计算机应用数学

Quaternijkon

计算机应用数学课程笔记

September 10, 2024

目录

1. 随机游走与马尔可夫链

1.1. 引言

Definition

随机游走 在有向图上：从一个起始顶点生成一系列顶点，每次随机选择一个出边，沿着这条边移动到一个新的顶点，并重复这个过程。 正式定义如下：

$$p(t)P = p(t+1)$$

其中，$p(t)$ 是一个行向量，它的每个分量表示在时间 t 时每个顶点的概率质量分布，P 是所谓的转移矩阵，$P_{i,j}$ 是游走从顶点 i 选择顶点 j 的概率。

Example

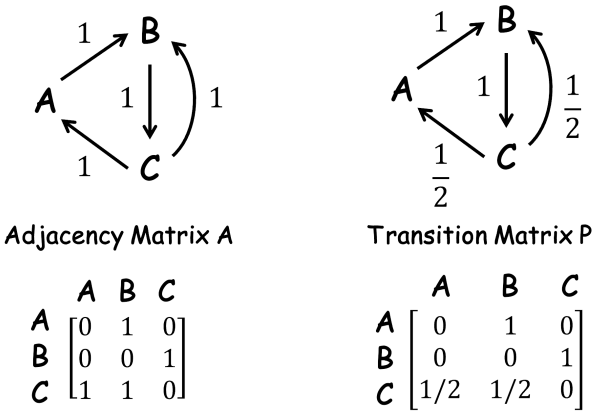


图 1　邻接矩阵与转移矩阵

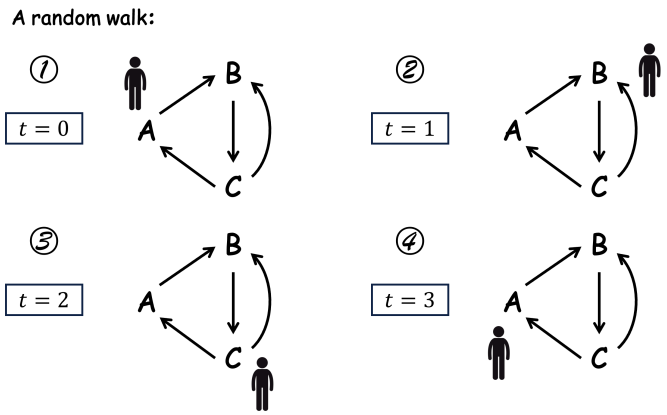


图 2　随机游走

Markov 链

有限的状态集合。

p_{xy}：从状态x到状态y的转移概率，$\sum_y p_{xy} = 1$。

Markov 链可以表示为有向图，其中从顶点 x 到顶点 y 的权重为p_{xy}。

Random walk	Markov chain
图 Graph	随机过程 Stochastic process
顶点 Vertex	状态 State
强连通 Strongly connected	持续 Persistent
非周期的 Aperiodic	非周期的 Aperiodic
强连通且非周期的 Strongly connected and aperiodic	遍历的 Ergodic
无向图 Undirected graph	时间可逆的 Time reversible

表 1 随机游走与马尔可夫链

我们将在本节中介绍以下内容：

- 示例：PageRank 和 Markov 决策过程。
- 平稳分布。
- 收敛性。
- Markov 过程。

Example

PageRank

将网页看作一个图：每个网页是一个顶点，超链接是边。

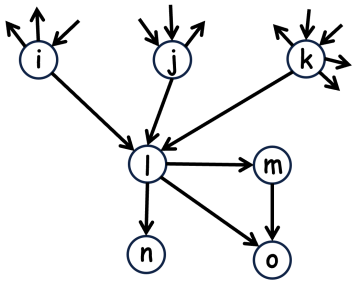
目标：根据重要性对网页进行排序。

Insight

一个网页的链接越多，它就越重要。

将入链看作投票，著名网站有更多的入链。

此外，来自重要网页的链接权重更大。



$$r_L = \frac{r_i}{3} + \frac{r_j}{2} + \frac{r_k}{4}$$

$$r_j = \sum_{i \to j} \frac{r_i}{d_i} \qquad (*)$$

图 3

随机邻接矩阵

d_i 是节点 i 的度。

如果 $i \to j$，则 $M_{ji} = \frac{1}{d_i}$。

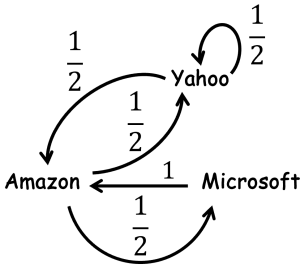
排序向量

r_i 是页面 i 的重要性得分。

公式 (*) 可以重写为：

$r = M \cdot r$

Example



$$M = \begin{array}{c} \\ Y \\ A \\ M \end{array} \begin{array}{c} \begin{array}{ccc} Y & A & M \end{array} \\ \begin{bmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1 \\ 0 & 1/2 & 0 \end{bmatrix} \end{array}$$

$$\begin{bmatrix} Y \\ A \\ M \end{bmatrix} = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

图 4

第一次迭代

$$\begin{bmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{1}{6} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}$$

$$PR(Y)^1 = \frac{1}{2}PR(Y)^0 + \frac{1}{2}PR(A)^0 = \frac{1}{2} \cdot \frac{1}{3} + \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{3}$$

$$PR(A)^1 = \frac{1}{2}PR(Y)^0 + 1 \cdot PR(M)^0 = \frac{1}{2} \cdot \frac{1}{3} + \frac{1}{3} = \frac{3}{6}$$

$$PR(M)^1 = \frac{1}{2}PR(A)^0 = \frac{1}{2} \cdot \frac{1}{3} = \frac{1}{6}$$

第二次迭代

$$\begin{bmatrix} \frac{5}{12} \\ \frac{1}{3} \\ \frac{1}{4} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \cdot \begin{bmatrix} \frac{1}{3} \\ \frac{1}{2} \\ \frac{1}{6} \end{bmatrix}$$

...

收敛

$$\begin{bmatrix} \frac{2}{5} \\ \frac{2}{5} \\ \frac{1}{5} \end{bmatrix} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{2} & 0 & 1 \\ 0 & \frac{1}{2} & 0 \end{bmatrix} \cdot \begin{bmatrix} \frac{2}{5} \\ \frac{2}{5} \\ \frac{1}{5} \end{bmatrix}$$

Markov 过程（Markov 决策过程）

$$\mathcal{S}, \mathcal{A}, \mathcal{R}, \mathbb{P}, \gamma$$

$\mathcal{S}$：状态集合。

$\mathcal{A}$：动作集合。

$\mathcal{R}$：在状态s下执行动作a的奖励$r(s,a)$。

$\mathbb{P}$：在状态 s 下执行动作 a 后转移到下一个状态 s' 的转移概率 $P(s' \mid s, a)$。

γ：折扣因子。

MDP（Markov 决策过程）：

$t = 0$初始状态$s_0 \sim p(s_0)$
对于$t = 0$到结束：
    执行动作a_t
    获得奖励$r_t \sim R(\cdot \mid s_t, a_t)$
    获得下一个状态$s_{t+1} \sim P(\cdot \mid s_t, a_t)$
    代理获得奖励r_t和状态s_{t+1}

算法 1　Markov 决策过程

目标：最大化长期奖励（累计奖励）$\sum_{t \geq 0} D^t r_t$。

Example

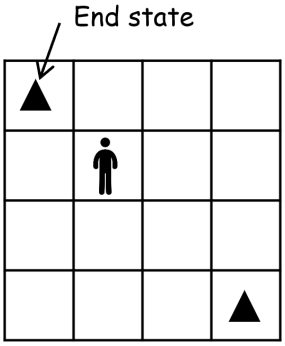


图 5

动作集合 = {左, 右, 上, 下} 到达空白格的奖励 → 使用最小化的动作数到达终点状态。

1.2. 平稳分布 Stationary Distribution

设 p_t 是随机游走经过 t 步后的概率分布。通过以下公式定义长期平均概率分布 a_t：

$$a_t = \frac{1}{t}(p_0 + p_1 + \cdots + p_{t-1})$$

Markov 链的基本定理：

对于一个连通的 Markov 链，它收敛于一个极限概率向量 x，满足：

$$XP = x; \sum_i x_i = 1 \Rightarrow X[P - I, 1] = [0, 1]$$

引理 1.3.1 设 P 是一个连通的 Markov 链的转移概率矩阵。通过在矩阵 $P - I$ 上增加一列 1 的列构造出的 $n \times (n + 1)$ 矩阵 $A = [P - I, 1]$ 的秩为 n。

证明：作业

定理 1.3.2 设 P 是连通 Markov 链的转移概率矩阵，则存在一个唯一的概率向量 π 满足 $\pi P = \pi$。此外，对于任何初始分布，$\lim_{t \to \infty} a_t$ 存在且等于 π。

证明：考虑 a_t 和 a_{t+1} 的差，$a_t - a_{t+1} = a_t P$：

$$
\begin{aligned}
a_tP - a_t &= \frac{1}{t}[p_0P + p_1P + \cdots + p_{t-1}P] - \frac{1}{t}[p_0 + p_1 + \cdots + p_{t-1}] \\
&= \frac{1}{t}[p_1 + p_2 + \cdots + p_t] - \frac{1}{t}[p_0 + p_1 + \cdots + p_{t-1}] \\
&= \frac{1}{t}(p_t - p_0)
\end{aligned}
$$

因此，$b_t = a_tP - a_t$ 满足 $\mid b_t \mid \leq \frac{2}{t}$，并且当 $t \to \infty$ 时趋于 0。

根据引理 1.3.1，$A = [P - I, 1]$ 的秩为 n。由于 A 的所有行和为 0，$n \times n$ 矩阵 B 中除了最后一列以外的所有列是可逆的。

令 c_t 由 $b_t = a_tP - a_t$ 去掉第一列得到，使得 $a_tB = [c_t, 1]$。

因此 $a_t \to [c_t, 1] \to [0, 1]$ 并且 $a_t \to [0, 1]B^{-1}$。

因此 $a_t \to \pi$，我们得出 π 是一个概率向量。

由于 $a_t[P - I] = b_t \to 0$，我们得到 $\pi[P - I] = 0$。

由于 A 的秩为 n，这是唯一的解，如所要求的。

引理 1.3.3 对于在强连通图上的随机游走，若边上带有概率，向量 π 满足 $\pi_x p_{xy} = \pi_y p_{yx}$ 对于所有 x和y，且 $\sum_x \pi_x = 1$，那么 π 是随机游走的平稳分布。

证明： $\pi_x p_{xy} = \pi_y p_{yx}$，两边求和，$\pi_x = \sum_y \pi_y p_{yx}$，因此 (pi) 满足 $\pi = \pi P$。(By Theorem 1.3.2 ...)

1.3. 无向图上随机游走的收敛性

下一个问题：游走需要多长时间开始反映 Markov 过程的平稳概率？

示例： 这需要很长时间才能收敛。游走很难通过图的两个部分之间的窄通道。

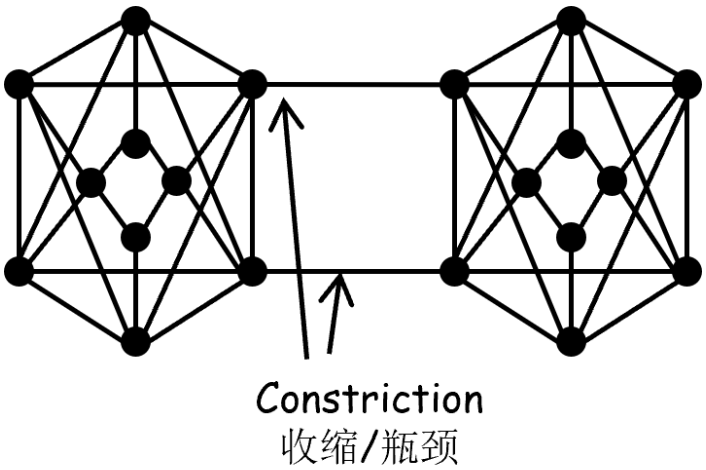


图 6

我们在下面定义了 Markov 链的收缩的一个组合度量，称为归一化导通率。

定义 1.3.1 设 $\varepsilon > 0$。Markov 链的 $\varepsilon - mixing$时间是最小的整数 t，使得对于任何初始分布 P_0，第 t 步的平均概率分布与平稳分布之间的 1-范数距离最多为 ε。

$$
\mid a_t - \pi \mid \leq \varepsilon
$$

定义 1.3.2 对于一个顶点子集 S，令 $\pi(S)$ 表示 $\sum_{x \in S} \pi_x$。归一化导通率定义为：

$$\Phi(S)=\frac{\sum_{(x,y)\in(S,\bar{S})}\pi_x p_{xy}}{\min(\pi(S),\pi(\bar{S}))}$$

其中，$\bar{S}=V-S$ 。$\pi(S)$ 是平稳分布下，Markov 链处于某状态属于 S 的概率。

定义 1.3.3 Markov 链的归一化导通率，记作 Φ ，定义为：

$$\Phi=\min_{S}\Phi(S)$$

定理 1.3.4. 在无向图上，随机游走的 $\varepsilon-mixing$时间为：

$$\Phi\left(\frac{\ln\left(\frac{1}{\pi_{\min}}\right)}{\Phi^2\varepsilon^3}\right)$$

其中，$\pi_{\min}$ 是任何状态的最小平稳概率。

使用归一化导通率证明收敛性。

接下来，我们应用定理 1.3.4 通过一些例子说明归一化导通率如何限制收敛速度。

① **一个一维的格子**

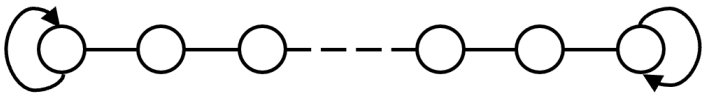


图 7

n 个顶点路径，两端都有自环。

平稳概率在所有顶点上是均匀的 $\frac{1}{n}$。

具有最小归一化导通率的集合是：
- 具有 $\pi\leq\frac{1}{2}$ 的集合；
- 包含前 $\frac{n}{2}$ 个顶点的集合。

从集合 S 到集合 $\bar{S}$ 的边的总导通率是：

$\pi_m p_{m,m+1}=\Omega\left(\frac{1}{n}\right),(m=\frac{n}{2})$

$\pi(S)=\frac{1}{2}$

因此，$\Phi(\bar{S})=2\pi_m p_{m,m+1}=\Omega\left(\frac{1}{n}\right)$

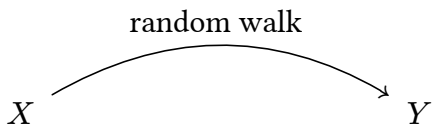
根据定理 1.3.4，对于 $\varepsilon=\frac{1}{100}$，经过 $O(n^2\log n)$ 步之后，$||a_t-\pi||\leq\frac{1}{100}$。

此图没有快速收敛性。

1.4. 4. 单位边权重的无向图上的随机游走

我们使用这种特殊类型的图来回答以下问题：

- 随机游走从 x 到达 y 的期望时间是多少？
- 从 x 到 y 并返回的期望时间是多少？
- 到达每个顶点的期望时间是多少？



① **命中时间**

h_{xy} —— 也称为发现时间。

引理 1.3.5. 从路径上的一个端点开始随机游走，穿过有 n 个顶点的路径到达另一端的期望时间是 $\Theta H(n^2)$。

证明：

$$h_{12} = 1$$

$$h_{i,i+1} = \frac{1}{2} + \frac{1}{2}(1 + h_{i-1,i+1})$$

$$= 1 + \frac{1}{2}(h_{i-1,i} + h_{i,i+1})$$

$$= 2 + h_{i-1,i}$$

因此，$h_{i,i+1} = 2i - 1$ ，$2 \le i \le n-1$

要从 1 走到 n，

$$h_{1,n} = \sum_{i=1}^{n-1} h_{i,i+1}$$

$$= \sum_{i=1}^{n-1}(2i-1)$$

$$= 2\sum_{i=1}^{n-1} i - \sum_{i=1}^{n-1} 1$$

$$= 2\frac{n(n-1)}{2} - (n-1)$$

$$= (n-1)^2$$

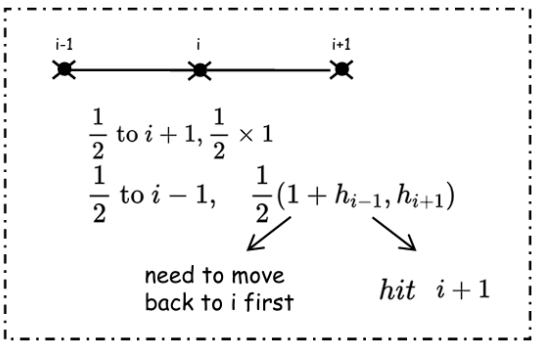


图 8

引理 1.3.6 设随机游走从顶点 1 到顶点 n，在包含 n 个顶点的链中。令 $t(i)$ 为在顶点 i 停留的期望时间。那么：

$$t(i) = \begin{cases} n-1, & i = 1 \\ 2(n-i), & 2 \le i \le n-1 \\ 1, & i = n \end{cases}$$

证明 现在 $t(n) = 1$，因为游走到达 n 时会停止。当游走到达 $n-1$ 时，一半的时间它会继续走向 n。因此，$t(n-1) = 2$。对于 $3 \le i \le n-1$，

$$t(i) = \frac{1}{2}[t(i-1) + t(i+1)]$$

$$t(1) = \frac{1}{2}t(2) + 1$$
$$t(2) = t(1) + \frac{1}{2}t(3)$$

因此我们得到

$t(i+1) = 2t(i) - t(i-1)$

因此，$t(i) = 2(n-i)$ 对于 $3 \leq i \leq n-1$。

$t(2) = 2(n-2)$，$t(1) = n-1$。

因此，在顶点停留的总时间是

$$n-1+2(1+2+...+n-2)+1 = (n-1+1+2\frac{(n-1)(n-2)}{2}+1) = (n-1)^2+1$$

这比 h_{1n} 多出 1。

② **往返时间**

$$commute(x,y) = h_{xy} + h_{yx}$$

③ **覆盖时间**

$Cover(x,G) \rightarrow$ 从顶点 x 开始的随机游走到达每个顶点至少一次的期望时间。

$$Cover(G) = \max_x Cover(x,G)$$

定理 1.3.7. 设 G 是一个有 n 个顶点和 m 条边的图。覆盖时间 $Cover(G)$ 的上界为 $4m(n-1)$。

证明. 进行一次从某个顶点 Z 开始的深度优先搜索。T 是结果生成的深度优先搜索生成树。深度优先搜索覆盖每个顶点。注意，生成树中的每条边在两个方向上都被遍历了两次。

$$Cover(Z,G) \leq \sum_{(x,y)\in T,(y,x)\in T} h_{xy}$$

推论. 如果 x 和 y 是相邻的，则 $h_{xy} + h_{yx} \leq 2m$，其中 m 是边的数量。该推论表明 $h_{xy} \leq 2m$。由于深度优先搜索树中有 $n-1$ 条边，并且每条边都被遍历两次，$Cover(Z) \leq 4m(n-1)$。因此，$Cover(G) \leq 4m(n-1)$。

1.5. 更多关于 Markov 的内容

△ **一个简单的 Markov 链** $\langle S, P\rangle$ S: 状态, P: 概率

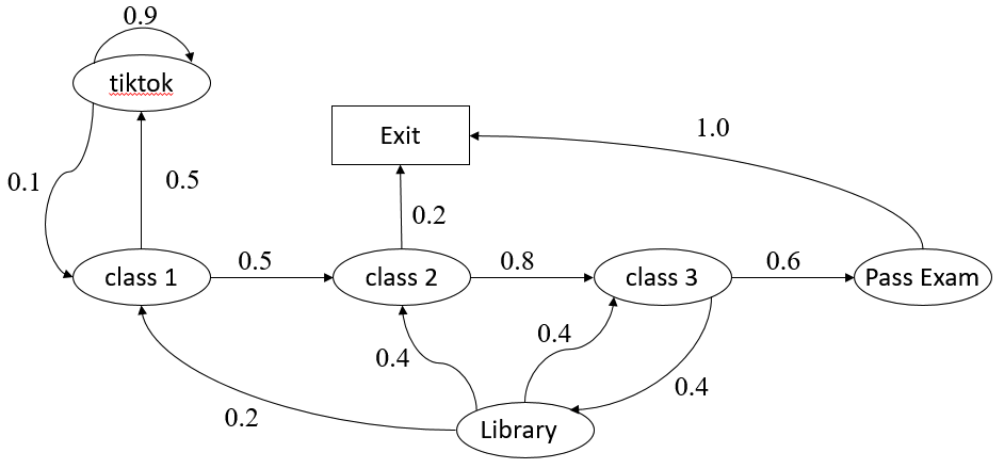


图 9

大量路径的例子：

$C_1, C_2, C_3, pass$

$C_1, TikTok, C_1, C_2, C_3, pass$

$C_1, C_2, C_3, Library, C_2, pass$

.....

△ **Markov 奖励过程** $\langle S, P, R, \gamma \rangle$ R: 奖励, γ: 折扣因子

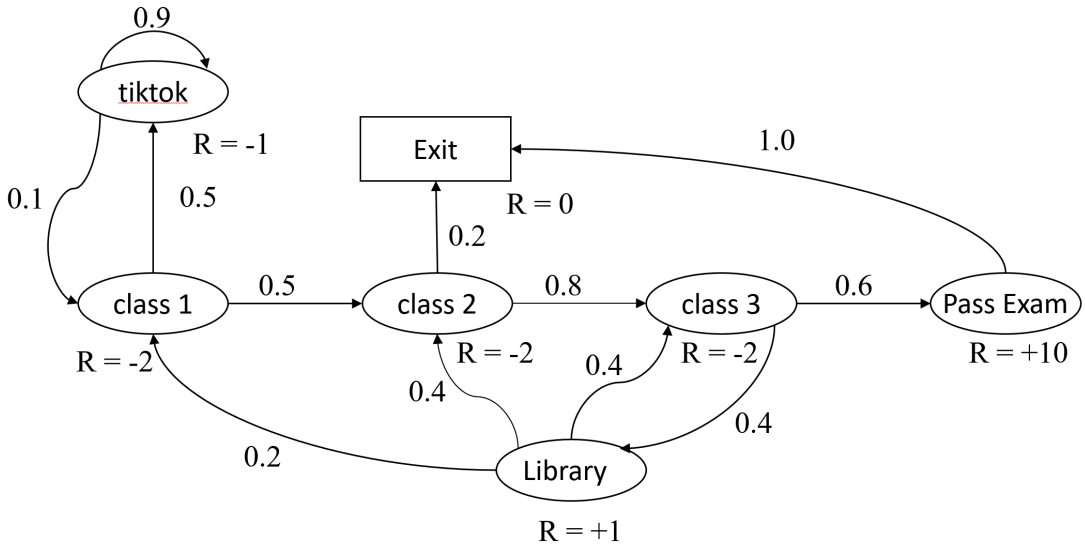


图 10

总奖励

$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$

$= \sum_{k=0}^{\infty} \gamma^k R_{t+k+1}$

状态的价值函数

$V(s) = \mathbb{E}[G_t \mid S_t = s]$

= 从该状态开始的期望奖励，即不同路径的平均奖励。

路径： $C_1, C_2, C_3, Pass, Exit$

$S_1 = C_1$ 且 $\gamma = 1/2$

$V_{C_1} = -2 - 2 \cdot \frac{1}{2} - 2 \cdot \frac{1}{4} - 10 \cdot \frac{1}{8} = -2.25$

路径： $C_1, TikTok, TikTok, C_1, C_2, Exit$

$V_{C_1} = -2 - 1 \cdot \frac{1}{2} - 1 \cdot \frac{1}{4} - 2 \cdot \frac{1}{8} - 2 \cdot \frac{1}{16} = -3.125$

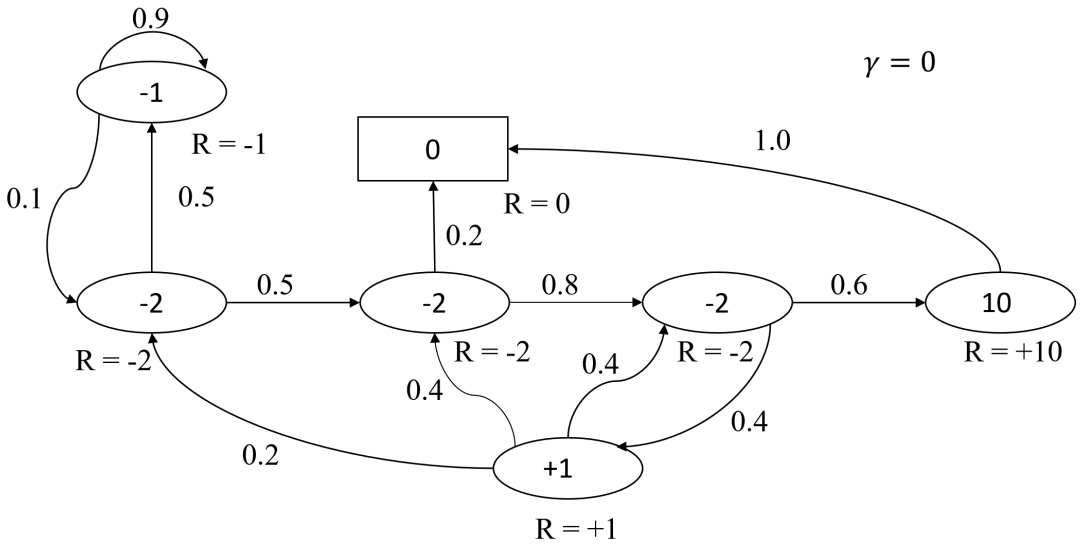


图 11

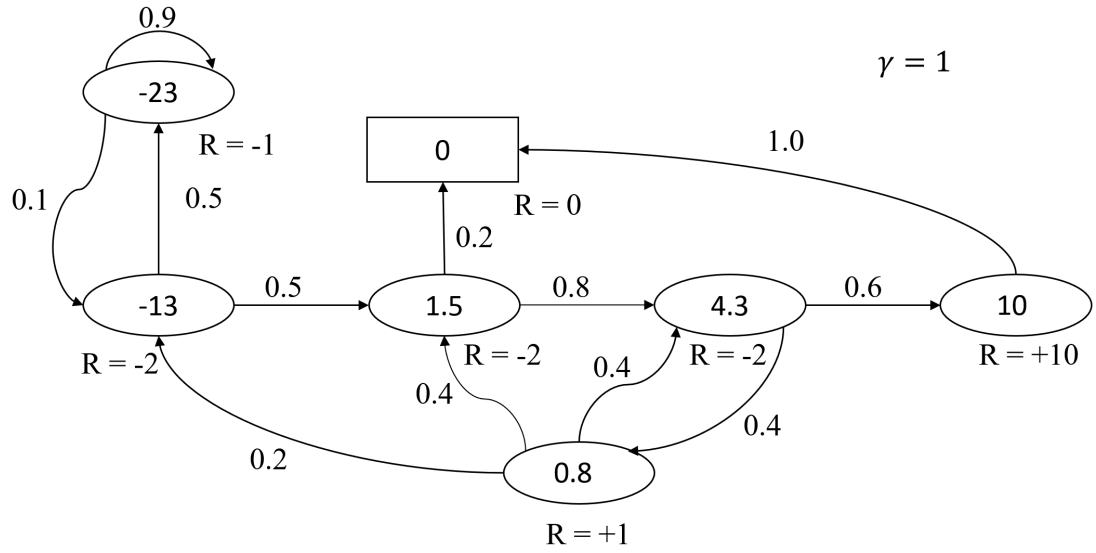


图 12

Bellman 期望方程

$$
\begin{aligned}
V(s) &= \mathbb{E}[G_t \mid s_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid s_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} \cdots) \mid s_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma G_{t+1} \mid s_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma V(s_{t+1}) \mid s_t = s]
\end{aligned}
$$

使用 s' 表示 $t+1$ 的可能状态，

$$
V(s) = R_s + \gamma \sum_{s' \in S} P_{ss'} V(s')
$$

对于 class 3

$$
4.3 = -2 + 0.6 \times 10 + 0.4 \times 0.8
$$

△ **Markov 决策过程** $\langle S, A, P, R, \gamma \rangle$ A：动作

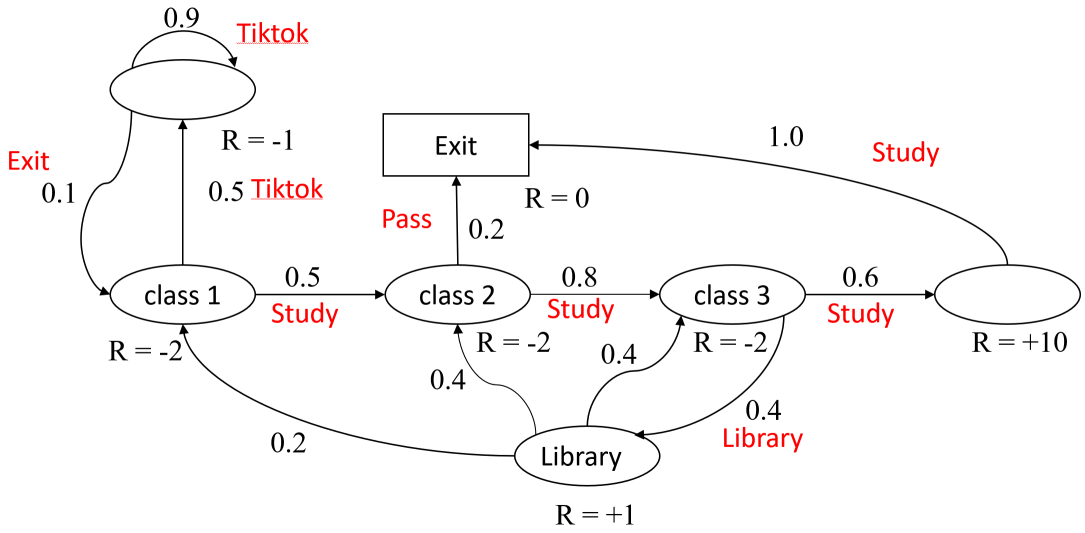


图 13

策略：采取动作的概率分布。

$$
\pi(a \mid s) = \mathbb{P}[A_t = a \mid S_t = s]
$$

给定一个 MDP $M = \langle S, A, P, R, \gamma \rangle$ 和一个策略 π。 序列 $S_1, S_2, \cdots$ 是一个 Markov 过程 $\langle S, p^\pi \rangle$。

状态和奖励序列 $S_1, R_2, S_2, \cdots$ 是一个 Markov 过程 $\langle S, P^\pi, R^\pi, \gamma \rangle$。

在策略 π 下，从状态 s 转移到 s' 的概率是：

$$
P^\pi_{ss'} = \sum_{a \in A} \pi(a \mid s) P^a_{ss'}
$$

在策略 π 下，状态 s 的奖励是：

$$
R^\pi_s = \sum_{a \in A} \pi(a \mid s) R^a_s
$$

价值函数：

$$V_\pi(s) = \mathbb{E}_\pi[G_t \mid S_t = s]$$

策略价值函数：

$$q_\pi(s,a) = \mathbb{E}_\pi[G_t \mid S_t = s, A_t = a]$$

Bellman 方程：

$$\begin{aligned} V_\pi(s) &= \mathbb{E}_\pi[R_{t+1} + \gamma v_\pi(s_{t+1}) \mid s_t = s] \\ q_\pi(s,a) &= \mathbb{E}_\pi[R_{t+1} + \gamma q_\pi(s_{t+1}, A_{t+1}) \mid s_t = s, A_t = a] \\ \text{因此} \quad V_\pi(s) &= \sum_{a \in \mathcal{A}} \pi(a \mid s) q_\pi(s,a) \end{aligned}$$

且 $q_\pi(s) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a V_\pi(s')$

因此，

$$q_\pi(s) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a \sum_{a' \in A} \pi(a' \mid s') q_\pi(s', a')$$

最优价值函数

$$V_*(s) = \max_\pi V_\pi(s)$$

最优动作-价值函数

$$q_*(s,a) = \max_\pi q_\pi(s,a)$$

我们可以使用 $q_*(s,a)$ 来得到最优策略 π：

$$\pi_*(a \mid s) = \begin{cases} 1 \text{ if } a = \arg\max_{a \in A} q_*(s,a) \\ 0 \text{ otherwise} \end{cases}$$

2. 距离和散度

欧几里得距离(Euclidean Distance)
曼哈顿距离(Manhattan Distance)
闵可夫斯基距离(Minkowski Distance)
余弦距离(Cosine Distance)
} 点距离(Point Distance)

汉明距离(Hamming Distance)
编辑距离(Edit Distance)
} 字符串距离(String Distance)

2.1. 点距离 Point Distance

物体之间的距离。

假设数据点来自 $M \subseteq \mathbb{R}^d$ 或 $M \subseteq \{0,1\}^d$

度量：

$D: M \times M \to \mathbb{R}$ 当且仅当对于所有 $x, y, z \in M$

$\Delta\ D(x,y) = 0 \Leftrightarrow x = y$

$\Delta\ D(x,y) = D(y,x)$

$\Delta\ D(x,z) \leq D(x,y) + D(y,z)$ – 三角不等式

注意 $D(x,y) \geq 0$

证明：

$D(x,y) + D(y,x) \geq D(x,x)$

由此得出 $2D(x,y) \geq D(x,x) \geq 0$ 并且 $D(x,y) \geq 0$

我们称 D 为距离函数。 D 可以用于聚类等。

I. 欧几里得距离

$$D_{l_2}(x,y) = \parallel x - y \parallel_2 = \sqrt{\sum_{i=1}^{d} \mid x_i - y_i \mid^2}$$

II. 曼哈顿距离

$$D_{l_1}(x,y) = \sum_{i=1}^{d} \mid x_i - y_i \mid$$

III. 明科夫斯基距离

$$D_{l_p}(x,y) = \parallel x - y \parallel_p = \left(\sum_{i=1}^{d} \mid x_i - y_i \mid^p \right)^{\frac{1}{p}}$$

它是欧几里得距离和曼哈顿距离的广义形式。

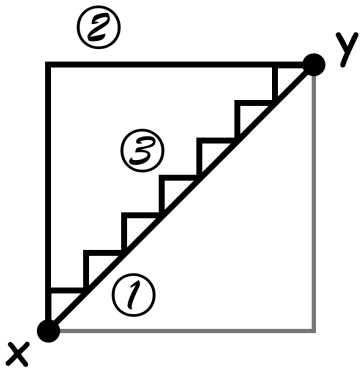


图 14

1. Euclidean 直线距离
2. Manhattan 出租车距离
3. Equivalent Manhattan

IV. 标准化欧氏距离

每个分量都有相同的均值和方差。

$$x^* = \frac{x - m}{s}$$

$$d(x,y) = \sqrt{\sum_{i=1}^{d} \left(\frac{x_i - y_i}{s_i} \right)^2}$$

2.2. 字符串距离 String Distance

I. 汉明距离

符号不同的位置数。

将一个字符串转换成另一个字符串所需的最少替换次数。

$c = a \oplus b$，其中 a 和 b 具有相同的长度。计算 c 中有多少个“1”。在网络中广泛使用。

II. 编辑距离 计算将一个字符串转换为另一个字符串所需的最小操作次数。操作符：

- 插入一个符号。
- 删除一个符号。
- 将符号 x 替换为符号 y（$y \neq x$），例如：$uxv \to uyv$。

它是汉明距离的广义形式。

$$lev(a,b) = \begin{cases} |\,a\,| & \text{if } |\,b\,| = 0 \\ |\,b\,| & \text{if } |\,a\,| = 0 \\ lev(\text{tail}(a),\ \text{tail}(b)) & \text{if head}(a) = \text{head}(b) \\ 1 + \min\{lev(\text{tail}(a), b), lev(a,\ \text{tail}(b)), lev(\text{tail}(a),\ \text{tail}(b))\} & \text{otherwise} \end{cases}$$

2.3. 集合距离 Set Distance

Jaccard 距离与相似度。

Jaccard 相似度

$$J(A,B) = \frac{\mid A \cap B \mid}{\mid A \cup B \mid}$$

Jaccard 距离

$$J_S(A,B) = 1 - J(A,B) = \frac{\mid A \cup B \mid - \mid A \cap B \mid}{\mid A \cup B \mid}$$

2.4. 变量与分布之间的距离 Distance between variables and distributions

熵

定义 1.5.1 随机变量 X 的（Shannon）熵为：

$$\begin{aligned} H[X] &= -\sum_x P(X=x)\log P(X=x) \quad \text{(离散情况)} \\ &= -E[\log P(X)] \end{aligned}$$

给定 Y 的条件熵 X：

$$\begin{aligned} H[X \mid Y] &= \sum_y P(Y=y) \sum_x P(X=x \mid Y=y) \log P(X=x \mid Y=y) \\ &= -E[\log P(X \mid Y)] \\ &= H[X,Y] - H[Y] \end{aligned}$$

Shannon 熵的性质：

1. $H[X] \geq 0$
2. $H[X] = 0$，如果存在 x_0：$X = x_0$
3. 如果 X 可以取 $n < \infty$ 个不同值（具有正概率），则 $H[X] \leq \log n$ 如果 X 是均匀分布的，$H[X] = \log n$。
4. $H[X] + H[Y] \geq H[X,Y]$ 当且仅当 X 和 Y 独立时取等号。
5. $H[X,Y] \geq H[X]$
6. $H[X \mid Y] \geq 0$ 当且仅当 X 在几乎所有 Y 给定的情况下是常数时取等号。
7. $H[X \mid Y] \leq H[X]$ 当且仅当 X 独立于 Y 时取等号。
8. $H[f(X)] \leq H[X]$ 对于任何可测函数 f，当且仅当 f 可逆时取等号。

引理 1.5.1（Shannon 熵的链式法则）

设 $X_1, X_2, ..., X_n$ 是在同一概率空间上的离散值随机变量，则：

$H[X_1, X_2, ..., X_n] = H[X_1] + \sum_{i=2}^{n} H[X_i \mid X_1, X_2, ..., X_{i-1}]$

定义 1.5.2（Shannon 熵的一般情况）

相对于参考测度 ρ，分布为 μ 的随机变量 X 的 Shannon 熵为：

$$H_\rho[x] = -E_\mu\left[\log \frac{d\mu}{d\rho}\right]$$

II. 交叉熵

两个概率分布 (p) 和 (q) 之间的交叉熵，基于相同的事件集，衡量的是使用针对估计概率分布 (q) 优化的编码方案来识别事件时，所需的平均比特数，而不是使用真实分布 (p)。

给定真实分布 (p)，使用非真实分布 (q) 指定策略略微消除系统不确定性所需付出努力的大小。

假设 (p) 是真实分布，(q) 是估计的（非真实）分布。使用 (p) 来识别一个事件时，所需的平均比特数为：

$H(p) = -\sum_{i=1}^{n} p_i \log p_i$

而使用 (q) 来表示该数值：

$H(p,q) = -\sum_{i=1}^{n} p_i \log q_i$

$= \sum_{i=1}^{n} p_i \log \frac{1}{q_i}$ ← 离散情况下的交叉熵

对于连续情况：

$H(p,q) = E_p[\log q]$

$= -\int_x p(x) \log q(x) dx$

应用： 交叉熵损失函数与逻辑回归。

真实概率 (p_i) 是真实标签，而给定的分布 (q_i) 是模型当前预测值。

考虑一个二元回归模型。在逻辑回归中，概率由逻辑函数 $g(z) = \frac{1}{1+e^{-z}}$ 给出，其中 z 是输入 x 的线性函数。

输出为 1 的概率为：

$q_{y=1} = \hat{y} = g(w \cdot x) = \frac{1}{1+e^{-w \cdot x}}$

$q_{y=0} = 1 - \hat{y}$

从定义我们可以得出：

$p \in \{y, 1-y\}, \quad y \in \{1, 0\}, \hat{y} = \frac{1}{1+e^{-w \cdot x}}$

我们使用交叉熵来衡量 (p) 和 (q) 之间的差异，

$$H(p,q) = -\sum_i p_i \log q_i = -y \log \hat{y} - (1-y) \log(1-\hat{y})$$

逻辑损失有时称为交叉熵损失或对数损失。

逻辑回归的交叉熵损失的梯度与线性回归中平方误差损失的梯度相同。

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{1p} \\ 1 & x_{21} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \dots & x_{np} \end{pmatrix} \in \mathbb{R}^{n \times (p+1)}$$

$$\hat{y}_i = f(x_{i1}, \dots, x_{ip}) = \frac{1}{1 + \exp(-\beta_0 - \beta_1 x_{i1} - \dots - \beta_p x_{ip})}$$

$$L(\beta) = -\sum_{i=1}^{N} [y_i \log \hat{y}_i + (1-y_i) \log(1-\hat{y}_i)]$$

那么：

$$\frac{\partial}{\partial\beta}L(\beta)=X^T(\hat{Y}-Y)$$

证明：

$$\frac{\partial}{\partial\beta_0}\ln\frac{1}{1+e^{-\beta_0+k_0}}=\frac{e^{-\beta_0+k_0}}{1+e^{-\beta_0+k_0}}$$

$$\frac{\partial}{\partial\beta_0}\ln\left(1-\frac{1}{1+e^{-\beta_0+k_0}}\right)=\frac{-1}{1+e^{-\beta_0+k_0}}$$

$$\frac{\partial L(\beta)}{\partial\beta}=-\sum_{i=1}^{N}\left[y_i\cdot\frac{e^{-\beta_0+k_0}}{1+e^{-\beta_0+k_0}}-(1-y_i)\cdot\frac{1}{1+e^{-\beta_0+k_0}}\right]$$

$$=-\sum_{i=1}^{N}[y_i-\hat{y}_i]=\sum_{i=1}^{N}(\hat{y}_i-y_i)$$

$$\frac{\partial}{\partial\beta_1}\ln\frac{1}{1+e^{-\beta_1x_{i1}+k_1}}=\frac{x_{i1}e^{k_1}}{e^{\beta_1x_{i1}}+e^{k_1}}$$

$$\frac{\partial}{\partial\beta_1}\ln\left(1-\frac{1}{1+e^{-\beta_1x_{i1}+k_1}}\right)=\frac{-x_{i1}e^{\beta_1x_{i1}}}{e^{\beta_1x_{i1}}+e^{k_1}}$$

$$\frac{\partial L(\beta)}{\partial\beta_1}=-\sum_{i=1}^{N}x_{i1}(y_i-\hat{y}_i)=\sum_{i=1}^{N}x_{i1}(\hat{y}_i-y_i)$$

III. 相对熵或 K-L 散度

给定两个概率分布 (P) 和 (Q)，

$$D_{KL}(P\parallel Q)=\sum_{x\in\mathcal{X}}p(x)\log\frac{P(x)}{Q(x)}$$

注意： 该值不对称！

对于连续情况：

$$D_{KL}(P\parallel Q)=\int_{-\infty}^{\infty}p(x)\log\frac{p(x)}{q(x)}d(x)$$

更一般地说，如果 (P) 和 (Q) 是可测空间 (mathcal{X}) 上的概率测度，并且 (P) 相对于 (Q) 是绝对连续的，则：

$$D_{KL}(P\parallel Q)=\int_{x\in\mathcal{X}}\log\frac{P(dx)}{Q(dx)}P(dx)$$

示例：

对于 (P)：

- 二项分布，参数 (p = 0.4)，(N = 2)

对于 (Q)：

- 均匀分布，(p = 1/3)

对应的概率表：

$$\begin{array}{rccc} \mathcal{X}: & 0 & 1 & 2 \\ P(x): & \frac{9}{25} & \frac{12}{25} & \frac{4}{25} \\ Q(x): & \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \end{array}$$

$$\begin{aligned} D_{KL}(P \parallel Q) &= \sum_{x \in \mathcal{X}} P(x) \ln \frac{P(x)}{Q(x)} \\ &= \frac{9}{25} \ln\left(\frac{9/25}{1/3}\right) + \frac{12}{25} \ln\left(\frac{12/25}{1/3}\right) + \frac{4}{25} \ln\left(\frac{4/25}{1/3}\right) \\ &= \frac{1}{25}(32 \ln 2 + 55 \ln 3 - 50 \ln 5) \\ &= 0.0853 \end{aligned}$$

$$\begin{aligned} D_{KL}(Q \parallel P) &= \sum_{x \in \mathcal{X}} Q(x) \ln \frac{Q(x)}{P(x)} \\ &= \frac{1}{3} \ln\left(\frac{1/3}{9/25}\right) + \frac{1}{3} \ln\left(\frac{1/3}{12/25}\right) + \frac{1/3}{\ln}\left(\frac{1/3}{4/25}\right) \\ &= 0.0975 \end{aligned}$$

应用：贝叶斯更新

KL 散度可以用于衡量从先验分布 $P(x)$ 到后验分布 $p(x \mid I)$ 中的信息增益。

如果发现某些新事实 $Y = y$，可以通过贝叶斯定理将后验分布从 $p(x \mid I)$ 更新为新的后验分布 $p(x \mid y, I)$，具体表达为：

$$p(x \mid y, I) = \frac{p(y \mid x, I) p(x \mid I)}{p(y \mid I)}$$

$$D_{KL}(p(x \mid y, I) \parallel p(x \mid I)) = \sum_{x} p(x \mid y, I) \log\left(\frac{p(x \mid y, I)}{p(x \mid I)}\right)$$

IV. Jensen-Shannon (JS) 散度

JS 散度是 KL 散度的对称和平滑版本。

$$D_{JS} = \frac{1}{2} KL(p \parallel \frac{p+q}{2}) + \frac{1}{2} KL(q \parallel \frac{p+q}{2})$$

应用：生成对抗网络（GAN）

为了学习生成器的分布 p_g 在数据 x 上的分布，我们定义了输入噪声变量的先验分布 $p_z(Z)$。表示一个从 $G(z; \theta_g)$ 到数据空间的映射，其中 G 是由参数 θ_g 表示的可微函数，该函数由多层感知机表示。

我们定义了第二个多层感知机 $D(X; \theta_d)$。训练 D 以最大化对真实样本和生成样本的正确分类概率。同时训练 G 以最小化 $\log(1 - D(G(z)))$。换句话说，D 和 G 进行双人极小极大博弈，目标函数为 $V(G, D)$。

$$\min_G \max_D V(D,G) = E_{X\sim P_{\text{data}}(X)}[\log D(X)] + E_{z\sim p_z(z)}[\log(1-D(G(z)))]$$

我们首先考虑对任何给定的 G 来优化 D。

引理 1.5.2 对固定的 G，最优的 D 是：

$$D_G^*(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_g(x)}$$

证明：对给定的 G，D 的训练准则是最大化以下目标函数：

$$\begin{aligned} V(G,D) &= \int_x p_{\text{data}}(x)\log(D(x))dx + \int_z p_z(z)\log(1-D(G(z)))dz \\ &= \int_x p_{\text{data}}(x)\log(D(x)) + p_g(x)\log(1-D(x))dx \end{aligned}$$

对于任意 $(a,b)\in\mathbb{R}^2\setminus\{0,0\}$，$y\to a\log(y)+b\log(1-y)$ 在 $[0,1]$ 中的最大值为 $\frac{a}{a+b}$。

D 的训练目标可以解释为最大化估计条件概率 $P(Y=y\mid x)$ 的对数似然函数，其中 Y 表示 x 是否来自 p_{data}（$y=1$）或来自 p_g（$y=0$）。

以下是图片内容翻译成中文，数学公式已按要求放入中：

—

极小极大变为：

$$\begin{aligned} C(G) &= \max_D V(G,D) \\ &= E_{x\sim p_{\text{data}}}[\log D_G^*(x)] + E_{z\sim p_z}[\log(1-D_G^*(G(z)))] \\ &= E_{x\sim p_{\text{data}}}[\log D_G^*(x)] + E_{x\sim p_g}[\log(1-D_G^*(x))] \\ &= E_{x\sim p_{\text{data}}}\left[\log\frac{p_{\text{data}}(x)}{p_{\text{data}}(x)+p_g(x)}\right] + E_{x\sim p_g}\left[\log\frac{p_g(x)}{p_{\text{data}}(x)+p_g(x)}\right] \end{aligned}$$

定理 1.5.3

$C(G)$ 的全局最小值当且仅当 $p_g = p_{\text{data}}$ 时取得。在这个点上，$C(G)$ 的值为 $-\log 4$。

证明：对于 $p_g = p_{\text{data}}$，$D_G^*(x) = \frac{1}{2}$。

因此，我们发现 $C(G) = \log\frac{1}{2} + \log\frac{1}{2} = -\log 4$。

为了证明这是 $C(G)$ 能达到的最优值，且仅当 $p_g = p_{\text{data}}$ 时达到，观察到：

$$E_{x\sim p_{\text{data}}}[-\log 2] + E_{x\sim p_g}[-\log 2] = -\log 4$$

通过从 $C(G) = V(D_G^*, G)$ 中减去该表达式，我们得到：

$$\begin{aligned} C(G) &= -\log(4) + D_{KL}\left(p_{\text{data}} \parallel \frac{p_{\text{data}}+p_g}{2}\right) + D_{KL}\left(p_g \parallel \frac{p_{\text{data}}+p_g}{2}\right) \\ &= -\log(4) + 2\cdot D_{JS}(p_{\text{data}} \parallel p_g) \end{aligned}$$

由于 JS 散度总是非负的，并且只有在两者相等时才为零，我们证明 $c^* = -\log(4)$ 且唯一的解是 $p_g = p_{\text{data}}$。

V. Wasserstein 距离（推土机距离）

如果 p 和 q 非常不同，即它们彼此距离很远且没有重叠，那么它们的 KL 散度没有意义，$J-S$ 散度是常数，因此梯度变为 0。

$$w(p,q) = \inf_{\gamma \in \Gamma(u,v)} \left(E_{(x,y)\sim\gamma} d(x,y)^p\right)^{1/p}$$

其中 $\Gamma(u,v)$ 是所有 u 和 v 的耦合集合，$W_\infty(u,v)$ 定义为 $\lim_{p\to+\infty} W_p(u,v)$。

W-距离也可用于比较离散和连续分布。

应用：Wasserstein GAN

为什么 Wasserstein 距离比 JS 或 KL 更好？

假设我们有两个概率分布 P, Q。

$\forall (x,y) \in P, x = 0, y \sim U(0,1)$

$\forall (x,y) \in Q, x = \theta, 0 \le \theta \le 1 \text{ and } y \sim U(0,1)$

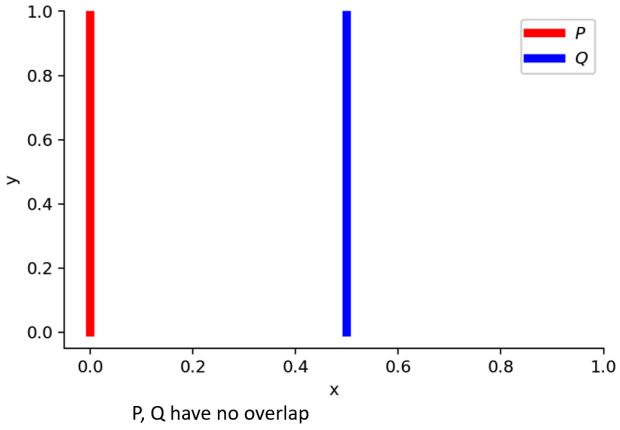


图 15　P, Q have no overlap

When $\theta \neq 0$:

$$D_{KL}(P \parallel Q) = \sum_{x=0; y\sim U(0,1)} 1 \cdot \log \frac{1}{0} = +\infty$$

$$D_{KL}(Q \parallel P) = \sum_{x=\theta; y\sim U(0,1)} 1 \cdot \log \frac{1}{0} = +\infty$$

$$D_{JS}(P,Q) = \frac{1}{2}\left(\sum_{x=0; y\sim U(0,1)} 1 \cdot \log \frac{1}{\frac{1}{2}} + \sum_{x=0; y\sim U(0,1)} 1 \cdot \log \frac{1}{\frac{1}{2}}\right) = \log 2$$

$W(P,Q) = |\ \theta\ |$

当 $\theta = 0$ 时，P, Q 完全重叠：

$$D_{KL}(P \parallel Q) = D_{KL}(Q \parallel P) = D_{JS}(P, Q) = 0$$
$$W(P, Q) = 0 = \mid \theta \mid$$

只有 W 提供了平滑的度量。

使用 W-距离作为 GAN 的损失函数。

在 $\Pi(p_r, p_g)$ 中穷尽所有可能的联合分布以计算 $\inf_{\gamma \sim \Pi(p_r, p_g)}$ 是不可行的。

基于 Kantorovich-Rubinstein 对偶性：

$$W(p_r, p_g) = \frac{1}{K} \sup_{\parallel f \parallel_L \leq K} \Big(E_{x \sim p_r}[f(x)] - E_{x \sim p_g}[f(x)] \Big)$$

其中 sup 是 inf 的相对概念。

现在我们想要衡量最小上界（最大值）。