Hidden Markov Model
隐马尔可夫模型

杨猛，郑伟诗

https://cse.sysu.edu.cn/content/2970

中山大学

机器智能与先进计算
教育部重点实验室

模拟真实世界的现象

- 可以用来检测一个新建的水坝溢流的频率（取决于连续下雨的天数）。为建立这个模型，可以从下面的雨天（R）和晴天（S）开始：

时间序列Times Series

- 随机过程$\{X_1, X_2, \dots\}$, $X_i \in \mathcal{X}$
  - $\mathcal{X}$称为状态空间，我们假设$\mathcal{X} = \{1,2, \dots, N\}$
  - 假设对所有的i， $\mathcal{X}$都相同
  - 假设只处理时间序列，即i代表时间
  - 随机性
- 目的是希望“过去”对“现在”有帮助
  - 即如果有对$X_1, \dots, X_{t-1}$的了解，能帮助确定X_t
  - Formally, $P(X_t | X_{1:t-1})$ vs. $P(X_t)$

Markov Property

- Curse of dimensionality
  - $P(X_2|X_1)$需要多少存储空间才能指定？
  - $P(X_3|X_2,X_1)$需要多少存储空间才能指定？
  - $P(X_t|X_{1:t-1})$需要多少存储空间才能指定？
    - N^t！
- Markov Property马尔科夫性质
  - 限定：$P(X_t|X_{1:t-1}) = P(X_t|X_{t-1})$，含义是？
  - 无记忆性memoryless
  - 这个假设有效吗？
  - 好处是什么？

人物介绍

- Andrey Markov
- https://en.wikipedia.org/wiki/Markov_chain

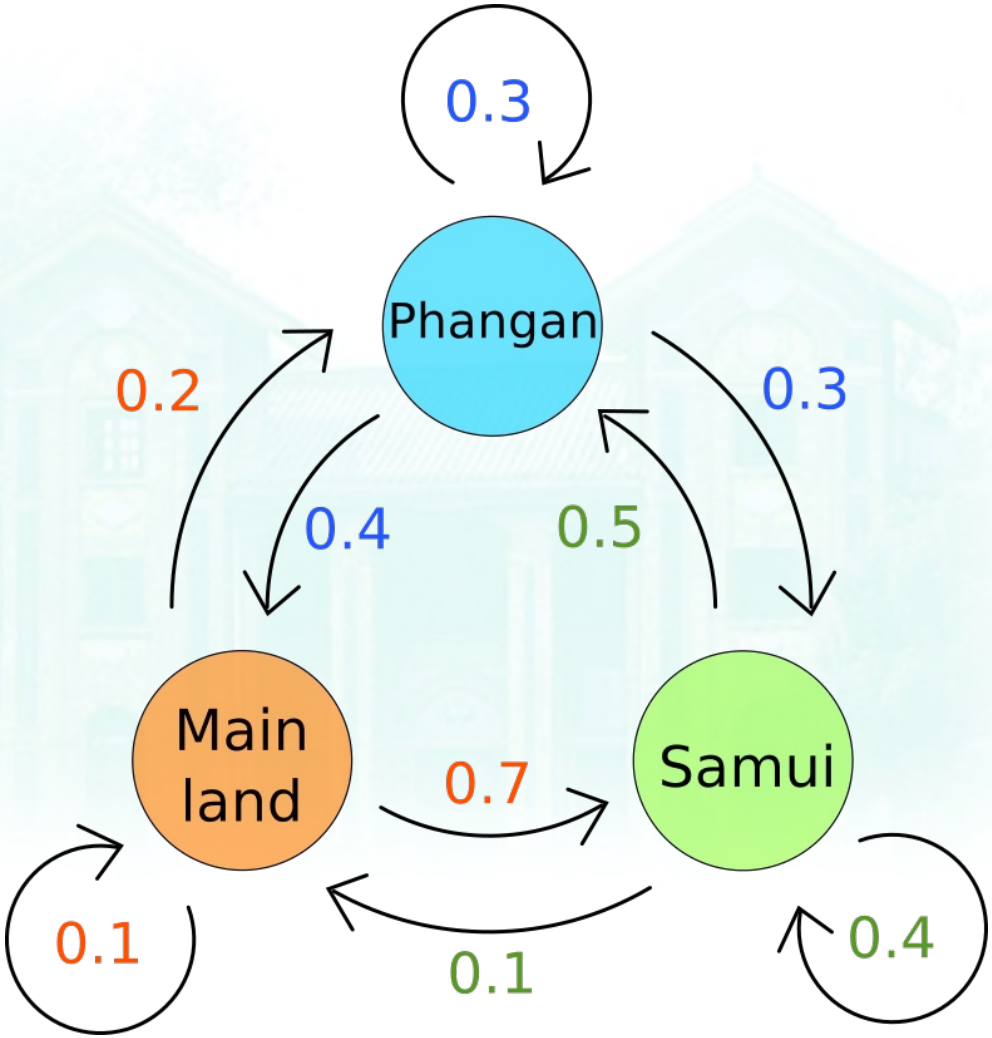
Computational finance
Speech synthesis
Cryptanalysis

Speech recognition
Part-of-speech tagging
Handwriting recognition

Speech synthesis
Single-molecule kinetic analysis
Machine translation

Markov Chain马尔科夫链

- Markov chain (discrete-time Markov chain or DTMC)



Markov Chain马尔科夫链

- Markov chain (discrete-time Markov chain or DTMC)

	A	B
A	P(A|A):0.50	P(B|A):0.50
B	P(A|B):0.50	P(B|B):0.50

状态空间中有A和B两种状态。共4种可能的转换。

1. 在A时，可以过渡到B或留在A。
2. 在B时，可以过渡到A或者留在B。

在图中，从**任意状态**到**任意状态**的转移概率是0.5。

人们会通过使用“转移矩阵”来计算转移概率。状态空间的每个状态都会出现在表格中的一列或者一行中。矩阵的每个单元格指明了从**行状态**转换到**列状态**的概率。

状态空间新增一个状态，矩阵将对应增加一行和一列，向现有的列和行中添加一个单元格。**这意味着当我们向马尔可夫链添加状态时，单元格的数量会呈二次方增长。**因此，**转换矩阵**就起到了很大的作用。

Markov Chain马尔科夫链

- Markov chain (discrete-time Markov chain or DTMC)

马尔科夫链的一个作用是用计算机模拟现实世界中的现象。
例如，可以用来检测一个新建的水坝溢流的频率（取决于连续下雨的天数）。
为建立这个模型，可以从下面的雨天（R）和晴天（S）开始：

表述这种模拟天气的方法就是：“有一半的天数是下雨天。所以模拟中的每一天都有50%的概率是下雨的。”这个规则在模拟中所产生的序列如下：

你注意到上面的序列和原来的不太一样了吗?第二个序列似乎具有跳跃性，而第一个(真实数据)似乎具有“粘性”。在真实的数据中，如果某一天是晴天，那么第二天也很可能是晴天。

Markov Chain马尔科夫链

- Markov chain (discrete-time Markov chain or DTMC)

通过两个状态的马尔可夫链来消除这种“粘性”。
当马尔科夫链处于状态“R”时，它保持在该状态的概率是0.9，状态改变的概率是0.1。同样，“S”状态保持不变的概率是0.9，过渡到“R”状态的概率是0.1。

S S S R R R S R R R R R R S S S

在许多需要对大规模的现象做研究的工作人员手中，马尔科夫链的作用可以变得非常强大。例如，谷歌用于确定搜索结果顺序的算法，称为PageRank，就是一种马尔可夫链。

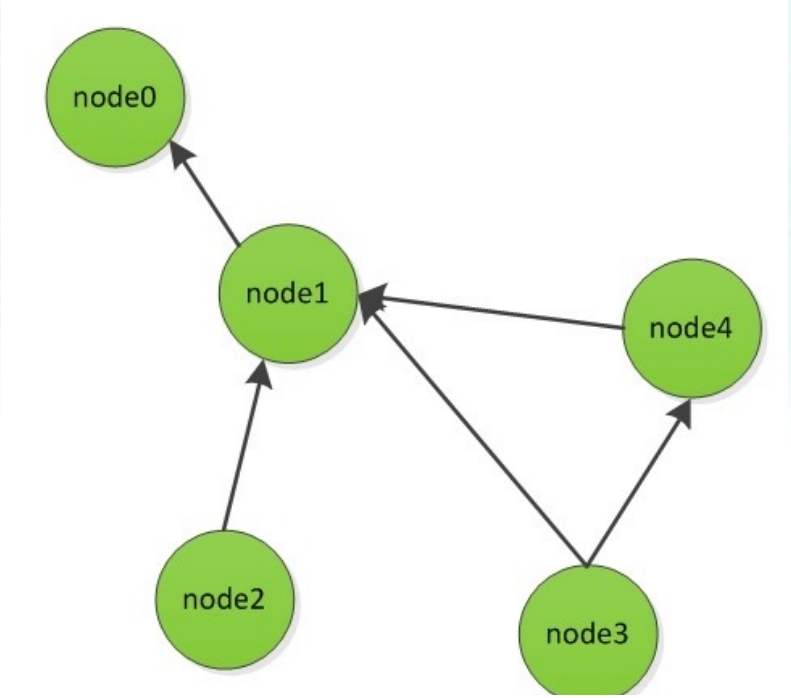
PageRank

- 随着时代的发展，网页变得越来越多，用人工识别的方式对网页进行识别变得越来越不现实。

- Google的两名创始人Larry Page与Sergey Brin开始对网页排序问题进行研究，依据学术界评判论文的重要程度的方法，查看论文引用次数，将这种方法用到了网页排序中，PageRank算法就产生了。

PageRank 搜索算法——马氏链视角

- 下是一个有向图，包含了5个节点，以及5条边。边的起点是一个网页、论文或者人，终点指向的是起点所引用的网页、论文或者人。Node1节点引用node0节点，表示前者从后者获取信息、知识、权力或者财富。



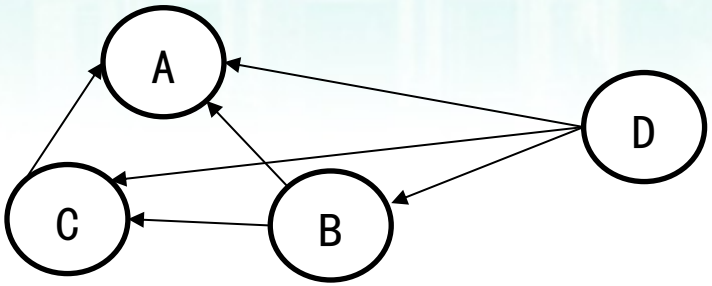
PageRank——马氏链视角

- 假设一个由4个网页组成的群体：A，B，C和D。如果所有页面都只链接至A，那么A的PR（PageRank）值将是B，C及D的Pagerank总和。

$$PR(A) = PR(B) + PR(C) + PR(D)$$

- 假设B链接到A和C，C只链接到A，并且D链接到全部其他的3个页面。一个页面总共只有一票。所以B给A和C每个页面半票。以同样的逻辑，D投出的票只有三分之一算到了A的PageRank上。

$$PR(A) = \frac{PR(B)}{2} + \frac{PR(C)}{1} + \frac{PR(D)}{3}$$



完整的PageRank——马氏链视角

问题：由于存在一些出链为0，也就是那些不链接任何其他网页的网， 也称为孤立网页，使得很多网页能被访问到

假设有N个页面，对于一个页面A，那么它的PR值为：

$$PR(A) = (1-d)/N + d\sum_{i=1}^{n}\frac{PR(T_i)}{C(T_i)}$$

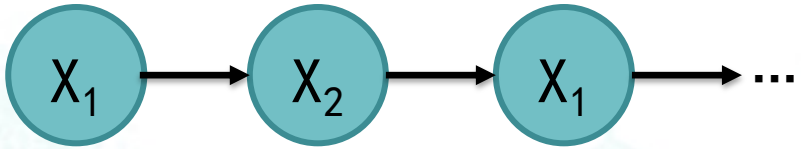
- d 为阻尼系数，其意义是，在任意时刻，用户到达某页面后并继续向后浏览的概率。

PageRank——马氏链视角

- 经过迭代计算至收敛，最终页面得到的PR值为该页面的分数。
- 网站被按照他们的PageRank算法分数从大到小排序，那些位于序列顶端的网站被认为是值得信赖的。在这些网站汇总标签为可信赖的网站被选作种子。
- PageRank算法高排名的页面都有很高的入度，说明指向一个网页的重要网页越多这个页面就越重要。

可视化和形式化

- 可视化：



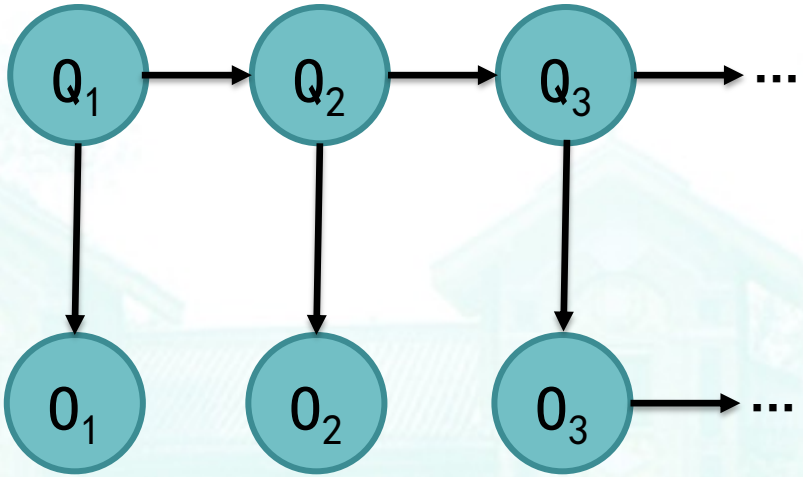
  - 注意填充的变量表示观察值（即随机变量值已知）
- 那么，如何形式化定义DTMC？需要哪些量？
  - 系统初始化Initialization: $P(X_1)$或者$X_1 = x_1$
  - Transition probability: $A = P(X_{t+1}|X_t)$
  - 还需要别的吗？
  - 两次运行结果会一样吗？

转移概率矩阵

- Transition probability matrix转移概率矩阵
  - A是一个$N \times N$的矩阵
  - $A_{ij} = P(X_t = j | X_{t-1} = i)$
  - 行和为1！

- 如果运行足够久（$t \to \infty$），那么X_t的分布在很多情况下将稳定下来，叫Stationary distribution，记为$\boldsymbol{\pi}$
  - $\boldsymbol{\pi} = A\boldsymbol{\pi}$

隐马尔可夫模型HMM



形式化

- Q：隐变量(hidden variable)，不可观测的状态
- N：number of states 状态数，N个可能的状态为$\{S_1, \ldots, S_N\}$
- $O(o)$：观察值(observations)，M个可能的观察值$\{V_1, V_2, \ldots, V_M\}$
- T：时间序列的长度
- $\boldsymbol{\pi}$：初始化，$\pi_j = P(Q_1 = S_j)$
- A：transition probability matrix，$A_{ij} = P(Q_{t+1} = S_j | Q_t = S_i)$
- B：emission probability 发出观察值的概率
  - $b_j(k) = \Pr(O_t = V_k | Q_t = S_j)$
  - 假设B不随时间变化，当未知状态为j时观察到为k的概率
  - 那么，j，k的取值范围是？B的行和是？

HMM中要解决的问题

- 怎样设计状态？-- 自动学习？
- 怎样设计观察值？--根据问题的特点和实践反复设计
- 与具体问题无关的
  - 指定一个HMM需要的所有参数：$\lambda = (\boldsymbol{\pi}, A, B)$
  - 问题1：Evaluation估值
  - 问题2：Decoding解码
  - 问题3：Learning学习

Problem 1. Evaluation

- 输入
  - 一个完全指定的HMM模型，即$\lambda = (\boldsymbol{\pi}, A, B)$已知
  - 一个完全观测的输出序列$O_1 O_2 \cdots O_T$，或$\boldsymbol{O} = O_{1:T}$
- 输出
  - $P(\boldsymbol{O}|\lambda)$ - 含义是?
  - 在这个模型λ中观察到特定输出$\boldsymbol{O}$的概率
- 作用是?
  - 可以看出此模型对该观察序列的成绩score
  - 可以用来从多个模型中选择最适合的模型

Problem 1. Evaluation

假设状态已知

- 已知$\lambda, o_{1:T}$，求$P(o_{1:T}|\lambda)$
- 若假设oracle已告知所有的隐变量的值$q_{1:T}$
  - $\Pr(o_{1:T}|\lambda, q_{1:T}) = \prod_{i=1}^{T} \Pr(o_t|q_t, \lambda) = \prod_{i=1}^{T} b_{q_i}(o_i)$
  - 证明？含义？
  - λ的存在只是表明概率的大小是基于该模型参数计算的，可以去除而不影响计算

Problem 1. Evaluation

一种naïve的计算方法

- 那么隐变量序列$q_{1:T}$的可能性多大呢？
  - $\Pr(q_{1:T}|\lambda) = \pi_{q_1} A_{q_1 q_2} A_{q_2 q_3} \cdots A_{q_{T-1} q_T}$
  - 含义？
- 用全概率公式对所有可能的$q_{1:T}$求和可以得到$\Pr(o_{1:T}|\lambda)$
  - $\Pr(o_{1:T}|\lambda) = \sum_{all\ Q} \Pr(o_{1:T}|\lambda, q_{1:T}) \Pr(q_{1:T}|\lambda)$，复杂度？
  - $O(T \times N^T)$

Problem 1. Evaluation

那么，如何快速计算？

动态规划！

只看最后一步（$t = T$），该如何计算?

1. 最后一步（$t = T$）时一共可能有N种状态：$q_T = S_1, \ldots, S_N$，其概率$\Pr(o_{1:T-1}, Q_T = S_i \,|\lambda) = ?$
2. 若最后一步状态为S_i，那么观察到输出o_T的概率是多少？
3. 所求的值是多少？（全概率公式）

$$\Pr(o_{1:T}|\lambda) = \sum_{i=1}^{N} \Pr(o_{1:T-1}, Q_T = S_i|\lambda)b_i(o_T)$$

只限于最后一步吗？

Problem 1. Evaluation

如何计算$\Pr(o_{1:T-1}, Q_T = S_i|\lambda)$?

- 有N种可能，即$T-1$时刻状态为$q_{T-1} = S_j$, $j = 1,2,\ldots,N$，然后通过概率A_{ji}转移
- 全概率公式，again!

$$\Pr(o_{1:T-1}, Q_T = S_i|\lambda) = \sum_{j=1}^{N} \Pr(o_{1:T-1}, Q_{T-1} = S_j|\lambda) A_{ji}$$

Problem 1. Evaluation

快速计算小结

- $\Pr(o_{1:T}|\lambda) = \sum_{i=1}^{N} \Pr(o_{1:T-1}, Q_T = S_i|\lambda)b_i(o_T) = \sum_{i=1}^{N}\left(b_i(o_T)\sum_{j=1}^{N} \Pr\left(o_{1:T-1}, Q_{T-1} = S_j|\lambda\right)A_{ji}\right)$
- 红色部分是什么？
  - 一个规模小一点的相同问题（$T-1$）
  - 但是需要对所有j的可能取值计算
  - 正如DTW中一样，可以通过动态规划解决，但是需要解决比原问题更多数目的小规模子问题
  - 但是，复杂的是，目前牵涉两个数值而不是一个：$\Pr(o_{1:T-1}, Q_T = S_i|\lambda)$和$P(o_{1:T}|\lambda)$
  - 计算的方向应该是什么？

Problem 1. Evaluation

动态规划算法（前向forward算法）

- $P(o_{1:T}|\lambda) = \sum_{i=1}^{N} \Pr(o_{1:T-1}, Q_T = S_i|\lambda)b_i(o_T) = \sum_{i=1}^{N}\left(b_i(o_T)\sum_{j=1}^{N}\Pr\left(o_{1:T-1}, Q_{T-1} = S_j|\lambda\right)A_{ji}\right)$
- 定义
  - $\alpha_t(i) = P(o_{1:t}, Q_t = S_i|\lambda)$ - 含义是?
  - Initialization: $\alpha_1(i) = \pi_i b_i(o_1), \quad 1 \le i \le N$
  - Induction: For $1 \le t \le T-1$

$$\alpha_{t+1}(i) = \left[\sum_{j=1}^{N} \alpha_t(j)A_{ji}\right] b_i(o_{t+1}), \quad 1 \le i \le N$$

  - Termination (output): $\Pr(o_{1:T}|\lambda) = \sum_{i=1}^{N} \alpha_T(i)$

Problem 1. Evaluation

后向算法backward algorithm

- 定义$\beta_t(i) = \Pr(o_{t+1:T}|Q_t = S_i, \lambda)$
  - 若在时刻t状态为S_i，将来观测到$o_{t+1:T}$的概率
- 初始化：$\beta_T(i) = 1,\ \ 1 \le i \le N$
- 反向更新：$t = T-1, T-2, \ldots, 2, 1$

$$\beta_t(i) = \sum_{j=1}^{N} A_{ij} b_j(o_{t+1}) \beta_{t+1}(j), \qquad 1 \le i \le N$$

- 输出：$\beta_1(i) = \Pr(o_{2:T}|Q_1 = S_i, \lambda)$

$$P(o_{1:T}|\lambda) = \sum_{i=1}^{N} \pi_i b_i(o_1) \beta_1(i)$$

Problem 2: Decoding

- 输入
  - 一个完全指定的HMM模型，即$\lambda = (\boldsymbol{\pi}, A, B)$已知
  - 一个完全观测的输出序列$O_1 O_2 \cdots O_T$，或$\boldsymbol{O} = O_{1:T}$
  - 某个标准criterion
- 输出
  - 一个完全指定的隐变量序列$X_{1:T}$的值
- 作用是？
  - 如，语音识别中状态可能有实际意义（各音节）
    - ❖唯一吗？
  - 可以用来观察模型结构，优化模型

Problem 2: Decoding

发现“最好”的隐变量值

- 标准1：对于每个时刻，发现其后验概率最大的状态
  - 定义$\gamma_t(i) = \Pr(Q_t = S_i | o_{1:T}, \lambda)$，当观测到输出为$o_{1:T}$时，时刻$t$时隐变量为第$i$个状态的后验概率
  - 那么，对于一个输出序列$o_{1:T}$，选择

$$q_t = \underset{1 \le i \le N}{\operatorname{argmax}} \gamma_t(i), \qquad t = 1,2,\dots,T$$

  - 可能出现什么问题？
  - 不存在这样的路径$q_{1:T}$

Problem 2: Decoding

怎样计算γ

- $\alpha_t(i)\beta_t(i) = P(o_{1:T}, Q_t = S_i|\lambda)$
  - 为什么？

- 贝叶斯定理

$$\gamma_t(i) = \Pr(Q_t = S_i|o_{1:T}, \lambda) = \frac{\Pr(o_{1:T}, Q_t = S_i|\lambda)}{\Pr(o_{1:T}|\lambda)} = \frac{\alpha_t(i)\beta_t(i)}{\Pr(o_{1:T}|\lambda)}$$

  - $\Pr(o_{1:T}|\lambda) = \sum_{i=1}^{N} \alpha_t(i)\beta_t(i)$ for any t!
  - 三种计算方法计算$P(o_{1:T}|\lambda)$了

- 或者 1) $\gamma_i = \alpha_t(i)\beta_t(i)$ 2) L1 normalize: $\gamma_i \leftarrow \frac{\gamma_i}{\sum_i \gamma_i}$

Problem 2: Decoding

寻找最大概率的路径

- 一共有N^T种可能的路径，有些的概率可能为0
  - 比如通过准则1得到的路径
  - 那么，如果寻找所有可能路径里面概率最大的那个呢？

$$q_{1:T} = \underset{Q_{1:T}}{\operatorname{argmax}} \Pr(Q_{1:T}|o_{1:T},\lambda) = \underset{Q_{1:T}}{\operatorname{argmax}} \Pr(Q_{1:T},o_{1:T}|\lambda)$$

- Naïve的方法复杂性是N^T，有没有更好的方法？
  - Viterbi方法

Problem 2: Decoding

Viterbi decoding

- $q_{1:T} = \underset{Q_{1:T}}{\operatorname{argmax}} \Pr(Q_{1:T}, o_{1:T}|\lambda)$
- 定义更多的子问题

$$\delta_t(i) = \max_{Q_{1:t-1}} \Pr(Q_{1:t-1}, Q_t = S_i, o_{1:t}|\lambda)$$

  - 含义：当限定两个条件1）前t个时刻的输出为$o_{1:t}$，2）第t个时刻的隐状态为第i个状态的时候，最佳路径所能取得的最大概率
  - 怎么取得q_t?
    - 用另外一个变量$\psi_t(i)$做记录
  - 怎么从t进展到$t+1$?

Problem 2: Decoding

两个步骤

- 从t进展到$t+1$
  - $\delta_{t+1}(i) = \max_j\left(\left[\delta_t(j)A_{ji}\right]b_i(o_{t+1})\right)$
  - $\delta_{t+1}(i)$是概率，如果只需要发现概率最大那个状态，$b_i(o_{t+1})$?
- 所以在时刻$t+1$，需要用另外一个变量$\psi_t(i)$记录最大概率的路径在时刻t是哪一个状态
  - $\psi_{t+1}(i) = \operatorname{argmax}_{1\le j\le N}\left(\left[\delta_t(j)A_{ji}\right]\right)$

Problem 2: Decoding

Viterbi算法

- 初始化：$\delta_1(i) = \pi_i b_i(o_1),\ \ \psi_1(i) = 0,\ \ 1 \le i \le N$
- 递归：$2 \le t \le T,\ 1 \le i \le N$

$$\delta_t(i) = \max_{1 \le j \le N} \left(\left[\delta_{t-1}(j) A_{ji} \right] b_i(o_t) \right)$$

$$\psi_t(i) = \operatorname*{argmax}_{1 \le j \le N} \left(\left[\delta_{t-1}(j) A_{ji} \right] \right)$$

- 输出：
  - 最大概率：$P^* = \max_{1 \le i \le n} \delta_T(i)$
  - 时刻T的最佳路径变量：$q_T^* = \operatorname*{argmax}_{1 \le i \le N} (\delta_T(i))$
  - 时刻$T-1, T-2, \dots, 2, 1$的最佳路径变量：$q_t^* = \psi_{t+1}(q_{t+1}^*)$

Problem 2: Decoding

分析

- 问题1的动态规划$\alpha_{t+1}(i) = \sum_{j=1}^{N} \alpha_t(j) A_{ji}$
- 问题2的动态规划$\delta_t(i) = \max_j \left(\left[\delta_{t-1}(j) A_{ji} \right] b_i(o_t) \right)$
- 最重要的操作分别是sum-product和max-product
  - 其复杂性均为N^2T
  - 和naïve方法的TN^T比较，极其巨大的速度提高

Problem 3: Learning

学习系统的参数

- 发现$\lambda = (A, B, \boldsymbol{\pi})$，使得对于固定的$N$，$T$，和观察值$\boldsymbol{O}$，似然(likelihood) $P(\boldsymbol{O}|\lambda)$最大
  - 目前没有方法能发现全局最优的解
  - 常用的方法是Baum-Welch算法，发现一个局部最优的解

Problem 3: Learning

- 输入
  - 网络结构，状态数、输出数
  - 若干观测序列$\{\boldsymbol{O}\}$
- 输出
  - 最优的参数$\lambda = (\boldsymbol{\pi}, A, B)$使得$P(\{\boldsymbol{O}\}|\lambda)$最大
- 作用
  - 显而易见
  - 最重要的问题
  - 有时候一个足够长的观测序列就够了

$$\xi_t(i,j) = Pr(Qt = Si, Qt+1 = Sj|o1{:}T, \lambda) = \frac{\alpha_t(i)A_{ij}b_j(o_{t+1})\beta_{t+1}(j)}{\Pr(o1{:}T|\lambda)}$$

Baum-Welch算法

- Baum-Welch算法
- 1：初始化参数$\lambda^{(1)}$（例如随机地）
- 2：$r \leftarrow 1$
- 3：**while** 似然尚未收敛 **do**
- 4：　　对所有 $t(1 \le t \le T)$ 和所有 $i(1 \le i \le N)$，使用前向过程基于 $\lambda^{(r)}$计算 $\alpha_t(i)$
- 5：　　对所有 $t(1 \le t \le T)$ 和所有 $i(1 \le i \le N)$，使用后向过程基于$\lambda^{(r)}$计算 $\beta_t(i)$
- 6：　　对所有 $t(1 \le t \le T)$ 和所有 $i(1 \le i \le N)$，根据公式计算 $\gamma_t(i)$
- 7：　　对所有 $t(1 \le t \le T-1)$ 和所有$i,j\ (1 \le i,j \le N)$，根据表 12.1中的公式计算$\xi_t(i,j)$
- 8：　　更新参数为$\lambda^{(r+1)}$
- $$\pi_i^{(r+1)} = \gamma_1(i) \qquad 1 \le i \le N$$
- $$A_{ij}^{(r+1)} = \frac{\sum_{t=1}^{T-1} \xi_t(i,j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \qquad 1 \le i,j \le N$$
- $$b_j^{(r+1)}(k) = \frac{\sum_{t=1}^{T}[[o_t=k]]\gamma_t(j)}{\sum_{t=1}^{T} \gamma_t(j)} \qquad 1 \le j \le N \quad 1 \le k \le M$$
- 9：　　$r \leftarrow r+1$
- 10：**end while**

怎样在模式识别中发挥更大作用

- 语音识别

语音识别的目的是将声音信号映射为文字信息，如何实现这种映射？

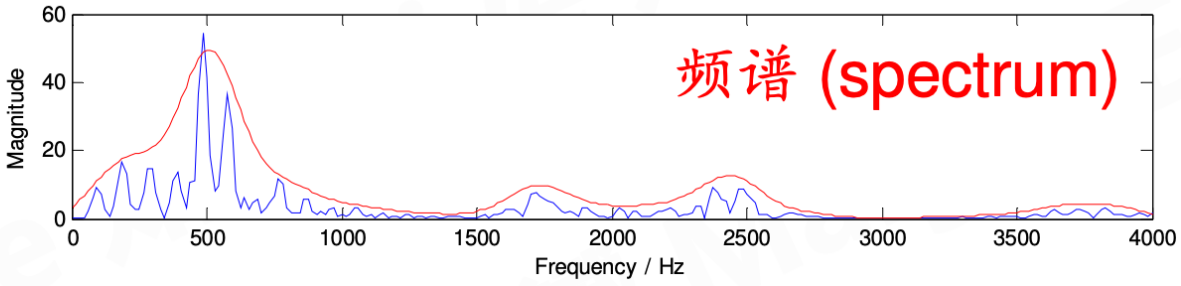
分帧：声音实际上是一种波，要对声音进行分析，需要对声音分帧，也就是把声音切开成一小段一小段，每小段称为一帧。分帧操作一般不是简单的切开，而是使用移动窗函数来实现，帧与帧之间一般有交叠。

怎样在模式识别中发挥更大作用

声学特征提取

- 分帧后，语音变成了很多小段，根据人耳的生理特性，把每一帧波形变成一个多维向量，可以简单地理解为这个向量包含了这帧语音的内容信息。为什么转化为向量，因为数据驱动的模型和算法基本都是从数据向量或者矩阵开始。下图为例，声音信号变成了12行（假设声学特征是12维）、N列的矩阵，每一帧都用一个12维的向量表示，色块的颜色深浅表示向量值的大小。

傅里叶变换



怎样在模式识别中发挥更大作用

HMM建模（隐马尔可夫模型）

1. 求值：给定模型参数和语音，求P(语音|模型)
   - 把P(语音，对齐|模型)对所有对齐方式求和
   - 动态规划算法：Forward algorithm

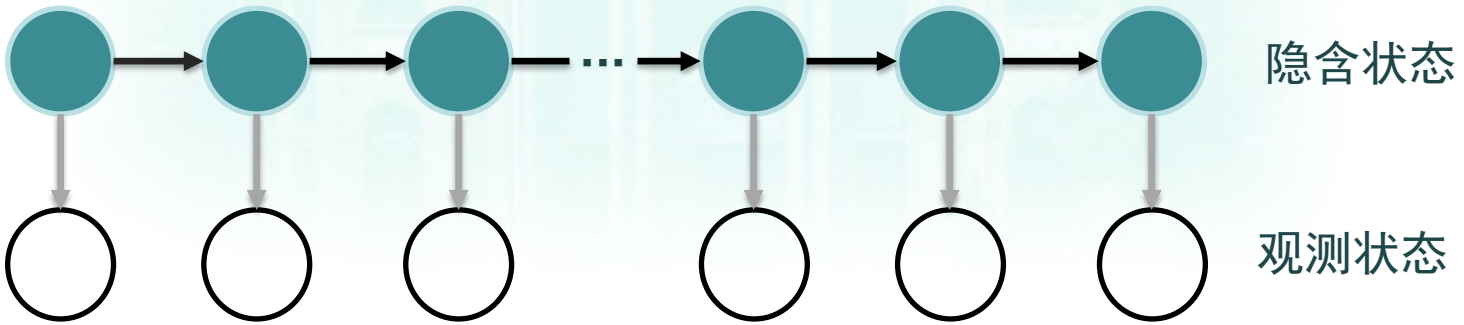
2. 解码:给定模型参数和语音，求最佳对齐方式
   - 动态规划算法：Viterbi decoding
   - 最佳对齐方式的概率，可以作为总概率的近似

3. 训练：给定模型参数和语音，求参数模型
   - 帧就是观测序列
   - 状态就是状态序列
   - 状态变化存在转移概率，状态与观测之间存在观测概率

这就构成了一个典型的HMM模型，知道每帧语音对应的状态，得到语音识别结果。

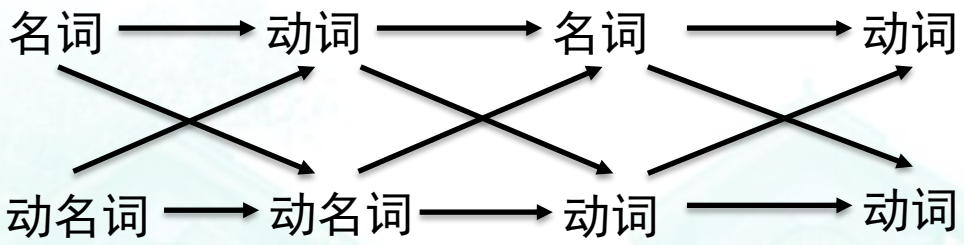
怎样在模式识别中发挥更大作用

- HMM用于NLP词性标注
- 对句子“教授喜欢画画”进行词性标注，分词之后的结果可能是“教授/喜欢/画/画”，“教授”词性可以是名词和动名词，“喜欢”词性可以是动词和动名词，“画”词性可以是名词和动词，画成图可以表示为：



怎样在模式识别中发挥更大作用

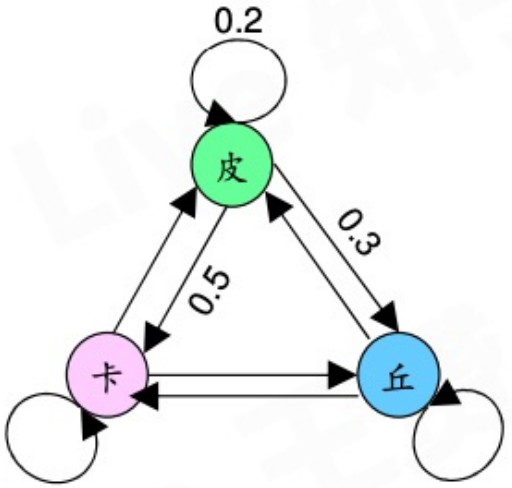
“教授喜欢画画”

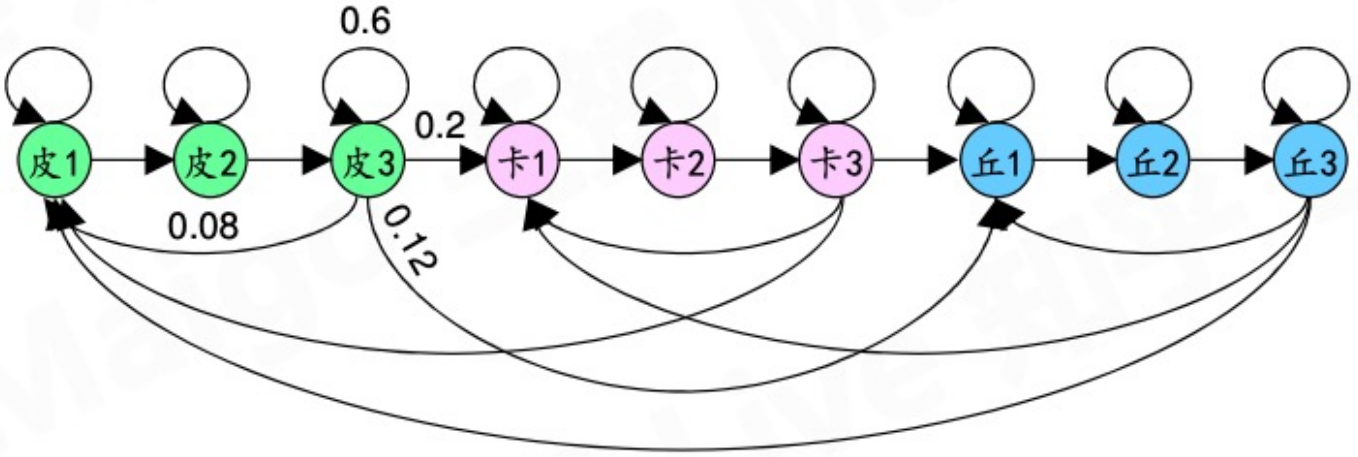


- 隐马是个生成模型，生成的过程是先生成状态节点，根据状态节点再生成观测节点。
- 首先生成“**教授**”词性是“**名词**”，然后生成词“**教授**”；
- 根据“**教授**”的词性节点“**名词**”生成“**喜欢**”的词性节点“**动词**”，然后生成词“**喜欢**”；
- 根据“**喜欢**”的词性“**动词**”生成“**画**”的词性“**动词**”，然后生成词“**画**”。

怎样在模式识别中发挥更大作用

- Bigram是马尔可夫模型
  - 下一个词只与当前词有关
  - 模型是遍历的，不是单向的
- 可与单词的声学模型复合
  - 得到一门语言的HMM

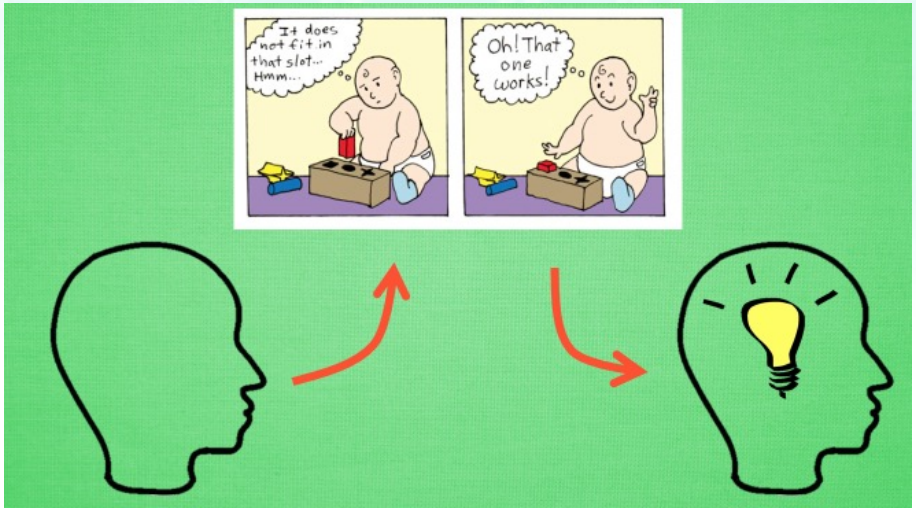




强化学习模型与马氏链

强化学习概念

- **Reinforcement learning (RL)** is an area of machine learning inspired by *behaviorist psychology*, concerned with how software *agents* ought to take *actions* in an *environment* so as to maximize some notion of cumulative *reward*. ——维基百科



强化学习基础

- 基于马尔可夫决策——MDP$< S, A, R, P, \gamma >$
  - S状态集，A动作集，R奖励函数，P状态转换概率，γ折扣因子
  - ✓ 目标：Agent最大化累积奖励G（Cumulative Reward）期望$E[G]$

$$G = R(s_0, a_0) + \gamma R(s_1, a_1) + \gamma^2 R(s_2, a_2) + \cdots$$

$$s_0 \xrightarrow{a_0} s_1 \xrightarrow{a_1} s_2 \xrightarrow{a_2} s_3 \xrightarrow{a_3} \dots$$

强化学习基础

- 策略(policy)是指任意一个从状态到行动的映射函数 $\pi: S \rightarrow A$
  - 价值函数：$V^{\pi}(s) = E[G|s_0 = s, \pi]$
  - 贝尔曼方程：$\mathrm{V}^{\pi}(\mathrm{s}) = \mathrm{R}(\mathrm{s},\mathrm{a}) + \gamma \sum_{s' \epsilon S} P_{s\pi(s)}(s') V^{\pi}(s')$
  - ✓ 价值迭代
    - ➢ 最优价值函数：$\mathrm{V}^{*}(s) = \max_{\pi} V^{\pi}(s)$
  - 策略迭代
    - ➢ 最优策略$\pi^{*}(s) = arg \max_{a \epsilon A} \sum_{s' \epsilon S} P_{sa}(s') V^{*}(s')$
  - 对于所有状态s与策略π，有：$\mathrm{V}^{*}(s) = V^{\pi^{*}}(s) \geq \mathrm{V}^{\pi}(\mathrm{s}).$
- 模型（model）
  - 预测agent行动后的状态

Diffusion模型概述

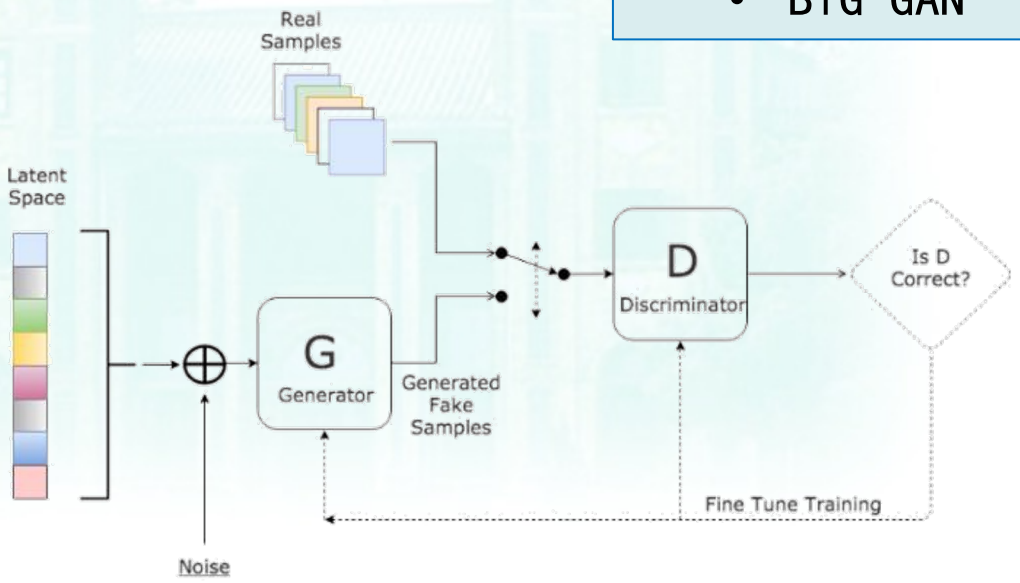
-从Markov链过渡到Diffusion模型

前言-对抗式生成网络(GAN)

- **起源**：生成模型缺乏固定的指标进行训练
- **动机**：零和博弈与纳什均衡
- **方法**：生成-对抗-再生成
- **结构**：生成器与判别器
- **缺陷**：不够稳定

应用：
- 图像生成
- 超分辨率
- 音频

后续改进：
- Conditional GANs
- DC GAN
- BiG GAN

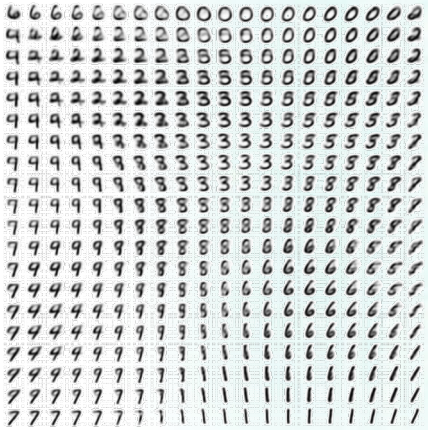


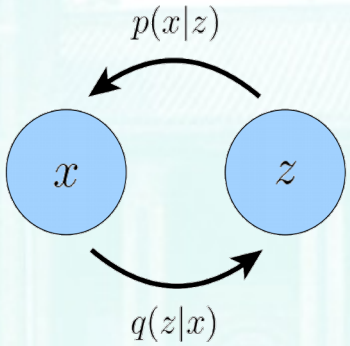
损失函数不是纯粹的最小化

$$\min_G \max_D V(D,G) = \mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}(\boldsymbol{x})}[\log D(\boldsymbol{x})] + \mathbb{E}_{\boldsymbol{z}\sim p_{\boldsymbol{z}}(\boldsymbol{z})}[\log(1 - D(G(\boldsymbol{z})))]$$

前言-变分自编码器VAE

- 起源：数据的真实分布很难计算
- 动机：贝叶斯进行后验近似估计
- 方法：重建图像并优化ELBO
- 结构：编码器与解码器
- 缺陷：过于连续导致图像边界模糊





$$\log p(\boldsymbol{x}) = \log \int p(\boldsymbol{x}, \boldsymbol{z}) d\boldsymbol{z}$$

$$= \log \int \frac{p(\boldsymbol{x}, \boldsymbol{z}) q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} d\boldsymbol{z}$$

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

$$\geq \mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log \frac{p(\boldsymbol{x}, \boldsymbol{z})}{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \right]$$

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

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

$$= \underbrace{\mathbb{E}_{q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x})} \left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}|\boldsymbol{z}) \right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q_{\boldsymbol{\phi}}(\boldsymbol{z}|\boldsymbol{x}) \parallel p(\boldsymbol{z}))}_{\text{prior matching term}}$$

重建图像与原始图像的差距

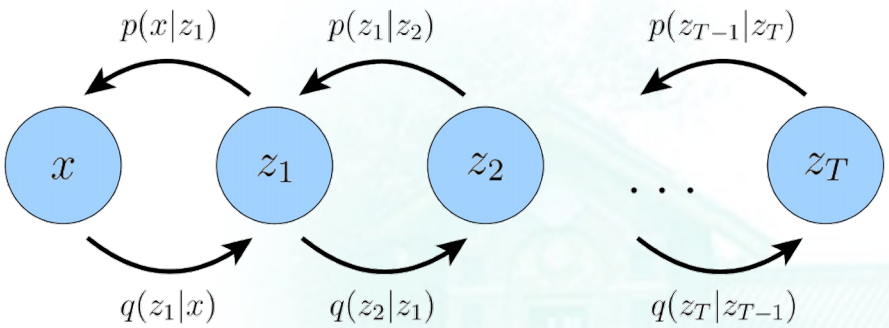
学习得到的分布q与先验假设p的相似

这导致了VAE的Loss中KL散度一项拥有解析解，而重构项只需要MSE就可以计算！

与GAN的区别：GAN中，数据的**真实分布未知**；
而在VAE中，假设了数据服从Gaussian分布，**只是方差均值未知**；

前言-Markov假设的HMVAE

VAE难以生成精致图片->用很多个VAE迭代生成可以解决。



x是原始数据，Z_1是第一层隐空间表征，. . . , Z_T是最终的高斯分布

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

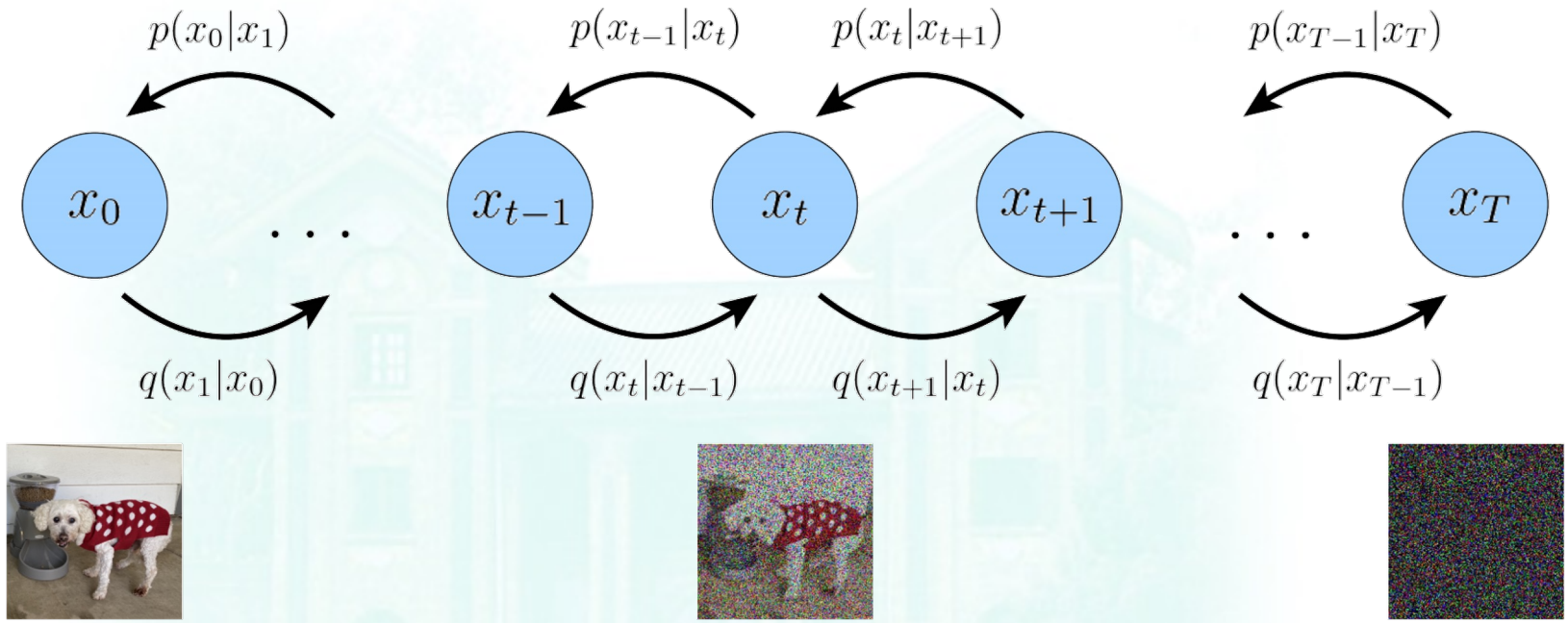
为什么是Markov？ 我需要这样的链来生成精致图片，但是我又不想要解析解过于复杂

仍然存在的问题？ 虽然用Markov链的性质消除了许多过长的条件概率，
但是因为先验扰动q仍然是含参的，进行优化仍然复杂。

那么能否消除先验扰动q的参数呢？

扩散概率模型Diffusion Probabilistic Model

答案：可以！

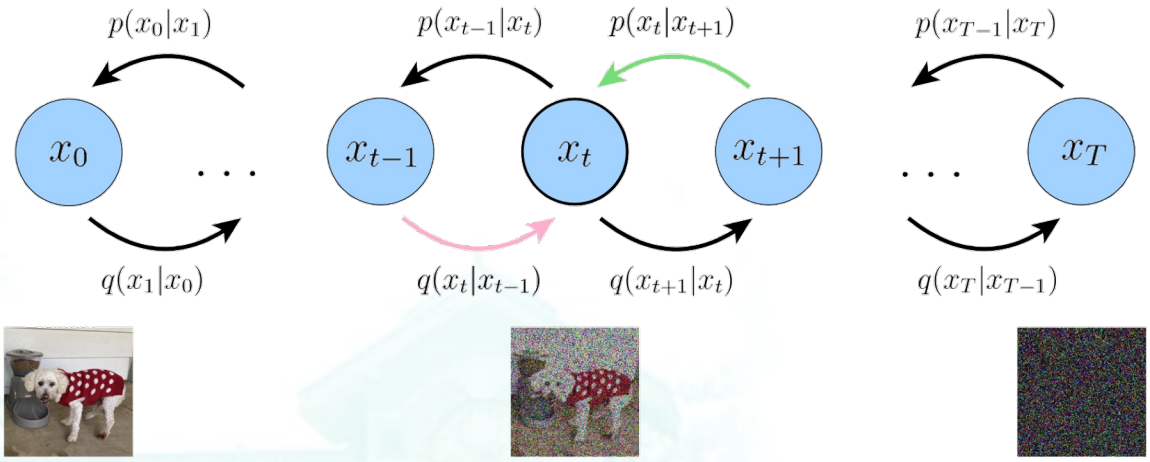


将HMVAE中，先验过程分布的**方差**和**均值**也设定为已知的，通常是一个高斯分布

GAN（分布未知，均值与方差未知）=>VAE（先验分布已知，均值方差未知）=>Diffusion（可解析的先验分布）

扩散概率模型Diffusion Probabilistic Model

具体怎么优化？



DPM的对数似然为：

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

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

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

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

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)} \left[\log p_\theta(\boldsymbol{x}_0|\boldsymbol{x}_1) \right]}_{\text{reconstruction term}} - \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{T-1}|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_{T-1}) \parallel p(\boldsymbol{x}_T)) \right]}_{\text{prior matching term}}$$

$$- \sum_{t=1}^{T-1} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_{t-1},\boldsymbol{x}_{t+1}|\boldsymbol{x}_0)} \left[D_{\mathrm{KL}}(q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}) \parallel p_\theta(\boldsymbol{x}_t|\boldsymbol{x}_{t+1})) \right]}_{\text{consistency term}}$$

正向过程$q(x_t|x_{t-1})$ => 对前一步的结果加噪声

逆向过程$p(x_t|x_{t+1})$ => 根据已加噪声进行恢复

怎么解决？

问题：虽然ELBO是可解析的形式，然而迭代项存在$x_{t-1}x_tx_{t+1}$，**这会导致次优的解**，这不是一个好的做法

扩散概率模型Diffusion Probabilistic Model

贝叶斯！ 代入原始的对数似然中

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

可得：

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

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

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

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

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

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

$$= \underbrace{\mathbb{E}_{q(\boldsymbol{x}_1|\boldsymbol{x}_0)}\left[\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0|\boldsymbol{x}_1)\right]}_{\text{reconstruction term}} - \underbrace{D_{\mathrm{KL}}(q(\boldsymbol{x}_T|\boldsymbol{x}_0)\parallel p(\boldsymbol{x}_T))}_{\text{prior matching term}} - \sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0)\parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

在原始ELBO的优化推导中引入贝叶斯后，消除了迭代项中的$x_{t-1}x_tx_{t+1}$ 之间的计算

$$\log \prod_{2}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{\frac{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)q(\mathbf{x}_t|\mathbf{x}_0)}{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}} = \log \prod_{2}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)} * \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}$$

$$= \log \prod_{2}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)} + \log \prod_{2}^{T} \frac{q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{q(\mathbf{x}_t|\mathbf{x}_0)}$$

$$= \log \prod_{2}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)} + \log \frac{\prod_2^T q(\mathbf{x}_{t-1}|\mathbf{x}_0)}{\prod_2^T q(\mathbf{x}_t|\mathbf{x}_0)}$$

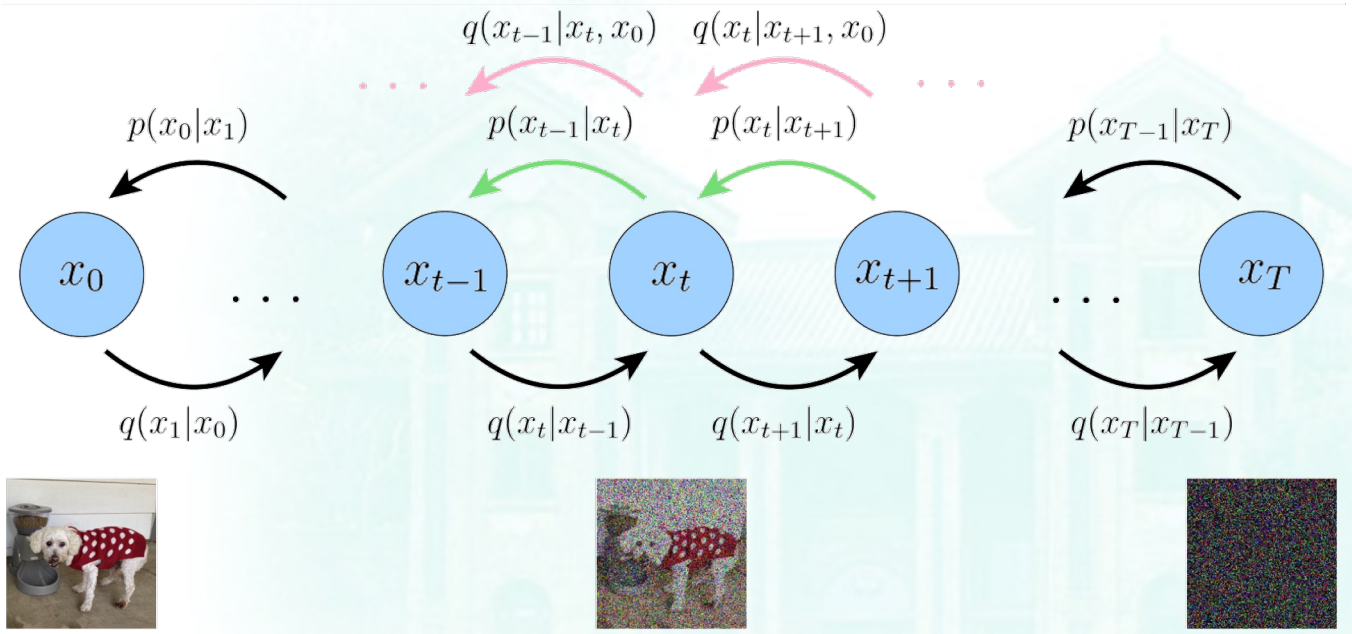
$$= \log \prod_{2}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)}$$

$$+ \log \frac{\cancel{q(\mathbf{x}_{T-1}|\mathbf{x}_0)} * \cancel{q(\mathbf{x}_{T-2}|\mathbf{x}_0)} * \cancel{q(\mathbf{x}_{T-3}|\mathbf{x}_0)} * \ldots * \cancel{q(\mathbf{x}_3|\mathbf{x}_0)} * q(\mathbf{x}_{2-1}|\mathbf{x}_0)}{q(\mathbf{x}_T|\mathbf{x}_0) * \cancel{q(\mathbf{x}_{T-1}|\mathbf{x}_0)} * \cancel{q(\mathbf{x}_{T-2}|\mathbf{x}_0)} * \ldots * \cancel{q(\mathbf{x}_3|\mathbf{x}_0)} * \cancel{q(\mathbf{x}_2|\mathbf{x}_0)}}$$

$$= \log \prod_{2}^{T} \frac{p_\theta(\mathbf{x}_{t-1}|\mathbf{x}_t)}{q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0)} + \log \frac{q(\mathbf{x}_1|\mathbf{x}_0)}{q(\mathbf{x}_T|\mathbf{x}_0)}$$

扩散概率模型Diffusion Probabilistic Model

直观上的改变 => Loss可以被计算，并有最优解



$$\sum_{t=2}^{T} \underbrace{\mathbb{E}_{q(\boldsymbol{x}_t|\boldsymbol{x}_0)}\left[D_{\mathrm{KL}}(q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) \parallel p_{\boldsymbol{\theta}}(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t))\right]}_{\text{denoising matching term}}$$

=> 我们需要求出来，
得到精确的Loss解析形式

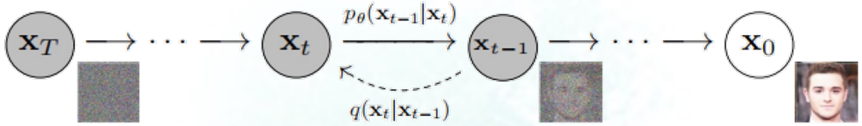
注意：目前，仍然仅需要一个Markov假设就能使如此复杂的过程成立。

扩散概率模型Diffusion Probabilistic Model

$$
\begin{aligned}
q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t,\boldsymbol{x}_0) &= \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1},\boldsymbol{x}_0)q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \\
&= \frac{\mathcal{N}(\boldsymbol{x}_t;\sqrt{\alpha_t}\boldsymbol{x}_{t-1},(1-\alpha_t)\mathbf{I})\mathcal{N}(\boldsymbol{x}_{t-1};\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0,(1-\bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\boldsymbol{x}_t;\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0,(1-\bar{\alpha}_t)\mathbf{I})} \\
&\propto \exp\left\{-\left[\frac{(\boldsymbol{x}_t-\sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{2(1-\alpha_t)}+\frac{(\boldsymbol{x}_{t-1}-\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_{t-1})}-\frac{(\boldsymbol{x}_t-\sqrt{\bar{\alpha}_t}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_t)}\right]\right\} \\
&= \exp\left\{-\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]\right\} \\
&= \exp\left\{-\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)}{1-\bar{\alpha}_{t-1}}+C(\boldsymbol{x}_t,\boldsymbol{x}_0)\right]\right\} \\
&\propto \exp\left\{-\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]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[(\frac{\alpha_t}{1-\alpha_t}+\frac{1}{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]\right\} \\
&= \exp\left\{-\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]\right\} \\
&= \exp\left\{-\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]\right\} \\
&= \exp\left\{-\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]\right\} \\
&= \exp\left\{-\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]\right\} \\
&= \exp\left\{-\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)(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\boldsymbol{x}_{t-1}\right]\right\} \\
&= \exp\left\{-\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]\right\} \\
&\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}\mathbf{I}}_{\boldsymbol{\Sigma}_q(t)})
\end{aligned}
$$

扩散概率模型Diffusion Probabilistic Model

模型训练 尽管DPM的原理已经在2015年就给出了完整的数学证明，但直到2020年，才广泛进入到研究者的视野，即：Denoising Diffusion Probabilistic Models



1. 说明任意时刻的数据x_t服从于高斯分布

$$
\begin{aligned}
\boldsymbol{x}_t &= \sqrt{\alpha_t}\boldsymbol{x}_{t-1} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t}\left(\sqrt{\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2}\right) + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t-\alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}^*_{t-2} + \sqrt{1-\alpha_t}\boldsymbol{\epsilon}^*_{t-1} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\sqrt{\alpha_t-\alpha_t\alpha_{t-1}}^2 + \sqrt{1-\alpha_t}^2}\,\boldsymbol{\epsilon}_{t-2} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{\alpha_t-\alpha_t\alpha_{t-1}+1-\alpha_t}\boldsymbol{\epsilon}_{t-2} \\
&= \sqrt{\alpha_t\alpha_{t-1}}\boldsymbol{x}_{t-2} + \sqrt{1-\alpha_t\alpha_{t-1}}\boldsymbol{\epsilon}_{t-2} \\
&= \dots \\
&= \sqrt{\prod_{i=1}^{t}\alpha_i}\boldsymbol{x}_0 + \sqrt{1-\prod_{i=1}^{t}\alpha_i}\boldsymbol{\epsilon}_0 \\
&= \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}_0 \\
&\sim \mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\,\mathbf{I})
\end{aligned}
$$

2. 说明Loss中的迭代项是高斯分布

$$
\begin{aligned}
q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_t, \boldsymbol{x}_0) &= \frac{q(\boldsymbol{x}_t|\boldsymbol{x}_{t-1}, \boldsymbol{x}_0)q(\boldsymbol{x}_{t-1}|\boldsymbol{x}_0)}{q(\boldsymbol{x}_t|\boldsymbol{x}_0)} \\
&= \frac{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\alpha_t}\boldsymbol{x}_{t-1}, (1-\alpha_t)\mathbf{I})\mathcal{N}(\boldsymbol{x}_{t-1}; \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0, (1-\bar{\alpha}_{t-1})\mathbf{I})}{\mathcal{N}(\boldsymbol{x}_t; \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0, (1-\bar{\alpha}_t)\mathbf{I})} \\
&\propto \exp\left\{-\left[\frac{(\boldsymbol{x}_t - \sqrt{\alpha_t}\boldsymbol{x}_{t-1})^2}{2(1-\alpha_t)} + \frac{(\boldsymbol{x}_{t-1} - \sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_{t-1})} - \frac{(\boldsymbol{x}_t - \sqrt{\bar{\alpha}_t}\boldsymbol{x}_0)^2}{2(1-\bar{\alpha}_t)}\right]\right\} \\
&= \exp\left\{-\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]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{(-2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1} + \alpha_t\boldsymbol{x}^2_{t-1})}{1-\alpha_t} + \frac{(\boldsymbol{x}^2_{t-1} - 2\sqrt{\bar{\alpha}_{t-1}}\boldsymbol{x}_{t-1}\boldsymbol{x}_0)}{1-\bar{\alpha}_{t-1}} + C(\boldsymbol{x}_t, \boldsymbol{x}_0)\right]\right\} \\
&\propto \exp\left\{-\frac{1}{2}\left[-\frac{2\sqrt{\alpha_t}\boldsymbol{x}_t\boldsymbol{x}_{t-1}}{1-\alpha_t} + \frac{\alpha_t\boldsymbol{x}^2_{t-1}}{1-\alpha_t} + \frac{\boldsymbol{x}^2_{t-1}}{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]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[(\frac{\alpha_t}{1-\alpha_t} + \frac{1}{1-\bar{\alpha}_{t-1}})\boldsymbol{x}^2_{t-1} - 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]\right\} \\
&= \exp\left\{-\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}^2_{t-1} - 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]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{\alpha_t - \bar{\alpha}_t + 1 - \alpha_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}^2_{t-1} - 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]\right\} \\
&= \exp\left\{-\frac{1}{2}\left[\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\boldsymbol{x}^2_{t-1} - 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]\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right)\left[\boldsymbol{x}^2_{t-1} - 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]\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{1-\bar{\alpha}_t}{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}\right)\left[\boldsymbol{x}^2_{t-1} - 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]\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\frac{1}{\frac{(1-\alpha_t)(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}}\right)\left[\boldsymbol{x}^2_{t-1} - 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]\right\} \\
&\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}\mathbf{I}}_{\Sigma_q(t)})
\end{aligned}
$$

扩散概率模型Diffusion Probabilistic Model

3. 根据两个高斯分布的KL可解，求出具体的解析形式

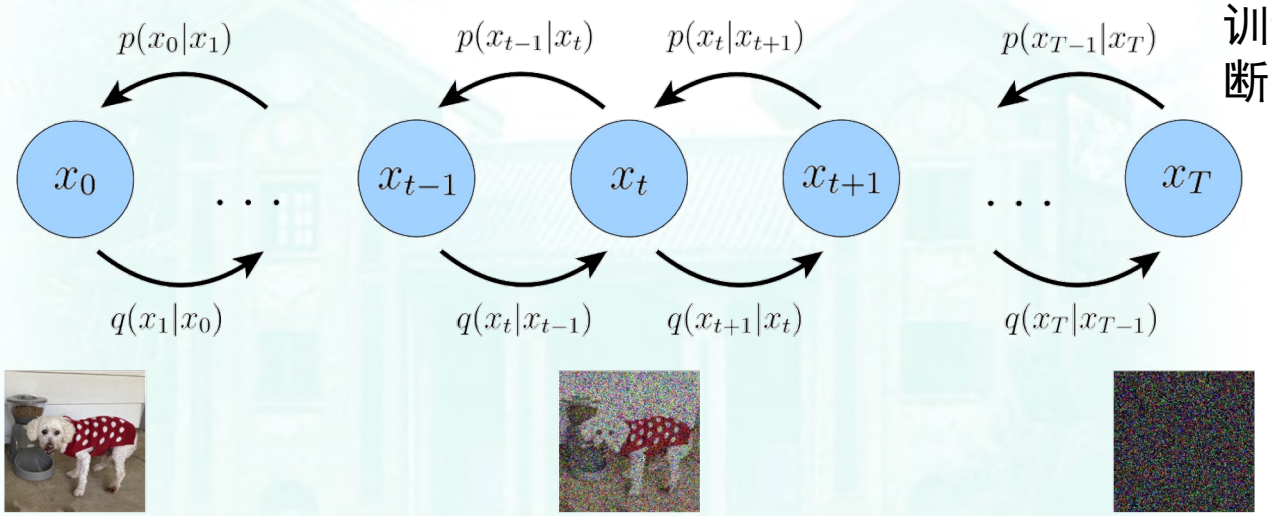
$$L_{t-1} = \mathbb{E}_q\left[\frac{1}{2\sigma_t^2}\|\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0) - \boldsymbol{\mu}_\theta(\mathbf{x}_t,t)\|^2\right] + C$$

$$D_{\mathrm{KL}}(\mathcal{N}(\boldsymbol{x};\boldsymbol{\mu}_x,\boldsymbol{\Sigma}_x)\,\|\,\mathcal{N}(\boldsymbol{y};\boldsymbol{\mu}_y,\boldsymbol{\Sigma}_y)) = \frac{1}{2}\left[\log\frac{|\boldsymbol{\Sigma}_y|}{|\boldsymbol{\Sigma}_x|} - d + \mathrm{tr}(\boldsymbol{\Sigma}_y^{-1}\boldsymbol{\Sigma}_x) + (\boldsymbol{\mu}_y-\boldsymbol{\mu}_x)^T\boldsymbol{\Sigma}_y^{-1}(\boldsymbol{\mu}_y-\boldsymbol{\mu}_x)\right]$$

$$\tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0) = \frac{\sqrt{\alpha_t}(1-\bar{\alpha}_{t-1})}{1-\bar{\alpha}_t}\mathbf{x}_t + \frac{\sqrt{\bar{\alpha}_{t-1}}\beta_t}{1-\bar{\alpha}_t}\mathbf{x}_0$$ 配完化简后的结果，但是我哪有X0啊现在

之前咱们说Xt可以由X0计算得到，现在逆一下 $\mathbf{x}_0 = \frac{1}{\sqrt{\bar{\alpha}_t}}(\mathbf{x}_t - \sqrt{1-\bar{\alpha}_t}\mathbf{z}_t)$

$$\tilde{\mu}_t = \frac{1}{\sqrt{a_t}}(x_t - \frac{\beta_t}{\sqrt{1-\bar{a}_t}}z_t)$$ 最终结果



训练网络不断拟合噪声z

4. 求和，简化后=>

$$L_{\text{simple}}(\theta) := \mathbb{E}_{t,\mathbf{x}_0,\boldsymbol{\epsilon}}\left[\left\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t)\right\|^2\right]$$

扩散概率模型Diffusion Probabilistic Model

Algorithm 1 Training

1: **repeat**
2: $\mathbf{x}_0 \sim q(\mathbf{x}_0)$
3: $t \sim \text{Uniform}(\{1, \dots, T\})$
4: $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
5: Take gradient descent step on
$$\nabla_\theta \left\| \boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\boldsymbol{\epsilon}, t) \right\|^2$$
6: **until** converged

Algorithm 2 Sampling

1: $\mathbf{x}_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
2: **for** $t = T, \dots, 1$ **do**
3: $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ if $t > 1$, else $\mathbf{z} = \mathbf{0}$
4: $\mathbf{x}_{t-1} = \frac{1}{\sqrt{\alpha_t}} \left(\mathbf{x}_t - \frac{1-\alpha_t}{\sqrt{1-\bar{\alpha}_t}} \boldsymbol{\epsilon}_\theta(\mathbf{x}_t, t) \right) + \sigma_t \mathbf{z}$
5: **end for**
6: **return** $\mathbf{x}_0$

$$x_t = \sqrt{\alpha_t} x_{t-1} + \sqrt{1-\alpha_t} \epsilon^*_{t-1}$$

Jonathan Ho et al. Denoising Diffusion Probabilistic Models. arXiv:2006.11239

扩散概率模型Diffusion Probabilistic Model

Stable Diffusion应用示例

Prompt: A silver mech horse running in a dark valley, in the night, Beeple, Kaino University, high-definition picture, unreal engine, cyberpunk

应用场景

- 图像生成
- 多模态
- 语音、文本等领域
- 例如：DALLE2、Imagen、GLIDE

缺陷与其他改进

- 需要迭代采样，速度相对来说较慢；目前已经有了不少对应的改进方法，例如DDIM和ANALYTIC-DPM等。
- 需要整个过程中，输入输出的大小保持相同维度；目前还没有比较好的方法。
- 加噪过程的最终分布是假设的标准高斯噪声，在风格转换之类的条件生成中受到挑战；已有了不少改进，例如Cold Diffusion已经取消了标准高斯的假设。
- 因为前两者的限制导致计算量过大，例如Stable Diffusion据说使用了4000台 A100显卡集群，用了一个月时间。