

$$= \int q(z_t) \int q(z_{t-1}|z_t) \log \frac{q(z_{t-1}|z_t)}{p(z_{t-1}|z_t)} dz_{t-1} dz_t \quad (6.4)$$

$$= - \underbrace{\int q(z_t) \int q(z_{t-1}|z_t) \log p(z_{t-1}|z_t) dz_{t-1} dz_t}_{\text{Cross Entropy}} + \underbrace{\int q(z_t) \int q(z_{t-1}|z_t) \log q(z_{t-1}|z_t) dz_t}_{\text{Is Constant}} \quad (6.5)$$

式6.2的积分没有闭合的形式，不能直接优化。可使用蒙特卡罗(Monte Carlo)积分近似计算，新的目标函数如下：

$$loss = - \iint q(z_{t-1}, z_t) \log p(z_{t-1}|z_t) dz_{t-1} dz_t \quad (6.6)$$

$$\approx - \sum_{i=0}^N \log p(Z_{t-1}^i | Z_t^i) \quad \text{where } (Z_{t-1}^i, Z_t^i) \sim q(z_{t-1}, z_t) \quad (6.7)$$

上述的样本(Z_{t-1}^i, Z_t^i)服从联合概率分布 $q(z_{t-1}, z_t)$ ，可通过祖先采样的方式采样得到。具体方式如下：通过正向转换的方式(式4.1~4.4)，逐步采样 $X, Z_1, Z_2 \dots Z_{t-1}, Z_t$ ，然后留下(Z_{t-1}, Z_t)作为一个样本。但这种采样方式比较慢，可利用 $q(z_t|x)$ 概率分布已知的特点(式4.8)加速采样，先从 $q(x)$ 采样 X ，然后由 $q(z_{t-1}|x)$ 采样 Z_{t-1} ，最后由 $q(z_t|z_{t-1})$ 采样 Z_t ，于是得到一个样本(Z_{t-1}, Z_t)。

可能有些人会有疑问，式6.3的形式跟DPM[1]和DDPM[2]论文里的形式不太一样。实际上，这两个目标函数是等价的，下面给出证明。

对于一致项(Consistent Term)，证明如下：

$$loss = - \iint q(z_{t-1}, z_t) \log p(z_{t-1}|z_t) dz_{t-1} dz_t \quad (6.8)$$

$$= - \iint \int q(x) q(z_{t-1}, z_t | x) dx \log p(z_{t-1}|z_t) dz_{t-1} dz_t \quad (6.9)$$

$$= \underbrace{\iint \int q(x) q(z_{t-1}, z_t | x) \log q(z_{t-1}|z_t, x) dx dz_{t-1} dz_t}_{\text{This Term Is Constant And Is Denoted As } C_1} \quad (6.10)$$

$$- \iint \int q(x) q(z_{t-1}, z_t | x) \log p(z_{t-1}|z_t) dx dz_{t-1} dz_t - C_1 \quad (6.11)$$

$$= \iint \int q(x) q(z_{t-1}, z_t | x) \log \frac{q(z_{t-1}|z_t, x)}{p(z_{t-1}|z_t)} dx dz_{t-1} dz_t - C_1 \quad (6.12)$$

$$= \iint q(x) q(z_t | x) \int q(z_{t-1}|z_t, x) \log \frac{q(z_{t-1}|z_t, x)}{p(z_{t-1}|z_t)} dz_{t-1} dz_t - C_1 \quad (6.13)$$

$$= \iint q(x) q(z_t | x) KL(q(z_{t-1}|z_t, x) \| p(z_{t-1}|z_t)) dz_t dx - C_1 \quad (6.14)$$

$$\propto \iint q(x) q(z_t | x) KL(q(z_{t-1}|z_t, x) \| p(z_{t-1}|z_t)) dz_t dx \quad (6.15)$$

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， $q(z_{t-1}|z_t)$ 也是固定概率分布，具体形式由 $q(x)$ 及系数 α 确定。

对于重构项(Reconstruction Term)，可通过类似的方式证明：

$$loss = - \int q(z_1) \overbrace{\int q(x|z_1) \log p(x|z_1) dx dz_1}^{\text{Cross Entropy}} \quad (6.16)$$

$$= - \iint q(z_1, x) \log p(x|z_1) dx dz_1 \quad (6.17)$$

$$= - \int q(x) \int q(z_1|x) \log p(x|z_1) dz_1 dx \quad (6.18)$$

因此，式6.1的目标函数与DPM的目标函数是等价的。

根据一致项证明的结论，以及交叉熵与KL散度的关系，可得出一个有趣的结论：

$$\min_p \int q(z_t) KL(q(z_{t-1}|z_t) \| p(z_{t-1}|z_t)) dz_t \iff \min_p \iint q(z_t) q(x|z_t) KL(q(z_{t-1}|z_t, x) \| p(z_{t-1}|z_t)) dx dz_t \quad (6.19)$$

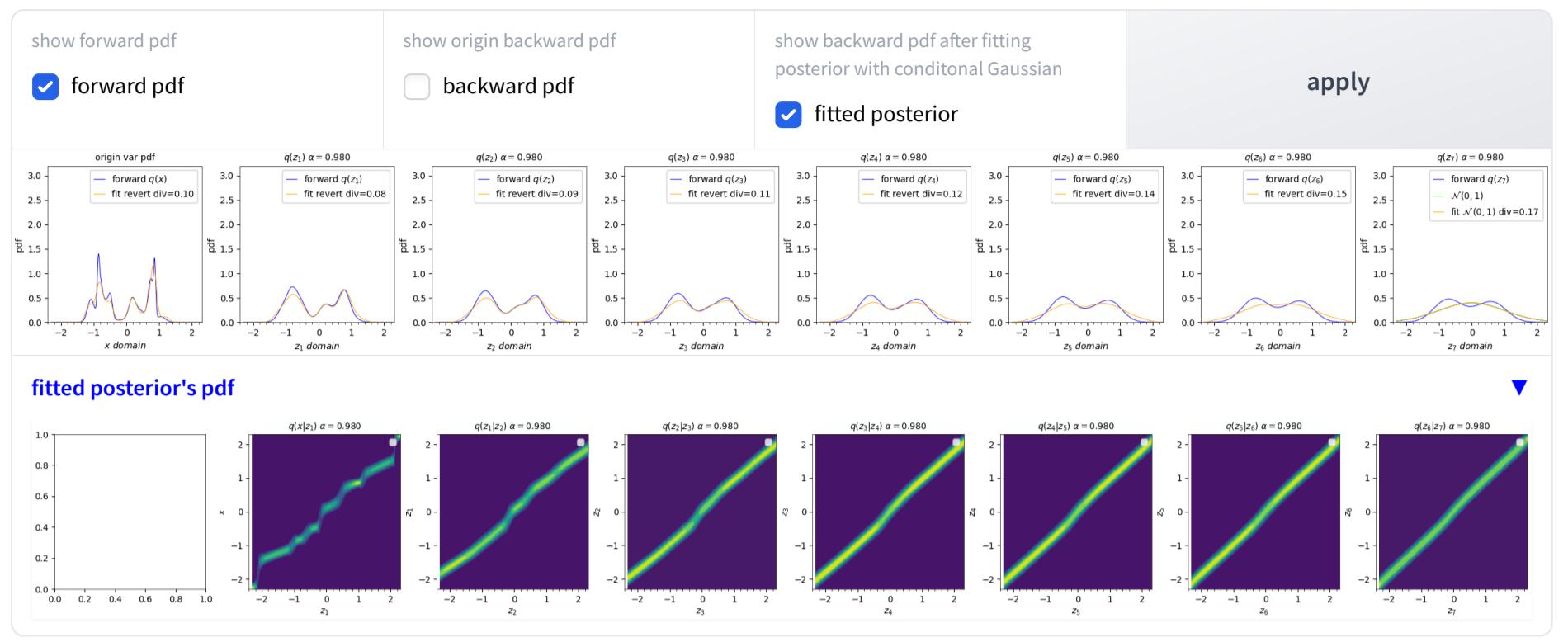
比较左右两边的式子，可以看出，右边的目标函数比左边的目标函数多了一个条件变量 X ，同时也多了一个关于 X 积分，并且以 $q(x|z_t)$ 作为积分的加权系数。

依照类似的思路，可推导出一个更通用的关系：

$$\min_p KL(q(z) \| p(z)) \iff \min_p \int q(x) KL(q(z|x) \| p(z)) dx \quad (6.20)$$

关于此结论的详细推导，可见Appendix A。

Demo 3.3 - Fitting Posterior with Conditional Gaussian Model



7. Posterior Transform

Non-expanding mapping and Stationary Distribution

$$q(x) = \int q(x, z) dz = \int q(x|z) q(z) dz \quad (7.1)$$

根据Appendix B的Corollary 1和Corollary 2可知，后验概率变换是一个non-expanding mapping。也就是说，对任意的两个概率分布 $q_{i1}(z)$ 和 $q_{i2}(z)$ ，经过后验概率变换后得到 $q_{o1}(x)$ 和 $q_{o2}(x)$ ， $q_{o1}(z)$ 和 $q_{o2}(z)$ 的距离总是小于或等于 $q_{i1}(x)$ 和 $q_{i2}(x)$ 的距离。这里的距离可使用KL Divergence或Total Variance或度量。

$$d(q_{o1}(z), q_{o2}(z)) \leq d(q_{i1}(x), q_{i2}(x)) \quad (7.2)$$

根据Appendix B的分析可知，在大多数情况，上述的等号并不会成立。并且，当 α 越小时(噪声越多)， $d(q_{o1}, q_{o2})$ 会越小于 $d(q_{i1}, q_{i2})$ 。

读者可查看Demo 4.1，左侧三个图呈现一个变换的过程，左1图是任意的数据分布 $q(x)$ ，左3图是变换后的概率分布，左2图是后验概率分布。可更改随机种子生成新的数据分布，调整 α 值引入不同程度的噪声。左侧最后两个图展示变换的“压缩性质”，左4图展示随机生成的两个输入分布，同时给出其距离度量值 div_{in} ；左5图展示经过变换后的两个输出分布，输出分布之间的距离标识为 div_{out} 。读者可改变输入的随机种子，切换不同的输入。可在图中看到，对于任意的输入， div_{in} 总是小于 div_{out} 。另外，也可改变 α 的值，将会看到， α 越小(噪声越大)， $\frac{div_{out}}{div_{in}}$ 的比值也越小，即收缩率越大。

根据Appendix C的分析可知：后验概率变换可视为markov chain的一步跳转，并且，当 $q(x)$ 和 α 满足一些条件时，此markov chain会收敛于唯一的稳态分布。另外，通过大量实验发现，稳态分布与数据分布 $q(x)$ 非常相似，当 α 越小时，稳态分布与 $q(x)$ 越相似。特别地，根据第5节的结论，当 $\alpha \rightarrow 0$ 时，经过一步变换后，输出分布即是 $q(x)$ ，所以稳态分布必定是 $q(x)$ 。

读者可看Demo 4.2，此部分展示迭代收敛的例子。选择合适的迭代次数，点中“apply iteration transform”，将逐步画出迭代的过程，每个子图均会展示各自变换后的输出分布(绿色曲线)，收敛的参考点分布 $q(x)$ 以蓝色曲线画出，同时给出输出分布与 $q(x)$ 之间的距离 $dist$ 。可以看出，随着迭代的次数增加，输出分布与 $q(x)$ 越来越相似，并最终会稳定在 $q(x)$ 附近。对于较复杂的分布，可能需要较多迭代的次数或者较大的噪声。迭代次数可以设置为上万步，但会花费较长时间。

对于一维离散的情况， $q(x|z)$ 将离散成一个矩阵(记为 $Q_{x|z}$)， $q(z)$ 离散成一个向量(记为 q_i)，积分操作 $\int q(x|z)q(z)dz$ 将离散成“矩阵-向量”乘法操作，所以后验概率变换可写成

$$q_o = Q_{x|z} q_i \quad 1 \text{ iteration} \quad (7.3)$$

$$q_o = Q_{x|z} Q_{x|z} q_i \quad 2 \text{ iteration} \quad (7.4)$$

$$\dots \quad (7.5)$$

$$q_o = (Q_{x|z})^n q_i \quad n \text{ iteration}$$

于是，为了更深入地理解变换的特点，Demo 4.2也画出矩阵 $(Q_{x|z})^n$ 的结果。从图里可以看到，当迭代趋向收敛时，矩阵 $(Q_{x|z})^n$ 的行向量将变成一个常数向量，即向量的各分量都相等。在二维密度图里将表现为一条横线。

对于一维离散的markov chain，收敛速度与转移概率矩阵的第二大特征值的绝对值($|\lambda_2|$)反相关， $|\lambda_2|$ 越小，收敛速度越快。经过大量的实验发现， α 与 $|\lambda_2|$ 有着明确的线性关系， α 越小， $|\lambda_2|$ 也越小。所以， α 越小(噪声越大)，收敛速度越快。特别地，当

$\alpha \rightarrow 0$ 时, 由第3节的结论可知, 各个 z 对应的后验概率分布趋向一致, 而由文献[21]的Theorem 21可知, $|\lambda_2|$ 小于任意两个 z 对应的后验概率分布的L1距离, 所以, 可知 $|\lambda_2| \rightarrow 0$ 。

Anti-noise Capacity In Restoring Data Distribution

由上面的分析可知, 在大多数情况下, "后验概率变换"是一个收缩映射, 所以存在如下的关系:

$$d(q(x), q_o(x)) < d(q(z), q_i(z)) \quad (7.12)$$

其中, $q(z)$ 是理想的输入分布, $q(x)$ 理想的输出分布, $q(x) = \int q(x|z)q(z)dz$, $q_i(z)$ 是任意的输入分布, $q_o(x)$ 是变换后的输出分布, $q_o(x) = \int q(x|z)q_i(z)dz$ 。

上式表明, 输出的分布 $q_o(x)$ 与理想输出分布 $q(x)$ 之间的距离总会**小于**输入分布 $q_i(z)$ 与理想输入分布 $q(x)$ 的距离。所以, "后验概率变换"天然具备一定的抵抗噪声能力。这意味着, 在恢复 $q(x)$ 的过程中(第5节), 哪怕输入的"末尾分布 $q(z_T)$ "存在一定的误差, 经过一系列变换后, 输出的"数据分布 $q(x)$ "的误差也会比输入的误差更小。

具体可看Demo 3.2, 通过增加"noise ratio"的值可以向"末尾分布 $q(z_T)$ "添加噪声, 点击"apply"按钮将逐步画出恢复的过程, 恢复的分布以**红色曲线**画出, 同时也会通过JS散度标出误差的大小。将会看到, 恢复的 $q(x)$ 的误差总是小于 $q(z_T)$ 的误差。

由上面的讨论可知, α 越小(即变换过程中使用的噪声越大), 收缩映射的收缩率越大, 相应地, 抗噪声的能力也越强。特别地, 当 $\alpha \rightarrow 0$ 时, 抗噪声能力无限大, 不论多大噪声的输入, 输出都为 $q(x)$ 。

Markov Chain Monte Carlo Sampling

在DPM模型中, 通常是通过Ancestral Sampling的方式进行采样。由上面的分析可知, 当 α 足够小时, 后验概率变换会收敛于 $q(x)$, 所以, 可通过Markov Chain Monte Carlo的方式进行采样。如图7.1所示。图中 α 代表一个较大的噪声的后验概率变换, 较大的噪声使稳态分布更接近于数据分布 $q(x)$, 但由第3节可知, 较大噪声的后验变换不利于拟合, 所以把较大噪声的后验概率变换分成多个小噪声的后验概率变换。

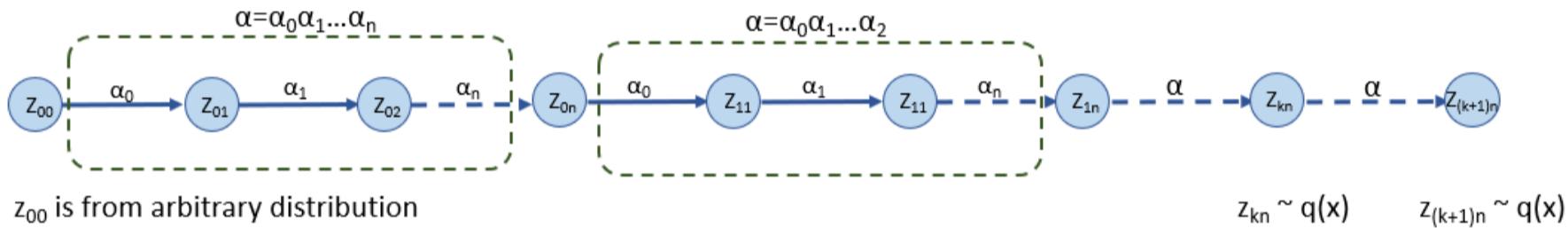
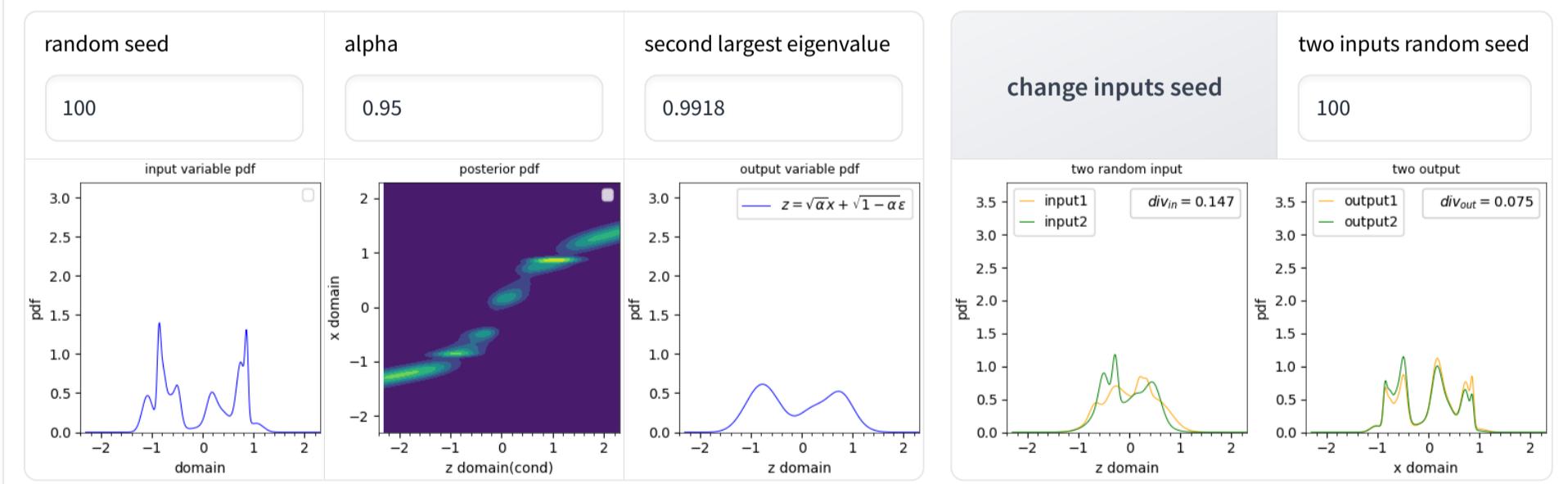
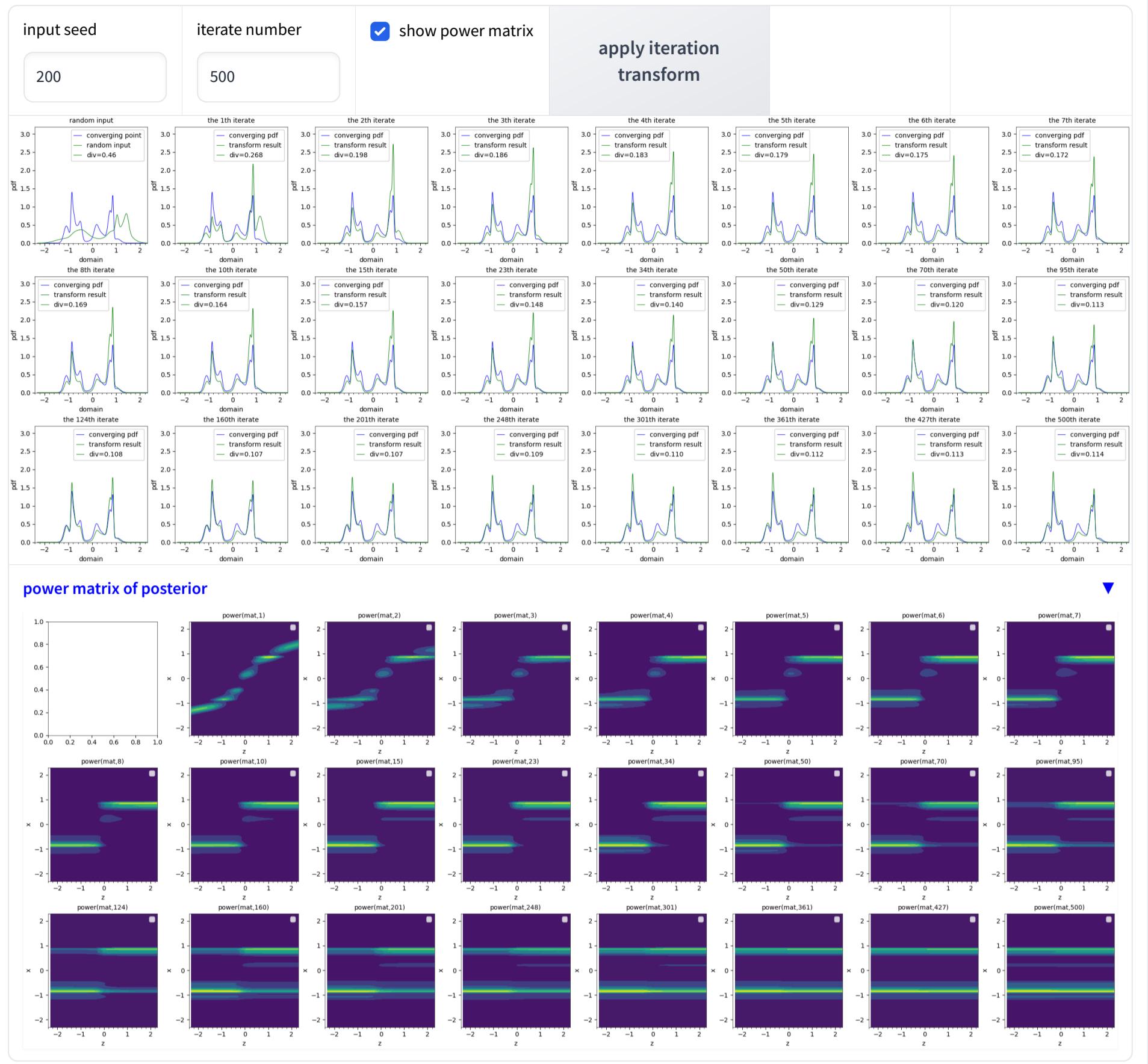


Figure 7.1: Markov Chain Monte Carlo Sampling

Demo 4.1 - Posterior Transform is a Contraction Mapping



Demo 4.2 - Posterior Transform Have a Converging Point



8. Can the data distribution be restored by deconvolution?

在第1节中提到，式2.1的变换可分为两个子变换，第一个子变换为“线性变换”，第二个为“加上独立高斯噪声”。线性变换相当于对概率分布进行拉伸变换，所以存在逆变换。“加上独立高斯噪声”相当于对概率分布执行卷积操作，卷积操作可通过逆卷积恢复。所以，理论上，可通过“逆线性变换”和“逆卷积”从最终的概率分布 $q(z_T)$ 恢复数据分布 $q(x)$ 。

但实际上，会存在一些问题。由于逆卷积对误差极为敏感，具有很高的输入灵敏度，很小的输入噪声就会引起输出极大的变化[11][12]。而在扩散模型中，会使用标准正态分布近似代替 $q(z_T)$ ，因此，在恢复的起始阶段就会引入噪声。虽然噪声较小，但由于逆卷积的敏感性，噪声会逐步放大，影响恢复。

另外，也可以从另一个角度理解“逆卷积恢复”的不可行性。由于前向变换的过程(式4.1~4.4)是确定的，所以卷积核是固定的，因此，相应的“逆卷积变换”也是固定的。由于起始的数据分布 $q(x)$ 可以是任意的分布，所以，通过一系列固定的“卷积正变换”，可以将任意的概率分布转换成近似 $\mathcal{N}(0, I)$ 的分布。如“逆卷积变换”可行，则意味着，可用一个固定的“逆卷积变换”，将 $\mathcal{N}(0, I)$ 分布恢复成任意的数据分布 $q(x)$ ，这明显是一个悖论。同一个输入，同一个变换，不可能会有多个输出。

Appendix A Conditional KL Divergence

本节主要介绍KL散度与条件KL散度之间的关系。在正式介绍之前，先简单介绍熵和条件熵的定义，以及两者之间存在的不等式关系，为后面的证明作准备。

熵及条件熵

对于任意两个随机变量 Z, X , 熵(Entropy)定义如下[16]:

$$H(Z) = \int -p(z) \log p(z) dz \quad (\text{A.1})$$

条件熵(Conditional Entropy)的定义如下[17]:

$$H(Z|X) = \int p(x) \overbrace{\int -p(z|x) \log p(z|x) dz}^{\text{Entropy}} dx \quad (\text{A.2})$$

两者存在如下的不等式关系:

$$H(Z|X) \leq H(Z) \quad (\text{A.3})$$

也就是说, 条件熵总是小于或者等于熵, 当且仅当 X 与 Z 相互独立时, 两者相等。此关系的证明可看文献[17]。

KL散度及条件KL散度

仿照条件熵定义的方式, 引入一个新定义, 条件KL散度, 记为 KL_C 。由于KL散度的定义是非对称的, 所以存在两种形式, 如下:

$$KL_C(q(z|x)\|\mathbf{p}(z)) = \int q(x)KL(q(z|x)\|\mathbf{p}(z))dx \quad (\text{A.4})$$

$$KL_C(q(z)\|\mathbf{p}(z|x)) = \int \mathbf{p}(x)KL(q(z)\|\mathbf{p}(z|x))dx \quad (\text{A.5})$$

与条件熵类似, 两种形式的条件KL散度也都存在类似的不等式关系:

$$KL_C(q(z|x)\|\mathbf{p}(z)) \geq KL(q(z)\|\mathbf{p}(z)) \quad (\text{A.6})$$

$$KL_C(q(z)\|\mathbf{p}(z|x)) \geq KL(q(z)\|\mathbf{p}(z)) \quad (\text{A.7})$$

也就是说, 条件KL散度总是大于或者等于KL散度, 当且仅当 X 与 Z 相互独立时, 两者相等。

下面对式A.5和式A.6的结论分别证明。

对于式A.6, 证明如下:

$$KL_C(q(z|x)\|\mathbf{p}(z)) = \int q(x)KL(q(z|x)\|\mathbf{p}(z))dx \quad (\text{A.8})$$

$$= \iint q(x)q(z|x) \log \frac{q(z|x)}{\mathbf{p}(z)} dz dx \quad (\text{A.9})$$

$$= -\overbrace{\iint -q(x)q(z|x) \log q(z|x) dz dx}^{\text{Conditional Entropy } H_q(Z|X)} - \iint q(x)q(z|x) \log \mathbf{p}(z) dz dx \quad (\text{A.10})$$

$$= -H_q(Z|X) - \int \left\{ \int q(x)q(z|x) dx \right\} \log \mathbf{p}(z) dz \quad (\text{A.11})$$

$$= -H_q(Z|X) + \overbrace{\int -q(z) \log p(z) dz}^{\text{Cross Entropy}} \quad (\text{A.12})$$

$$= -H_q(Z|X) + \int q(z) \left\{ \log \frac{q(z)}{\mathbf{p}(z)} - \log q(z) \right\} dz \quad (\text{A.13})$$

$$= -H_q(Z|X) + \int q(z) \log \frac{q(z)}{\mathbf{p}(z)} dz + \overbrace{\int -q(z) \log q(z) dz}^{\text{Entropy } H_q(Z)} \quad (\text{A.14})$$

$$= KL(q(z)\|\mathbf{p}(z)) + \overbrace{H_q(Z) - H_q(Z|X)}^{\geq 0} \quad (\text{A.15})$$

$$\leq KL(q(z)\|\mathbf{p}(z)) \quad (\text{A.16})$$

其中式A.15应用了"条件熵总是小于或者等于熵"的结论。于是, 得到式A.6的关系。

对于式A.7，证明如下：

$$KL(q(z) \| p(z)) = \int q(z) \log \frac{q(z)}{p(z)} dz \quad (\text{A.15})$$

$$= \int q(z) \log \frac{q(z)}{\int p(z|x)p(x)dx} dz \quad (\text{A.16})$$

$$= \int p(x) dx \int q(z) \log q(z) dz - \int q(z) \log \int p(z|x)p(x) dx dz \quad \int p(x) dx = 1 \quad (\text{A.17})$$

$$\leq \iint p(x) q(z) \log q(z) dz dx - \int q(z) \int p(x) \log p(z|x) dx dz \quad \text{jensen inequality} \quad (\text{A.18})$$

$$= \iint p(x) q(z) \log q(z) dz dx - \iint p(x) q(z) \log p(z|x) dz dx \quad (\text{A.19})$$

$$= \iint p(x) q(z) (\log q(z) - \log p(z|x)) dz dx \quad (\text{A.20})$$

$$= \iint p(x) q(z) \log \frac{q(z)}{p(z|x)} dz dx \quad (\text{A.21})$$

$$= \int p(x) \left\{ \int q(z) \log \frac{q(z)}{p(z|x)} dz \right\} dx \quad (\text{A.22})$$

$$= \int p(x) KL(q(z) \| p(z|x)) dx \quad (\text{A.23})$$

$$= KL_C(q(z) \| p(z|x)) \quad (\text{A.24})$$

于是，得到式A.7的关系。

从式A.15可得出另外一个重要的结论。

KL散度常用于拟合数据的分布。在此场景中，数据潜在的分布用 $q(z)$ 表示，参数化的模型分布用 $p_{\theta}(z)$ 表示。在优化的过程中，由于 $q(z|x)$ 和 $q(x)$ 均保持不变，所以式A.15中的 $H(Z) - H(Z|X)$ 为一个常数项。于是，可得到如下的关系

$$\min_{p_{\theta}} KL(q(z) \| p_{\theta}(z)) \iff \min_{p_{\theta}} \int q(x) KL(q(z|x) \| p_{\theta}(z)) dx \quad (\text{A.25})$$

把上述的关系与Denoised Score Matching[18]作比较，可发现一些相似的地方。两者均引入一个新变量 X ，并且将拟合的目标分布 $q(z)$ 代替为 $q(z|x)$ 。代替后，由于 $q(z|x)$ 是条件概率分布，所以，两者均考虑了所有的条件，并以条件发生的概率 $q(x)$ 作为权重系数执行加权和。

$$\min_{p_{\theta}} \frac{1}{2} \int q(z) \left\| \psi_{\theta}(z) - \frac{\partial q(z)}{\partial z} \right\|^2 dz \iff \min_{p_{\theta}} \int q(x) \overbrace{\frac{1}{2} \int q(z|x) \left\| \psi_{\theta}(z) - \frac{\partial q(z|x)}{\partial z} \right\|^2 dz}^{\text{Score Matching of } q(z|x)} dx \quad (\text{A.26})$$

上述加权和的操作有点类似于"全概率公式消元"。

$$q(z) = \int q(z, x) dx = \int q(x) q(z|x) dx \quad (\text{A.27})$$

Appendix B When does the Posterior Approximate to Gaussian ?

由式3.4可知， $q(x|z)$ 有如下的形式

$$q(x|z) = \text{Normalize} \left(\frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(x-\mu)^2}{2\sigma^2} q(x) \right) \quad \text{where } \mu = \frac{z}{\sqrt{\alpha}} \quad \sigma = \sqrt{\frac{1-\alpha}{\alpha}} \quad (\text{B.1})$$

$$\propto \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(x-\mu)^2}{2\sigma^2}}_{\text{GaussFun}} q(x) \quad (\text{B.2})$$

下面证明，如果满足如下两个假设， $q(x|z)$ 近似于高斯分布。

- 假设在GaussFun的支撑集内， $q(x)$ 是线性变化的。以GaussFun的均值为中心，对 $q(x)$ 进行泰勒展开。由泰勒展开的性质可知，当GaussFun的标准差 σ 足够小时，上述假设可以满足。

$$q(x) \approx q(\mu) + \nabla_x q(\mu)(x - \mu) \quad \text{where } q(\mu) \triangleq q(x) \Big|_{x=\mu} \quad \nabla_x q(\mu) \triangleq \nabla_x q(x) \Big|_{x=\mu} \quad (\text{B.3})$$

$$= q(\mu) \left(1 + \frac{\nabla_x q(\mu)}{q(\mu)} (x - \mu) \right) \quad (\text{B.4})$$

$$= q(\mu) \left(1 + \nabla_x \log q(\mu)(x - \mu) \right) \quad \text{where } \nabla_x \log q(\mu) \triangleq \nabla_x \log q(x) \Big|_{x=\mu} \quad (\text{B.5})$$

- 假设在GaussFun的支撑集内， $\log(1 + \nabla_x \log q(\mu)(x - \mu))$ 可近似为 $\nabla_x \log q(\mu)(x - \mu)$ 。对 $\log(1 + y)$ 进行泰勒展开，由泰勒展开的性质可知，当 $\|y\|_2$ 较小时， $\log(1 + y)$ 可近似为 y 。当 σ 足够小时， $\|x - \mu\|_2$ 将较小， $\nabla_x \log q(\mu)(x - \mu)$ 也将较小，

所以上述假设可以满足。一般情况下，当 $\nabla_x \log q(\mu)(x - \mu) < 0.1$ 时，近似的误差较小，可忽略。

$$\log(1 + y) \approx \log(1 + y) \Big|_{y=0} + \nabla_y \log(1 + y) \Big|_{y=0} (y - 0) \quad (\text{B.6})$$

$$= y \quad (\text{B.7})$$

利用上面的两个假设，可对 $q(x|z)$ 进行如下的推导：

$$q(x|z) \propto \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(x - \mu)^2}{2\sigma^2} q(x) \quad (\text{B.8})$$

$$\approx \frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(x - \mu)^2}{2\sigma^2} q(\mu) (1 + \nabla_x \log q(\mu)(x - \mu)) \quad (\text{B.9})$$

$$= \frac{q(\mu)}{\sqrt{2\pi}\sigma} \exp \left(\frac{-(x - \mu)^2}{2\sigma^2} + \log (1 + \nabla_x \log q(\mu)(x - \mu)) \right) \quad (\text{B.10})$$

$$\approx \frac{q(\mu)}{\sqrt{2\pi}\sigma} \exp \left(\frac{-(x - \mu)^2}{2\sigma^2} + \nabla_x \log q(\mu)(x - \mu) \right) \quad (\text{B.11})$$

$$= \frac{q(\mu)}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(x - \mu)^2 - 2\sigma^2 \nabla_x \log q(\mu)(x - \mu)}{2\sigma^2} \right) \quad (\text{B.12})$$

$$= \frac{q(\mu)}{\sqrt{2\pi}\sigma} \exp \left(-\frac{(x - \mu - \sigma^2 \nabla_x \log q(\mu))^2}{2\sigma^2} + \frac{(\sigma^2 \nabla_x \log q(\mu))^2}{2\sigma^2} \right) \quad (\text{B.13})$$

$$= \exp \left(-\frac{(x - \mu - \sigma^2 \nabla_x \log q(\mu))^2}{2\sigma^2} \right) \underbrace{\frac{q(\mu)}{\sqrt{2\pi}\sigma} \exp \left(\frac{1}{2} (\sigma^2 \nabla_x \log q(\mu))^2 \right)}_{\text{const}} \quad (\text{B.14})$$

其中，式B.9应用了假设1的结论，式B.11应用了假设2的结论。

式B.14中的const项是常数项，不会影响函数的形状。另外，由上面可知， $q(x|z)$ 具有自归一化的功能，所以， $q(x|z)$ 是一个高斯概率密度函数，均值为 $\mu + \sigma^2 \nabla_x \log q(\mu)$ ，方差为 σ^2 。

Appendix C Posterior Transform is a Non-expanding Mapping

Corollary 1

以KL Divergence为度量，markov chain的转移变换是non-expanding的[23]，即

$$KL(p(x), q(x)) \leq KL(p(z), q(z)) \quad (\text{C.1})$$

其中， $p(z)$ 和 $q(z)$ 是任意的概率密度函数， $r(x|z)$ 是markov chain的转移概率密度函数， $p(x) = \int r(x|z)p(z)dz$ ， $q(x) = \int r(x|z)q(z)dz$ 。

证明：

对于 $p(x, z)$ 和 $q(x, z)$ 的KL divergence，存在如下的关系：

$$KL(p(x, z), q(x, z)) = \iint p(x, z) \log \frac{p(x, z)}{q(x, z)} dx dz \quad (\text{C.2})$$

$$= \iint p(x, z) \log \frac{p(x)p(x|z)}{q(x)q(x|z)} dx dz \quad (\text{C.3})$$

$$= \iint p(x, z) \log \frac{p(x)}{q(x)} dx dz + \iint p(x, z) \log \frac{p(x|z)}{q(x|z)} dx dz \quad (\text{C.4})$$

$$= \iint p(x, z) dz \log \frac{p(x)}{q(x)} dx + \int p(z) \int p(x|z) \log \frac{p(x|z)}{q(x|z)} dx dz \quad (\text{C.5})$$

$$= KL(p(x), q(x)) + \int p(z) KL(p(x|z), q(x|z)) dz \quad (\text{C.6})$$

类似地，调换Z和X的顺序，可得到下面的关系：

$$KL(p(x, z), q(x, z)) = KL(p(z), q(z)) + \int p(x) KL(p(z|x), q(z|x)) dx \quad (\text{C.7})$$

比较两个关系式，可得：

$$KL(p(z), q(z)) + \int p(x) KL(p(z|x), q(z|x)) dx = KL(p(x), q(x)) + \int p(z) KL(p(x|z), q(x|z)) dz \quad (\text{C.8})$$

由于 $q(x|z)$ 和 $p(x|z)$ 都是markov chain的转移概率密度，均等于 $r(x|z)$ ，所以 $\int p(z) KL(p(x|z), q(x|z)) dz$ 等于0。于是，上式简化为：

$$KL(p(x), q(x)) = KL(p(z), q(z)) - \int p(x) KL(p(z|x), q(z|x)) dx \quad (\text{C.9})$$

由于KL divergence总是大于或者等于0，所以，加权和 $\int p(x) KL(p(z|x), q(z|x)) dx$ 也是大于等于0。于是，可得：

$$KL(p(x), q(x)) \leq KL(p(z), q(z)) \quad (\text{C.10})$$

上式等号成立的条件是 $\int p(x)KL(p(z|x), q(z|x))dx$ 等于0，这要求对不同的条件 x ， $p(z|x)$ 与 $q(z|x)$ 均要相等。在大多数情况下，当 $p(z)$ 和 $q(z)$ 不同时， $p(z|x)$ 也和 $q(z|x)$ 不同。这意味着，在大多数情况下，有

$$KL(p(x), q(x)) < KL(p(z), q(z)) \quad (\text{C.11})$$

Corollary 2

以Total Variance(L1 distance)为度量，markov chain的转移变换是non-expanding，即

$$\|p(x) - q(x)\|_1 \leq \|p(z) - q(z)\|_1 \quad (\text{C.12})$$

其中， $p(z)$ 和 $q(z)$ 是任意的概率密度函数， $r(x|z)$ 是markov chain的转移概率密度函数， $p(x) = \int r(x|z)p(z)dz$ ， $q(x) = \int r(x|z)q(z)dz$ 。

证明：

$$\|p(x) - q(x)\|_1 = \int |p(x) - q(x)| dx \quad (\text{C.13})$$

$$= \int \left| \int r(x|z)p(z)dz - \int r(x|z)q(z)dz \right| dx \quad (\text{C.14})$$

$$= \int \left| \int r(x|z)(p(z) - q(z))dz \right| dx \quad (\text{C.15})$$

$$\leq \iint r(x|z) |(p(z) - q(z))| dz dx \quad (\text{C.16})$$

$$= \iint r(x|z) dx |(p(z) - q(z))| dz \quad (\text{C.17})$$

$$= \int |(p(z) - q(z))| dz \quad (\text{C.18})$$

$$= \|p(z) - q(z)\|_1 \quad (\text{C.19})$$

其中，式C.16应用了绝对值不等式，式C.18利用了 $r(x|z)$ 是概率分布的性质。

证明完毕。

图C.1展示了一个一维随机变量的例子，可以更直观地理解上述推导的过程。

上述等式的成立的条件是：各个绝对值括号内的非零项均是同样的符号。如图C.1(a)，包含5个绝对值括号，每个对应一行，每个括号内有5项，当且仅当每行各个非零项同号时，上述的等式才成立。如果出现不同号的情况，则会导致 $\|p(x) - q(x)\|_1 < \|p(z) - q(z)\|_1$ 。不同号出现的数量与转移概率矩阵的非零元素有关，一般情况下，非零元素越多，不同号出现的数量会越多。

在后验概率变换中，一般情况下，当 α 越小(噪声越多)时，转移概率密度函数会有越多的非零元素，如图C.2(a)所示；当 α 越大(噪声越小时)，转移概率密度函数会有越少的非零元素，如图C.2(b)所示。

所以，有这么一个规律：当 α 越小时，则会导致 $\|p(x) - q(x)\|_1$ 越小于 $\|p(z) - q(z)\|_1$ ，也就是说，这个变换的压缩率越大。

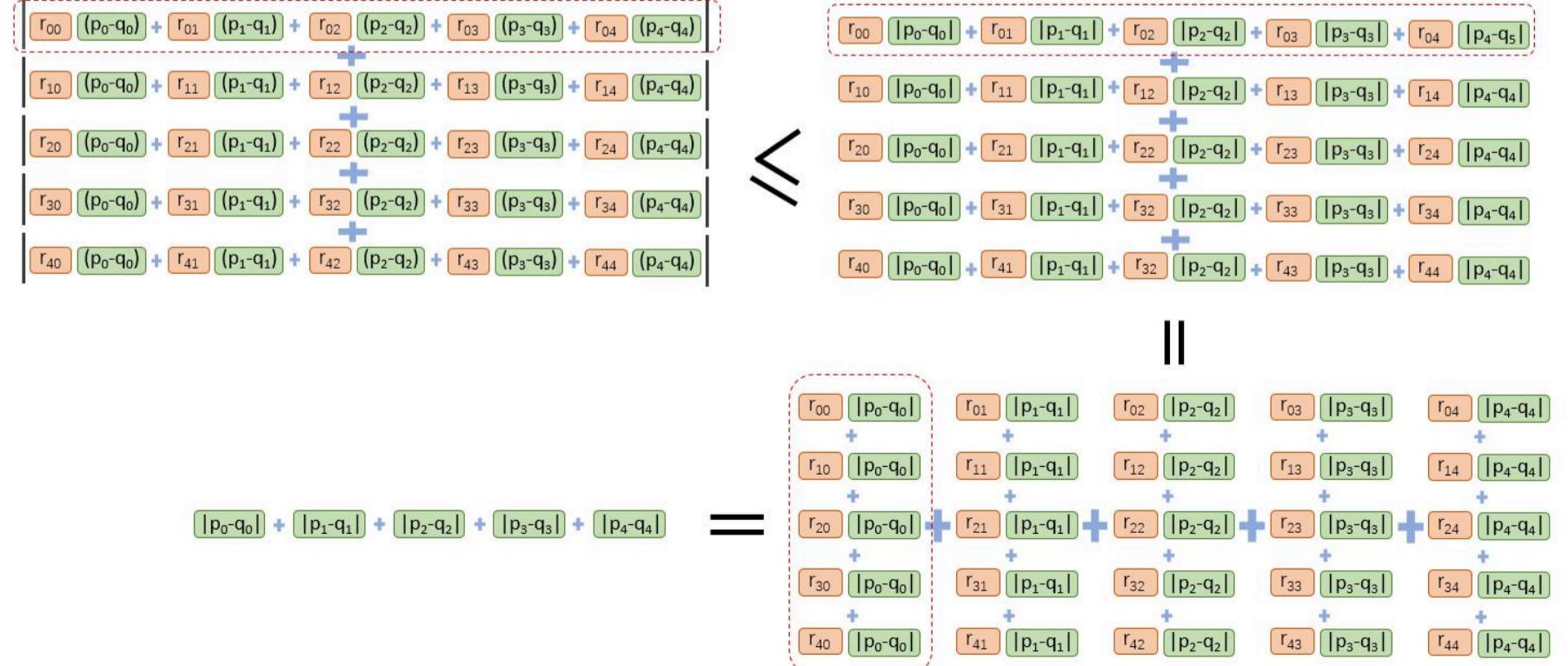


Figure C.1: Non-expanding under L1 norm

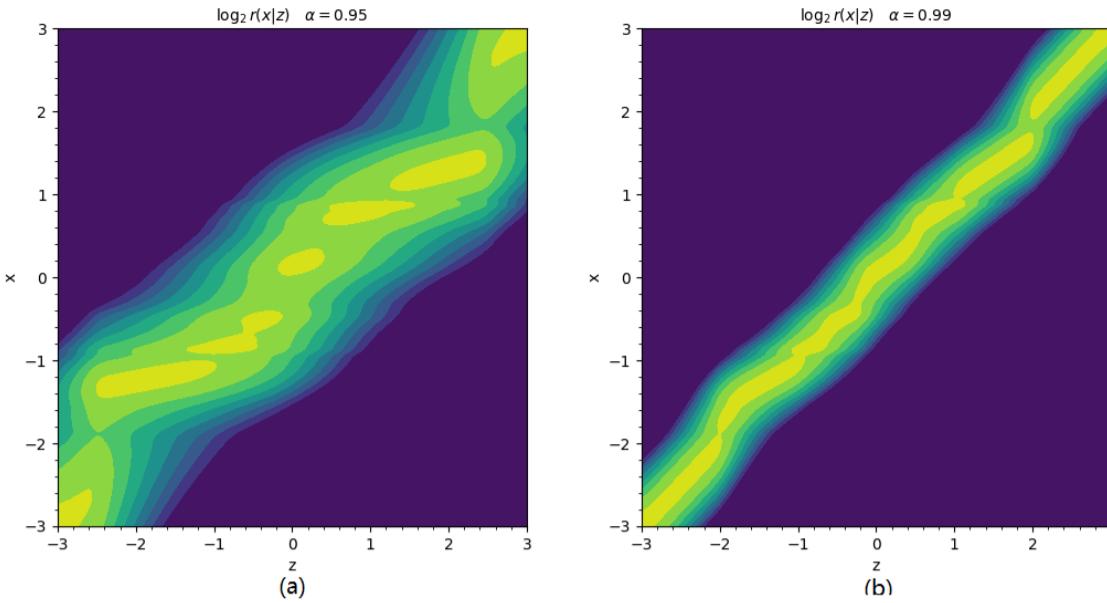


Figure C.2: More non-zero elements as α gets smaller

Appendix D Posterior Transform Converges to the Unique Stationary Distribution



根据文献[\[19\]](#)Theorem 3的结论，**非周期(aperiodic)****不可约(irreducible)**的markov chain会收敛于惟一的**稳态分布**。

下面将表明，当满足一定的条件时，后验概率变换是一个非周期不可约的markov chain的转移概率密度函数。

为了表述方便，下面以一个更通用的形式来描述扩散模型的前向变换。

$$Z = \sqrt{\alpha}X + \sqrt{\beta}\epsilon \quad (\text{D.1})$$

由第1节可知， $\sqrt{\alpha}X$ 会对 X 的概率密度函数执行缩放，所以 α 控制着缩放的强度， β 控制着添加噪声的大小。当 $\beta = 1 - \alpha$ 时，上述的变换与式1.1一致。

新变换对应的后验概率分布的形式如下：

$$q(x|z=c) = \text{Normalize} \left(\overbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp \frac{-(x-\mu)^2}{2\sigma^2}}^{\text{GaussFun}} q(x) \right) \quad (\text{D.2})$$

where $\mu = \frac{c}{\sqrt{\alpha}}$ $\sigma = \sqrt{\frac{\beta}{\alpha}}$ c is a fixed value

当 $\beta = 1 - \alpha$ 时，上述的变换与式3.4一致。

为了表述方便，下面以 $g(x)$ 表示式D.2中GaussFun。

由于 $\sqrt{\alpha}X$ 会缩放 X 的概率密度函数 $q(x)$ ，这会使分析转移概率密度函数 $q(x|z)$ 的非周期性和不可约性变得更复杂。所以，为了分析方便，先假设 $\alpha = 1$ ，后面再分析 $\alpha \neq 1$ 且 $\beta = 1 - \alpha$ 的情况。

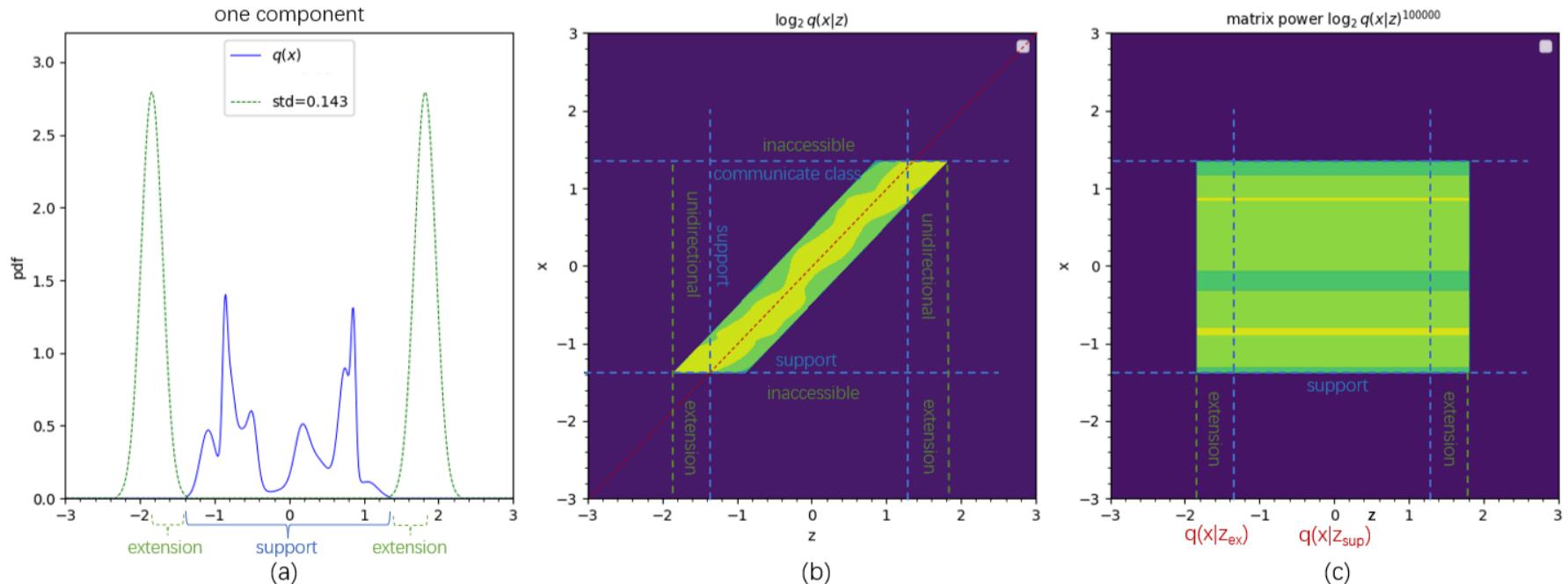


Figure D.1: Only one component in support

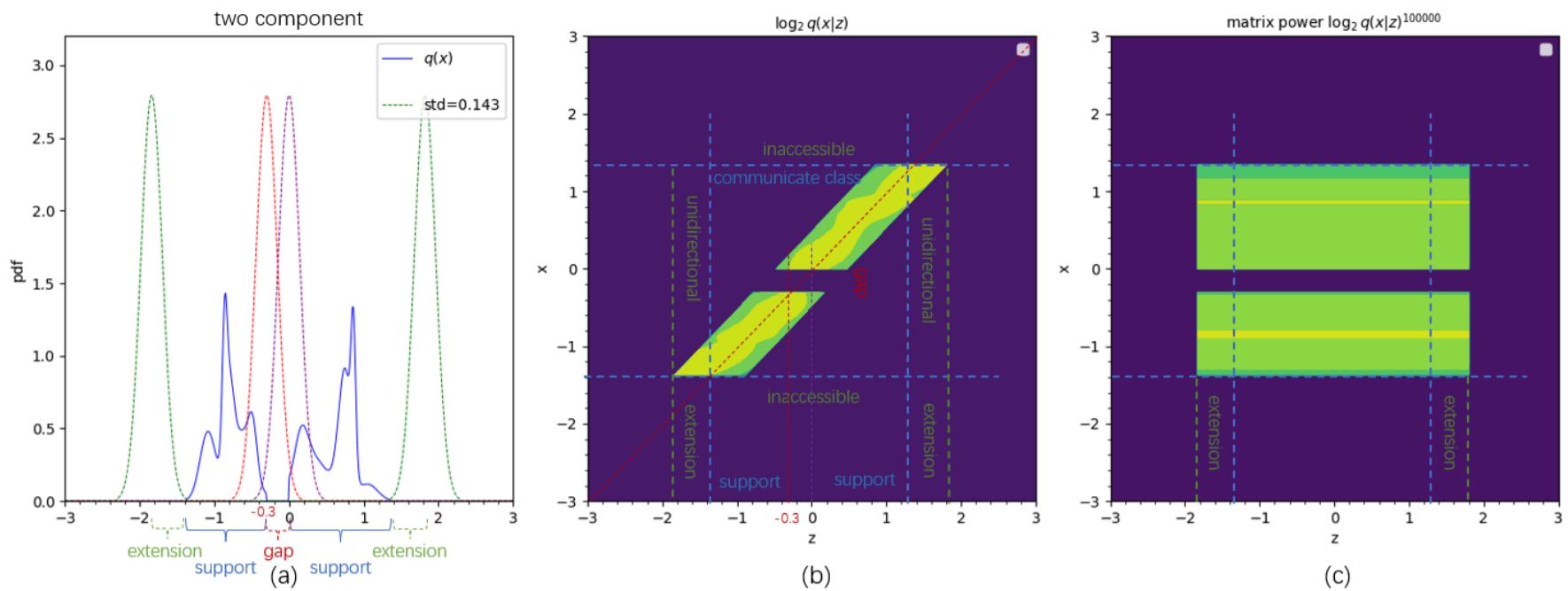


Figure D.2: One component which can communicate with each other

$$\alpha = 1$$

当 $\alpha = 1$ 时，如果 $q(x)$ 和 β 满足下面两个条件之一，则 $q(x|z)$ 对应的markov chain是非周期且不可约的。

1. 如果 $q(x)$ 的支撑集只存在一个connected component。
2. 如果 $q(x)$ 的支撑集存在多个connected component，但各个connected component之间的距离小于 3σ 。也就是说，间隙能被 $g(x)$ 的有效区域的半径所覆盖。

证明如下：

1. 对 $q(x)$ 支撑集内的任意点 c ，当 $z = c$ 和 $x = c$ 时， $q(x = c) > 0$ ；由式D.2可知， $g(x)$ 的中心位于 c ，所以 $g(x)$ 在 $x = c$ 处也大于0。于是，根据式D.2中相乘的关系可知， $q(x = c|z = c) > 0$ 。因此， $q(x|z)$ 对应的markov chain是非周期的。

对 $q(x)$ 支撑集内的任意点 c ，当 $z = c$ 时， $g(x)$ 的中心位于 c ，所以存在一个以 c 为中心的超球($\|x - c\|_2 < \delta$)，在此超球内， $q(x|z = c) > 0$ ，也就是说，状态 c 可以访问(access)附近的其它状态。由于支撑集内每个状态都具有此性质，所以，整个支撑集内的状态构成一个Communicate Class[14]。因此， $q(x|z)$ 对应的markov chain是不可约的。

所以，满足条件1的markov chain是非周期和不可约的。可看图D.1的例子，其展示了单个connected component的例子。

2. 当 $q(x)$ 支撑集存在多个connected component时，markov chain可能存在多个communicate class。但当各间隙小于 $g(x)$ 的3倍标准差时，那各个component的状态将可互相访问(access)，因此， $q(x|z)$ 对应的markov chain也只存在一个communicate class，与条件1的情况相同。所以，满足条件2的markov chain是非周期和不可约的。

可看图d2的例子，其展示了多个connected component的例子。

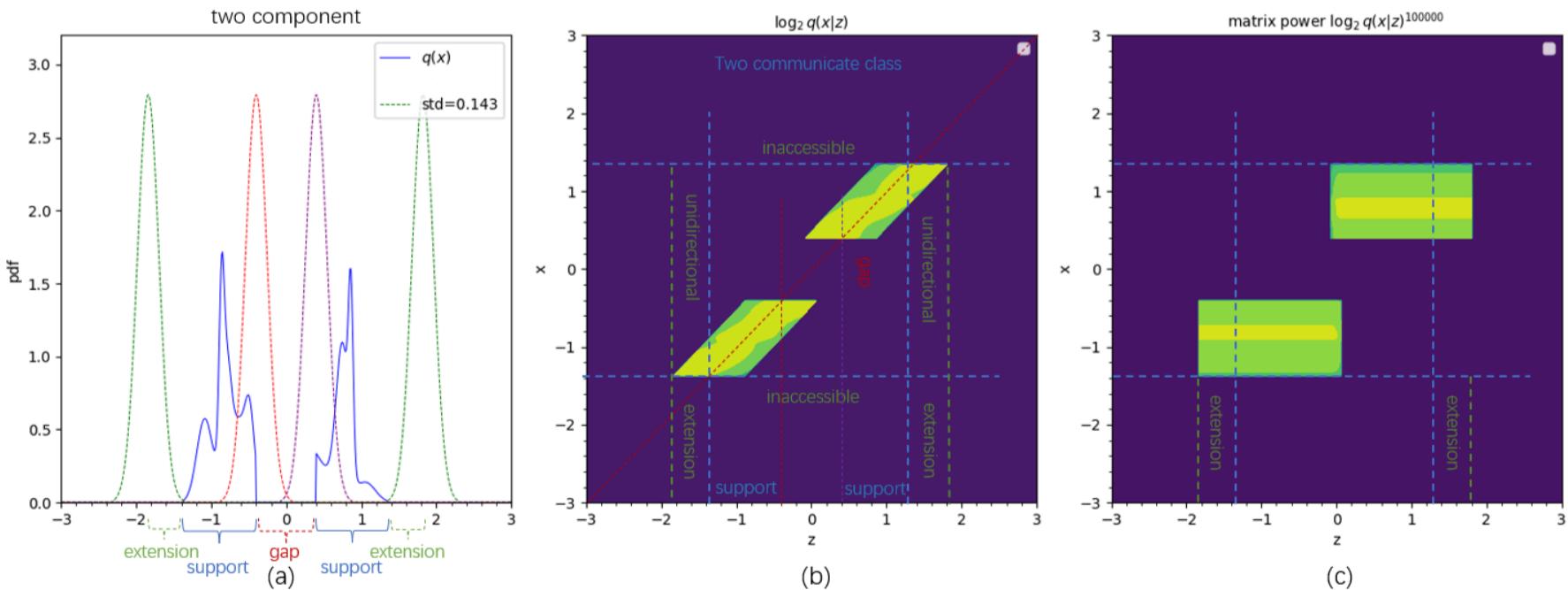


Figure D.3: Two component which cannot communicate with each other

$$\alpha \neq 1$$

当 $\alpha \neq 1$ 时，对 $q(x)$ 支撑集内的任意点 c ，由式D.2可知， $g(x)$ 的中心不再是 c ，而是 $\frac{c}{\sqrt{\alpha}}$ 。也就是说 $g(x)$ 的中心会偏离 c ，偏离的距离为 $\|c\|(\frac{1-\sqrt{\alpha}}{\sqrt{\alpha}})$ 。可以看出， $\|c\|$ 越大，偏离越多。具体可看图D.4(c)和图D.4(d)的例子，在图D.4(d)中，当 $z = 2.0$ ， $g(x)$ 的中心明显偏离 $x = 2.0$ 。本文将此现象称之为**中心偏离现象**。

中心偏离现象将会影响markov chain一些状态的性质。

当偏离的距离明显大于 3σ 时， $g(x)$ 在 $x = c$ 及其附近可能均为零，于是， $q(x = c|z = c)$ 将可能等于0，并且在 $x = c$ 附近 $q(x|z = c)$ 也可能等于0。所以，状态 c 不一定可访问附近的状态。这一点与 $\alpha = 1$ 的情况不同。具体可图D.5的例子，绿色曲线是 $z = 6.0$ 的

$g(x)$, 黄线曲线是 $q(x|z = 6.0)$, 由于 $g(x)$ 的中心偏离 $x = 6.0$ 太多, 导致 $q(x = 6.0|z = 6.0) = 0$ 。

当偏离的距离明显小于 3σ 时, $g(x)$ 在 $x = c$ 及其附近均不为零, 于是, $q(x = c|z = c)$ 将不等于0, 并且在 $x = c$ 附近 $q(x|z = c)$ 也不等于0。所以, 状态 c 可访问附近的状态, 并且是非周期的。

当 c 满足什么要求时, $g(x)$ 中心的偏离距离会小于 3σ 呢?

$$\|c\|\left(\frac{1-\sqrt{\alpha}}{\sqrt{\alpha}}\right) < 3\frac{\sqrt{\beta}}{\sqrt{\alpha}} \Rightarrow \|c\| < 3\frac{\sqrt{\beta}}{1-\sqrt{\alpha}} \quad (\text{D.3})$$

由上可知, 存在一个上限, 只要 $\|c\|$ 小于这个上限, 可保证偏离量小于 3σ 。

当 $\beta = 1 - \alpha$ 时, 上式变为

$$\|c\| < 3\frac{\sqrt{1-\alpha}}{1-\sqrt{\alpha}} \quad (\text{D.4})$$

$3\frac{\sqrt{1-\alpha}}{1-\sqrt{\alpha}}$ 与 α 有着严格的单调递减的关系。

当 $\alpha \in (0, 1)$ 时,

$$3\frac{\sqrt{1-\alpha}}{1-\sqrt{\alpha}} > 3 \quad (\text{D.5})$$

根据上面的分析, 可总结出以下的结论:

1. 如果 $q(x)$ 的支撑集只存在一个connected component, 并且支撑集的点离原点的距离均小于 $3\frac{\sqrt{1-\alpha}}{1-\sqrt{\alpha}}$, 那么 $q(x|z)$ 对应的markov chain是非周期和不可约的。
2. 如果 $q(x)$ 的支撑集存在多个connected component, 由于 $g(x)$ 的中心偏离效应, 准确判断两个component之间是否可以互相访问变得更加复杂, 这里不再详细分析。但下面给出一个保守的结论: 如果支撑集的点离原点的距离均小于1, 并且各个connected component之间的间隙均小于 2σ , 那么 $q(x|z)$ 对应的markov chain是非周期和不可约的。

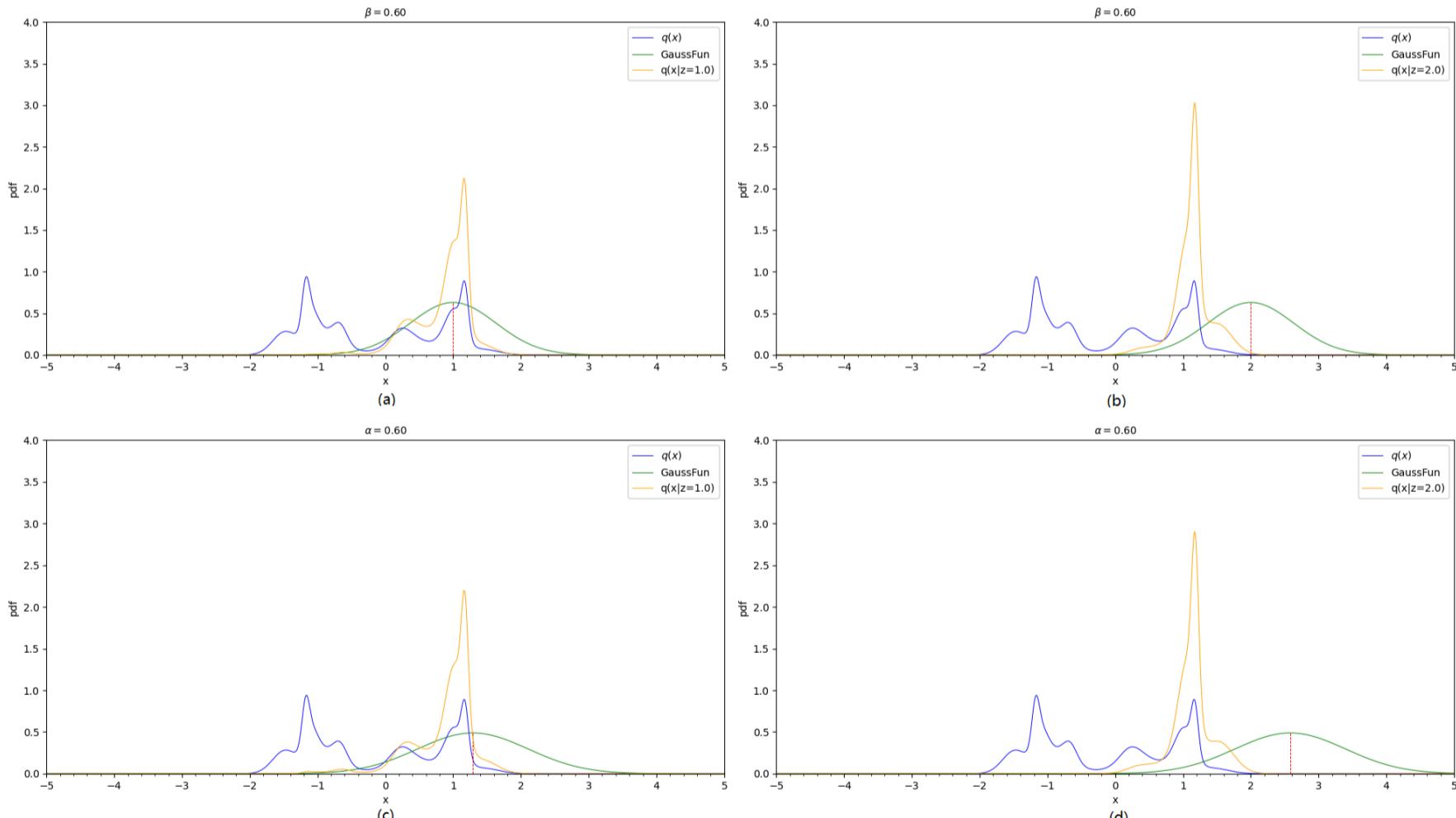


Figure D.4: Center Deviation of the GaussFun

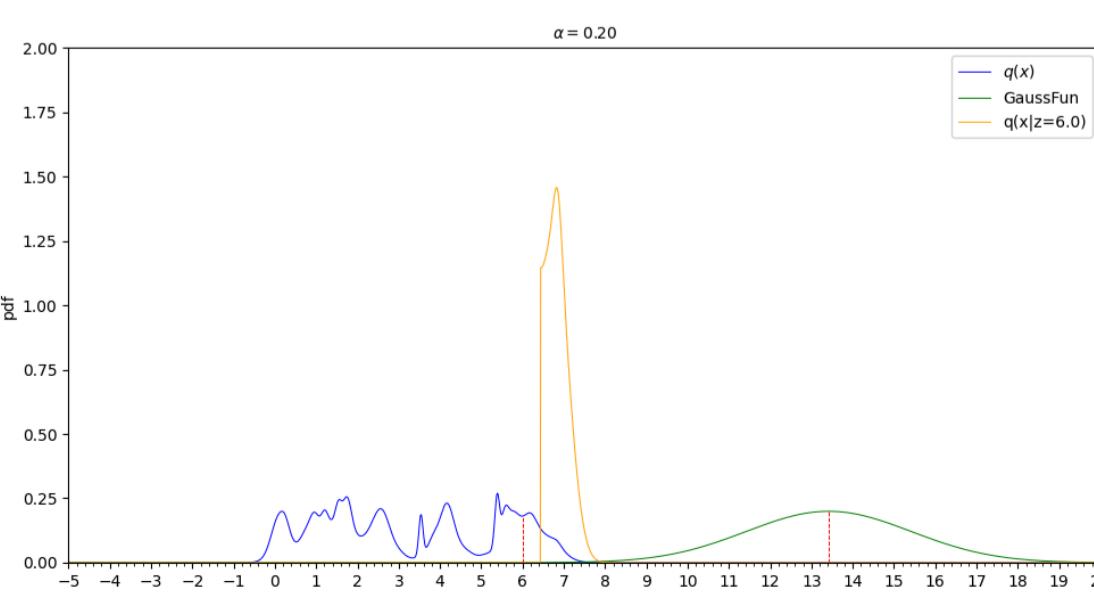


Figure D.5: Deviation is More Than 3σ

Reference

[1] Deep Unsupervised Learning Using Nonequilibrium Thermodynamics

[2] Denoising Diffusion Probabilistic Models

[3] Linear Transformations of Random Variable

[4] Sums and Convolution

[5] Banach fixed-point theorem

[6] Contraction mapping

[7] Fundamental Limit Theorem for Regular Chains

[8] Markov Chain:Basic Theory - Proposition 6

[9] A Converse to Banach's Fixed Point Theorem and its CLS Completeness

[10] Cross-entropy minimization

[11] Deconvolution Using Frequency-Domain Division

[12] deconvolution-by-division-in-the-frequency-domain

[13] Markov Chain:Basic Theory - Theorem 7

[14] Markov Chain:Basic Theory - Definition 4

[15] Variational Diffusion Models

[16] Entropy

[17] Conditional Entropy

[18] A Connection Between Score Matching and Denoising autoencoders

[19] Markov Chain:Basic Theory - Theorem 3

[20] Markov Chains and Mixing Times, second edition - 12.2 The Relaxation Time

[21] Non-negative Matrices and Markov Chains - Theorem 2.10

[22] Pattern Recognition and Machine Learning - 11.2. Markov Chain Monte Carlo

[23] Elements of Information Theory Elements - 2.9 The Second Law of Thermodynamics

About

APP: 本Web APP是使用Gradio开发，并部署在HuggingFace。由于资源有限(2核，16G内存)，所以可能会响应较慢。为了更好地体验，建议从[github](#)复制源代码，在本地机器运行。本APP只依赖Gradio, SciPy, Matplotlib。

Author: 郑镇鑫，资深视觉算法工程师，十年算法开发经历，曾就职于腾讯京东等互联网公司，目前专注于视频生成(类似Sora)。

Email: blair.star@163.com。