河 北 工 业 大 学

毕业设计说明书

作　　者：作者姓名 学　　号：学号

学　　院：学院

系 (专业)：专业

题　　目：论文标题

指导者：指导老师姓名　指导老师职称

评阅者：

2xxx 年 x 月 x 日

毕业设计中文摘要

论文题目

摘要:

论文的摘要是对论文研究内容和成果的高度概括。摘要应对论文所研究的问题及其研究目的进行描述，对研究方法和过程进行简单介绍，对研究成果和所得结论进行概括。摘要应具有独立性和自明性，其内容应包含与论文全文同等量的主要信息。使读者即使不阅读全文，通过摘要就能了解论文的总体内容和主要成果。

论文摘要的书写应力求精确、简明。切忌写成对论文书写内容进行提要的形式，尤其要避免“第 1 章⋯⋯；第 2 章⋯⋯；⋯⋯”这种或类似的陈述方式。

关键词是为了文献标引工作、用以表示全文主要内容信息的单词或术语。关键词不超过 5 个，每个关键词中间用分号分隔。

关键词: 关键词 1 关键词 2 关键词 3

毕业设计外文摘要

Thesis Title

ABSTRACT

An abstract of a dissertation is a summary and extraction of research work and contributions. Included in an abstract should be description of research topic and research objective, brief introduction to methodology and research process, and summary of conclusion and contributions of the research. An abstract should be characterized by independence and clarity and carry identical information with the dissertation. It should be such that the general idea and major contributions of the dissertation are conveyed without reading the dissertation.

An abstract should be concise and to the point. It is a misunderstanding to make an abstract an outline of the dissertation and words "the first chapter", "the second chapter" and the like should be avoided in the abstract.

Keywords are terms used in a dissertation for indexing, reflecting core information of the dissertation. An abstract may contain a maximum of 5 keywords, with semi-colons used in between to separate one another.

Keywords: keyword 1, keyword 2, keyword 3

目录

1　绪论

1.1　研究背景与意义

在当今时代，人工智能技术与我们日常生活的联系逐渐紧密。智能手机、智能家居等智能设施让我们的生活更加便利，影视作品、新闻报道中也常常出现与人工智能有关内容。理查德 · 费曼曾经说过：“我无法创造我所不理解的事物”，对生成模型的研究有助于让人工智能更加“理解”现实世界。中国画[chinesepainting] 是我国传统的绘画形式，其在内容和艺术创作上，体现了古人对自然、社会及与之相关联的政治、哲学、宗教、道德、文艺及自然等方面的认识。使用生成模型算法进行中国画创作，有助于继承和弘扬我国优秀传统文化，让中国画在新时代焕发新的光彩。

1.2　国内外研究现状

1950 年，艾伦 · 图灵提出著名的“图灵测试”[machinery1950computing]，给出了判定机器是否具有“智能”的试验方法，即机器是否能够模仿人类的思维方式来“生成”内容继而与人交互。某种程度上来说，人工智能从那时起就被给寄予了用于内容创造的期许[aigcwhitebook]。最初，人们尝试通过编码，将具体的规则赋予人工智能，但面临的各种困难使人们转变策略。不再是直接提供人为设定的规则，而是为人工智能提供数据资料，让人工智能从数据资料中“学习”所需的规则，也就是“机器学习”[goodfellow2016deep]。随着计算机运算能力的增加与数据量的增长，在机器学习中，让计算模型通过多个中间层来逐渐理解数据潜在表示的“深度学习”[lecun2015deep]，逐渐在计算机视觉、自然语言处理、语音识别等领域取得了突破。

深度学习与生成模型的结合，让人工智能生成内容迎来了新时代，生成内容百花齐放，效果逐渐逼真直至人类难以分辨。

1.2.1　2014–2017 变分自编码器和生成对抗网络时代

2013 年，变分自编码器[kingma2013auto] 的提出使人们认识到，生成模型不仅可以生成简单的数字图像，还可以生成像人脸一样更复杂的图像。生成的图像可以随着隐空间取值的变化而变化。2014 年，生成对抗模型[goodfellow2020generative] 的提出，为人们提供了一种通过对抗结构来解决生成模型问题的全新思路。在随后三年内，人们对生成对抗

网络进行了扩展与改进，如对基础模型结构进行改进的 DCGAN[radford2015unsupervised]，对损失函数进行改进的 Wassertein GAN[arjovsky2017wasserstein]，对训练过程进行改进的 ProGAN[karras2017progressive]，还有使用生成对抗网络在新的领域进行应用，如图像与图像的转化[isola2017imagezhu2017unpaired] 以及音乐的生成[dong2018musegan]。

在同时期，变分自编码器也有了一些重要的改进，如 VAE-GAN[larsen2016autoencoding] 和随后的 VQ-VAE[razavi2019generating]，以及在强化学习上的应用[ha2018world]。

在文本生成领域，自回归模型如 LSTMs 和 GRUs 仍然是主流研究方向，相同的自回归思路也应用在图像生成领域，如 PixelRNN[van2016pixel] 和 PixelCNN[van2016conditional] 引入了一种新的生成图像的方式。也有一些其他生成图像方式的尝试如 RealNVP[dinh2016density] 的出现，为之后规范化流模型打下了基础。

2017 年，Transformer[vaswani2017attention] 的提出使生成模型的研究转向 Transformers。

1.2.2 2018–2019 Transformer **时代**

注意力机制是 Transformer 模型的核心，其使自回归模型不再需要循环层。随着只有 Transformer 解码器的 GPT[radford2018improving] 以及只有 Transformer 编码器的 BERT[devlin2018bert] 的提出，Transformer 很快成为了研究的主流方向。之后，越来越大的模型在很多生成文本的任务上取得了很好的效果。此外，Transformer 也在音乐生成上取得了不错的效果[huang2018music]。

在这两年，一些重要的生成对抗网络模型的提出也巩固了其在图像生成领域的地位。如 SAGAN[zhang2019self] 和 BigGAN[brock2018large] 将注意力机制引入生成对抗模型，取得了很好的效果，StyleGAN[karras2019style] 和随后的 StyleGAN2[karras2020analyzing] 向人们展示了如何控制生成图像的风格与内容。

另一方面，基于分数的模型 NCSN[song2019generative] 的提出，为扩散模型的出现打下了基础。

1.2.3 2020–2023 **大模型时代**

在此时期，一些模型将以往不同种类的模型混合，并在此基础上进一步改进。如 VQ-GAN[esser2021taming] 将生成对抗网络的判别器引入 VQ-VAE 模型结构，Vision Transformer[dosovitskiy2020image] 将 Transformer 应用在图像领域。2022 年，StyleGAN-XL[sauer2022stylegan] 对 StyleGAN 的进一步改进，使其能够在更大的数据集上生成更高分辨率的图像。

于 2020 年提出的两个模型——DDPM[ho2020denoising] 和 DDIM[song2020denoising]，为

之后的大规模图像生成模型打下基础。在图像生成领域，突然出现的扩散模型成为了生成对抗网络的竞争对手。扩散模型不仅可以生成高质量图像，相比于生成对抗网络也更容易训练。

在 2020 年同一时期，拥有 1750 亿参数的 GPT-3[brown2020language] 发布，其可以生成关于任何主题的文本内容。在 2022 年，OpenAI 在 GPT-3 基础上发布网页应用 ChatGPT，让用户可以与 ChatGPT 自然地交流。

在 2021 年到 2022 年，一些在大规模语料库上训练的 Transformer 的变体模型也涌现而出，如微软和英伟达发布的 Megatron-Turing NLG[smith2022using]，谷歌发布的 LaMDA[thoppilan2022lamda] 等。

在过去两年，由于 Transformer 系列模型在文本生成领域的成功，以及扩散模型在图像生成领域的应用，人们开始关注多模态内容生成的研究，如使用文本生成图像。如 OpenAI 在 2021 年发布的模型 DALL.E[ramesh2021zero], 随后的 GLIDE[nichol2021glide] 和 DALL.E 2[ramesh2022hierarchical], 谷歌发布的 Imagen[saharia2022photorealistic]、Parti[yu2022scaling] 和 MUSE 以及 Stability AI 发布的 Stable Diffusion[rombach2022high] 等。

1.3 研究内容与创新点

生成模型算法近年来发展较为迅速，但仍缺少资料对生成模型进行较为系统的探索与研究。如今主流生成模型有规范化流模型、自回归模型、基于能量的模型、变分自编码器、扩散模型和生成对抗网络，各类模型最先进的算法都可以取得很好的效果，但生成模型为什么这样设计，其真正原理是什么，各种生成模型之间有什么关系，理解这些问题有助于设计更好的生成模型，以达到更好的生成效果。本文为首次对各种生成模型原理进行介绍，并将各种框架纳入同一体系的，以有助于研究人员对生成模型进一步研究。

此外，中国画为传统艺术形式，但近年来相比于西方绘画技术的普及，中国画逐渐式微。本文对生成模型在中国画生成方面进行探索，以期望使用人工智能技术生成高质量中国画，继承与弘扬中华传统文化。

1.4 论文组织结构

第一章首先介绍研究背景与意义, 随后介绍早期人工智能生成模型情况, 再将近年来生成模型发展分为三个阶段, 2014–2017 为变分自编码器和生成对抗网络时代, 2018–2019 为 Transformer 时代 2020–2023 为大模型时代。随后介绍研究内容与创新点与论

文组织结构。

第二章首先介绍生成模型算法基本理论与生成模型用途，再根据如何表示或近似似然函数将生成模型分为三类，分别为可求解的显示密度模型、近似估计的显示密度模型和隐式密度模型。其中可求解的显示密度模型与近似估计的显示密度模型都属于显式密度模型。如图1.1所示，可求解密度模型包含规范化流模型和自回归模型，近似估计密度模型包含基于能量的模型、变分自编码器和扩散模型。2.2对各显式密度模型进行了详细介绍，2.3对隐式密度模型生成对抗网络进行了详细介绍，最后，介绍了生成模型评价方法。

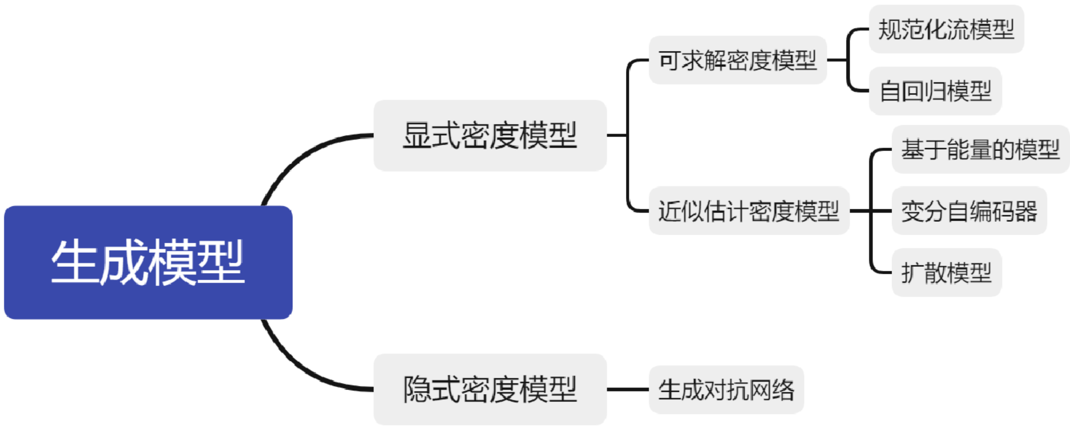

图 1.1 生成模型分类

第三章使用扩散模型进行中国画生成，根据所使用数据集与训练超参数不同，生成不同类型的中国画。

最后对各生成模型算法与中国画生成进行总结与展望。

此外，为保持文中推理证明的完整性，在附录中添加证明所用基础知识。

2 基本理论与方法

2.1 生成模型介绍

生成模型可以通过在数据集上训练，学习数据集中样本的概率分布p_{data}，用生成模型所获得的p_{model} 来近似p_{data}，训练完成获得p_{model} 之后，可以通过从p_{model} 中采样来生成与数据集中相似的样本。

生成模型不仅仅可以应用在近些年有较大突破的图片生成与聊天机器人，还包括很多更广泛的应用[goodfellow2016nips]，如：

- 生成模型的训练和采样，有助于表示和运用高维度概率分布。高维度概率分布在数学和工程领域应用广泛。
- 生成模型可以与强化学习相结合，如进行未来决策或对环境的模拟。
- 生成模型可以辅助半监督学习，如对无标签数据进行标注。
- 生成模型可以应用于多模态领域，如对同一输入，不仅仅生成文字，同时也可以生成图像。
- 从本质上来说，很多任务都需要从一些概率分布中进行采样，如提高图像分辨率，艺术创作，图像与图像之间的转换等。

很多生成模型应用极大似然估计的原理。极大似然估计的基本思想是，定义一个由参数$\boldsymbol{\theta}$ 确定的对概率分布的估计$p_{model}(\boldsymbol{x};\boldsymbol{\theta})$，之后, 对训练集定义似然函数

$$\prod_{i=1}^{m} p_{model}(\boldsymbol{x}^{(i)};\boldsymbol{\theta}) \tag{2.1}$$

式2.1中，m 为该训练集样本数，$\boldsymbol{x}^{(i)}$ 为训练集中样本。

$$\begin{aligned}
\boldsymbol{\theta}^* &= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{model}(\boldsymbol{x}^{(i)};\boldsymbol{\theta}) && (2.2)\\
&= \arg\max_{\boldsymbol{\theta}} \log \prod_{i=1}^{m} p_{model}(\boldsymbol{x}^{(i)};\boldsymbol{\theta}) && (2.3)\\
&= \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p_{model}(\boldsymbol{x}^{(i)};\boldsymbol{\theta}) && (2.4)
\end{aligned}$$

选取参数$\boldsymbol{\theta}^*$ 使似然函数式2.1, 取得最大值, 为方便计算, 相比于将原似然函数进行最大化，可以将其转化之对数空间，这样可以将求积变为求和。在式2.3中，应用了对数函数为单调递增函数, 不改变参数最大值取值的性质。获得的生成模型即为$p_{model}(\boldsymbol{x};\boldsymbol{\theta})$。

图2.1展示了一维极大似然估计过程，极大似然估计从训练集中采样，增大样本所在位置概率的同时保证概率密度总积分为 1。

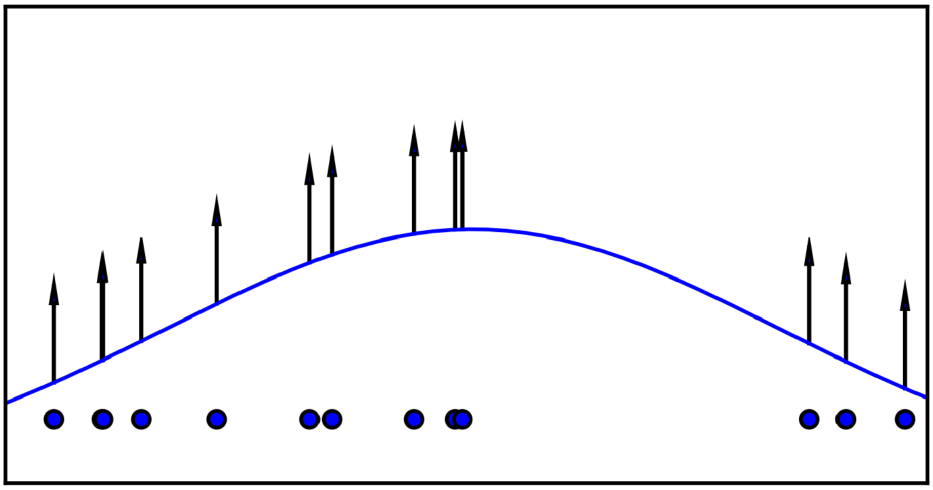

图 2.1 一维极大似然估计过程

从另一方面来说，对参数$\boldsymbol{\theta}^*$ 进行极大似然估计与最小化生成模型与样本数据的实际概率分布的 KL 散度等价，即：

$$\boldsymbol{\theta}^* = \arg\max_{\boldsymbol{\theta}} \ D_{KL}(p_{data}(\boldsymbol{x})||p_{model}(\boldsymbol{x};\boldsymbol{\theta})) \tag{2.5}$$

可以根据如何表示或近似似然函数对生成模型进行分类。具体地说，主要可以分为以下三类：

- 可求解的显示密度模型，直接定义概率密度函数$p_{model}(\boldsymbol{x};\boldsymbol{\theta})$，且要求其似然函数可直接求解。如规范化流模型和自回归模型。
- 近似估计的显示密度模型，直接定义概率密度函数$p_{model}(\boldsymbol{x};\boldsymbol{\theta})$，且要求其似然函数可求得近似解。如基于能量的模型、变分自编码器和扩散模型。
- 隐式密度模型，不定义概率密度函数，而通过其他间接方式对概率分布p_{data} 进行学习。如生成对抗网络。

2.2 显式密度模型

显式密度模型直接定义概率密度函数$p_{model}(\boldsymbol{x};\boldsymbol{\theta})$，之后即可以根据极大似然估计进行模型拟合。显式密度模型存在的主要问题是：很难设计既可以表示样本分布又容易求解的模型。有两种方法可以解决这个问题：

1. 设计易求解的模型结构，在此基础上提高模型表达能力，即可求解密度模型。
2. 设计模型后，通过近似方法求解似然函数，即近似估计密度模型。

2.2.1 可求解密度模型

规范化流模型

规范化流模型通过一系列可逆变换方程，将简单的概率分布逐渐转化为复杂的概率分布，以希望能够拟合数据样本的概率分布。

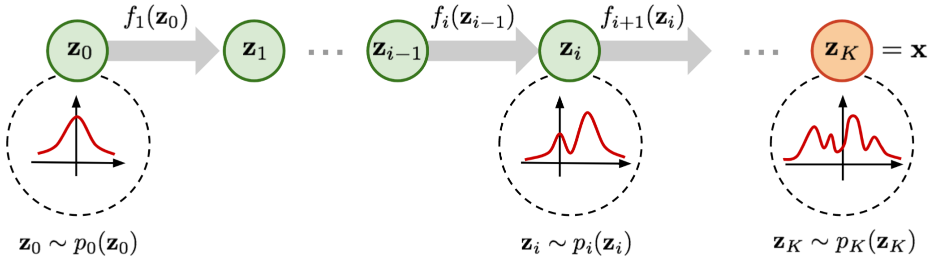

图 2.2 规范化流模型示意图

如图2.2所示，

$$\boldsymbol{z}_{i-1} \sim p_{i-1}(\boldsymbol{z}_{i-1}) \tag{2.6}$$

$$\boldsymbol{z}_i = f_i(\boldsymbol{z}_{i-1}) \tag{2.7}$$

由反函数定理，

$$\boldsymbol{z}_{i-1} = f_i^{-1}(\boldsymbol{z}_i) \tag{2.8}$$

由附录式??，

$$p_i(\boldsymbol{z}_i) = p_{i-1}(f_i^{-1}(\boldsymbol{z}_i)) \left|\det \frac{df_i^{-1}}{d\boldsymbol{z}_i}\right| \tag{2.9}$$

为获得$p_i(\boldsymbol{z}_i)$ 与$p_{i-1}(\boldsymbol{z}_{i-1})$ 之间的关系，对式2.9进行变形：

$$\begin{aligned} p_i(\boldsymbol{z}_i) &= p_{i-1}(f_i^{-1}(\boldsymbol{z}_i)) \left|\det \frac{df_i^{-1}}{d\boldsymbol{z}_i}\right| && (2.10) \\ &= p_{i-1}(\boldsymbol{z}_{i-1}) \left|\det \left(\frac{df_i}{d\boldsymbol{z}_{i-1}}\right)^{-1}\right| \quad \text{（根据式??）} && (2.11) \\ &= p_{i-1}(\boldsymbol{z}_{i-1}) \left|\det \frac{df_i}{d\boldsymbol{z}_{i-1}}\right|^{-1} \quad \text{（根据式A.10）} && (2.12) \end{aligned}$$

对式2.12进行对数化得：

$$\log p_i(\boldsymbol{z}_i) = \log p_{i-1}(\boldsymbol{z}_{i-1}) - \log \left|\det \frac{df_i}{d\boldsymbol{z}_{i-1}}\right| \tag{2.13}$$

连续应用式2.13，对$\boldsymbol{x}$ 不断变换可得其关于简单概率分布变量$\boldsymbol{z}$ 的表达式，即：

$$\boldsymbol{x} = \boldsymbol{z}_K = f_K \circ f_{K-1} \circ \cdots \circ f_1(\boldsymbol{z}_0) \tag{2.14}$$

$$\log p(\boldsymbol{x}) = \log p_K(\boldsymbol{z}_K) = \log p_{K-1}(\boldsymbol{z}_{K-1}) - \log \left|\det \frac{df_K}{d\boldsymbol{z}_{K-1}}\right| \tag{2.15}$$

$$= \log p_{K-2}(\boldsymbol{z}_{K-2}) - \log \left|\det \frac{df_{K-1}}{d\boldsymbol{z}_{K-2}}\right| - \log \left|\det \frac{df_K}{d\boldsymbol{z}_{K-1}}\right| \tag{2.16}$$

$$= \cdots \tag{2.17}$$

$$= \log p_0(\boldsymbol{z}_0) - \sum_{i=1}^{K} \log \left|\det \frac{df_i}{d\boldsymbol{z}_{i-1}}\right| \tag{2.18}$$

所谓规范化流模型中的流指的是一系列随机变量之间的代换，即不断应用$\boldsymbol{z} = f_i(\boldsymbol{z}_{i-1})$。规范化流是指由一系列分布$p_i$ 组成的完整链式过程[weng2018flow]。

由上述计算过程，转换函数f_i 需要满足两个条件：

- 具备可逆性质；
- 雅可比矩阵容易计算。

自回归模型

自回归模型将生成问题简化成顺序问题，即通过之前的顺序值来预测下一个值。一般来说，自回归模型对于高维数据x 将其联合概率分布化为条件概率的乘积的形式：

$$p(\boldsymbol{x}) = p(x_1, \ldots, x_n) = \prod_{i=1}^{n} p(x_i|x_1, \ldots, x_{i-1}) \tag{2.19}$$

对条件概率的模拟比直接对联合概率分布进行建模更加容易。

具体而言，为便于求解可以假定每一个变量只依赖于不超过一定数量的变量，比如两个变量，即：

$$p(\boldsymbol{x}) = p(x_1)p(x_2|x_1) \prod_{d=3}^{D} p(x_d|x_{d-1}, x_{d-2}) \tag{2.20}$$

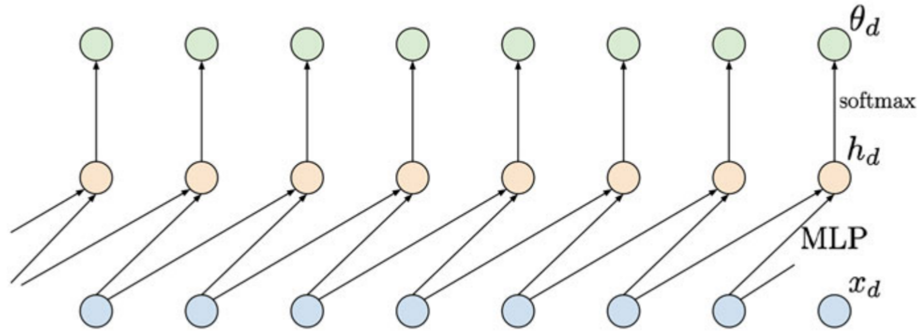

图 2.3 自回归模型依赖两变量示意图

可以使用多层感知机预测x_d 的概率分布，图2.3表示使用多层感知机进行预测x_d 的概率分布。最下面的蓝色结点表示输入，中间层橘色结点表示多层感知机对前两个

数据处理后的输出，最后绿色点表示通过归一化指数函数输出的概率$p(x_d|x_{d-1},x_{d-2})$，即θ_d。但是，这种假设，即每一个变量只依赖于不超过一定数量的变量，对模型造成了很大的限制。通过循环神经网络可以对以往信息进行更长期的保存**Jakub2022deep**，即：

$$p(x_d|\boldsymbol{x}_{<d}) = p(x_d|RNN(x_{d-1},h_{d-1})) \tag{2.21}$$

式2.21中，$h_d = RNN(x_d, h_{d-1})$，h_d 可以看作对所有历史信息的保存，可称为隐环境。图2.4表示使用循环神经网络 RNN 预测x_d 的概率分布。最下面的蓝色结点表示输入，中间层橘色结点表示循环神经网络对前两个输入数据与隐环境处理后的输出，最后绿色结点表示通过归一化指数函数后输出的概率$p(x_d|x_{d-1},x_{d-2})$，即θ_d。

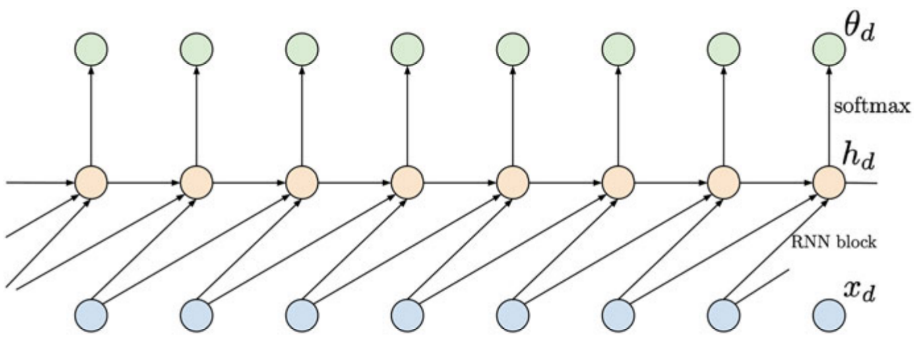

图 2.4 循环神经网络示意图

此外，transformer 也属于自回归模型，transformer 通过自注意力机制来对历史信息进行处理。

2.2.2 近似估计密度模型

基于能量的模型

基于能量的模型的诞生受到对物理系统建模的启发——一个事件的概率可以由玻尔兹曼分布式A.69表示**lippe2022uvadlc**。如果一个神经网络$E_\theta(\boldsymbol{x})$ 只有一个输出神经元，θ 表示神经网络的参数，$\boldsymbol{x}$ 表示神经网络的输入，其输出结果为实值标量，那么有：

$$q_\theta(\boldsymbol{x}) = \frac{\exp(-E_\theta(\boldsymbol{x}))}{Z_\theta}\ ,\ 其中Z_\theta = \begin{cases} \int_{\boldsymbol{x}} \exp(-E_\theta(\boldsymbol{x}))\,d\boldsymbol{x} & \boldsymbol{x}为连续变量 \\ \sum_{\boldsymbol{x}} \exp(-E_\theta(\boldsymbol{x})) & \boldsymbol{x}为离散变量 \end{cases} \tag{2.22}$$

式2.22中，指数函数保证所得概率大于 0，在$E_\theta(\boldsymbol{x})$ 前添加负号以表示E_θ 为能量函数：样本点概率取值越高则其能量越低，概率取值越低则具有更高的能量。Z_θ 是归一化项用来保证概率密度积分或和为 1，如式2.23所示：

$$\int_{\boldsymbol{x}} q_\theta(\boldsymbol{x})\,d\boldsymbol{x} = \int_{\boldsymbol{x}} \frac{\exp(-E_\theta(\boldsymbol{x}))}{\int_{\hat{\boldsymbol{x}}} \exp(-E_\theta(\hat{\boldsymbol{x}}))\,d\hat{\boldsymbol{x}}}\,d\boldsymbol{x} = \frac{\int_{\boldsymbol{x}} \exp(-E_\theta(\boldsymbol{x}))\,d\boldsymbol{x}}{\int_{\hat{\boldsymbol{x}}} \exp(-E_\theta(\hat{\boldsymbol{x}}))\,d\hat{\boldsymbol{x}}} = 1 \tag{2.23}$$

式2.22与式2.23中，$q_\theta(\boldsymbol{x})$ 为$p(\boldsymbol{x})$ 的近似，通过训练，可以使$q_\theta(\boldsymbol{x})$ 逐渐接近$p(\boldsymbol{x})$。

对于基于能量的模型，E_θ 可以根据需要灵活选择。由于归一化常数Z_θ 未知，无法直接计算对数似然函数，且由于Z_θ 不一定能保证不变，不可直接对未进行归一化的概率$\exp(-E_\theta(\boldsymbol{x}_{train}))$ 极大化，即不一定保证训练样本点相比其他数据点有更高的概率出现。对于极大似然函数的计算，可以通过对比散度方法来进行近似。通过比较不同数据点的似然来进行计算，根据附录式$A.45$ 可得负对数似然函数，

$$\nabla_\theta Loss(q_\theta(\boldsymbol{x})) = -\mathbb{E}_{p(\boldsymbol{x})}\left[\nabla_\theta \log q_\theta(\boldsymbol{x})\right] \tag{2.24}$$

$$= \mathbb{E}_{p(\boldsymbol{x})}\left[\nabla_\theta E_\theta(\boldsymbol{x})\right] - \mathbb{E}_{q_\theta(\boldsymbol{x})}\left[\nabla_\theta E_\theta(\boldsymbol{x})\right] \tag{2.25}$$

式2.24即为需要最小化的损失函数。对于式2.25，从直观上理解，第一项表示最小化数据集中样本点的能量，以增加其概率；第二项表示最大化随机生成的样本点的能量，以减小其概率。图2.5中，f_θ 表示$\exp(-E_\theta(\boldsymbol{x}))$，通过训练，可以降低数据集中样本的能量函数值，增大随机生成样本的能量函数值。

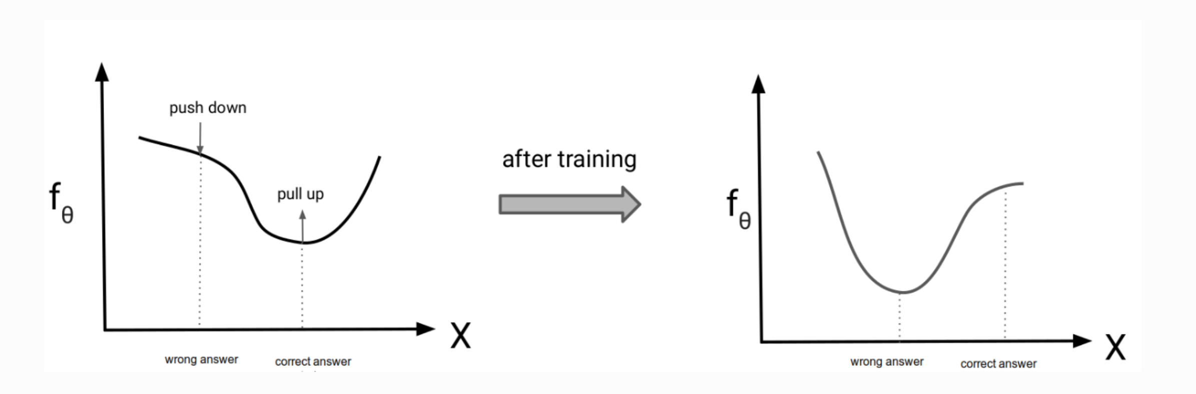

图 2.5 对比散度训练前后对比

式2.25中，需要对概率分布$q_\theta(\boldsymbol{x})$ 采样。可以应用朗志万动力学与马尔科夫链蒙特卡洛结合的方法采样，即从随机数据点开始，根据能量函数的梯度$\nabla_{\boldsymbol{x}} E_\theta(\boldsymbol{x})$，使数据点的值向能量函数降低的方向移动。此外，为避免陷入局部最小值，需要在每一次更新梯度时添加噪声$\omega \backsim \mathcal{N}(0,\sigma)$。理论上讲，根据梯度进行足够多次数据点取值更新，即可以准确地从概率分布$q_\theta(\boldsymbol{x})$ 采样，但在实际中，通常将马尔科夫链的链长限制为K，即进行K 次取值更新。算法1为从基于能量的模型采样的算法。

变分自编码器

自编码器是一个包含了两部分的神经网络，其编码器可以将原始高维输入映射到低维隐变量空间，解码器可以将隐变量空间中的低维表示恢复成原始输入[weng2018VAE]。通过使用自编码器，可以对数据进行更高效地压缩。

算法 1 从基于能量的模型采样

从高斯分布或均匀分布中采样$\tilde{\boldsymbol{x}}^0$

for 采样步骤$k = 1$ 到K **do**

  $\tilde{\boldsymbol{x}}^k \leftarrow \tilde{\boldsymbol{x}}^{k-1} - \eta \nabla_{\boldsymbol{x}} E_\theta(\tilde{\boldsymbol{x}}^{k-1}) + \omega$，其中$\omega \sim \mathcal{N}(0, \sigma)$

end for

$\boldsymbol{x}_{sample} \leftarrow \tilde{\boldsymbol{x}}^K$

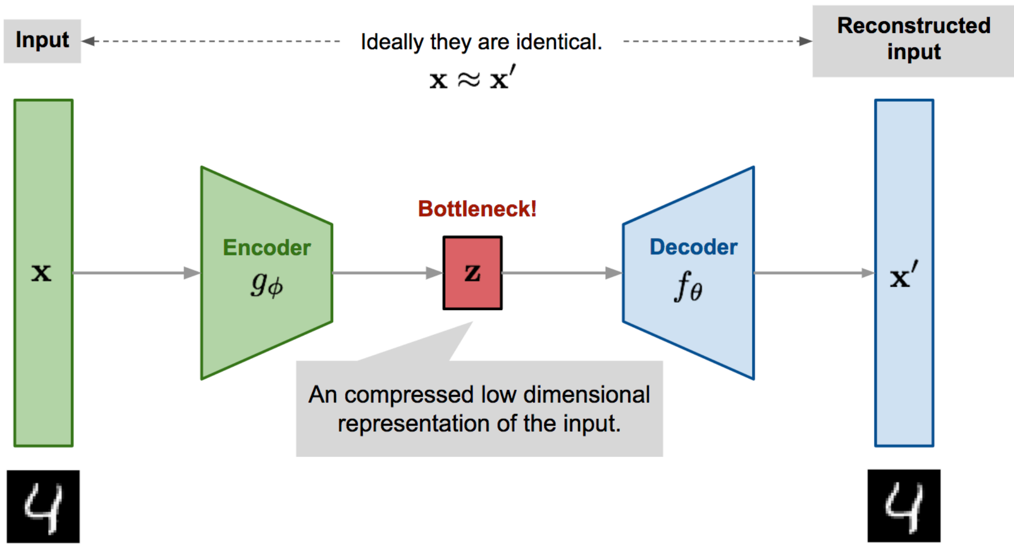

图 2.6 自编码器示意图

图2.6中，$\boldsymbol{x}$ 表示样本，$\boldsymbol{x}^{'}$ 表示对原始样本的重构，$\boldsymbol{z}$ 表示样本在隐变量空间的表示，$g_\phi(.)$ 表示由参数ϕ 确定的编码器函数，$f_\theta(.)$ 表示由参数θ 确定的解码器函数。

$$\boldsymbol{z} = g_\phi(\boldsymbol{x}) \tag{2.26}$$

$$\boldsymbol{x}^{'} = f_\theta(g_\phi(\boldsymbol{x})) \tag{2.27}$$

参数(θ, ϕ) 可以通过比较原始样本值$\boldsymbol{x}$ 与重构样本值$\boldsymbol{x}^{'}$ 来学习，如使用均方误差：

$$L_{AE}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}^{(i)} - f_\theta(g_\phi(\boldsymbol{x}))) \tag{2.28}$$

为避免过拟合，降噪自编码器**vincent2008extracting** 对原始样本添加随机噪声：

$$\hat{\boldsymbol{x}}^{(i)} \backsim \mathcal{M}_{\mathcal{D}}(\hat{\boldsymbol{x}}^{(i)} \mid \boldsymbol{x}^{(i)}) \tag{2.29}$$

$$L_{DAE}(\theta, \phi) = \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{x}^{(i)} - f_\theta(g_\phi(\hat{\boldsymbol{x}}^{(i)}))) \tag{2.30}$$

式2.29中，$\hat{\boldsymbol{x}}^{(i)}$ 表示对原始样本值$\boldsymbol{x}$ 添加随机噪声或者进行掩码处理后的值，$\mathcal{D}$ 为原始样本所在数据集，$\mathcal{M}_{\mathcal{D}}$ 表示从原始样本到$\hat{\boldsymbol{x}}^{(i)}$ 的映射。式2.30与式2.28相比，不再

将重构样本与原始样本$\boldsymbol{x}$ 进行比较，而是与添加随机噪声或进行掩码处理后的$\hat{\boldsymbol{x}}^{(i)}$ 进行比较。图2.7为降噪自编码器结构图。

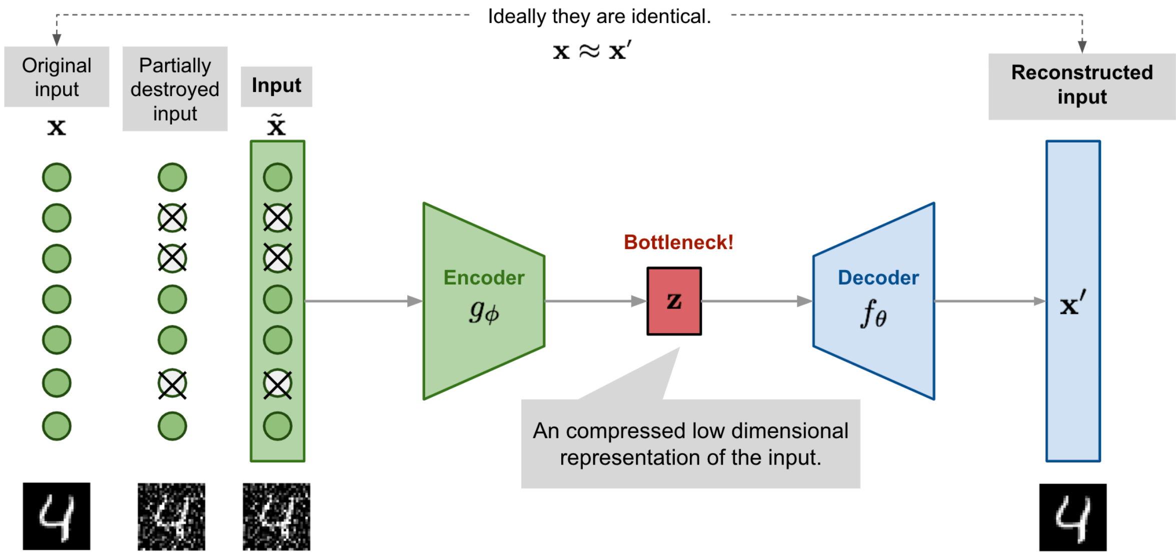

图 2.7 降噪自编码器结构

相比于自编码器和降噪自编码器，变分自编码器并不将样本映射到一个固定的向量而是映射到一个由参数θ 确定的概率分布p_θ。输入样本$\boldsymbol{x}$ 和隐变量$\boldsymbol{z}$ 的关系可以定义为：

- 先验概率$p_\theta(\boldsymbol{z})$；
- 似然函数$p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$；
- 后验概率$p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$。

若知道概率分布p_θ 的真实参数值θ^*，则生成样本点某一维度的值$\boldsymbol{x}^{(i)}$（如图片样本的一个像素值）需要两个步骤：

1. 从先验分布$p_{\theta^*}(\boldsymbol{z})$ 中采样获得$\boldsymbol{z}^{(i)}$；
2. 从条件概率$p_{\theta^*}(\boldsymbol{x} \mid \boldsymbol{z} = \boldsymbol{z}^{(i)})$ 中生成值$\boldsymbol{x}^{(i)}$。

最优参数值θ^* 可以通过最大化生成真实数据样本的概率获得，即：

$$\theta^* = \arg\max_\theta \prod_{i=1}^{n} p_\theta(\boldsymbol{x}^{(i)}) \tag{2.31}$$

取对数可得：

$$\theta^* = \arg\max_\theta \sum_{i=1}^{n} \log p_\theta(\boldsymbol{x}^{(i)}) \tag{2.32}$$

式2.32中，$p_\theta(\boldsymbol{x}^{(i)})$ 可以表示为：

$$p_\theta(\boldsymbol{x}^{(i)}) = \int p_\theta(\boldsymbol{x}^{(i)} \mid \boldsymbol{z}) p_\theta(\boldsymbol{z}) d\,\boldsymbol{z} \tag{2.33}$$

但由于某些隐变量无法积分或计算开销巨大，式2.33无法直接计算。可用证据下界式A.57来近似对数似然函数。使用证据下界可以获得对后验概率$p_\theta(\boldsymbol{z} \mid \boldsymbol{x})$的近似，即由参数$\phi$ 确定的$q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$。图2.8中，近似函数$q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$定义了编码器，条件概率$p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$定义了解码器。

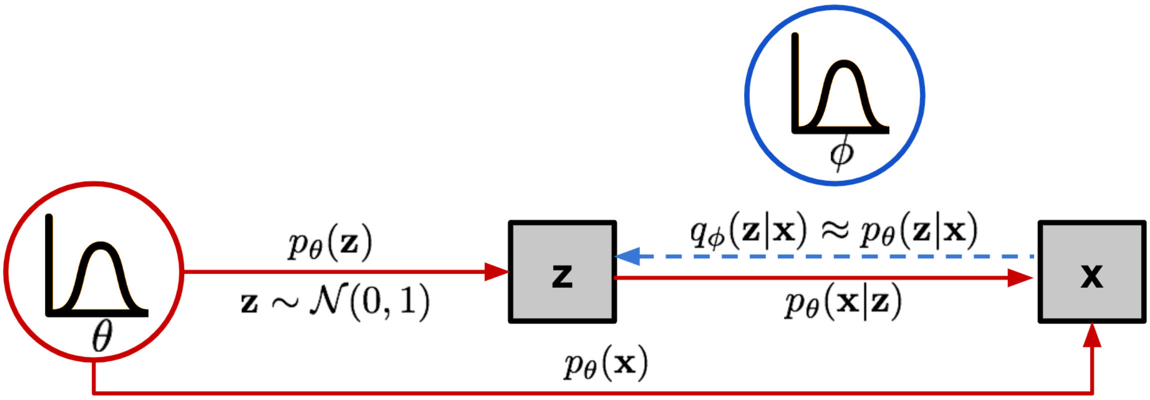

图 2.8 使用 ELBO 近似变分自编码器对数似然函数

对证据下界式A.57变形可得：

$$\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})}\right] = \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p_\theta(\boldsymbol{x} \mid \boldsymbol{z})p(\boldsymbol{z})}{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})}\right] \tag{2.34}$$

$$= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\right] + \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{z})}{q_\phi(\boldsymbol{z} \mid \boldsymbol{x})}\right] \tag{2.35}$$

$$= \underbrace{\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\right]}_{\text{重构项}} - \underbrace{D_{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \,||\, p(\boldsymbol{z}))}_{\text{先验匹配项}} \tag{2.36}$$

式2.36中，重构项度量了解码器相对于变分分布$q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$的重构似然函数$p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$，从而使学到的分布可以对隐变量有效地建模，以便于可以通过$p_\theta(\boldsymbol{x} \mid \boldsymbol{z})$进行原样本重构。先验匹配项度量了学到的变分分布与隐变量先验分布的相似度，从而保证编码器能够真正学到隐变量分布而不是记忆原样本到隐变量的映射。最大化证据下界等价于最大化重构项和最小化先验匹配项。

变分自编码器中，通常编码器都为具有对角协方差矩阵的多变量高斯分布，隐变量的先验分布为多变量标准高斯分布：

$$q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{\mu}_\phi(\boldsymbol{x}, \boldsymbol{\sigma}_\phi^2(\boldsymbol{x})\boldsymbol{I})) \tag{2.37}$$

$$p(\boldsymbol{z}) = \mathcal{N}(\boldsymbol{z}; \boldsymbol{0}, \boldsymbol{I}) \tag{2.38}$$

因此，由式2.37和式2.38可以计算式2.36中的先验匹配项。而式2.36中的重构项可

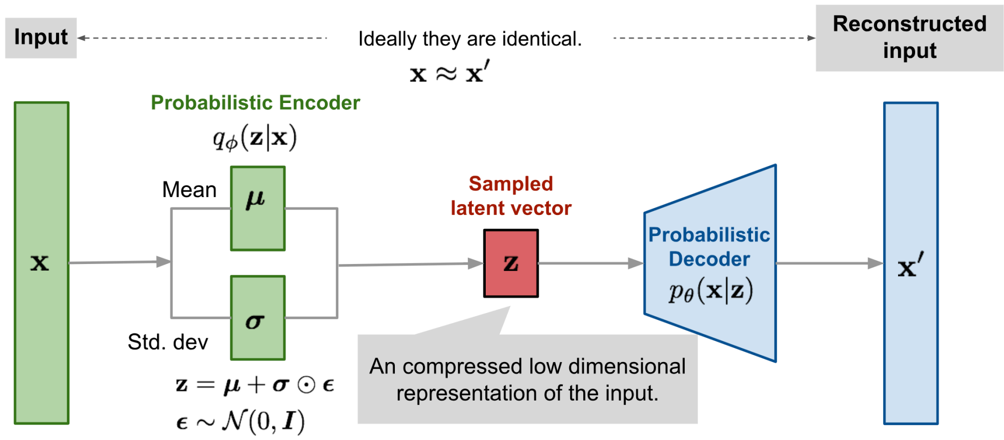

图 2.9 变分自编码器

以由蒙特卡洛方法进行估计，即：

$$\underset{\phi,\theta}{\arg\max}\ \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p_\theta(\boldsymbol{x} \mid \boldsymbol{z})\right] - D_{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \mid\mid p(\boldsymbol{z})) \tag{2.39}$$

$$\approx \underset{\phi,\theta}{\arg\max} \sum_{l=1}^{L} \log p_\theta(\boldsymbol{x} \mid \boldsymbol{z}^{(l)}) - D_{KL}(q_\phi(\boldsymbol{z} \mid \boldsymbol{x}) \mid\mid p(\boldsymbol{z})) \tag{2.40}$$

式2.40中，$\boldsymbol{z}^{(l)}{}_{l=1}^{L}$ 为关于数据集中每个样本$\boldsymbol{x}$ 从$q_\phi(\boldsymbol{z} \mid \boldsymbol{x})$ 中获得的采样。

但由于通过随机过程获得的采样$\boldsymbol{z}^{(l)}$ 不可微，无法计算梯度，根据重参数方法式A.66可得：

$$\boldsymbol{z} = \boldsymbol{\mu}_\phi(\boldsymbol{x}) + \boldsymbol{\sigma}_\phi(\boldsymbol{x}) \odot \boldsymbol{\epsilon} \qquad \text{其中} \epsilon \sim \mathcal{N}(\boldsymbol{\epsilon}; \boldsymbol{0}, \boldsymbol{I}) \tag{2.41}$$

式2.41中，$\odot$ 为对应元素分别相乘。图2.10中，重参数方法使损失梯度不可反向传播的过程变为可以反向传播。

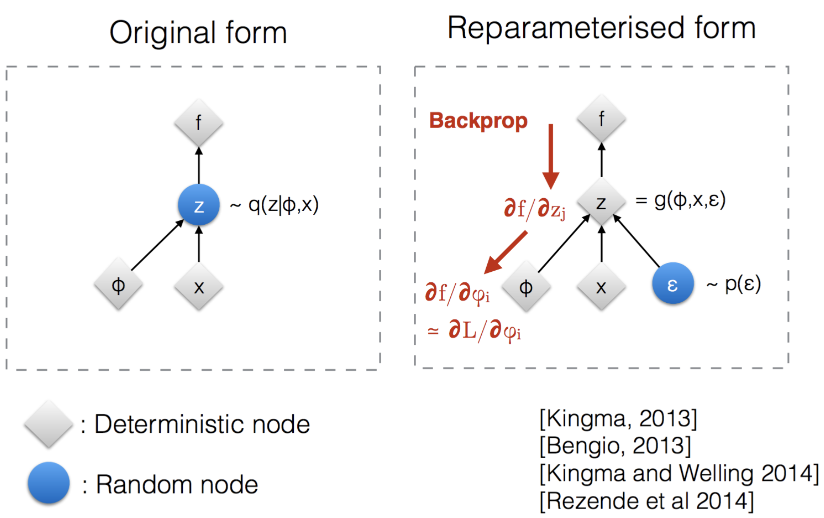

图 2.10 重参数方法示意图

式2.40中，变分自编码器使用重参数方法和蒙特卡洛方法，对证据下界关于ϕ 和θ 同时进行优化。训练完成后，即可直接从隐变量分布$p(\boldsymbol{z})$ 中进行采样，再经由解码器获得生成样本$\boldsymbol{x}'$。

层级变分自编码器是对变分自编码器的扩展[Luo2022UnderstandingDM]，相比于变分自编码器，层级变分自编码器可以有多层隐变量。层级变分自编码器假设浅层隐变量由更深层的隐变量生成。对于有T 层隐变量的层级变分自编码器，每一个隐变量都依赖于先前所有隐变量，而其一个特例为马尔可夫层级变分自编码器。马尔可夫层级变分自编码器中，生成过程为马尔可夫链，解码时每一个隐变量$\boldsymbol{z}_t$ 只依赖于之前一个隐变量$\boldsymbol{z}_{t+1}$，图2.11为马尔可夫层级变分自编码器示意图。

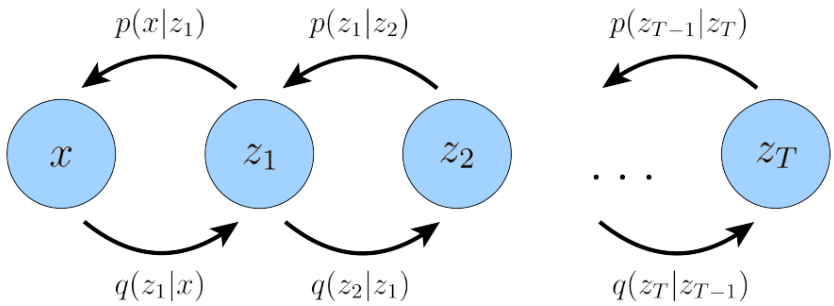

图 2.11 马尔可夫层级变分自编码器示意图

马尔可夫层级变分自编码器的联合概率分布可以表示为:

$$p(\boldsymbol{x}, \boldsymbol{z}_{1:T}) = p(\boldsymbol{x}_T)p_\theta(\boldsymbol{x} \mid \boldsymbol{z}_1)\prod_{t=2}^{T} p_\theta(\boldsymbol{z}_{t-1} \mid \boldsymbol{z}_t) \tag{2.42}$$

其后验概率可以表为:

$$q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}) = q_\phi(\boldsymbol{z}_1 \mid \boldsymbol{x})\prod_{t=2}^{T} q_\phi(\boldsymbol{z}_t \mid \boldsymbol{z}_{t-1}) \tag{2.43}$$

证据下界可以扩展为:

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log \int p(\boldsymbol{x}, \boldsymbol{z}_{1:T}) d\, z && \text{（根据式A.49）} && (2.44)\\
&= \log \int p(\boldsymbol{x}, \boldsymbol{z}_{1:T}) \frac{q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{x})}{q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{x})} d\, z && && (2.45)\\
&= \log \int \frac{p(\boldsymbol{x}, \boldsymbol{z}_{1:T}) q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{x})}{q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{x})} d\, z && && (2.46)\\
&= \log \int q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{x}) \frac{p(\boldsymbol{x}, \boldsymbol{z}_{1:T})}{q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{x})} d\, z && && (2.47)\\
&= \log \mathbb{E}_{q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\left[\frac{p(\boldsymbol{x}, \boldsymbol{z}_{1:T})}{q_\phi(\boldsymbol{z}_{1:T} \mid \boldsymbol{x})}\right] && \text{（根据期望定义式A.21）} && (2.48)
\end{aligned}$$

$$\geq \mathbb{E}_{q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z}_{1:T})}{q_\phi(\boldsymbol{z}_{1:T}\mid\boldsymbol{x})}\right] \qquad \text{（根据杰森不等式??）} \tag{2.49}$$

将式2.42与2.43代入式2.49可得：

$$\mathbb{E}_{q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z}_{1:T})}{q_\phi(\boldsymbol{z}_{1:T}\mid\boldsymbol{x})}\right] = \mathbb{E}_{q_\phi(\boldsymbol{z}_{1:T}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{z}_T)p_\theta(\boldsymbol{x}\mid\boldsymbol{z}_1)\prod_{t=2}^{T}p_\theta(\boldsymbol{z}_{t-1}\mid\boldsymbol{z}_t)}{q_\phi(\boldsymbol{z}_1\mid\boldsymbol{x})\prod_{t=2}^{T}q_\phi(\boldsymbol{z}_t\mid\boldsymbol{z}_{t-1})}\right] \tag{2.50}$$

在扩散模型中，式2.50可以进一步分解更具有意义的成分。

此外，由于自编码器与降噪自编码器并不直接定义似然函数，可以将自编码器和降噪自编码器视为隐式密度模型。而变分自编码器由于使用证据下界对似然函数进行近似，变分自编码器为近似估计的显示密度模型。

扩散模型

扩散模型，是近年来在生成图像领域最有影响力的模型之一。在很多评价基准上，扩散模型已经取得了相比生成对抗网络更好的成绩。就如 2017——2020 年生成对抗网络的流行，在 2022 年，扩散模型也被广泛地用于各项生成任务。

2015 年，扩散模型被从物理学中借鉴引入至深度学习领域[sohl2015deep]。2019 年，基于分数的模型 NCSN[song2019generative] 的提出，以及 2020 年对基于分数的模型训练方法的改进，都为扩散模型的诞生打下了良好的基础。

扩散模型可以理解为加一定限制的马尔可夫层级变分自编码器：

1. 隐变量维度与样本维度相同；
2. 每一时间（层级）隐变量编码器的结构为预先定义的线性高斯模型，并非学习获得，即将前一层隐变量值作为高斯分布的均值；
3. 隐变量编码器高斯分布的参数随着时间（层级）的变化而变化，且保证最终的隐变量分布为标准高斯分布。

图2.12为扩散模型示意图，$\boldsymbol{x}_0$ 表示真实样本，如自然图像，$\boldsymbol{x}_T$ 表示完全的高斯噪声，$\boldsymbol{x}_t$ 表示真实样本$\boldsymbol{x}_0$ 添加噪声后所得的隐变量。$q(\boldsymbol{x}_t\mid\boldsymbol{x}_{t-1})$ 为将前一状态作为均值的高斯分布。

由第一个假设，将真实样本和隐变量均由$\boldsymbol{x}_t$ 表示，其中$t=0$ 表示真实样本，$t\in[1,T]$ 表示相应的t 层隐变量，则扩散模型的后验与马尔可夫层级变分自编码器相似，根据式2.43得：

$$q(\boldsymbol{x}_{1:T}\mid\boldsymbol{x}_0)=\prod_{t=1}^{T}q(\boldsymbol{x}_t\mid\boldsymbol{x}_{t-1}) \tag{2.51}$$

由第二个假设，编码器中隐变量服从以前一隐变量为均值的高斯分布，与马尔可夫结构变分自编码器不同的是，编码器在每一时间步t 的结构并非学习获得，而

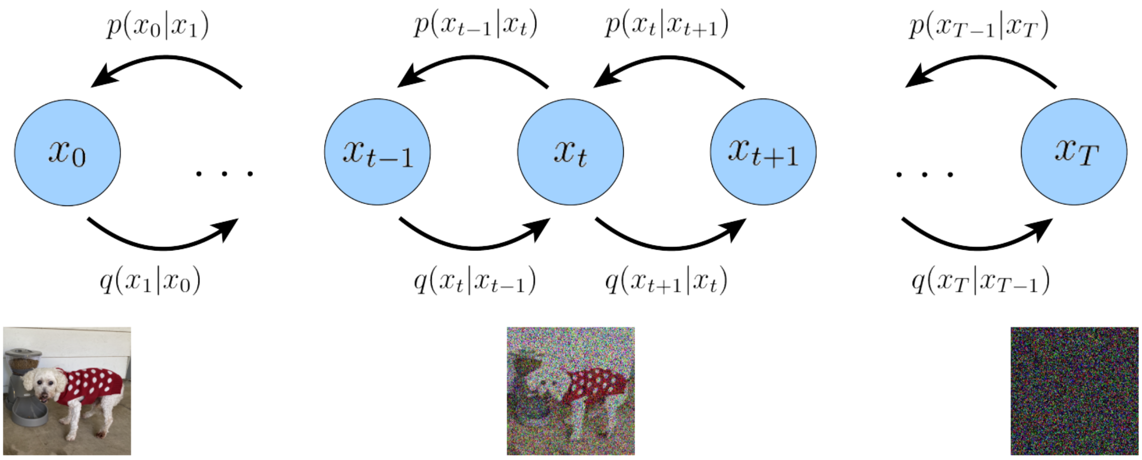

图 2.12 扩散模型示意图

是预先设定的线性高斯模型，均值和方差可以设为超参数[ho2020denoising]，或经学习获得[kingma2023variational]。设高斯编码器均值为$\boldsymbol{\mu}_t(\boldsymbol{x}_t) = \sqrt{\alpha_t}\boldsymbol{x}_{t-1}$，方差为$\boldsymbol{\Sigma}_t(\boldsymbol{x}_t) = (1-\alpha_t)\boldsymbol{I}$，以使隐变量的方差大小相近。为使模型更灵活，$\alpha_t$ 的值可以随着层级深度t 的变化而变化。编码转换过程可以表示为：

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}(\boldsymbol{x}_{t-1}), (1-\alpha_t)\boldsymbol{I}) \tag{2.52}$$

由第三个假设，预先定义或可学习获得的α_t 可以随着时间的变换而变化，且最深层隐变量$p(\boldsymbol{x}_T)$ 的分布为标准高斯分布，则马尔可夫层级变分自编码器的联合概率分布式2.42可以改写为：

$$p(\boldsymbol{x}_{0:T}) = p(\boldsymbol{x}_T)\prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t) \tag{2.53}$$

式2.53中，$p(\boldsymbol{x}_T) = \mathcal{N}(\boldsymbol{x}_T; \boldsymbol{0}, \boldsymbol{I})$。假设三表明，扩散模型逐渐向图片添加噪声直至图片变为完全的高斯噪声，即如图2.12所示。

相比于马尔可夫层级变分自编码器，编码器分布$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ 不再由参数ϕ 定义，而完全服从预先定义了均值与方差的高斯分布。因此，为能够生成新的样本点，扩散模型更注重对条件概率$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ 的学习。对扩散模型优化后，采样过程即为首先从高斯噪声$p(\boldsymbol{x}_T)$ 中采样，随后连续地进行T 次去噪变换$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$，最终获得新的样本点$\boldsymbol{x}_0$。

类比于层级变分自编码器，扩散模型可以通过最大化证据下界来优化，即：

$$\log p(\boldsymbol{x}) = \log \int p(\boldsymbol{x}_{0:T}) d\,\boldsymbol{x}_{1:T} \tag{2.54}$$

$$= \log \int p(\boldsymbol{x}_{0:T}) \frac{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} d\,\boldsymbol{x}_{1:T} \tag{2.55}$$

$$= \log \int q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0) \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} d\, \boldsymbol{x}_{1:T} \tag{2.56}$$

$$= \log \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} \right] \tag{2.57}$$

$$\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)} \right] \tag{2.58}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) \prod_{t=1}^{T} p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{\prod_{t=1}^{T} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \tag{2.59}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1) \prod_{t=2}^{T} p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \tag{2.60}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1) \prod_{t=1}^{T-1} p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \prod_{t=1}^{T-1} q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \tag{2.61}$$

$$= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p(\boldsymbol{x}_T) p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \prod_{t=1}^{T-1} \frac{p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \tag{2.62}$$

$$\begin{aligned} &= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} [\log p(\boldsymbol{x}_T)] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] \\ &\quad + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\sum_{t=1}^{T-1} \log \frac{p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \end{aligned} \tag{2.63}$$

$$\begin{aligned} &= \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} [\log p(\boldsymbol{x}_T)] + \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] \\ &\quad + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)} \left[\log \frac{p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \end{aligned} \tag{2.64}$$

$$\begin{aligned} &= \mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} [\log p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)] + \mathbb{E}_{q(\boldsymbol{x}_{T-1},\boldsymbol{x}_T|\boldsymbol{x}_0)} \left[\log \frac{p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)}{q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1})} \right] \\ &\quad + \sum_{t=1}^{T-1} \mathbb{E}_{q(\boldsymbol{x}_{T-1},\boldsymbol{x}_t,\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[\log \frac{p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})} \right] \end{aligned} \tag{2.65}$$

$$\begin{aligned} &= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} [\log p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)]}_{\text{重构项}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \| p(\boldsymbol{x}_T))]}_{\text{先验匹配项}} \\ &\quad - \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} [D_{KL}(q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) \| p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}))]}_{\text{一致检验项}} \end{aligned} \tag{2.66}$$

式2.66中：

- $\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} [\log p_\theta(\boldsymbol{x}_0 \mid \boldsymbol{x}_1)]$ 项为重构项，是对样本点关于第一层隐变量的对数条件概率的预测，与一般变分自编码器式2.36中的重构项类似，可由蒙特卡洛模拟近似计算与优化。
- $\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)}[D_{KL}(q(\boldsymbol{x}_T \mid \boldsymbol{x}_{T-1}) \| p(\boldsymbol{x}_T))]$ 项为先验匹配项，当最后一层隐变量的概率

分布为高斯分布时达到最小，先验匹配项由于没有需要学习的参数，因此无需进行优化，假设当T 足够大，则最后一层隐变量的概率分布为高斯分布，先验匹配项的值为0。

- $\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)}\left[D_{KL}(q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})\|p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1}))\right]$ 项为一致检验项，一致检验项使扩散模型的前向过程和反向过程在$\boldsymbol{x}_t$ 保持一致，即降噪过程获得的$p_\theta(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})$ 应与高斯分布$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ 尽可能一致。

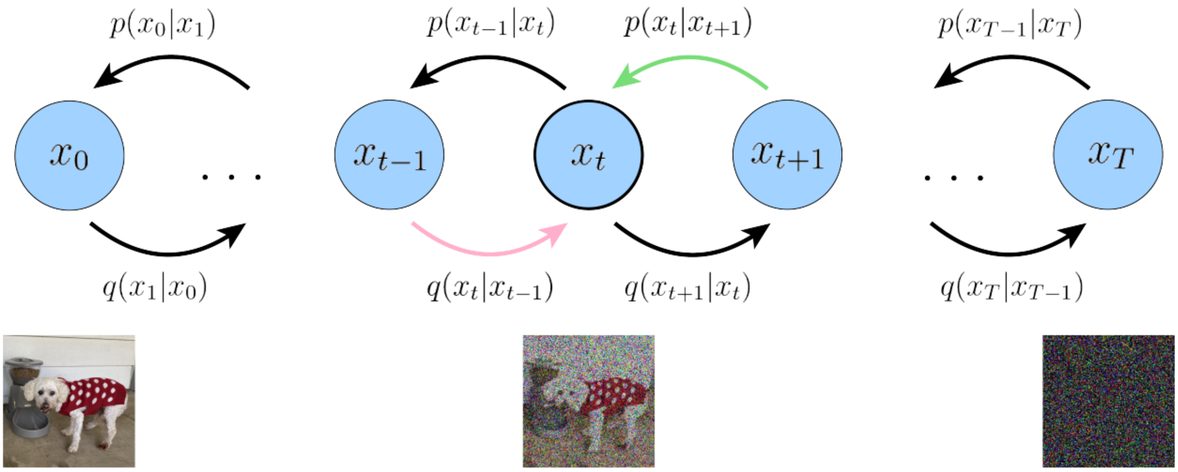

图 2.13 扩散模型证据下界一致检验项示意图

图2.13中，对于中间层隐变量$\boldsymbol{x}_t$，可以通过减小绿色箭头所代表的条件概率$p(\boldsymbol{x}_t \mid \boldsymbol{x}_{t+1})$ 与粉色箭头代表的高斯分布$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1})$ 的差异来优化，由于需要对所有时间步骤t 进行优化，扩散模型的优化时间取决于一致检验项。

式2.66将证据下界分解为不同的期望项，因此可以使用蒙特卡洛模拟来获得近似解，但由于一致检验项在每个时间步计算关于两个随机变量$\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t+1}$ 的期望，其蒙特卡洛估计的方差可能相比每个时间步只关于一个随机变量的项更高。此外，由于需要对$T-1$ 步一致检验项进行求和，当T 越大时，证据下界估计值的方差也越大。

由于马尔可夫性质，$\boldsymbol{x}_t$ 不依赖于$\boldsymbol{x}_0$，可将编码过程改写为$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0)$，则根据贝叶斯定理可得：

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0)} \tag{2.67}$$

根据式2.67，可由式2.58得：

$$\log p(\boldsymbol{x}) \tag{2.68}$$

$$\geq \mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log \frac{p(\boldsymbol{x}_{0:T})}{q(\boldsymbol{x}_{1:T} \mid \boldsymbol{x}_0)}\right] \tag{2.69}$$

$$=\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log\frac{p(\boldsymbol{x}_T)\prod_{t=1}^{T}p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t)}{\prod_{t=1}^{T}q(\boldsymbol{x}_t\mid\boldsymbol{x}_{t-1})}\right] \tag{2.70}$$

$$=\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log\frac{p(\boldsymbol{x}_T)p_\theta(\boldsymbol{x}_0\mid\boldsymbol{x}_1)\prod_{t=2}^{T}p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t)}{q(\boldsymbol{x}_1\mid\boldsymbol{x}_0)\prod_{t=2}^{T}q(\boldsymbol{x}_t\mid\boldsymbol{x}_{t-1})}\right] \tag{2.71}$$

$$=\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log\frac{p(\boldsymbol{x}_T)p_\theta(\boldsymbol{x}_0\mid\boldsymbol{x}_1)\prod_{t=2}^{T}p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t)}{q(\boldsymbol{x}_1\mid\boldsymbol{x}_0)\prod_{t=2}^{T}q(\boldsymbol{x}_t\mid\boldsymbol{x}_{t-1},\boldsymbol{x}_0)}\right] \tag{2.72}$$

$$=\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log\frac{p(\boldsymbol{x}_T)p_\theta(\boldsymbol{x}_0\mid\boldsymbol{x}_1)}{q(\boldsymbol{x}_1\mid\boldsymbol{x}_0)}+\log\prod_{t=2}^{T}\frac{p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t)}{q(\boldsymbol{x}_t\mid\boldsymbol{x}_{t-1},\boldsymbol{x}_0)}\right] \tag{2.73}$$

$$=\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log\frac{p(\boldsymbol{x}_T)p_\theta(\boldsymbol{x}_0\mid\boldsymbol{x}_1)}{q(\boldsymbol{x}_1\mid\boldsymbol{x}_0)}+\log\prod_{t=2}^{T}\frac{p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t)}{\frac{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)q(\boldsymbol{x}_t|\boldsymbol{x}_0)}{q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}}\right] \tag{2.74}$$

$$=\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log\frac{p(\boldsymbol{x}_T)p_\theta(\boldsymbol{x}_0\mid\boldsymbol{x}_1)}{q(\boldsymbol{x}_1\mid\boldsymbol{x}_0)}+\log\frac{q(\boldsymbol{x}_1\mid\boldsymbol{x}_0)}{q(\boldsymbol{x}_T\mid\boldsymbol{x}_0)}+\log\prod_{t=2}^{T}\frac{p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{x}_0)}\right] \tag{2.75}$$

$$=\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log\frac{p(\boldsymbol{x}_T)p_\theta(\boldsymbol{x}_0\mid\boldsymbol{x}_1)}{q(\boldsymbol{x}_T\mid\boldsymbol{x}_0)}+\sum_{t=2}^{T}\log\frac{p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{x}_0)}\right] \tag{2.76}$$

$$\begin{aligned}=&\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log p_\theta(\boldsymbol{x}_0\mid\boldsymbol{x}_1)\right]+\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log\frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T\mid\boldsymbol{x}_0)}\right]\\&+\sum_{t=2}^{T}\mathbb{E}_{q(\boldsymbol{x}_{1:T}|\boldsymbol{x}_0)}\left[\log\frac{p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{x}_0)}\right]\end{aligned} \tag{2.77}$$

$$\begin{aligned}=&\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_\theta(\boldsymbol{x}_0\mid\boldsymbol{x}_1)\right]+\mathbb{E}_{q(\boldsymbol{x}_T|\boldsymbol{x}_0)}\left[\log\frac{p(\boldsymbol{x}_T)}{q(\boldsymbol{x}_T\mid\boldsymbol{x}_0)}\right]\\&+\sum_{t=2}^{T}\mathbb{E}_{q(\boldsymbol{x}_t,\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}\left[\log\frac{p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t)}{q(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{x}_0)}\right]\end{aligned} \tag{2.78}$$

$$\begin{aligned}=&\underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_\theta(\boldsymbol{x}_0\mid\boldsymbol{x}_1)\right]}_{\text{重构项}}-\underbrace{D_{KL}(q(\boldsymbol{x}_T\mid\boldsymbol{x}_0)\|p(\boldsymbol{x}_T))}_{\text{先验匹配项}}\\&-\sum_{t=2}^{T}\underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{KL}(q(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t))\right]}_{\text{降噪匹配项}}\end{aligned} \tag{2.79}$$

式2.79相比式2.66，每一项均为最多只关于一个随机变量的期望。类比对式2.66的解释，式2.79有如下解释：

- $\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_\theta(\boldsymbol{x}_0\mid\boldsymbol{x}_1)\right]$ 项为重构项，与变分自编码器证据下界中的重构项类似。可以蒙特卡洛模拟方法近似估计与优化。
- $D_{KL}(q(\boldsymbol{x}_T\mid\boldsymbol{x}_0)\|p(\boldsymbol{x}_T))$ 项为先验匹配项，表明不断添加噪声后的图像与高斯分布的匹配程度。先验匹配项在扩散模型假设下为0。
- $\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{KL}(q(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t))\right]$ 为降噪匹配项，扩散模型用降噪变换$p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t)$ 来近似真实降噪变换$q(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{x}_0)$。通过最小化降噪匹配项，可

以使两个降噪变换逐渐接近。

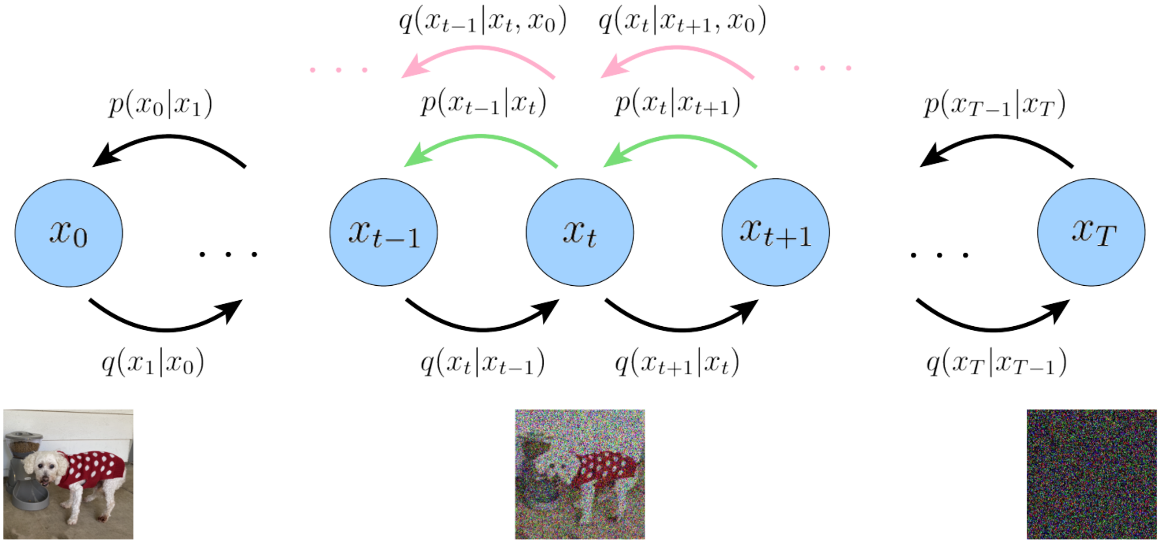

图 2.14 扩散模型证据下界降噪匹配项单变量期望示意图

图2.14为使用式2.79优化扩散模型示意图，使用贝叶斯定理计算真实降噪过程$q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)$，并计算其关于近似降噪过程$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ 的 KL 散度，即尽可能使绿色箭头所代表分布与粉色箭头所代表分布一致。

KL 散度项$D_{KL}(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \| p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t))$ 由于需要同时对多个时间步的编码器进行学习，难以进行最小化。为使 KL 散度项更易优化，根据贝叶斯定理有：

$$q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) = \frac{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0) q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0)}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)} \tag{2.80}$$

根据式2.52，可得：

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0) = q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}(\boldsymbol{x}_{t-1}), (1-\alpha_t)\boldsymbol{I}) \tag{2.81}$$

由于扩散模型编码器为线性高斯模型，根据重参数方法，对于$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t, \boldsymbol{x}_{t-1})$ 可得：

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}\text{其中，} \epsilon \sim \mathcal{N}(\boldsymbol{\epsilon}; \boldsymbol{0}, \boldsymbol{I}) \tag{2.82}$$

对于$\boldsymbol{x}_{t-1} \sim q(\boldsymbol{x}_{t-1}, \boldsymbol{x}_{t-2})$ 可得：

$$\boldsymbol{x}_{t-1} = \sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}\text{其中，} \epsilon \sim \mathcal{N}(\boldsymbol{\epsilon}; \boldsymbol{0}, \boldsymbol{I}) \tag{2.83}$$

可以使用重参数方法对$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)$ 进行递归推导，假设有$2T$ 个随机噪声变量$\{\boldsymbol{\epsilon}_t^*, \boldsymbol{\epsilon}_t\}_{t=0}^{T} \overset{iid}{\sim} \mathcal{N}(\boldsymbol{\epsilon}; \boldsymbol{0}, \boldsymbol{I})$，则对于任意采样$\boldsymbol{x}_t \sim q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)$，有：

$$\boldsymbol{x}_t = \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}_{t-1}^* \tag{2.84}$$

$$= \sqrt{\alpha_t}(\sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2}) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \tag{2.85}$$

$$= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \tag{2.86}$$

$$= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\sqrt{\alpha_t - \alpha_t\alpha_{t-1}}^2 + \sqrt{1-\alpha_t}^2}\boldsymbol{\epsilon}_{t-2} \qquad \text{根据式A.19} \tag{2.87}$$

$$= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t - \alpha_t\alpha_{t-1} + 1 - \alpha_t}\boldsymbol{\epsilon}_{t-2} \tag{2.88}$$

$$= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1 - \alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2} \tag{2.89}$$

$$= \ldots \tag{2.90}$$

$$= \sqrt{\prod_{i=1}^{t}\alpha_i}\boldsymbol{x}_0 + \sqrt{1 - \prod_{i=1}^{t}\alpha_i}\boldsymbol{\epsilon}_0 \tag{2.91}$$

$$= \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0 \tag{2.92}$$

$$\sim \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\boldsymbol{I}) \tag{2.93}$$

与式2.93类似，有:

$$\boldsymbol{x}_{t-1} \sim \mathcal{N}(\boldsymbol{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0, (1-\bar{\alpha}_{t-1})\boldsymbol{I}) \tag{2.94}$$

由式2.80、式2.81、式2.93和式2.94，可得:

$$q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \tag{2.95}$$

$$= \frac{q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}, \boldsymbol{x}_0)q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_0)}{q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)} \tag{2.96}$$

$$\propto \exp\{-\frac{1}{2}\left[\frac{(\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{1-\alpha_t} + \frac{(\boldsymbol{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0)^2}{1-\bar{\alpha}_{t-1}} - \frac{(\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0)^2}{1-\bar{\alpha}_t}\right]\} \tag{2.97}$$

$$= \exp\{-\frac{1}{2}\left[\frac{-2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1} + \alpha_t\boldsymbol{x}_{t-1}^2}{1-\alpha_t} + \frac{(\boldsymbol{x}_{t-1}^2 - 2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0)^2}{1-\bar{\alpha}_{t-1}} + C(\boldsymbol{x}_t, \boldsymbol{x}_0)\right]\} \tag{2.98}$$

$$\propto \exp\{-\frac{1}{2}\left[-\frac{2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1}}{1-\alpha_t} + \frac{\alpha_t\boldsymbol{x}_{t-1}^2}{1-\alpha_t} + \frac{\boldsymbol{x}_{t-1}^2}{1-\bar{\alpha}_{t-1}} - \frac{2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right]\} \tag{2.99}$$

$$= \exp\{-\frac{1}{2}\left[\left(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\} \tag{2.100}$$

$$= \exp\{-\frac{1}{2}\left[\frac{\alpha_t(1-\bar{\alpha}_{t-1}) + 1 - \alpha_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\} \tag{2.101}$$

$$= \exp\{-\frac{1}{2}\left[\frac{\alpha_t - \bar{\alpha}_t + 1 - \alpha_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\} \tag{2.102}$$

$$= \exp\{-\frac{1}{2}\left[\frac{1 - \bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}_{t-1}^2 - 2\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)\boldsymbol{x}_{t-1}\right]\} \tag{2.103}$$

$$= \exp\{-\frac{1}{2}\left(\frac{1 - \bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right)\left[\boldsymbol{x}_{t-1}^2 - 2\frac{\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t} + \frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)}{\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}}\boldsymbol{x}_{t-1}\right]\} \tag{2.104}$$

$$
\begin{aligned}
&= \exp\{-\frac{1}{2}\left(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right) \\
&\qquad \left[\boldsymbol{x}_{t-1}^2 - 2\frac{\left(\frac{\sqrt{\alpha_t}\boldsymbol{x}_t}{1-\alpha_t}+\frac{\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0}{1-\bar{\alpha}_{t-1}}\right)(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_{t-1}\right]\} &\quad (2.105)\\
&= \exp\{-\frac{1}{2}\left(\frac{1}{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}}\right)\left[\boldsymbol{x}_{t-1}^2 - 2\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t+\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}\boldsymbol{x}_{t-1}\right]\} \\
& &\quad (2.106)\\
&\propto \mathcal{N}(\boldsymbol{x}_{t-1}; \underbrace{\frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t+\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\boldsymbol{x}_0}{1-\bar{\alpha}_t}}_{\mu_q(\boldsymbol{x}_t,\boldsymbol{x}_0)}, \underbrace{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{I}}_{\boldsymbol{\Sigma}_q(t)}) &\quad (2.107)
\end{aligned}
$$

式2.98中，$C(\boldsymbol{x}_t, \boldsymbol{x}_0)$ 为由$\boldsymbol{x}_t, \boldsymbol{x}_0, \alpha$ 组成的相对于$\boldsymbol{x}_{t-1}$ 的常数项，可与式2.106经过配方法式??，获得式2.107。即$\boldsymbol{x}_{t-1} \sim q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)$ 为正态分布，均值$\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)$ 为$\boldsymbol{x}_t, \boldsymbol{x}_0$ 的函数，方差$\boldsymbol{\Sigma}_q(t)$ 为α 的函数，α 为已知的关于时间步的确定量，可以为超参数或由神经网络学习获得。根据式2.107，可得：

$$
\Sigma_q(t) = \sigma_q^2(t)\boldsymbol{I} = \frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{I} \tag{2.108}
$$

为使近似降噪转换$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ 与真实降噪转换$q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0)$ 更加接近，可以使$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ 服从高斯分布。此外，由于α 在转换过程中为固定值，可以令近似降噪转换的方差与真实降噪转换的方差相同，即$\boldsymbol{\Sigma}_q(t) = \sigma_q^2(t)\boldsymbol{I}$。而对于高斯分布的均值，由于真实降噪转换的高斯分布均值$\mu_q(\boldsymbol{x}_t, \boldsymbol{x}_0)$ 为$\boldsymbol{x}_t, \boldsymbol{x}_0$ 的函数，近似降噪转换$p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)$ 却不含有$\boldsymbol{x}_0$，可将近似降噪转换高斯分布的均值设为$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t, t)$。

根据两高斯分布 KL 散度计算式??，由于近似降噪转换的方差与真实降噪转换的方差相同，最小化 KL 散度即转换为最小化两分布的均值，即：

$$
\begin{aligned}
&\arg\min_\theta D_{KL}(q(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t, \boldsymbol{x}_0) \| p_\theta(\boldsymbol{x}_{t-1} \mid \boldsymbol{x}_t)) &\quad (2.109)\\
&= \arg\min_\theta D_{KL}(\mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_q, \boldsymbol{\Sigma}_q(t)) \| \mathcal{N}(\boldsymbol{x}_{t-1}; \boldsymbol{\mu}_\theta, \boldsymbol{\Sigma}_q(t))) &\quad (2.110)\\
&= \arg\min_\theta \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_q(t)|}{|\boldsymbol{\Sigma}_q(t)|} - d + tr(\boldsymbol{\Sigma}_q(t)^{-1}\boldsymbol{\Sigma}_q(t)) + (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)^\intercal \boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)\right] \\
& &\quad (2.111)\\
&= \arg\min_\theta \frac{1}{2}\left[\log 1 - d + d + (\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)^\intercal \boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)\right] &\quad (2.112)\\
&= \arg\min_\theta \frac{1}{2}\left[(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)^\intercal \boldsymbol{\Sigma}_q(t)^{-1}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)\right] &\quad (2.113)\\
&= \arg\min_\theta \frac{1}{2}\left[(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)^\intercal \sigma_q^2(t)\boldsymbol{I}^{-1}(\boldsymbol{\mu}_\theta - \boldsymbol{\mu}_q)\right] &\quad (2.114)
\end{aligned}
$$

$$=\arg\min_{\theta} \frac{1}{2(\sigma_q^2(t))}\left[\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_q\|_2^2\right] \tag{2.115}$$

式2.115中，$\boldsymbol{\mu}_q$ 为$\boldsymbol{\mu}_q(\boldsymbol{x}_t,\boldsymbol{x}_0)$ 的缩写，$\boldsymbol{\mu}_\theta$ 为$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t)$ 的缩写。由式2.107，真实降噪转换分布的均值为：

$$\boldsymbol{\mu}_q(\boldsymbol{x}_t,\boldsymbol{x}_0)=\frac{\alpha_t(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t+\sqrt{\bar{\alpha}_{t-1}}1-\alpha_t\boldsymbol{x}_0}{1-\bar{\alpha}_t} \tag{2.116}$$

为使近似降噪转换的均值$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t)$ 与真实降噪转换的均值$\boldsymbol{\mu}_q(\boldsymbol{x}_t,\boldsymbol{x}_0)$ 的均值尽可能一致，由于近似降噪转换的均值$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t)$ 也包含$\boldsymbol{x}_t$，可令两均值具有相似形式，即：

$$\boldsymbol{\mu}_\theta(\boldsymbol{x}_t,t)=\frac{\alpha_t(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t+\sqrt{\bar{\alpha}_{t-1}}1-\alpha_t\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)}{1-\bar{\alpha}_t} \tag{2.117}$$

式2.117中，$\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)$ 为由噪声图像$\boldsymbol{x}_t$ 和时间索引t 预测原始样本$\boldsymbol{x}_0$ 的神经网络。则优化问题可转化为：

$$\arg\min_{\theta} D_{KL}(q(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t)) \tag{2.118}$$

$$=\arg\min_{\theta} D_{KL}(\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_q,\boldsymbol{\Sigma}_q(t))\|\mathcal{N}(\boldsymbol{x}_{t-1};\boldsymbol{\mu}_\theta,\boldsymbol{\Sigma}_q(t))) \tag{2.119}$$

$$=\arg\min_{\theta} \frac{1}{2(\sigma_q^2(t))}\left[\|\boldsymbol{\mu}_\theta-\boldsymbol{\mu}_q\|_2^2\right] \tag{2.120}$$

$$=\arg\min_{\theta} \frac{1}{2(\sigma_q^2(t))}\left[\left\|\frac{\alpha_t(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t+\sqrt{\bar{\alpha}_{t-1}}1-\alpha_t\boldsymbol{x}_0}{1-\bar{\alpha}_t}\right.\right. \tag{2.121}$$

$$\left.\left.-\frac{\alpha_t(1-\bar{\alpha}_{t-1})\boldsymbol{x}_t+\sqrt{\bar{\alpha}_{t-1}}1-\alpha_t\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)}{1-\bar{\alpha}_t}\right\|_2^2\right] \tag{2.122}$$

$$=\arg\min_{\theta} \frac{1}{2(\sigma_q^2(t))}\left[\left\|\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)}{1-\bar{\alpha}_t}-\frac{\sqrt{\bar{\alpha}_{t-1}}1-\alpha_t\boldsymbol{x}_0}{1-\bar{\alpha}_t}\right\|_2^2\right] \tag{2.123}$$

$$=\arg\min_{\theta} \frac{1}{2(\sigma_q^2(t))}\left[\left\|\frac{\sqrt{\bar{\alpha}_{t-1}}(1-\alpha_t)}{1-\bar{\alpha}_t}(\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0)\right\|_2^2\right] \tag{2.124}$$

$$=\arg\min_{\theta} \frac{1}{2(\sigma_q^2(t))}\frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{1-\bar{\alpha}_t{}^2}\left[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|_2^2\right] \tag{2.125}$$

由式2.125，对扩散模型的优化可以转化为对神经网络的训练，该神经网络可以从任意时刻噪声图像来预测真实样本。此外，对式2.79中的降噪匹配求和项，可以通过最小化所有时刻的期望η 来近似，即：

$$\arg\min_{\theta} \mathbb{E}_{t\sim U\{2,T\}}\left[\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{KL}(q(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t,\boldsymbol{x}_0)\|p_\theta(\boldsymbol{x}_{t-1}\mid\boldsymbol{x}_t))\right]\right] \tag{2.126}$$

噪声参数α_t 可设置为超参数，也可参与训练学习获得。可以使用以η 为参数的神经网络$\hat{\alpha}_{\boldsymbol{\eta}}(t)$ 来获得α_t，但由于每个时间步t 都需计算α_t，效率较低，可将式2.108代入

式2.125，可得：

$$\frac{1}{2(\sigma_q^2(t))}\frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{1-\bar{\alpha}_t{}^2}\left[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|_2^2\right] \tag{2.127}$$

$$=\frac{1}{2\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}}\frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{1-\bar{\alpha}_t{}^2}\left[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|_2^2\right] \tag{2.128}$$

$$=\frac{1}{2}\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{1-\bar{\alpha}_t{}^2}\left[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|_2^2\right] \tag{2.129}$$

$$=\frac{1}{2}\frac{\bar{\alpha}_{t-1}(1-\alpha_t)}{(1-\bar{\alpha}_{t-1})(1-\bar{\alpha}_t)}\left[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|_2^2\right] \tag{2.130}$$

$$=\frac{1}{2}\frac{\bar{\alpha}_{t-1}-\bar{\alpha}_t}{(1-\bar{\alpha}_{t-1})(1-\bar{\alpha}_t)}\left[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|_2^2\right] \tag{2.131}$$

$$=\frac{1}{2}\frac{\bar{\alpha}_{t-1}-\bar{\alpha}_t\bar{\alpha}_{t-1}+\bar{\alpha}_t\bar{\alpha}_{t-1}-\bar{\alpha}_t}{(1-\bar{\alpha}_{t-1})(1-\bar{\alpha}_t)}\left[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|_2^2\right] \tag{2.132}$$

$$=\frac{1}{2}\frac{\bar{\alpha}_{t-1}(1-\bar{\alpha}_t)-\bar{\alpha}_t(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_{t-1})(1-\bar{\alpha}_t)}\left[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|_2^2\right] \tag{2.133}$$

$$=\frac{1}{2}\left(\frac{\bar{\alpha}_{t-1}(1-\bar{\alpha}_t)}{(1-\bar{\alpha}_{t-1})(1-\bar{\alpha}_t)}-\frac{\bar{\alpha}_t(1-\bar{\alpha}_{t-1})}{(1-\bar{\alpha}_{t-1})(1-\bar{\alpha}_t)}\right)\left[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|_2^2\right] \tag{2.134}$$

$$=\frac{1}{2}\left(\frac{\bar{\alpha}_{t-1}}{1-\bar{\alpha}_{t-1}}-\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}\right)\left[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|_2^2\right] \tag{2.135}$$

根据式2.93，$q(\boldsymbol{x}_t \mid \boldsymbol{x}_0)$ 为服从高斯分布$\mathcal{N}(\boldsymbol{x}_t;\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0,(1-\bar{\alpha}_t)\boldsymbol{I})$，则根据信噪比定义式??可得：

$$SNR(t)=\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t} \tag{2.136}$$

根据式2.136可将式2.135改写为：

$$\frac{1}{2(\sigma_q^2(t))}\frac{\bar{\alpha}_{t-1}(1-\alpha_t)^2}{1-\bar{\alpha}_t{}^2}\left[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|_2^2\right] \tag{2.137}$$

$$=\frac{1}{2}\left(\frac{\bar{\alpha}_{t-1}}{1-\bar{\alpha}_{t-1}}-\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t}\right)\left[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|_2^2\right] \tag{2.138}$$

$$=\frac{1}{2}\left(SNR(t-1)-SNR(t)\right)\left[\|\hat{\boldsymbol{x}}_\theta(\boldsymbol{x}_t,t)-\boldsymbol{x}_0\|_2^2\right] \tag{2.139}$$

根据信噪比定义，信噪比越高，则包含更多的信息，信噪比越低，则噪声越多。扩散模型的信噪比随着时间步t 的增加而单调减小，即$\boldsymbol{x}_t$ 随着时间推移而逐渐包含更多的噪声，直至在$t=T$ 时变为标准高斯分布。

为简化式2.139，可使用神经网络来模拟任意时间的信噪比，并与扩散模型同时训练。即：

$$SNR(t)=\exp(-\omega_{\boldsymbol{\eta}}(t)) \tag{2.140}$$

式2.140中，$\omega_{\boldsymbol{\eta}}(t)$ 为单调递增函数，$-\omega_{\boldsymbol{\eta}}(t)$ 为单调递减函数，进行指数化以使信噪比为正数。此时，使用神经网络模拟任意时间信噪比，式2.126需要同时对$\boldsymbol{\eta}$ 进行优化。

此外，由式2.136、式2.140和式??，可得：

$$\frac{\bar{\alpha}_t}{1-\bar{\alpha}_t} = \exp(-\omega_{\boldsymbol{\eta}}(t)) \tag{2.141}$$

$$\bar{\alpha}_t = sigmoid(-\omega_{\boldsymbol{\eta}}(t)) \tag{2.142}$$

$$1-\bar{\alpha}_t = sigmoid(\omega_{\boldsymbol{\eta}}(t)) \tag{2.143}$$

式2.142和式2.143可用于式2.106中使用重参数方法从原始样本$\boldsymbol{x}_0$ 生成任意时刻添加噪声后的图像$\boldsymbol{x}_t$。

2.3 隐式密度模型

隐式密度模型不定义概率密度函数$p_{model}(\boldsymbol{x};\boldsymbol{\theta})$，不通过最大化似然函数来训练，而通过其他方式与实际概率密度函数p_{data} 进行交互。

2.3.1 生成对抗网络

生成对抗网络可以视为根据博弈论设计的一场游戏，包含两个机器学习模型，机器学习模型通常采用神经网络。

其中一个神经网络为生成器，其隐式地确定了概率分布p_{model}，生成器不需要对概率分布p_{model} 进行评估。根据一个先验分布$p(\boldsymbol{z})$ 以及生成函数方程$G(\boldsymbol{z};\theta^{(G)})$，生成器可以从概率分布$p_{model}$ 中采样，其中可学习参数$\theta^{(G)}$ 确定了生成器在博弈中采取的策略，先验分布$p(\boldsymbol{z})$ 经常为相对而言无特殊结构的分布，比如高维高斯分布或超立方体均匀分布，从这类分布中采样的$\boldsymbol{z}$ 为随机噪声。生成器的主要任务是通过学习将随机噪声$\boldsymbol{z}$ 转化为能够以假乱真的样本。

另一个神经网络为辨别器，辨别器可以通过辨别函数$D(\boldsymbol{x};\boldsymbol{\theta}^{(D)})$，判断一个样本$x$是从训练集中采样的真实样本或是由生成器生成的伪造样本。

在生成对抗网络游戏中，每一个玩家都有其本身的损失函数，生成器的损失函数为$J^G(\boldsymbol{\theta}^{(G)},\boldsymbol{\theta}^{(D)})$，辨别器的损失函数为$J^D(\boldsymbol{\theta}^{(G)},\boldsymbol{\theta}^{(D)})$。每一个玩家都希望能够最小化其损失函数。辨别器的损失函数使其能够更好地辨别一个样本是真实样本还是伪造样本，生成器的损失函数使其能够制造出可以迷惑辨别器的伪造样本，使辨别器将伪造样本误认为真实样本。

生成对抗网络可以形象的表示成警察和造假者的博弈，造假者尝试制造假币，而警察在尽力逮捕造假者的同时，还要保证正规货币的流通。随着警察和造假者的博弈，造假者的“造假”能力逐渐提高，直至警察无法分辨。对于生成对抗网络的训练过程，

可以比喻为造假者在警察中有内线，其可以向造假者提供警方辨别真假的方法，以让造假者制造更具有迷惑性的假币。

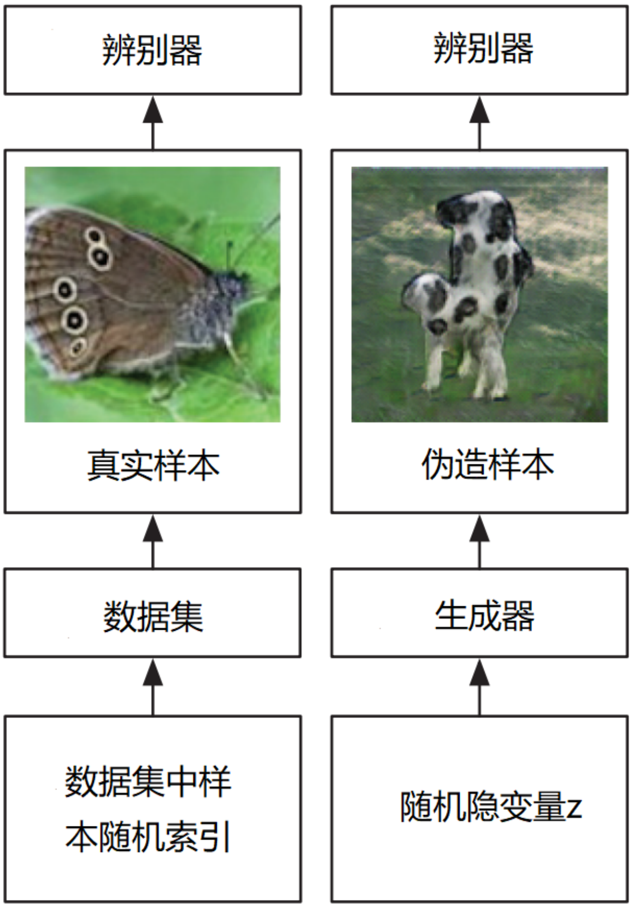

图 2.15 生成对抗网络训练过程

图2.15为生成对抗网络训练过程，生成对抗网络的训练包括两部分：对生成网络的训练和对辨别网络的训练。训练的过程包括不断地从数据集中获得真实样本和由生成网络生成伪造样本，辨别网络的训练与其他辨别深度神经网络的训练类似。在图2.15的左侧，向辨别器输入从数据集中获取的真实样本，其对应的标签为真；在图2.15的右侧，向辨别器输入由生成网络生成的伪造样本，其对应的标签为假。从隐变量先验分布中采样获得随机向量$\boldsymbol{z}$，再将该随机变量$\boldsymbol{z}$ 输入生成器，即可得伪造样本$\boldsymbol{x} = G(\boldsymbol{z})$。生成器生成函数$G$ 由神经网络表示，其可以将随机无结构的向量$\boldsymbol{z}$ 转换为伪造样本$\boldsymbol{x}$，且该伪造样本要尽可能与训练数据集中的样本在统计意义上难以分辨。通过反向传播算法，可以借助辨别器输出相对于对辨别器输入的梯度来训练生成器。生成网络训练的方向为，使其生成的伪造样本能够更多地被辨别器判定为真实样本。辨别器的训练与其他辨别网络训练类似，唯一区别是标记为伪造样本类别的概率分布，可以随着生成器的训练而不断变化。

2.4 生成模型评价指标

在生成模型研究中，为了证明一种方法比令一种方法更好，往往需要有一个评价标准，即生成模型评价指标。

2.4.1 图灵测试

为了让生成模型生成的图像更加真实，可以借鉴图灵测试，即让人们判断生成图像的质量与真实图像相比如何。但一些模型由于过拟合问题，可能仅仅对原始样本进行记忆，也可以生成真实度很高的图像。

2.4.2 图像质量评估分数

Inception 分数[salimans2016improved] 是为判断生成对抗网络生成图像的质量而提出的一种计算分数，生成图像的多样性和质量都会对影响 Inception 分数。Inception 分数通过计算标签关于生成图像的条件概率$p(y \mid \boldsymbol{x})$ 来评估图像质量。Inception 分数有以下假设:

- 某标签关于生成图像的条件分布$p(y \mid \boldsymbol{x})$ 应当具有较低的信息熵，即越真实的图像越有可能属于某一预设类别。
- 可以生成多种类别图像的模型，其边缘分布$\int p(y \mid \boldsymbol{x} = G(\boldsymbol{z}))d\,z$ 应当具备更高的信息熵，即生成的图像应当均匀地属于不同类别。

$$IS = \exp\left(\mathbb{E}_{\boldsymbol{x}\sim p_G} D_{KL}(p(y \mid \boldsymbol{x}) \mid\mid p(y))\right) \tag{2.144}$$

式2.144中，$\boldsymbol{x}$ 表示生成器生成的样本。$p(y \mid \boldsymbol{x})$ 表示标签y 关于生成样本$\boldsymbol{x}$ 的条件概率。$p(y)$ 表示关于标签y 的边缘分布。$D_{KL}(P \mid\mid Q)$ 表示 KL 散度，即式**??**。进行指数化以便于比较。Inception 分数越高，则图片生成的质量越好。

Inception 分数的不足之处是没有使用真实样本的统计性质。

FID[heusel2017gans]（Fréchet Inception Distance）使用在 ImageNet 数据集上预训练的 Inception V3 模型，计算生成图像和真实图像特征向量之间的距离，越低的分数表明两组图像越相似，分数为 0 则代表两组图像完全相同。假设两组图像在 Inception V3 特征层特征向量的均值分别为$\boldsymbol{m}_{model}, \boldsymbol{m}_{real}$，方差分别为$\boldsymbol{C}_{model}, \boldsymbol{C}_{real}$，则:

$$FID = \|\boldsymbol{m}_{model} - \boldsymbol{m}_{real}\|_2^2 + Tr(\boldsymbol{C}_{model} + \boldsymbol{C}_{real} - 2\sqrt{(\boldsymbol{C}_{model}\boldsymbol{C}_{real})}) \tag{2.145}$$

KID[binkowski2018demystifying]（Kernel Inception Distance）计算生成图像和真实图像特征向量之间最大平均距离的平方，特征向量也由与训练 Inception 模型获得。根据

式??，有

$$MMD^2(X,Y) = \frac{1}{m(m-1)}\sum_{i\neq j}^{m} k(x_i,x_j) + \frac{1}{n(n-1)}\sum_{i\neq j}^{n} k(y_i,y_j) - \frac{2}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} k(x_i,y_j) \tag{2.146}$$

式2.146中，m 是生成图像的样本量，n 是真实图像的样本量，每个x 和y 均是 Inception 网络的特征表示层的 2048 维向量，且

$$k(x,y) = \frac{1}{d}x^{\intercal}y + 1 \tag{2.147}$$

相比于 FID，KID 不假定特征向量概率分布为高斯分布[betzalel2022study]。

3 变分自编码器模型生成中国画

3.1 模型概述

3.2 模型结构

3.3 模型训练

3.4 实验结果

4 自回归模型生成中国画

4.1 模型概述

4.2 模型结构

4.3 模型训练

4.4 实验结果

致谢

衷心感谢导师 ××× 教授和 xx 系 ×× 副教授对本人的精心指导。他们的言传身教将使我终生受益。

在河北工业大学⋯⋯研究期间，承蒙教授热心指导与帮助，不胜感激。

感谢 ××××× 实验室主任 ××× 教授，以及实验室全体老师和同窗们学的热情帮助和支持！

本课题承蒙国家自然科学基金资助，特此致谢。

附录 A 基础知识

A.1 数学

A.1.1 线性代数

对角矩阵

除主对角线之外的元素皆为0 的矩阵。

矩阵乘法

如果A 是一个$l \times m$ 矩阵，B 是一个$m \times n$ 矩阵，则AB 是一个$l \times n$ 矩阵。

矩阵的迹

矩阵的迹为矩阵所有对角元素之和，即：

$$Tr(A) = \sum_i A_{i,i} \tag{A.1}$$

范数

范数可以表示向量的大小，范数定义为:

$$\|\boldsymbol{x}\|_p = \left(\sum_i \mid x_i \mid^p \right)^{\frac{1}{p}} \tag{A.2}$$

式A.2中，$p \in \mathbb{R}, p \geq 1$

L^1 范数为：

$$\|\boldsymbol{x}\|_1 = \sum_i \mid x_i \mid \tag{A.3}$$

最大范数为：

$$\|\boldsymbol{x}\|_\infty = \max_i \mid x_i \mid \tag{A.4}$$

矩阵的 Frobenius 范数为：

$$\|\boldsymbol{x}\|_F = \sqrt{\sum_{i,j} A_{i,j}^2} = \sqrt{Tr(AA^T)} \tag{A.5}$$

矩阵行列式

矩阵的行列式是一个关于方形矩阵（方阵）内所有元素的标量函数值。一个$n \times n$的矩阵M 的行列式为：

$$\det M = \det \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix} = \sum_{j_1 j_2 \cdots j_n} (-1)^{\tau(j_1 j_2 \cdots j_n)} a_{1j_1} a_{2j_2} \ldots a_{nj_n} \tag{A.6}$$

式A.6中，求和符号$\sum$ 的下标$j_1 j_2 \cdots j_n$ 表示集合$1, 2, \ldots, n$ 的全排列，即共有n 项；$\tau(j_1 j_2 \cdots j_n)$ 表示排列$j_1 j_2 \cdots j_n$ 的符号，排列的符号定义为式??。

方阵的行列式可以用来判断矩阵M 是否可逆：如果$\det M = 0$，那么矩阵不可逆。

矩阵乘积的行列式等于矩阵行列式的乘积，即：

$$\det(AB) = \det(A)\det(B) \tag{A.7}$$

可逆矩阵

如果一个$n \times n$ 的矩阵A 可逆，那么存在$n \times n$ 矩阵B 使得：

$$AB = BA = I_n \tag{A.8}$$

式A.8中，I_n 为单位矩阵，并且使用矩阵乘法。

不可逆的方阵又称为奇异矩阵或退化矩阵。

对于可逆矩阵有：

$$\det(M)\det(M^{-1}) = \det(M\dot{M}^{-1}) = \det(I) = 1 \tag{A.9}$$

因此，

$$\det(M^{-1}) = (\det(M))^{-1} \tag{A.10}$$

雅可比矩阵

假设某函数$\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$，从$\boldsymbol{x} \in \mathbb{R}^n$ 映射到向量$\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m$，这个函数的一阶偏导矩阵称为雅可比矩阵，是一个$m \times n$ 的矩阵。其第i 行，第j 列的值为$\boldsymbol{J}_{ij} = \frac{\partial f_i}{\partial x_i}$

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \tag{A.11}$$

A.1.2 概率论

概率密度函数

一个连续随机变量X 的概率可以由概率密度函数表示：

$$p(X \in A) = \int_A f_{\boldsymbol{X}}(x)dx \tag{A.12}$$

概率质量函数

一个离散随机变量X 的概率可以由概率质量函数表示：

$$p_X(x) = p(\{X = x\}) \tag{A.13}$$

正态分布

若实值随机变量X 服从正态分布（高斯分布），则其概率密度函数为：

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp(\frac{1}{2}(\frac{x-\mu}{\sigma})^2) \tag{A.14}$$

记为：$X \sim \mathcal{N}(\mu, \sigma^2)$

两高斯分布之和仍为高斯分布，假设X, Y 为两独立随机变量，且

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2) \tag{A.15}$$

$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2) \tag{A.16}$$

$$\tag{A.17}$$

则X, Y 之和仍为高斯分布，即：

$$Z = X + Y \tag{A.18}$$

$$Z \sim \mathcal{N}(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2) \tag{A.19}$$

期望

当$x \sim p(x)$，函数$f(x)$ 的期望为：

$$\mathbb{E}_{x\sim p}[f(x)] = \sum_x p(x)f(x) \tag{A.20}$$

或

$$\mathbb{E}_{x\sim p}[f(x)] = \int p(x)f(x)\,dx \tag{A.21}$$

协方差

随机变量X, Y 的协方差为：

$$cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \tag{A.22}$$

极大似然估计

假设θ 为未知参数（标量或者矢量），对于服从联合概率质量函数$p_X(\boldsymbol{x};\theta)$ 的一组观测向量$X = X_1, \ldots, X_n$，假设我们有X 的具体的观测值$\boldsymbol{x} = (x_1, \ldots, x_n)$。那么，其极大似然估计是未知参数$\theta$ 的一个取值，该取值能够使函数$p_X(x_1, \ldots, x_n;\theta)$ 取得最大值。

$$\hat{\theta}_n = \arg\max_{\theta}\ p_X(\boldsymbol{x};\theta) = \arg\max_{\theta}\ p_X(x_1, \ldots, x_n;\theta) \tag{A.23}$$

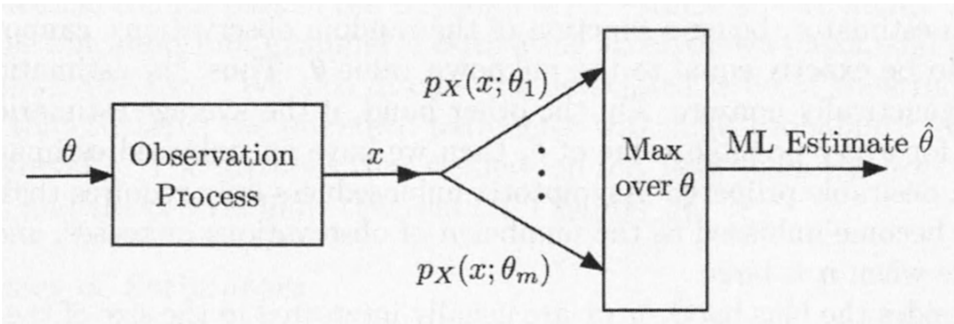

图 A.1 极大似然估计示意图

图A.1中，假设X 为离散变量，未知参数θ 可以从$\theta_1, \ldots, \theta_m$ 中选取。给定观测值$X = \boldsymbol{x}$，对于每个θ_i 取值，都可以计算$p_X(\boldsymbol{x};\theta_i)$。使函数$p_X(\boldsymbol{x};\theta)$ 取得最大值的θ_i 即为极大似然估计θ。

在很多情况下，都假定观测向量中的每一个X_i 为互相独立的，因此，似然函数通常可以改写为：

$$p_X(x_1, \ldots, x_n;\theta) = \prod_{i=1}^{n} p_{X_i}(x_i;\theta) \tag{A.24}$$

为了分析与计算方便，可以将其改写为对数似然函数：

$$\log p_X(x_1, \ldots, x_n;\theta) = \log \prod_{i=1}^{n} p_{X_i}(x_i;\theta) = \sum_{i=1}^{n} \log p_{X_i}(x_i;\theta) \tag{A.25}$$

当X 为连续变量时，由概率密度函数替换概率质量函数可得：

$$\log f_X(x_1, \ldots, x_n;\theta) = \log \prod_{i=1}^{n} f_{X_i}(x_i;\theta) = \sum_{i=1}^{n} \log f_{X_i}(x_i;\theta) \tag{A.26}$$

值得注意的是，对于X 的观测值$\boldsymbol{x}$，其似然函数$p_X(\boldsymbol{x};\theta)$ 并不是指未知参数取值为θ 的概率，而是指在未知参数取值为θ 时，X 的观测值为$\boldsymbol{x}$ 的概率。

边缘似然

边缘似然是似然函数在参数空间上的积分，表示生成观测样本的概率，因此，边缘似然也被称为模型证据，简称为证据。

全概率公式

全概率公式将对一复杂事件 A 的概率求解问题转化为了在不同情况下发生的简单事件的概率的求和问题。

若事件$B_1, B_2, \ldots, B_n$ 构成一个完备事件组且都有正概率，则有，

$$
\begin{aligned}
p(A) &= p(AB_1) + p(AB_2) + \cdots + p(AB_n) && \text{(A.27)}\\
&= p(A|B_1)p(B_1) + p(A|B_2)p(B_2) + \cdots + p(A|B_n)p(B_n) && \text{(A.28)}
\end{aligned}
$$

贝叶斯定理

若事件$B_1, B_2, \ldots, B_n$ 构成一个完备事件组且都有正概率，则有，

$$
\begin{aligned}
p(B_i|A) &= \frac{p(A|B_i)p(B_i)}{p(A)} && \text{(A.29)}\\
&= \frac{p(A|B_i)p(B_i)}{p(A|B_1)p(B_1) + p(A|B_2)p(B_2) + \cdots + p(A|B_n)p(B_n)} && \text{(A.30)}
\end{aligned}
$$

在式A.29与A.30中，B_i 通常表示一个命题，如“硬币正面朝上的次数占投掷次数的 50%”；A 通常表示事实，如“连续多次投掷硬币的结果”。$p(B_i)$ 表示B_i 的先验概率，先验概率为不考虑事实A 时，人们对事件B_i 的相信程度，其也包含了人们对于B_i 的先验知识。$p(A|B_i)$ 为似然函数，表示当命题B_i 发生时，事实A 发生的概率。“似然”表达了事件A 对命题B_i 的支撑程度。$p(B_i|A)$ 为B_i 的后验概率，表示考虑事实A 后，命题B_i 发生的概率。贝叶斯定理根据事实B_i，对先验概率$p(B_i)$ 进行更新。

贝叶斯推理

贝叶斯推理可以由先验概率和似然函数得出后验概率，其中先验概率和似然函数皆由统计模型关于观测数据产生。

$$
p(H \mid E) = \frac{p(E \mid H)p(H)}{p(E)} \tag{A.31}
$$

式A.31中：

- H 为假设，其概率受数据（以下称为证据）影响。
- $p(H)$ 为先验概率，是在获得数据E 前，对假设H 概率的估计，E 为当前证据。

- E 为证据，即未参与先验概率$p(H)$ 计算的数据。
- $p(H \mid E)$ 未后验概率，给出证据E 后，H 为真的概率，是关于H 的函数。
- $p(E \mid H)$ 为似然函数，给出假设H 后观测到E 的概率，是关于E 的函数。
- $p(E)$ 为边缘似然，也称为模型证据。表示从先验概率中获得观测样本的概率。

蒙特卡洛方法

蒙特卡洛方法是依赖于随机采样来获得数值结果的一种计算方法。其底层思想是，用随机方法来解决原理上具备确定性的问题。在物理和数学上，当其他方法不可用时，常常采用蒙特卡洛方法。蒙特卡洛方法主要应用于三类问题：优化、数值积分与从概率分布中采样。

当无法精确求和或者计算积分时，通常使用蒙特卡洛采样方法来近似。基本思想为，将其和或者积分视为某分布的期望，通过相应的计算来近似该期望。如：

$$s=\sum_{\boldsymbol{x}} p(\boldsymbol{x}) f(\boldsymbol{x})=E_p[f(\boldsymbol{x})] \tag{A.32}$$

$$s=\int p(\boldsymbol{x}) f(\boldsymbol{x})\, d\boldsymbol{x}=E_p[f(\boldsymbol{x})] \tag{A.33}$$

式A.32中，$p(\boldsymbol{x})$ 为随机变量$\boldsymbol{x}$ 的概率分布，式A.33中，$p(\boldsymbol{x})$ 为随机变量$\boldsymbol{x}$ 的概率密度。

随机微分方程

随机微分方程（Stochastic differential equation）是添加了一项或多项随机项的微分方程。

对比散度

用函数$f(x;\boldsymbol{\theta})$ 来为数据点的概率分布建模，其中，x 为模型的输入，$\boldsymbol{\theta}$ 为模型的参数，且要保证概率积分为 1 的性质，即：

$$p(x;\boldsymbol{\theta})=\frac{f(x;\boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \tag{A.34}$$

式A.34中，$Z(\boldsymbol{\theta})$ 为划分函数：

$$Z(\boldsymbol{\theta})=\int f(x;\boldsymbol{\theta})\, dx \tag{A.35}$$

假定数据点集合为$\boldsymbol{x}=x_1,\ldots,x_K$，则其似然函数为：

$$p(\boldsymbol{x};\boldsymbol{\theta})=\prod_{k=1}^{K} \frac{f(x_k;\boldsymbol{\theta})}{Z(\boldsymbol{\theta})} \tag{A.36}$$

极大化似然函数A.36等价于最小化负对数似然函数，即能量函数$E(\boldsymbol{x};\boldsymbol{\theta})$:

$$E(\boldsymbol{x};\boldsymbol{\theta}) = \log Z(\boldsymbol{\theta}) - \frac{1}{K}\sum_{i=1}^{K}\log f(x_i;\boldsymbol{\theta}) \tag{A.37}$$

对式A.37关于参数$\boldsymbol{\theta}$ 求偏导，即:

$$\frac{\partial E(\boldsymbol{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \frac{\partial\log Z(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - \frac{1}{K}\sum_{i=1}^{K}\frac{\partial\log f(x_i;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} \tag{A.38}$$

$$= \frac{\partial\log Z(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - \mathbb{E}_{p_{data}}\left[\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] \tag{A.39}$$

$$= \frac{1}{Z(\boldsymbol{\theta})}\frac{\partial Z(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} - \mathbb{E}_{p_{data}}\left[\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] \tag{A.40}$$

$$= \frac{1}{Z(\boldsymbol{\theta})}\frac{\partial\int f(x;\boldsymbol{\theta})\,dx}{\partial\boldsymbol{\theta}} - \mathbb{E}_{p_{data}}\left[\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] \tag{A.41}$$

$$= \frac{1}{Z(\boldsymbol{\theta})}\int\frac{\partial f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\,dx - \mathbb{E}_{p_{data}}\left[\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] \tag{A.42}$$

$$= \frac{1}{Z(\boldsymbol{\theta})}\int f(x;\boldsymbol{\theta})\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\,dx - \mathbb{E}_{p_{data}}\left[\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] \tag{A.43}$$

$$= \int p(x;\boldsymbol{\theta})\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\,dx - \mathbb{E}_{p_{data}}\left[\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] \tag{A.44}$$

$$= \mathbb{E}_{p(x;\boldsymbol{\theta})}\left[\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] - \mathbb{E}_{p_{data}}\left[\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] \tag{A.45}$$

式A.45即为对比散度的梯度，对于其等号右侧第一项

$$\mathbb{E}_{p(x;\boldsymbol{\theta})}\left[\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] \tag{A.46}$$

可以通过多次循环使用马尔科夫链蒙特卡洛采样来将训练集数据转化为从$p(x;\boldsymbol{\theta})$ 分布中的采样。假设$\boldsymbol{x}^n$ 表示对训练样本数据$\boldsymbol{x}$ 使用n 次马尔科夫链蒙特卡洛采样获得数据，可令$\boldsymbol{x}^0 = \boldsymbol{x}$，即得:

$$\frac{\partial E(\boldsymbol{x};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \mathbb{E}_{\boldsymbol{x}^\infty}\left[\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] - \mathbb{E}_{\boldsymbol{x}^0}\left[\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] \tag{A.47}$$

对于式A.47，在机器学习中，即使使用一次马尔科夫链蒙特卡洛采样也可以取得很好得效果，参数更新式为:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_t + \eta\left(\mathbb{E}_{\boldsymbol{x}^0}\left[\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right] - \mathbb{E}_{\boldsymbol{x}^1}\left[\frac{\partial\log f(x;\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]\right) \tag{A.48}$$

证据下界

显式密度模型需要表示出概率密度函数，假定观测变量$\boldsymbol{x}$ 和隐变量$\boldsymbol{z}$ 构成联合概率分布$p(\boldsymbol{x},\boldsymbol{z})$，概率密度函数$p(\boldsymbol{x})$ 可表示为:

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x},\boldsymbol{z})\,d\boldsymbol{z} \tag{A.49}$$

或使用概率链式法则：

$$p(\boldsymbol{x})=\frac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{z}\mid\boldsymbol{x})} \tag{A.50}$$

对于式A.49，复杂模型无法对所有隐变量$\boldsymbol{z}$ 进行积分；对于式A.50，因无法获得实际后验分布$p(\boldsymbol{z}\mid\boldsymbol{x})$ 而也无法直接计算。结合式A.49与式A.50，有：

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log\int p(\boldsymbol{x},\boldsymbol{z})d\,z && \text{（根据式A.49）} && \text{(A.51)}\\
&= \log\int p(\boldsymbol{x},\boldsymbol{z})\frac{q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}{q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}d\,z && && \text{(A.52)}\\
&= \log\int \frac{p(\boldsymbol{x},\boldsymbol{z})q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}{q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}d\,z && && \text{(A.53)}\\
&= \log\int q_\phi(\boldsymbol{z}\mid\boldsymbol{x})\frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}d\,z && && \text{(A.54)}\\
&= \log\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}\right] && \text{（根据期望定义式A.21）} && \text{(A.55)}\\
&\geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log\frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}\right] && \text{（根据杰森不等式??）} && \text{(A.56)}
\end{aligned}$$

可获得观测变量$\boldsymbol{x}$ 对数似然函数的证据下界（Evidence Lower BOund, ELBO）：

$$\mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log\frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}\mid\boldsymbol{x})}\right] \tag{A.57}$$

其中，$q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 是由参数ϕ 确定的近似变分分布，最大化证据下界即优化参数ϕ。最大化证据下界可以取得与极大似然估计相近的效果。

此外，对证据下界的另一种证明方式能够体现最大化证据下界的原因，

$$\begin{aligned}
\log p(\boldsymbol{x}) &= \log p(\boldsymbol{x})\int q_\phi(\boldsymbol{z}\mid\boldsymbol{x})d\,\boldsymbol{z} && \text{（}1=\textstyle\int q_\phi(\boldsymbol{z}\mid\boldsymbol{x})d\,\boldsymbol{z}\text{）} && \text{(A.58)}\\
&= \int q_\phi(\boldsymbol{z}\mid\boldsymbol{x})(\log p(\boldsymbol{x}))d\,\boldsymbol{z} && \text{（不改变积分值）} && \text{(A.59)}\\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log p(\boldsymbol{x})\right] && \text{（期望定义式A.21）} && \text{(A.60)}\\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log\frac{p(\boldsymbol{x},\boldsymbol{z})}{p(\boldsymbol{z}\mid\boldsymbol{x})}\right] && \text{（根据式A.50）} && \text{(A.61)}\\
&= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log\frac{p(\boldsymbol{x},\boldsymbol{z})q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z}\mid\boldsymbol{x})q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] && \text{（}1=\tfrac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\text{）} && \text{(A.62)}
\end{aligned}$$

$$= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] + \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{q_\phi(\boldsymbol{z}|\boldsymbol{x})}{p(\boldsymbol{z} \mid \boldsymbol{x})}\right] \quad \text{（期望拆分）} \tag{A.63}$$

$$= \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] + D_{KL}(q_\phi(\boldsymbol{z}|\boldsymbol{x}) \,||\, p(\boldsymbol{z} \mid \boldsymbol{x})) \quad \text{（散度定义式??）} \tag{A.64}$$

$$\geq \mathbb{E}_{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\left[\log \frac{p(\boldsymbol{x},\boldsymbol{z})}{q_\phi(\boldsymbol{z}|\boldsymbol{x})}\right] \quad \text{（根据式??）} \tag{A.65}$$

根据式A.64，对数似然函数$\log p(\boldsymbol{x})$ 即为证据下界与近似后验$q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 和真实后验$p(\boldsymbol{z} \mid \boldsymbol{x})$ 的 KL 散度，KL 散度项即为式A.56中被移除的项。由于对数似然函数$\log p(\boldsymbol{x})$ 与证据下界的差值仅为非负的 KL 散度项，因此，证据下界的值不可能超过对数似然函数$\log p(\boldsymbol{x})$ 的值。通过最小化 KL 散度项，即可使变分后验$q_\phi(\boldsymbol{z}|\boldsymbol{x})$ 更接近于真实后验$p(\boldsymbol{z} \mid \boldsymbol{x})$，但由于无法获得真实后验$p(\boldsymbol{z} \mid \boldsymbol{x})$，无法直接对 KL 散度项最小化。而注意到，式A.64左侧证据下界项中，$p(\boldsymbol{x},\boldsymbol{z})$ 与参数ϕ 无关，对其关于变量$\boldsymbol{z}$ 计算边缘概率所得$p(\boldsymbol{x})$ 也因此与参数ϕ 无关，即$\log p(\boldsymbol{x})$ 与参数ϕ 无关。因此，在改变参数ϕ 时，证据下界与 KL 散度项的和为定值，对证据下界项的最大化即代表了 KL 散度项的最小化。对证据下界项的优化程度可以体现模型对隐变量后验概率的拟合程度，证据下界项优化程度越高，近似后验则越接近真实后验。此外，由于证据下界是对模型证据$\log p(\boldsymbol{x})$ 的近似，模型经过训练后，证据下界也可以作为对观测数据或生成数据似然的估计。

重参数方法

从均值为μ，方差为σ 的正态分布$x \sim \mathcal{N}(x;\mu,\sigma^2)$ 中采样过程可以改写为

$$x = \mu + \sigma\epsilon \qquad \text{其中}\sigma \sim \mathcal{N}(\epsilon;0,1) \tag{A.66}$$

使用重参数方法，从任意高斯分布取样可以变为从标准高斯分布取样，即将标准高斯分布采样取值根据σ 对进行伸缩变换，再根据μ 进行平移变换。

A.2 物理学

A.2.1 玻尔兹曼分布

在统计力学与数学中，玻尔兹曼分布（也成为吉布斯分布）给出了一个系统处在特定状态的概率，其值与该状态的能量和该系统的热力学温度相关：

$$p_i \propto \exp(-\frac{\varepsilon_i}{kT}) \tag{A.67}$$

式A.69中，p_i 表示系统处在状态i 的概率，ε_i 式该状态的能量，k 为玻尔兹曼常数，T 为热力学温度。$\propto$ 表示其成正比。其比例系数为$\frac{1}{Q}$，其中

$$Q = \sum_{i=1}^{M} \exp(-\frac{\varepsilon_i}{kT}) \tag{A.68}$$

结合式A.67与式A.68的

$$p_i = \frac{1}{Q}\exp(-\frac{\varepsilon_i}{kT}) = \frac{\exp(-\frac{\varepsilon_i}{kT})}{\sum_{i=1}^{M}\exp(-\frac{\varepsilon_i}{kT})} \tag{A.69}$$