毕业设计答辩

概率生成及采样方法的一些探讨研究

黄海乘

指导老师：曾德炉

数学与应用数学

2021 年 6 月 2 日

1 选题背景及意义

2 论文框架与主要内容

3 研究方法及不足

4 结语

1 选题背景及意义

2 论文框架与主要内容

3 研究方法及不足

4 结语

选题背景

- 在实际工程或者模拟中，我们常常需要面临数据生成问题
- 一种常见的情况是，需要得到服从已知特定分布的数据
- 另一种常见的情况是，我们并不知道数据的分布，希望能让机器从已有的观测数据中学习到数据分布并产生相同分布的数据

选题背景

- 概率生成模型的两大基本框架：变分自编码器（VAE）、生成式对抗网络（GAN）
- 采样方法的发展：基本采样方法、MCMC 方法

选题意义

- 深入浅出地介绍 VAE 与 GAN 模型的理论以及展示实验结果，表明概率生成模型在数据生成问题中所展现的巨大潜力
- 结合具体分布探讨研究各种常见采样方法的不同特点，尝试提出改进方法，为面对具体问题时选择合适的采样方法提供理论支持与数据支持

1 选题背景及意义

2 论文框架与主要内容

概率生成模型的探讨研究

采样方法的探讨研究

小结

3 研究方法及不足

4 结语

1 选题背景及意义

2 论文框架与主要内容

概率生成模型的探讨研究

采样方法的探讨研究

小结

3 研究方法及不足

4 结语

基本理论

- **主要思想**：希望建立隐变量模型，依靠神经网络等“学习器”学习到隐变量的分布，进而利用隐变量去控制生成数据。
- **隐变量模型**：假设训练样本为 X，对于每一个数据点 x，都有相应的隐变量 $z \in \mathcal{Z}$ 控制。举个例子，假设模型想生成一个手写数字，那么在生成之前，模型需要做一些“潜在”的决定，例如数字的值、笔画的粗细等等，这些就是隐变量的含义。

VAE 基本理论

对于 VAE 模型，通过最大化后验概率来进行建模优化：

$$p(z|X) = \frac{p(X|z)p(z)}{p(X)} = \frac{p(X|z)p(z)}{\int_z p(X|z)p(z)\mathrm{d}z} \tag{2.1}$$

利用变分思想，用变分函数 $q(z)$ 去替代 $p(z|X)$，整理得到

$$\begin{aligned} &\log p(X) - \mathrm{KL}\left(q(z)\|p(z|X)\right) \\ =&\int q(z)\log p(X|z)\mathrm{d}z - \mathrm{KL}\left(q(z)\|p(z)\right) \end{aligned} \tag{2.2}$$

VAE 基本理论

优化目标：

$$\min \mathrm{KL}\left(q(z|X)||p(z)\right) \tag{2.3}$$

$$\max \int q(z|X)\log p(X|z)\mathrm{d}z \tag{2.4}$$

假设 $p(z)=\mathcal{N}(0,I)$ 以及 $q(z|X)=\mathcal{N}(\mu(X),\Sigma(X))$，再利用重参数化技巧，模型构建完成。

GAN 基本理论

GAN 同时训练两个神经网络，一个是生成器 G，一个是判别器 D。目标函数为：

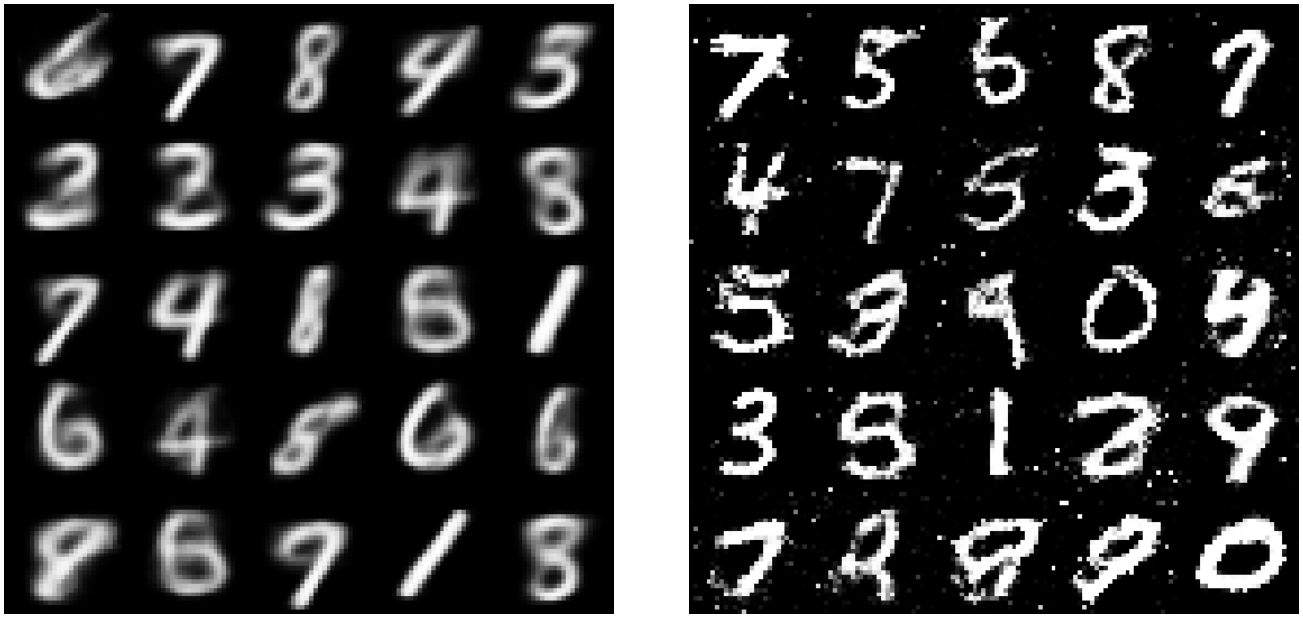
$$\begin{aligned}&\min_G \max_D V(G,D)\\ =&\mathbb{E}_{x\sim p_{data}(x)}\left[\log D(x)\right]+\mathbb{E}_{z\sim p_z(z)}\left[\log\left(1-D(G(z))\right)\right]\end{aligned} \tag{2.5}$$

通过理论证明可以得到：

- 最优判别器：$D^*(x)=\frac{p_{data}(x)}{p_{data}(x)+p_g(x)}$
- 全局最优解：$p_g=p_{data}$

MNIST 数据集上的实验

VAE 与 GAN 均使用神经网络全连接层，在 MNIST 数据集上编码实验结果如下：



a) VAE　　b) GAN

MNIST 数据集上的实验

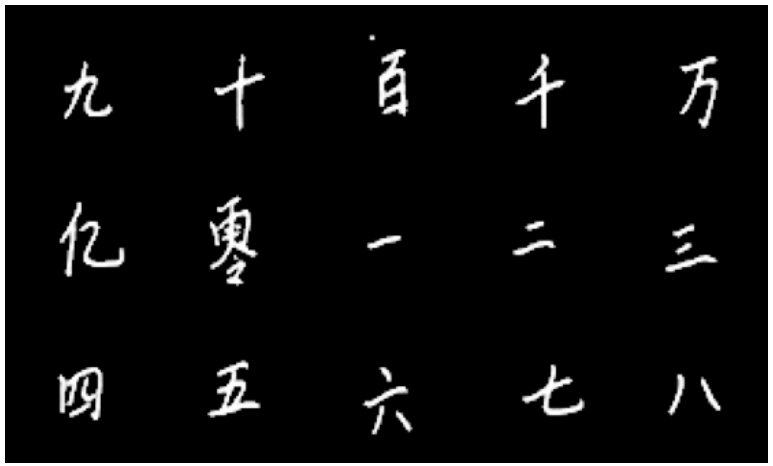
VAE 与 GAN 的对比

VAE	GAN
生成图片模糊	生成图片更逼真
定义了连续的隐变量空间	隐变量空间没有连续性
训练比较容易	训练比较困难

Handwritten Chinese Numbers 数据集上的实验

- 数据处理：加入平滑滤镜，再增强对比度

a) 处理前的数据



b) 处理后的数据

- 数据增强：随机旋转的度数范围为 20，水平平移和垂直平移以及缩放范围都是 0.2

Handwritten Chinese Numbers 数据集上的实验

利用卷积神经网络构建 VAE 模型，训练完成后的生成效果：

a) VAE 生成图 1

b) VAE 生成图 2

1 选题背景及意义

2 论文框架与主要内容

概率生成模型的探讨研究

采样方法的探讨研究

小结

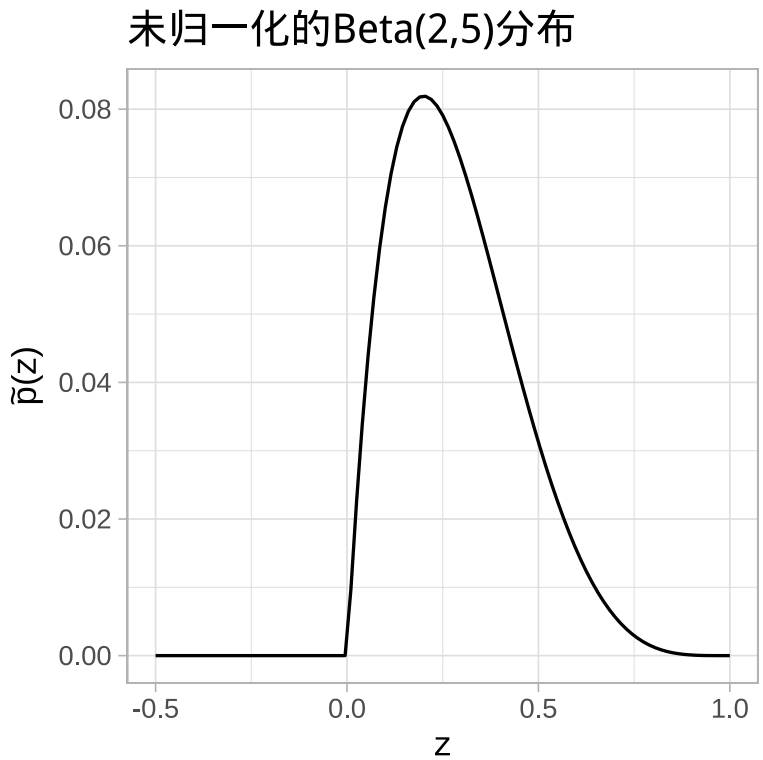
3 研究方法及不足

4 结语

基本采样方法

以具体分布 $\mathrm{Beta}(2,5)$ 探讨研究，未归一化密度函数为：

$$\tilde{p}(z) = \begin{cases} z(1-z)^4 & 0 \le z \le 1 \\ 0 & \text{其他} \end{cases}$$



基本采样方法

主要探讨内容：

- **拒绝采样**：提议分布、k 值对采样效率的影响
- **SIR 方法**：提议分布、采样样本数对采样效果的影响
- 以拒绝采样为例，探讨研究基本采样方法在高维样本空间下采样效率

基本采样方法

探讨研究结论：

拒绝采样	SIR 方法
提议分布与待采样分布的匹配度影响采样接受率	提议分布与待采样分布的匹配度影响采样效果
k 值大小影响采样接受率	采样样本数过小采样样本偏差大
高维空间下采样接受率指数下降，较难确定提议分布	较难确定提议分布

基本采样方法

确定提议分布的一种方法

确定提议分布的步骤：

- 确定分布族 $q(z;\theta)$
- 确定具体提议分布，即确定分布族参数 θ

论文提出一种在给定分布族中选择更好的提议分布的优化方法：

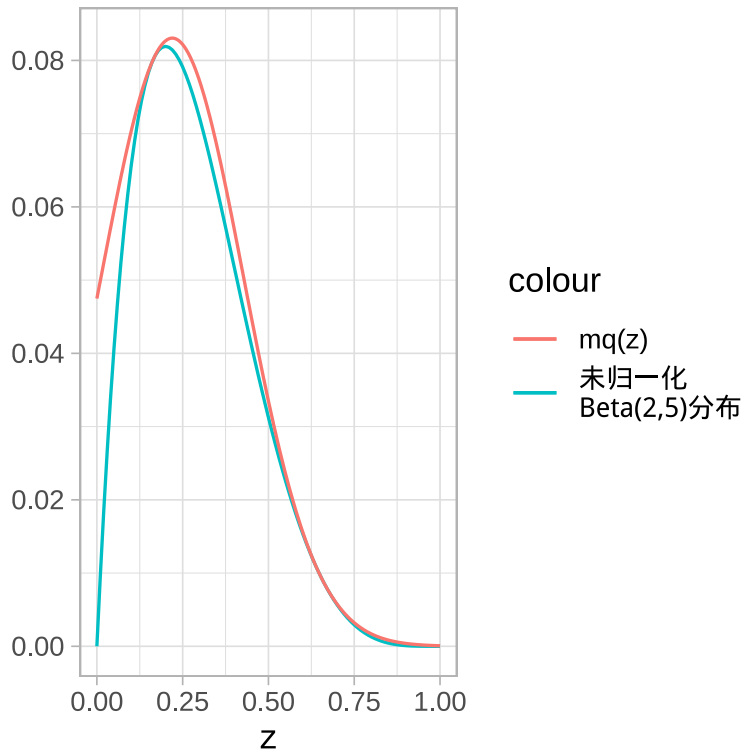
$$\min_{\theta,m} V(\theta,m)=\sum_i (m\cdot q(z_i;\theta)-\tilde{p}(z_i))^2+\lambda\cdot\max\left(\tilde{p}(z_i)-m\cdot q(z_i;\theta)\right) \tag{2.6}$$

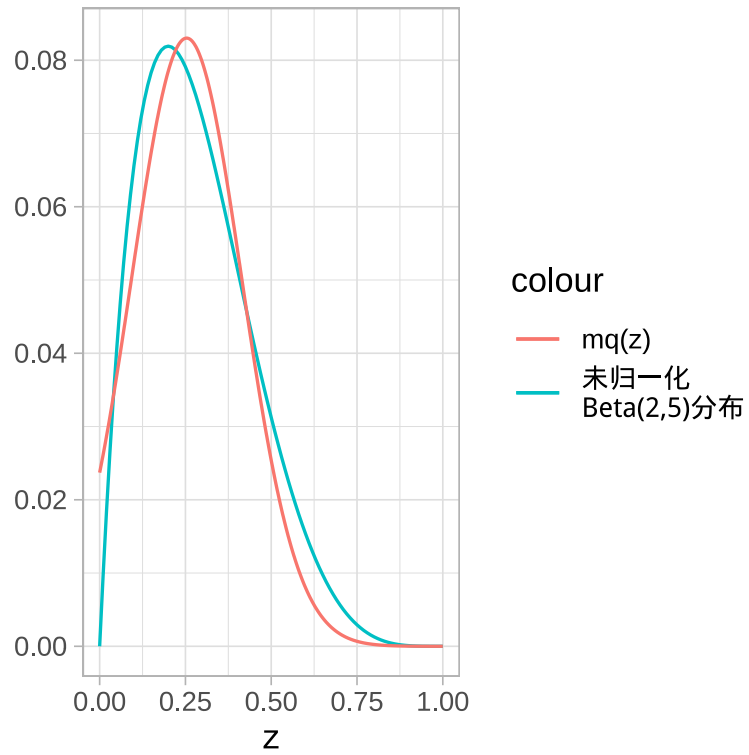
其中，$\{z_i\}$ 是离散化的样本空间，$\lambda\geq 0$。

基本采样方法

确定提议分布的一种方法

利用该方法确定拒绝采样（取 $\lambda = 100$）以及 SIR 方法（取 $\lambda = 0$）的提议分布分别为：





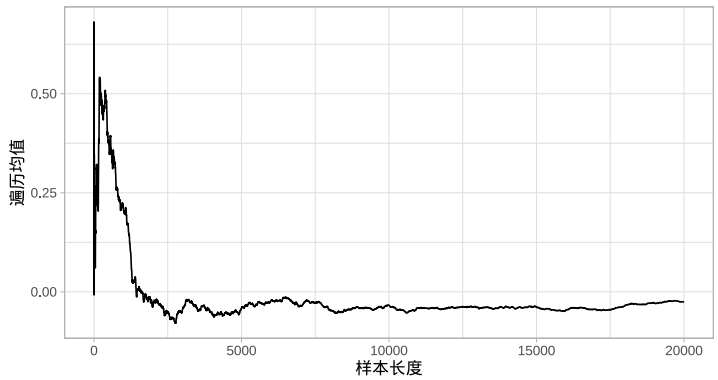
MCMC 方法

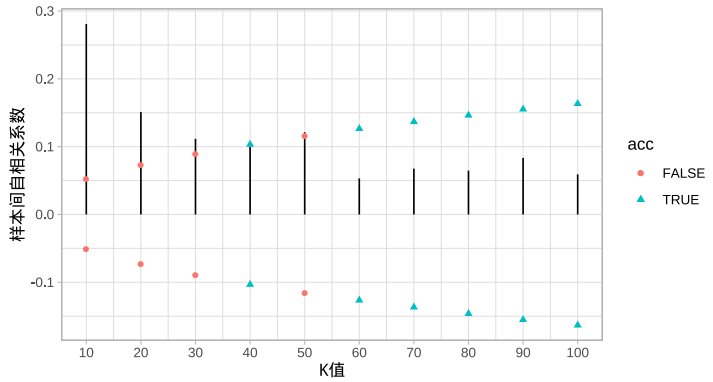
主要探讨内容：

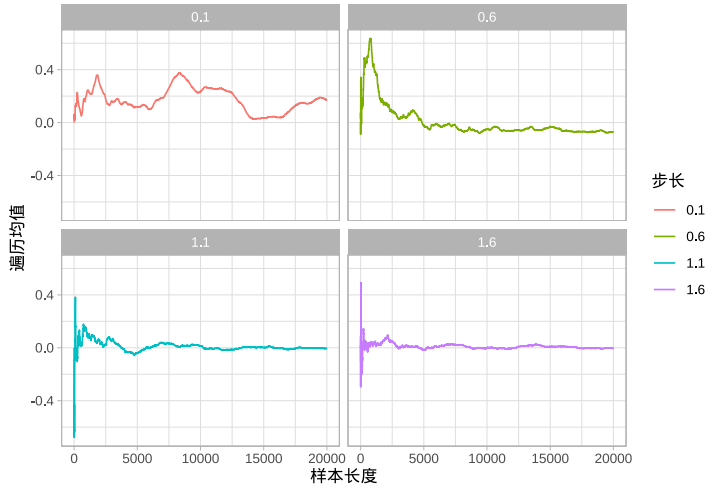
- **Metropolis-Hastings 算法：**利用以当前状态为中心、自定义步长的均匀分布族为提议分布，采样高斯分布，探讨研究老化期与相关性、提议分布的影响、高维空间的优势
- **Gibbs 采样：**采样效率、坐标轴轮换

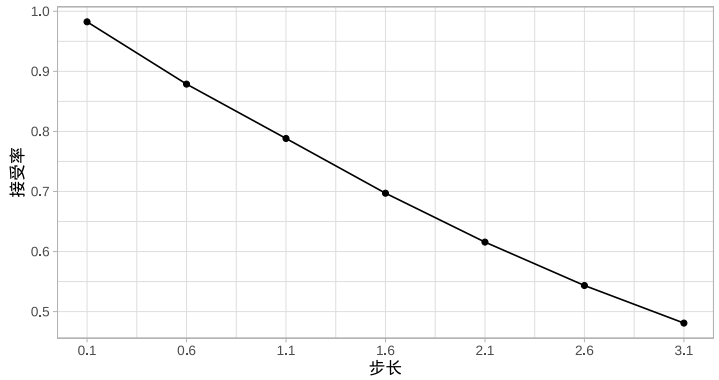
MCMC 方法

Metropolis-Hastings 算法探讨研究部分结果：









MCMC 方法

Metropolis-Hastings 算法探讨研究结论：

探讨研究内容	结论
确定老化期长度	利用遍历均值方法，观察图像波动情况
得到独立样本	分块重采样方法，利用各块样本均值 1 阶自相关系数确定份数
提议分布的影响	步长过小，收敛时间长；步长过大，接受率低
高维空间	接受率近似呈线性下降

1 选题背景及意义

2 论文框架与主要内容

概率生成模型的探讨研究

采样方法的探讨研究

小结

3 研究方法及不足

4 结语

概率生成模型与采样方法的一些对比：

	概率生成模型	采样方法
不同	依赖于数据	不需要数据
	不需要也无法表示具体概率分布	需要具体概率分布
	偏差较大，计算花费较少	偏差较小，计算花费较大
联系	都是数据生成问题的解决方法	
	概率生成模型是以采样方法为基础的	
	两者相结合的模型，如 DRS 与 MH-GAN	

1 选题背景及意义

2 论文框架与主要内容

3 研究方法及不足

4 结语

研究方法及不足

- **研究方法：**以文献研究为基础、以实验为主导的探讨性研究
- 不足：
  - 创新不足，特别是理论方法
  - 文献研究不足
  - 实验设计不足

1 选题背景及意义

2 论文框架与主要内容

3 研究方法及不足

4 结语

结语

希望各位老师能够提出宝贵意见！

Thanks!