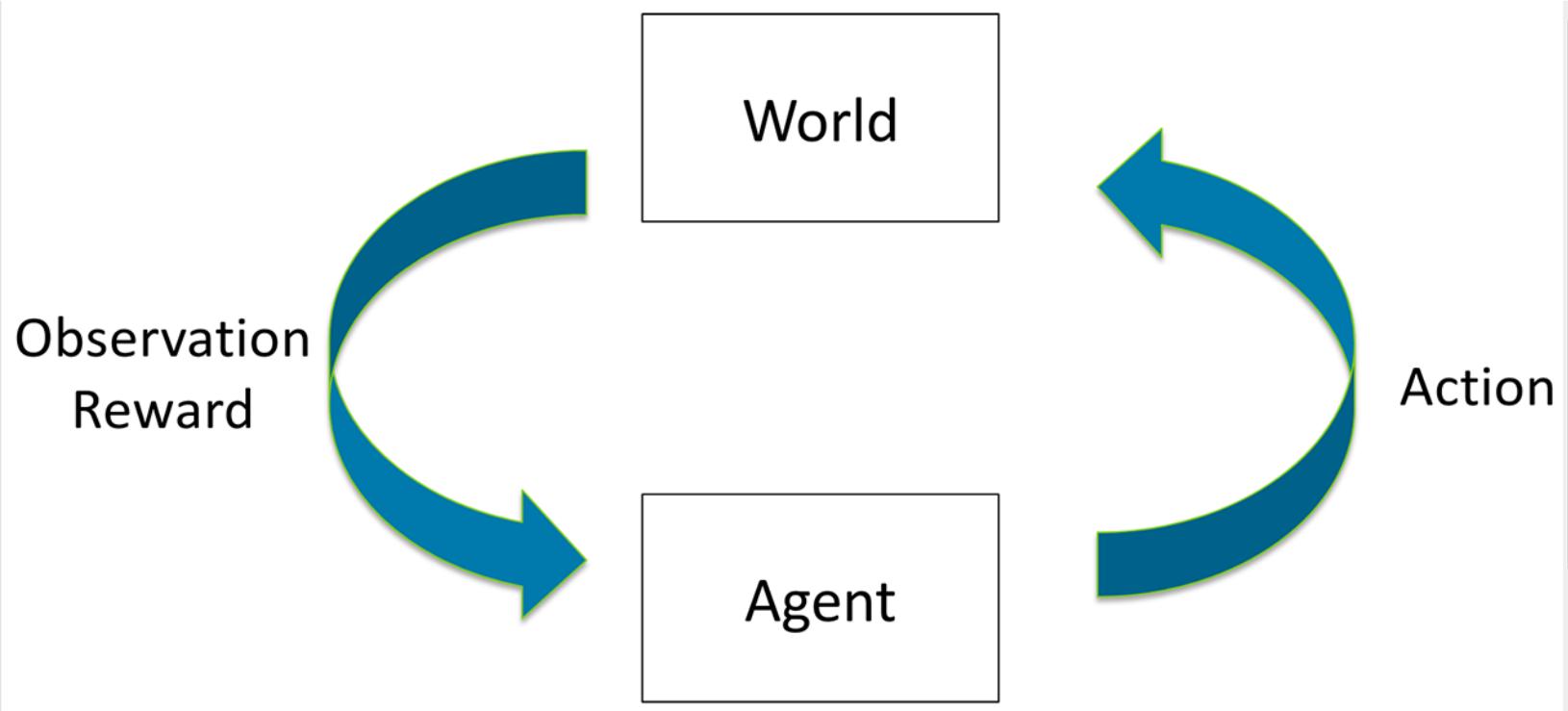


# 强化学习理论基础

RL858 1.09 版

何家志 整理  
讲课:Sound\_of\_wind  
2023 年 1 月 19 日



本讲义为我整理的强化学习理论基础的学习笔记, 仅供学习交流.<sup>1</sup>其主要来自于 bilibili 平台 up 主 Sound\_of\_wind<sup>2</sup>连载的《强化学习理论基础》系列视频. 该系列课程主要内容为强化学习相关的算法推导和定理证明, 特色是基本没有 gap, 证明的细节和动机讲解的非常清楚, 对于入坑 RL 的小白来说非常友好. 由于本人大四期间做了点有关 bandit 的磕盐, 深知强化学习理论方面的艰深, 因此将该课程的内容整理出 LATEX 讲义, 希望更多的人能看到如此宝藏课程. 特别感谢 up 主 Sound\_of\_wind 工作之余抽出时间制作课程视频, 我在与 up 主交流一些证明的细节问题与纰漏时都能得到极其耐心的回复, 给我提供了极大的帮助, 再一次特别感谢. 该课程预计涵盖内容:

- 基础知识 (集中不等式, 正定核与 RKHS, Minimax 定理, SVGD...)
- Bandit 问题 (UCB, Thompson 采样...)
- MDP 的定义以及各种变体
- 策略迭代, 值迭代以及理论保证
- MC, Q-Learning 以及相关理论保证
- REINFORCE 策略梯度以及理论保证
- Actor-Critic 以及相关理论保证
- 确定性策略梯度以及相关理论保证
- TRPO 相关理论
- 模仿学习算法及相关理论保证
- 逆强化学习及相关理论

<sup>1</sup>发现错误, 欢迎联系:[jiazhike@mail.ustc.edu.cn](mailto:jiazhike@mail.ustc.edu.cn)

<sup>2</sup> <https://space.bilibili.com/2374895>

- 最大熵强化学习理论及算法
- ... 其他内容



# 目录

<b>1 集中不等式</b>	<b>5</b>
1.1 尾概率, 马尔可夫不等式, 切比雪夫不等式 . . . . .	5
1.2 矩母函数与 Chernoff 不等式 . . . . .	8
1.3 次高斯性与 Hoeffding 界 . . . . .	11
1.4 条件期望与条件 Hoeffding 引理 . . . . .	17
1.5 鞍, 杜布分解与吾妻不等式 . . . . .	21
<b>2 最小最大定理及其证明</b>	<b>28</b>
<b>3 Bandit 问题</b>	<b>34</b>
3.1 Bandit 简介与遗憾分解引理 . . . . .	34
3.2 Explore-Then-Commit 算法 . . . . .	42
3.3 UCB 算法: 简介, 流程与公式推导 . . . . .	48
3.4 UCB1 算法的理论分析 . . . . .	50
3.5 贝叶斯定理简介及测度论角度的解释 . . . . .	58
3.6 共轭先验 (Conjugate Priors) . . . . .	66
3.7 贝叶斯 Bandit(Bayesian Bandits) . . . . .	71
3.8 汤普森采样 (Thompson Sampling) . . . . .	79
3.9 汤普森采样的贝叶斯遗憾分析 . . . . .	84
3.10 汤普森采样的频率派遗憾分析 . . . . .	96

# 1 集中不等式

## 1.1 尾概率, 马尔可夫不等式, 切比雪夫不等式

### 1. 尾概率 (Tail Probabilities)

问题 1. 假设  $X, X_1, X_2, \dots, X_n$  是一个 i.i.d. 随机变量序列, 假设其均值  $\mu = \mathbb{E}[X]$  及其方差  $\sigma^2 = \text{Var}[X]$  均存在. 若采用如下估计量来估计  $\mu$ :

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^n X_i$$

问:1. 该估计量是否是无偏估计?

2.  $\mu$  和  $\hat{\mu}$  之间大概相差多远?

首先看第一个问题, 直接套无偏估计定义即可

$$\mathbb{E}[\hat{\mu}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n} n \cdot \mu = \mu$$

对于第二个问题, 我们首先回顾方差的概念

方差的定义:  $\text{Var}[Z] := \mathbb{E}[(Z - \mathbb{E}[Z])^2]$ .

方差的性质: (1)  $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$  (注:  $X, Y$  必须相互独立)

(2)  $\text{Var}(kX) = k^2 \text{Var}(X)$ . 则

$$\text{Var}(\hat{\mu}) = \text{Var}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}(X_i) = \frac{1}{n^2} \cdot n \cdot \sigma^2 = \frac{\sigma^2}{n}$$

第二个问题比较微妙, 我们不知道  $\hat{\mu}$  和  $\mu$  相差多远该如何定义. 用方差来衡量  $\hat{\mu}$  和  $\mu$  的差距其实在现实应用中并不理想. 因为现实中人们往往关心  $\hat{\mu}$  严重偏离的概率, 我们在做统计估计时,  $\hat{\mu}$  和  $\mu$  接近时往往不会造成很严重的后果. 反过来, 如果  $\hat{\mu}$  和  $\mu$  相差非常大时, 往往会造成致命的后果. 现实中如果只从方差上来评估  $\hat{\mu}$  和  $\mu$  相差多远, 我们看不出估计量是多数时误差极低, 少数时错的很离谱; 还是多数时  $\hat{\mu}$  给出的估计都是可以接受的, 从而不会造成很严重的错误. 后者是我们理想中的估计量, 我们希望估计量可以犯一些小错误, 无论犯多少小错误都无所谓,

但是不要犯一些特别离谱的错误. 既然我们想度量犯离谱的错误的概率, 我们就要定义好什么样的错误属于难以接受的错误.

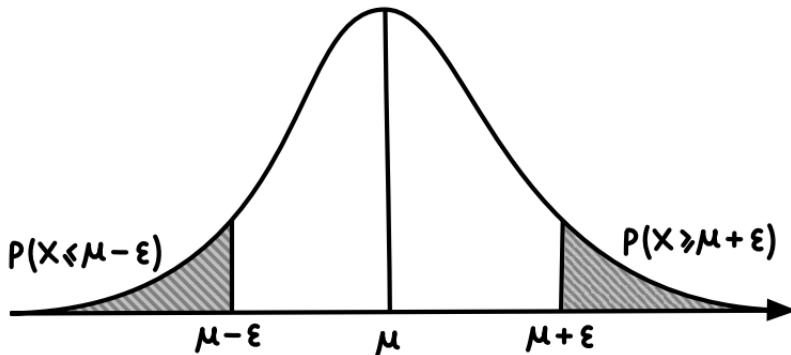


**定义 1.1.** 若  $X$  是一个均值为  $\mu$  的随机变量,  $\varepsilon$  是一个常数.

$\mathbb{P}(X \geq \mu + \varepsilon)$  称作 **右尾概率** (upper tail probability).

$\mathbb{P}(X \leq \mu - \varepsilon)$  称作 **左尾概率** (lower tail probability).

$\mathbb{P}(|X - \mu| \geq \varepsilon)$  称作 **双尾概率** (two - sided tail probability).



我们通常并不能显式算出尾概率, 但我们往往能算出他们的上界, 使得尾概率不超过那个界. 本节中介绍可以计算尾概率的界的不等式中最简单的类型.

**定理 1.2. (马尔可夫不等式).** 设  $(\Omega, \mathcal{A}, \mathbb{P})$  为概率空间,  $X$  为非负实值随机变量且  $X$  定义在  $\Omega$  上,  $\varepsilon > 0$  为一个常数, 则有:

$$\mathbb{P}(X \geq \varepsilon) \leq \frac{\mathbb{E}[X]}{\varepsilon}$$

证明. 设定义在  $\Omega$  上的函数  $s$ :

$$s(\omega) = \begin{cases} \varepsilon & \text{当 } X(\omega) \geq \varepsilon \\ 0 & \text{当 } X(\omega) < \varepsilon \end{cases} = \varepsilon \mathbb{I}_{\{\omega: X(\omega) \geq \varepsilon\}}(\omega)$$

则显然有  $0 \leq s(\omega) \leq X(\omega), \forall \omega \in \Omega$ .

$$\mathbb{E}[X] = \int_{\Omega} X(\omega) d\mathbb{P} \geq \int_{\Omega} s(\omega) d\mathbb{P} = \varepsilon \mathbb{P}(\{\omega \in \Omega : X(\omega) \geq \varepsilon\})$$

$$\mathbb{E}[X] \geq \varepsilon \mathbb{P}(X \geq \varepsilon)$$

因为  $\varepsilon > 0$ , 所以

$$\frac{\mathbb{E}[X]}{\varepsilon} \geq \mathbb{P}(X \geq \varepsilon)$$

□

马尔可夫不等式的证明很简单, 它的一个直接的推论就是 Chebyshev 不等式.

**推论 1.3. (Chebyshev不等式).** 对于任意随机变量  $X$  和常数  $\varepsilon > 0$  有:

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\text{Var}[X]}{\varepsilon^2}$$

证明. 将  $(X - \mathbb{E}[X])^2$  和  $\varepsilon^2$  代入马尔可夫不等式

$$\begin{aligned} \mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) &= \mathbb{P}((X - \mathbb{E}[X])^2 \geq \varepsilon^2) \\ &\leq \frac{\mathbb{E}[(X - \mathbb{E}[X])^2]}{\varepsilon^2} \end{aligned}$$

容易验证下边两个集合是相同的.

$$\{\omega : |X(\omega) - \mathbb{E}[X]| \geq \varepsilon\}$$

$$\{\omega : (X(\omega) - \mathbb{E}[X])^2 \geq \varepsilon^2\}$$

因此

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\text{Var}[X]}{\varepsilon^2}$$

□

## 1.2 矩母函数与 Chernoff 不等式

尽管 Chebyshev 不等式为我们提供了尾概率的上界, 但事实上这个界相对来说是比较松的。为了获得更紧的界, 一个直接的想法是去借鉴 Chebyshev 不等式的思路。Chebyshev 不等式是将随机变量  $(X - \mathbb{E}[X])^2$  代入马尔可夫不等式得到的。我们可以类似得到下面的界

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \frac{\mathbb{E}[|X - \mathbb{E}[X]|^k]}{\varepsilon^k}$$

代入不同的  $k$ , 不止只  
 $k=2$  的方差情况

我们可以通过  $k$  代入不同的数来得到不同的界, 选取一个最小的上界, 这个界肯定是比 Chebyshev 不等式提供的界紧的, 但缺点是实际计算的时候非常不方便。我们需要一个既足够紧, 又容易计算的界, Chernoff 界就满足了这两个要求, 它的不等式右侧是矩母函数, 矩母函数具有比较好的性质, 使得其更容易计算和处理。我们首先来看一下矩母函数的定义。

**定义 1.4. (矩母函数).** 假设  $X$  为一个随机变量, 若存在  $h > 0$  使得对于任意  $\lambda \in [0, h]$ ,  $\mathbb{E}[e^{\lambda X}]$  均存在, 则称  $X$  存在矩母函数 (Moment Generating Function, MGF)。记作  $M_X(\lambda)$ , 定义式为

$$M_X(\lambda) := \mathbb{E}[e^{\lambda X}] = \sum_{n=0}^{\infty} \frac{\lambda^n}{n!} \mathbb{E}[X^n]$$

$$\begin{aligned} k=0 \text{ 时}, D^0 &= 1 \\ 0! &= 1 \\ M_X^{(k)}(\lambda) &= \sum_{k=n}^{\infty} \mathbb{E}[X^k] \cdot \frac{D^{k-n}}{(k-n)!} \\ &= \mathbb{E}[X^{(k)}] \end{aligned}$$

矩母函数具有很多好的性质, 比如  $M_X^{(i)}(0) = \mathbb{E}[X^i]$ , 即矩母函数的第  $i$  阶导数在原点处的值等于  $X$  的  $i$  阶原点矩。随机变量的矩母函数不一定会存在的, 我们通常把矩母函数不存在的随机变量或分布称为是重尾的, 矩母函数存在的随机变量或分布则称为是轻尾的。有如下定义:

**定义 1.5. (重尾/轻尾).** 若随机变量  $X$  满足  $\mathbb{E}[e^{\lambda X}] = \infty, \forall \lambda > 0$ , 则称之为重尾(heavy tailed)。否则, 称之为 轻尾(light tailed)。

其实我们在机器学习中所接触的多数分布都是轻尾的, 一个重尾分布的例子是柯西分布。下面我们介绍 Chernoff 界, 其思路与 Chebyshev 界的思路类似, 我们把随机变量  $e^{\lambda(X-\mu)}$  代入马尔可夫不等式有

$$\mathbb{P}(X - \mu \geq \varepsilon) = \mathbb{P}(e^{\lambda(X-\mu)} \geq e^{\lambda\varepsilon}) \leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda\varepsilon}}$$

注意到这个式子对任意  $\lambda \in [0, h]$  均成立, 而每个  $\lambda$  都对应一个上界. 那么我们的  $\lambda$  应该取多少呢? 由于我们的目标是推尾概率的上界, 那么这个界是越紧越好. 出于这个思想, 我们需要对不等式右侧取下确界, 从而得到一个最紧的界, 即 Chernoff 界.

**定义 1.6. (Chernoff 界).** 对于任意随机变量  $X$ , 假设其均值存在且为  $\mu$ , 并且其矩母函数  $M_X(\lambda), \lambda \in [0, h]$  存在, 则  $X$  的 Chernoff 界 定义为:

$$\mathbb{P}(X - \mu \geq \varepsilon) \leq \inf_{\lambda \in [0, h]} \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda\varepsilon}}$$

注意到  $\mathbb{E}[e^{\lambda(X-\mu)}]$  是  $X - \mu$  的矩母函数, 我们计算一个随机变量的 Chernoff 界之前, 需要先计算其矩母函数.

**例 1.7.** 计算  $X \sim N(\mu, \sigma^2)$  的 Chernoff 界.

解.

$$\begin{aligned} M_X(\lambda) &= \mathbb{E}[e^{\lambda X}] \\ &= \int_{-\infty}^{\infty} e^{\lambda x} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\lambda x - \frac{(x-\mu)^2}{2\sigma^2}\right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{2\lambda x\sigma^2 - x^2 + 2x\mu - \mu^2}{2\sigma^2}\right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2 - 2x(\mu + \lambda\sigma^2) + \mu^2}{2\sigma^2}\right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{x^2 - 2x(\mu + \lambda\sigma^2) + (\mu + \lambda\sigma^2)^2 - (\mu + \lambda\sigma^2)^2 + \mu^2 - (\mu + \lambda\sigma^2)^2}{2\sigma^2}\right) dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{[x - (\mu + \lambda\sigma^2)]^2 + \mu^2 - \mu^2 - 2\mu\sigma^2\lambda - (\sigma^2\lambda)^2}{2\sigma^2}\right) dx \end{aligned}$$

其中蓝色部分我们凑出了高斯分布的指数部分，从而有

$$\begin{aligned}
 M_X(\lambda) &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{[x - (\mu + \lambda\sigma^2)]^2}{2\sigma^2} + \mu\lambda + \frac{\sigma^2\lambda^2}{2}\right) dx \\
 &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{[x - (\mu + \lambda\sigma^2)]^2}{2\sigma^2}\right) \underbrace{\exp\left(\mu\lambda + \frac{\sigma^2\lambda^2}{2}\right)}_{N(\mu+\lambda\sigma^2, \sigma^2) \text{ 的密度函数}} dx \\
 &= \exp\left(\mu\lambda + \frac{\sigma^2\lambda^2}{2}\right) \int_{-\infty}^{\infty} \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{[x - (\mu + \lambda\sigma^2)]^2}{2\sigma^2}\right)}_{N(\mu+\lambda\sigma^2, \sigma^2) \text{ 的密度函数}} dx \\
 &= \exp\left(\mu\lambda + \frac{\sigma^2\lambda^2}{2}\right)
 \end{aligned}$$

显然  $M_X(\lambda)$  对于任意  $\lambda \in \mathbb{R}$  均有定义

$$\inf_{\lambda \in \mathbb{R}} \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda\varepsilon}} = \inf_{\lambda \in \mathbb{R}} \frac{e^{\frac{\sigma^2\lambda^2}{2}}}{e^{\lambda\varepsilon}} = \inf_{\lambda \in \mathbb{R}} e^{\frac{\sigma^2\lambda^2}{2} - \lambda\varepsilon}$$

$$\begin{aligned}
 \lambda\sigma^2 - \varepsilon &= 0 \\
 \Rightarrow \lambda^* &= \frac{\varepsilon}{\sigma^2}
 \end{aligned}$$

从而有

$$\arg \min_{\lambda \in \mathbb{R}} e^{\frac{\sigma^2\lambda^2}{2} - \lambda\varepsilon} = \arg \min_{\lambda \in \mathbb{R}} \frac{\sigma^2\lambda^2}{2} - \lambda\varepsilon = \frac{\varepsilon}{\sigma^2}$$

取对数不影响单调性  
对入无导

代入  $\lambda = \frac{\varepsilon}{\sigma^2}$ , 因此有 Chernoff 界为

$$\inf_{\lambda \in \mathbb{R}} e^{\frac{\sigma^2\lambda^2}{2} - \lambda\varepsilon} = e^{\frac{\varepsilon^2}{2\sigma^2} - \frac{\varepsilon^2}{\sigma^2}} = e^{-\frac{\varepsilon^2}{2\sigma^2}}$$

故利用 Chernoff 界得到的高斯分布的尾概率界为



$$\mathbb{P}(X - \mu \geq \varepsilon) \leq e^{-\frac{\varepsilon^2}{2\sigma^2}}, \forall \varepsilon \geq 0$$

□

### 注记

很多满足这个 Chernoff 界的分布并不是高斯分布, 这就引出了次高斯分布.

### 1.3 次高斯性与 Hoeffding 界

**定义 1.8.** (次高斯性). 假设  $X$  是一个均值为  $\mu = \mathbb{E}[X]$  的随机变量, 若存在  $\sigma > 0$  使得

$$\mathbb{E}[\exp\{\lambda(X - \mu)\}] \leq \exp\left\{\frac{\sigma^2\lambda^2}{2}\right\}, \forall \lambda \in \mathbb{R}$$

则称它为  $\sigma$ -次高斯的, 其中  $\sigma$  称作次高斯参数.

**定理 1.9.** 若  $X$  为  $\sigma$ -次高斯随机变量, 则  $X$  满足

$$(1.10) \quad \mathbb{P}(X - \mu \geq \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right), \forall \varepsilon \geq 0$$

证明. 设  $\lambda \in (0, \infty)$ , 则易知

$$X - \mu \geq \varepsilon \iff \exp(\lambda(X - \mu)) \geq \exp(\lambda\varepsilon)$$

从而由事件的等价性

$$\begin{aligned} \mathbb{P}(X - \mu \geq \varepsilon) &= \mathbb{P}(\exp(\lambda(X - \mu)) \geq \exp(\lambda\varepsilon)) \\ &\leq \mathbb{E}[\exp(\lambda(X - \mu))] \exp(-\lambda\varepsilon) \quad (\text{应用Markov不等式}) \\ &\leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right) \exp(-\lambda\varepsilon) \quad (\text{次高斯性定义}) \\ &= \exp\left(\frac{\lambda^2\sigma^2}{2} - \lambda\varepsilon\right) \end{aligned}$$

$f(x) = \exp(x)$   
 $\hookrightarrow R \xrightarrow{(0, \infty)} \text{双射}$   
 $\text{单调增}$

$f^{-1}(x) = \frac{1}{\lambda} \ln(x)$   
 $\text{单调增}$

将  $\lambda = \frac{\varepsilon}{\sigma^2}$  代入上式, 得:

$$\mathbb{P}(X - \mu \geq \varepsilon) \leq \exp\left(-\frac{\varepsilon^2}{2\sigma^2}\right)$$

□

## 注记

次高斯性的定义使得其零均值化后的矩母函数是逐点小于高斯分布零均值化后的矩母函数，因此 Chernoff 界中求最小值的那个函数也逐点小于高斯分布对应的求最小值的函数。这就是次高斯随机变量可以使用高斯随机变量的尾概率界的原因。

次高斯随机变量在现实应用中广泛存在，一个典型的例子是所有有界随机变量都是次高斯的。我们在强化学习中每个时刻的奖励，每个状态的价值通常都是有界的。接下来介绍的 Hoeffding 引理保证了有界随机变量的次高斯性。 ★

**定理 1.11. (Hoeffding 引理)**. 设  $X$  是一个均值为  $\mu = \mathbb{E}[X]$  的随机变量，使得  $a \leq X \leq b$  几乎处处成立，则  $X$  是次高斯的，其次高斯参数为  $\sigma = \frac{b-a}{2}$ .

证明. 先证对于任意随机变量  $Y$ ，满足  $a \leq Y \leq b$  a.s.，有  $\text{Var}(Y) \leq \frac{(b-a)^2}{4}$ 。根据方差定义

$$\text{Var}(Y) = \mathbb{E}[(Y - \mathbb{E}[Y])^2]$$

$$= \min_t \mathbb{E}[(Y - t)^2] \quad \text{取最小值时, } t = \mathbb{E}[Y] \rightarrow \text{最大化到 } \frac{a+b}{2} \text{ 的距离}$$

$$\leq \mathbb{E}\left[\left(Y - \frac{a+b}{2}\right)^2\right] \leq \max_{y \in [a,b]} \left[\left(y - \frac{a+b}{2}\right)^2\right]$$

显然当  $y = a$  或  $y = b$  时取到最大值  $\frac{(b-a)^2}{4}$ 。所以  $\text{Var}(Y) \leq \frac{(b-a)^2}{4}$ 。设  $P$  为  $X$  的概率分布，定义  $\phi(\lambda) := \ln \mathbb{E}_P[e^{\lambda X}]$ ，则有

$$\phi'(\lambda) = \frac{(\mathbb{E}_P[e^{\lambda X}])'}{\mathbb{E}_P[e^{\lambda X}]} = \frac{\mathbb{E}_P[X e^{\lambda X}]}{\mathbb{E}_P[e^{\lambda X}]}$$

方差上界

$$\phi''(\lambda) = \frac{\mathbb{E}_P[X^2 e^{\lambda X}] \mathbb{E}_P[e^{\lambda X}] - \mathbb{E}_P[X e^{\lambda X}] \mathbb{E}_P[X e^{\lambda X}]}{\mathbb{E}_P[e^{\lambda X}]^2}$$

对称诱导

$$= \frac{\mathbb{E}_P[X^2 e^{\lambda X}]}{\mathbb{E}_P[e^{\lambda X}]} - \frac{\mathbb{E}_P[X e^{\lambda X}]^2}{\mathbb{E}_P[e^{\lambda X}]^2}$$

$$\int Q_\lambda d\lambda = 1$$

设  $Q_\lambda$  为一个关于  $X$  的分布 (Radon-Nikodym 导数)，定义为：

$$dQ_\lambda = \frac{e^{\lambda x}}{\mathbb{E}_P[e^{\lambda x}]} dP(x)$$

在概率测度  $P$  下,

$$\mathbb{E}_P[f(x)] = \int f(x) dP(x)$$

$$\mathbb{E}_P[X e^{\lambda X}] = \int x \cdot e^{\lambda x} dP(x)$$

在测度下对点  $x$  的“小概率增量”  
Sound\_of\_wind 讲课 何家志 码字

Reinforcement Learning

易验证  $Q_\lambda$  为概率测度. 则

$$\phi'(\lambda) = \frac{\mathbb{E}_P[X e^{\lambda X}]}{\mathbb{E}_P[e^{\lambda X}]} = \int x \frac{e^{\lambda x}}{\mathbb{E}_P[e^{\lambda X}]} dP(x) = \int x dQ_\lambda(x) = \mathbb{E}_{Q_\lambda}[X]$$

$$\phi''(\lambda) = \frac{\mathbb{E}_P[X^2 e^{\lambda X}]}{\mathbb{E}_P[e^{\lambda X}]} - \frac{\mathbb{E}_P[X e^{\lambda X}]^2}{\mathbb{E}_P[e^{\lambda X}]^2} = \mathbb{E}_{Q_\lambda}[X^2] - \mathbb{E}_{Q_\lambda}[X]^2 = \text{Var}_{Q_\lambda}[X]$$

对  $\phi(\lambda)$  在  $\lambda = 0$  处进行泰勒展开, 得:

$$\begin{aligned}\phi(\lambda) &= \phi(0) + \phi'(0)\lambda + \frac{1}{2}\phi''(\tilde{\lambda})\lambda^2, \quad \tilde{\lambda} \in [0, \lambda] \\ &= 0 + \mu\lambda + \frac{1}{2}\phi''(\tilde{\lambda})\lambda^2\end{aligned}$$

$$\phi'(0) = \frac{\mathbb{E}_P[X]}{\mathbb{E}_P[1]} = \mathbb{E}(X) = \mu$$

其中最后一项为拉格朗日余项. 因为  $\phi''(\tilde{\lambda}) = \text{Var}_{Q_{\tilde{\lambda}}}[X] \leq \frac{(b-a)^2}{4}$ , 所以

$$\phi(\lambda) \leq \mu\lambda + \frac{1}{2} \frac{(b-a)^2}{4} \lambda^2 = \boxed{\mu\lambda + \frac{\lambda^2(b-a)^2}{8}}$$



即有

$$\ln \mathbb{E}_P[e^{\lambda X}] \leq \mu\lambda + \frac{\lambda^2(b-a)^2}{8}$$

不等式两边取指数有

$$\mathbb{E}_P[e^{\lambda X}] \leq \exp\left(\mu\lambda + \frac{\lambda^2(b-a)^2}{8}\right)$$

不等式两边同乘  $e^{-\mu\lambda}$  有

$$\mathbb{E}_P[e^{\lambda X}] \underbrace{e^{-\mu\lambda}}_{\text{不包含 } X \text{ 的项}} \leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right)$$

整理得

$$\mathbb{E}_P[e^{\lambda(X-\mu)}] \leq \exp\left(\frac{\lambda^2 \left(\frac{b-a}{2}\right)^2}{2}\right)$$

刚好是  $X$  服从  $\frac{b-a}{2}$  为参数的次高斯分布的定义.

□

下面介绍一些次高斯随机变量的常用性质.

**定理 1.12.** 假设  $X$  是  $\sigma$ -次高斯的随机变量,  $X_1, X_2$  相互独立, 分别为  $\sigma_1, \sigma_2$ -次高斯, 则有:



- (1)  $\text{Var}[X] \leq \sigma^2$ .
- (2)  $\forall c$ , 有  $cX$  是  $|c|\sigma$ -次高斯的随机变量.
- (3)  $X_1 + X_2$  是  $\sqrt{\sigma_1^2 + \sigma_2^2}$ -次高斯的.

证明. (1) 设  $Y$  为一个随机变量, 定义为  $Y = X - \mathbb{E}[X]$ , 则显然  $\mathbb{E}[Y] = 0, \text{Var}[Y] = \text{Var}[X]$ . 根据次高斯的定义,  $Y$  也是次高斯随机变量, 且次高斯参数也是  $\sigma$ . 将  $Y$  的矩母函数泰勒展开, 得:

$$M_Y(\lambda) = M_Y(0) + \frac{M'_Y(0)}{1!}\lambda + \frac{M''_Y(0)}{2!}\lambda^2 + \frac{M^{(3)}_Y(\tilde{\lambda}_1)}{3!}\lambda^3, \quad \tilde{\lambda}_1 \in [0, \lambda]$$

其中最后一项为拉格朗日余项. 代入数值有

$$M_Y(\lambda) = 1 + 0 + \frac{1}{2}\text{Var}[Y]\lambda^2 + \frac{M^{(3)}_Y(\tilde{\lambda}_1)}{6}\lambda^3$$

由于  $Y$  是次高斯随机变量, 所以  $M_Y(\lambda) \leq \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$ . 设  $f(\lambda) = \exp\left(\frac{\lambda^2\sigma^2}{2}\right)$ , 则

$$\begin{aligned} f'(\lambda) &= \exp\left(\frac{\lambda^2\sigma^2}{2}\right)\sigma^2\lambda \\ f''(\lambda) &\exp\left(\frac{\lambda^2\sigma^2}{2}\right)\lambda^2\sigma^4 + \exp\left(\frac{\lambda^2\sigma^2}{2}\right)\sigma^2 \end{aligned}$$

将  $f(\lambda)$  在原点附近进行泰勒展开, 得:

$$\begin{aligned} f(\lambda) &= f(0) + \frac{f'(0)}{1!}\lambda + \frac{f''(0)}{2!}\lambda^2 + \frac{f^{(3)}(\tilde{\lambda}_2)}{3!}\lambda^3, \quad \tilde{\lambda}_2 \in [0, \lambda] \\ &= 1 + 0 + \frac{1}{2}\sigma^2\lambda^2 + \frac{f^{(3)}(\tilde{\lambda}_2)}{6}\lambda^3 \end{aligned}$$

根据次高斯性的定义,  $M_Y(\lambda) \leq f(\lambda), \forall \lambda \in \mathbb{R}$ . 代入泰勒展开式, 得:

$$1 + 0 + \frac{1}{2}\text{Var}[Y]\lambda^2 + \frac{M^{(3)}_Y(\tilde{\lambda}_1)}{6}\lambda^3 \leq 1 + 0 + \frac{1}{2}\sigma^2\lambda^2 + \frac{f^{(3)}(\tilde{\lambda}_2)}{6}\lambda^3$$

限制  $\lambda \neq 0$ , 不等号左右同除  $\lambda^2$ , 得

$$\frac{1}{2}\text{Var}[Y] + \frac{M^{(3)}_Y(\tilde{\lambda}_1)}{6}\lambda \leq \frac{1}{2}\sigma^2 + \frac{f^{(3)}(\tilde{\lambda}_2)}{6}\lambda$$

不等号两边同时令  $\lambda \rightarrow 0$ , 有

$$\text{Var}[Y] \leq \sigma^2$$

即  $\text{Var}[X] \leq \sigma^2$ .

(2) 因为  $X$  是  $\sigma$ -次高斯的, 根据次高斯性定义, 有:

$$\mathbb{E} [e^{\lambda(X-\mu)}] \leq e^{\frac{\lambda^2\sigma^2}{2}}, \quad \forall \lambda \in \mathbb{R}$$

因为  $\mathbb{E}[X] = \mu$ , 所以  $\mathbb{E}[cX] = c\mu$ . 故我们要证明下式:

$$\mathbb{E} [e^{\lambda(cX-c\mu)}] \leq e^{\frac{\lambda^2(|c|\sigma)^2}{2}} \quad \forall \lambda \in \mathbb{R}$$

设  $\lambda' = c\lambda$ , 则有:

$$\mathbb{E} [e^{\lambda(cX-c\mu)}] = \mathbb{E} [e^{\lambda'(X-\mu)}] \leq e^{\frac{(\lambda')^2\sigma^2}{2}} = e^{\frac{c^2\lambda^2\sigma^2}{2}} = e^{\frac{\lambda^2(|c|\sigma)^2}{2}} \quad \forall \lambda \in \mathbb{R}$$

因此,  $cX$  是  $|c|\sigma$ -次高斯的.

(3) 因为  $X_1$  是  $\sigma_1$ -次高斯的, 所以

$$\mathbb{E} [e^{\lambda(X_1-\mu_1)}] \leq \exp \left( \frac{\lambda^2\sigma_1^2}{2} \right)$$

因为  $X_2$  是  $\sigma_2$ -次高斯的, 所以

$$\mathbb{E} [e^{\lambda(X_2-\mu_2)}] \leq \exp \left( \frac{\lambda^2\sigma_2^2}{2} \right)$$

需要证明:

$$\mathbb{E} [\exp \{ \lambda((X_1 + X_2) - (\mu_1 + \mu_2)) \}] \leq \exp \left( \frac{\lambda^2(\sigma_1^2 + \sigma_2^2)}{2} \right)$$

$$\begin{aligned} & \mathbb{E} [\exp \{ \lambda((X_1 + X_2) - (\mu_1 + \mu_2)) \}] \\ &= \mathbb{E} [\exp \{ \lambda(X_1 - \mu_1) + \lambda(X_2 - \mu_2) \}] \\ &= \mathbb{E} [\exp \{ \lambda(X_1 - \mu_1) \} \cdot \exp \{ \lambda(X_2 - \mu_2) \}] \\ &= \mathbb{E} [\exp \{ \lambda(X_1 - \mu_1) \}] \cdot \mathbb{E} [\exp \{ \lambda(X_2 - \mu_2) \}] \quad (\text{独立性}) \\ &\leq \exp \left( \frac{\lambda^2\sigma_1^2}{2} \right) \exp \left( \frac{\lambda^2\sigma_2^2}{2} \right) = \exp \left( \frac{\lambda^2\sigma_1^2 + \lambda^2\sigma_2^2}{2} \right) \\ &= \exp \left( \frac{\lambda^2(\sqrt{\sigma_1^2 + \sigma_2^2})^2}{2} \right) \end{aligned}$$

所以  $X_1 + X_2$  是  $\sqrt{\sigma_1^2 + \sigma_2^2}$ -次高斯的. □

**定理 1.13. (Hoeffding界).** 若随机变量  $X_1, X_2, \dots, X_n$  互相独立, 且  $X_i$  的均值为  $\mu_i$ , 次高斯参数为  $\sigma_i$ , 则对于任意  $\varepsilon > 0$ , 有:

$$\mathbb{P} \left[ \sum_{i=1}^n (X_i - \mu_i) \geq \varepsilon \right] \leq \exp \left( -\frac{\varepsilon^2}{2 \sum_{i=1}^n \sigma_i^2} \right)$$

证明. 根据定理(1.12)的 (3),  $\sum_{i=1}^n X_i$  为  $\sqrt{\sum_{i=1}^n \sigma_i^2}$ -次高斯随机变量. 根据期望的线性性有

$$\mathbb{E} \left[ \sum_{i=1}^n X_i \right] = \sum_{i=1}^n \mathbb{E}[X_i] = \sum_{i=1}^n \mu_i$$

将  $\sum_{i=1}^n X_i$  的次高斯系数和均值分别代入(1.10)式, 即得出待证结论.  $\square$

**推论 1.14.** 若随机变量  $X_1, X_2, \dots, X_n$  互相独立, 且  $X_i \in [a, b], \forall i \in [n]$ . 则

$$\mathbb{P} \left[ \sum_{i=1}^n (X_i - \mu_i) \geq \varepsilon \right] \leq \exp \left( -\frac{2\varepsilon^2}{n(b-a)^2} \right)$$

**推论 1.15.** 若随机变量  $X_1, X_2, \dots, X_n$  互相独立, 且  $X_i \in [a_i, b_i], \forall i \in [n]$ , 则

$$\mathbb{P} \left[ \sum_{i=1}^n (X_i - \mu_i) \geq \varepsilon \right] \leq \exp \left( -\frac{2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2} \right)$$

证明. 因为  $X_i \in [a_i, b_i], \forall i \in [n]$ , 所以根据 Hoeffding 引理,  $X_i$  是  $\frac{b_i - a_i}{2}$ -次高斯的. 把  $X_i, \forall i \in [n]$  的次高斯参数代入 Hoeffding 界 (定理(1.13)) 即可.  $\square$

Hoeffding 界的最大弊端是必须要求独立性, 在对有监督学习的机器学习算法进行理论分析时, 通常可以使用 Hoeffding 界. 因为在有监督学习的框架中, 我们通常假设训练样本彼此是独立同分布的. 但对强化学习算法进行理论分析时, 其中涉及的随机变量就不一定是独立的, 许多随机变量下一时刻的值是依赖上一时刻

的值, 不再具备独立性, 无法使用 Hoeffding 界. 但只要它们具备一些性质, 我们就可以用吾妻不等式 (Azuma's inequality) 来求尾概率界, 这将在后面介绍.

**推论 1.16.** 若随机变量  $X_1, X_2, \dots, X_n$  互相独立, 且  $X_i \in [a_i, b_i], \forall i \in [n]$ . 则

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu_i) \geq \varepsilon\right) \leq \exp\left(-\frac{2n^2\varepsilon^2}{\sum_{i=1}^n (b_i - a_i)^2}\right)$$

证明. 因为

$$\mathbb{P}\left(\frac{1}{n} \sum_{i=1}^n (X_i - \mu_i) \geq \varepsilon\right) = \mathbb{P}\left(\sum_{i=1}^n (X_i - \mu_i) \geq n\varepsilon\right)$$

将上式代入推论(1.15)即可.  $\square$

## 1.4 条件期望与条件 Hoeffding 引理

为证明吾妻不等式 (Azuma's inequality), 我们需要先介绍条件 Hoeffding 引理.

**定义 1.17. (条件期望).** 给定一个概率空间  $(\Omega, \mathcal{F}, \mathbb{P})$ , 一个  $\sigma$  代数  $\mathcal{G} \subset \mathcal{F}$ , 一个  $\mathcal{F}$ -可测的随机变量  $X$  满足  $\mathbb{E}[|X|] < \infty$ , 则  $X$  在给定  $\mathcal{G}$  时的条件期望记作  $\mathbb{E}[X | \mathcal{G}]$ , 定义为  $L_1(\Omega, \mathcal{G}, \mathbb{P})$  中的唯一满足下式的元素:

$$\begin{aligned} \mathbb{E}[X \mathbb{I}_A] &= \mathbb{E}[\mathbb{E}[X | \mathcal{G}] \mathbb{I}_A] \quad \forall A \in \mathcal{G} \\ \iff \int_A X \, d\mathbb{P} &= \int_A \mathbb{E}[X | \mathcal{G}] \, d\mathbb{P} \quad \forall A \in \mathcal{G} \end{aligned}$$

条件期望有如下常用性质:

- (1) 线性:  $\mathbb{E}[aX + bY | \mathcal{F}] = a\mathbb{E}[X | \mathcal{F}] + b\mathbb{E}[Y | \mathcal{F}], \forall a, b \in \mathbb{R}$ .
- (2) 若  $X \geq 0$ , 则  $\mathbb{E}[X | \mathcal{F}] \geq 0$ .
- (3) 若  $X$  为  $\mathcal{F}$ -可测的, 则  $\mathbb{E}[XY | \mathcal{F}] = X\mathbb{E}[Y | \mathcal{F}]$ .
- (4) 若  $X$  与  $\mathcal{F}$  互相独立, 则  $\mathbb{E}[X | \mathcal{F}] = \mathbb{E}[X]$  a.s.

(5) 若  $\mathcal{F} \subset \mathcal{G}$ , 则  $\mathbb{E}[\mathbb{E}[X | \mathcal{G}] | \mathcal{F}] = \mathbb{E}[X | \mathcal{F}]$ .

(6)  $\mathbb{E}[\mathbb{E}[X | \mathcal{F}]] = \mathbb{E}[X]$ .

(7)  $\mathbb{E}[a | \mathcal{F}] = a, \forall a \in \mathbb{R}$ .

**性质 1.18.** 给定一个概率空间  $(\Omega, \mathcal{F}, \mathbb{P})$ , 及另一个  $\sigma$  代数  $\mathcal{G} \subset \mathcal{F}$ , 设  $\mathbb{Q}$  为另一个定义在  $\mathcal{F}$  上的概率测度, 且满足  $\frac{d\mathbb{Q}}{d\mathbb{P}} = L$ . 将使用  $\mathbb{Q}$  求期望记作  $\mathbb{E}_{\mathbb{Q}}[\cdot]$ , 使用  $\mathbb{P}$  求期望记作  $\mathbb{E}[\cdot]$ . 则:

$$(1) \frac{d\mathbb{Q}}{d\mathbb{P}} \Big|_{\mathcal{G}} = \mathbb{E}[L | \mathcal{G}]$$

(2) 若  $\mathbb{P} \ll \mathbb{Q}$ , 则有:  $\mathbb{E}_{\mathbb{Q}}[X | \mathcal{G}] = \frac{1}{\mathbb{E}[L | \mathcal{G}]} \mathbb{E}[LX | \mathcal{G}]$  a.s. (使用  $\frac{0}{0} = 0$  的约定).

证明. 先证性质 (1). 根据条件期望定义,  $\mathbb{E}[L | \mathcal{G}]$  是  $\mathcal{G}$ -可测的.  $\forall A \in \mathcal{G}$ ,

$$\begin{aligned} \mathbb{Q}(A) &= \mathbb{E}_{\mathbb{Q}}[\mathbb{I}_A] = \mathbb{E}[L\mathbb{I}_A] = \mathbb{E}[\mathbb{E}[L\mathbb{I}_A | \mathcal{G}]] \\ &= \mathbb{E}[\mathbb{E}[L | \mathcal{G}]\mathbb{I}_A] = \int_{\Omega} \mathbb{E}[L | \mathcal{G}]\mathbb{I}_A d\mathbb{P} \\ &= \int_A \mathbb{E}[L | \mathcal{G}] d\mathbb{P} \end{aligned}$$

根据 Radon-Nikodym 导数定义,  $\frac{d\mathbb{Q}}{d\mathbb{P}} \Big|_{\mathcal{G}} = \mathbb{E}[L | \mathcal{G}]$ .

然后证性质 (2), 根据条件期望定义, 只需证明:

$$\forall A \in \mathcal{G}, \quad \mathbb{E}_{\mathbb{Q}} \left[ \frac{1}{\mathbb{E}[L | \mathcal{G}]} \mathbb{E}[LX | \mathcal{G}]\mathbb{I}_A \right] = \mathbb{E}_{\mathbb{Q}}[X\mathbb{I}_A]$$

$\forall A \in \mathcal{G}$ , 有

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}} \left[ \frac{1}{\mathbb{E}[L | \mathcal{G}]} \mathbb{E}[LX | \mathcal{G}]\mathbb{I}_A \right] &= \mathbb{E}_{\mathbb{Q}|_{\mathcal{G}}} \left[ \frac{1}{\mathbb{E}[L | \mathcal{G}]} \mathbb{E}[LX | \mathcal{G}]\mathbb{I}_A \right] \\ &= \mathbb{E}_{\mathbb{Q}|_{\mathcal{G}}} \left[ \frac{d\mathbb{P}}{d\mathbb{Q}} \Big|_{\mathcal{G}} \mathbb{E}[LX | \mathcal{G}]\mathbb{I}_A \right] = \mathbb{E}_{\mathbb{P}|_{\mathcal{G}}} [\mathbb{E}[LX | \mathcal{G}]\mathbb{I}_A] = \mathbb{E}[\mathbb{E}[LX | \mathcal{G}]\mathbb{I}_A] \\ &= \mathbb{E}[\mathbb{E}[LX\mathbb{I}_A | \mathcal{G}]] = \mathbb{E}[LX\mathbb{I}_A] = \mathbb{E}_{\mathbb{Q}}[X\mathbb{I}_A] \end{aligned}$$

故  $\frac{1}{\mathbb{E}[L | \mathcal{G}]} \mathbb{E}[LX | \mathcal{G}]$  满足条件期望  $\mathbb{E}_{\mathbb{Q}}[X | \mathcal{G}]$  的定义, 即:

$$\frac{1}{\mathbb{E}[L | \mathcal{G}]} \mathbb{E}[LX | \mathcal{G}] = \mathbb{E}_{\mathbb{Q}}[X | \mathcal{G}] \quad \text{a.s.}$$

□

对于我们之前学过的 Hoeffding 引理, 其等价形式为: 对任意随机变量  $X$  满足  $a \leq X \leq b$  a.s. 均有:

$$\begin{aligned}\mathbb{E}[\exp(\lambda(X - \mathbb{E}[X]))] &\leq \exp\left(\frac{\lambda^2(b-a)^2}{8}\right) \\ \iff \ln \mathbb{E}[e^{\lambda X}] &\leq \lambda \mathbb{E}[X] + \frac{\lambda^2}{8}(b-a)^2\end{aligned}$$

下面介绍条件 Hoeffding 引理:

**定理 1.19. (条件Hoeffding引理).** 对于任意  $\sigma$  代数  $\mathcal{G}$ , 任意随机变量  $X$  满足  $A \leq X \leq B$  a.s., 其中  $A$  和  $B$  均为  $\mathcal{G}$ -可测随机变量, 则有:

$$\ln \mathbb{E}[e^{\lambda X} | \mathcal{G}] \leq \lambda \mathbb{E}[X | \mathcal{G}] + \frac{\lambda^2}{8}(B-A)^2 \quad \text{a.s., } \forall \lambda > 0$$

证明. 设  $\phi(\lambda) := \ln \mathbb{E}[e^{\lambda X} | \mathcal{G}]$ , 则:

$$\phi'(\lambda) = \frac{\mathbb{E}[X e^{\lambda X} | \mathcal{G}]}{\mathbb{E}[e^{\lambda X} | \mathcal{G}]}, \quad \phi''(\lambda) = \frac{\mathbb{E}[X^2 e^{\lambda X} | \mathcal{G}]}{\mathbb{E}[e^{\lambda X} | \mathcal{G}]} - \left(\frac{\mathbb{E}[X e^{\lambda X} | \mathcal{G}]}{\mathbb{E}[e^{\lambda X} | \mathcal{G}]}\right)^2$$

定义  $\mathbb{Q}_\lambda$  测度:  $\frac{d\mathbb{Q}_\lambda}{d\mathbb{P}} = L_\lambda = \frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X} | \mathcal{G}]}$ . 因为

$$\int_{\Omega} d\mathbb{Q}_\lambda = \int_{\Omega} L_\lambda d\mathbb{P} = \mathbb{E}[L_\lambda] = \mathbb{E}[E[L_\lambda | \mathcal{G}]] = 1$$

所以  $\mathbb{Q}_\lambda$  是概率测度. 显然  $\mathbb{P} \ll \mathbb{Q}_\lambda$ , 根据上面证过的性质 (2) 可知:

$$\mathbb{E}_{\mathbb{Q}_\lambda}[X | \mathcal{G}] = \frac{1}{\mathbb{E}[L_\lambda | \mathcal{G}]} \mathbb{E}[L_\lambda X | \mathcal{G}]$$

其中

$$\begin{aligned}\text{分母} &= \mathbb{E}\left[\frac{e^{\lambda X}}{\mathbb{E}[e^{\lambda X} | \mathcal{G}]} \middle| \mathcal{G}\right] = \frac{\mathbb{E}[e^{\lambda X} | \mathcal{G}]}{\mathbb{E}[e^{\lambda X} | \mathcal{G}]} = 1 \\ \text{分子} &= \mathbb{E}\left[\frac{X e^{\lambda X}}{\mathbb{E}[e^{\lambda X} | \mathcal{G}]} \middle| \mathcal{G}\right] = \frac{\mathbb{E}[X e^{\lambda X} | \mathcal{G}]}{\mathbb{E}[e^{\lambda X} | \mathcal{G}]}\end{aligned}$$

所以

$$\mathbb{E}_{\mathbb{Q}_\lambda}[X \mid \mathcal{G}] = \frac{\mathbb{E}[X e^{\lambda X} \mid \mathcal{G}]}{\mathbb{E}[e^{\lambda X} \mid \mathcal{G}]}$$

则  $\phi'(\lambda)$  和  $\phi''(\lambda)$  可化简为:

$$\begin{aligned}\phi'(\lambda) &= \mathbb{E}_{\mathbb{Q}_\lambda}[X \mid \mathcal{G}] \\ \phi''(\lambda) &= \mathbb{E}_{\mathbb{Q}_\lambda}[X^2 \mid \mathcal{G}] - (\mathbb{E}_{\mathbb{Q}_\lambda}[X \mid \mathcal{G}])^2 \\ &= \mathbb{E}_{\mathbb{Q}_\lambda}[X^2 \mid \mathcal{G}] - 2(\mathbb{E}_{\mathbb{Q}_\lambda}[X \mid \mathcal{G}])(\mathbb{E}_{\mathbb{Q}_\lambda}[X \mid \mathcal{G}]) + (\mathbb{E}_{\mathbb{Q}_\lambda}[X \mid \mathcal{G}])^2 \\ &= \mathbb{E}_{\mathbb{Q}_\lambda}[X^2 \mid \mathcal{G}] - 2\mathbb{E}_{\mathbb{Q}_\lambda}[X \mathbb{E}_{\mathbb{Q}_\lambda}[X \mid \mathcal{G}] \mid \mathcal{G}] + \mathbb{E}_{\mathbb{Q}_\lambda}[(\mathbb{E}_{\mathbb{Q}_\lambda}[X \mid \mathcal{G}])^2 \mid \mathcal{G}] \\ &= \mathbb{E}_{\mathbb{Q}_\lambda}[X^2 - 2X \mathbb{E}_{\mathbb{Q}_\lambda}[X \mid \mathcal{G}] + (\mathbb{E}_{\mathbb{Q}_\lambda}[X \mid \mathcal{G}])^2 \mid \mathcal{G}] \\ &= \mathbb{E}_{\mathbb{Q}_\lambda}[(X - \mathbb{E}_{\mathbb{Q}_\lambda}[X \mid \mathcal{G}])^2 \mid \mathcal{G}] \leq \mathbb{E}_{\mathbb{Q}_\lambda}[(X - Y)^2 \mid \mathcal{G}] \quad (Y \text{ 为 } \mathcal{G} - \text{ 可测随机变量})\end{aligned}$$

其中最后一个不等式证明如下:  $\forall \mathbb{Q}, \forall Y$  满足  $\mathcal{G}$ -可测,

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[(X - Y)^2 \mid \mathcal{G}] &= \mathbb{E}_{\mathbb{Q}}[(X - \mathbb{E}_{\mathbb{Q}}[X \mid \mathcal{G}] + \mathbb{E}_{\mathbb{Q}}[X \mid \mathcal{G}] - Y)^2 \mid \mathcal{G}] \\ &= \mathbb{E}_{\mathbb{Q}}[(X - \mathbb{E}_{\mathbb{Q}}[X \mid \mathcal{G}])^2 + (\mathbb{E}_{\mathbb{Q}}[X \mid \mathcal{G}] - Y)^2 \\ &\quad + 2(X - \mathbb{E}_{\mathbb{Q}}[X \mid \mathcal{G}])(\mathbb{E}_{\mathbb{Q}}[X \mid \mathcal{G}] - Y) \mid \mathcal{G}] \\ &= \mathbb{E}_{\mathbb{Q}}[(X - \mathbb{E}_{\mathbb{Q}}[X \mid \mathcal{G}])^2 \mid \mathcal{G}] + \mathbb{E}_{\mathbb{Q}}[(\mathbb{E}_{\mathbb{Q}}[X \mid \mathcal{G}] - Y)^2 \mid \mathcal{G}] \\ &\quad + \mathbb{E}_{\mathbb{Q}}[2(X - \mathbb{E}_{\mathbb{Q}}[X \mid \mathcal{G}])(\mathbb{E}_{\mathbb{Q}}[X \mid \mathcal{G}] - Y) \mid \mathcal{G}] \\ &\geq \mathbb{E}_{\mathbb{Q}}[(X - \mathbb{E}_{\mathbb{Q}}[X \mid \mathcal{G}])^2 \mid \mathcal{G}]\end{aligned}$$

因为  $A$  和  $B$  都是  $\mathcal{G}$ -可测的, 所以  $\frac{A+B}{2}$  也是  $\mathcal{G}$ -可测的. 令  $Y = \frac{A+B}{2}$ ,

$$\phi''(\lambda) \leq \mathbb{E}_{\mathbb{Q}_\lambda} \left[ \left( X - \frac{A+B}{2} \right)^2 \middle| \mathcal{G} \right] \quad \forall \lambda \geq 0$$

因为  $A \leq X \leq B$  a.s., 所以

$$\begin{aligned}A - \frac{A+B}{2} &\leq X - \frac{A+B}{2} \leq B - \frac{A+B}{2} \quad \text{a.s.} \\ \Rightarrow -\frac{B-A}{2} &\leq X - \frac{A+B}{2} \leq \frac{B-A}{2} \quad \text{a.s.}\end{aligned}$$

因为  $B \geq A$  a.s., 所以  $\frac{B-A}{2} \geq 0$ . 故  $\left| X - \frac{A+B}{2} \right| \leq \frac{B-A}{2}$ , 所以

$$\left( X - \frac{A+B}{2} \right)^2 \leq \left( \frac{B-A}{2} \right)^2$$

因此

$$\phi''(\lambda) \leq \left(\frac{B-A}{2}\right)^2 = \frac{(B-A)^2}{4}$$

对  $\phi$  进行泰勒展开, 得:

$$\begin{aligned}\phi(\lambda) &= \phi(0) + \phi'(0)\lambda + \frac{1}{2}\phi''(\tilde{\lambda})\lambda^2 \quad \tilde{\lambda} \in [0, 1] \\ &= \mathbb{E}[X \mid \mathcal{G}]\lambda + \frac{1}{2}\phi''(\tilde{\lambda})\lambda^2 \\ &\leq \mathbb{E}[X \mid \mathcal{G}]\lambda + \frac{1}{2}\frac{(B-A)^2}{4}\lambda^2\end{aligned}$$

所以

$$\ln \mathbb{E}[e^{\lambda X} \mid \mathcal{G}] \leq \mathbb{E}[X \mid \mathcal{G}]\lambda + \frac{\lambda^2}{8}(B-A)^2 \quad \forall \lambda > 0, \text{ a.s.}$$

□

## 1.5 鞅, 杜布分解与吾妻不等式

鞅 (Martingale) 和杜布分解 (Doob's Decomposition) 在强化学习乃至机器学习的理论分析中都占据了非常重要的地位, 无论是对 Bandit 问题, 强化学习问题还是随机优化问题进行理论分析时, 我们都离不开鞅的各种概念和性质. 杜布分解可以把一个随机过程分解为一个鞅和一个可料过程相加, 这种分解使得我们可以对问题分而治之. 对鞅的那一部分, 采用鞅的性质来处理; 对于可料过程那一部分, 采用可料过程的性质来处理.

本节的第三部分介绍吾妻不等式 (Azuma's inequality). Hoeffding 界是独立次高斯随机变量之和的尾概率界, 它的弊端也恰恰是要求这些随机变量必须是独立的, 而很多算法中遇到的随机变量并不具有这么好的独立性. 吾妻不等式将 Hoeffding 界中独立随机变量之和推广到了任意的上鞅, 下鞅和鞅上, 吾妻不等式的应用是非常广泛的. 它不仅应用在强化学习, 模仿学习等算法的理论分析中, 还被应用在深度学习的理论分析中, 尤其是随机梯度下降的理论分析也会利用到吾妻不等式.

下面我们首先从鞅的概念说起.

**定义 1.20.** (鞅). 设  $\{X_t\}_{t \in \mathbb{N}_+}$  为一个定义在  $(\Omega, \mathcal{F}, \mathbb{P})$  上的随机变量序列,  $\mathbb{F} = \{\mathcal{F}_t\}_{t \in \mathbb{N}_+}$  为一个  $\sigma$  域流. 假设  $\{X_t\}_{t \in \mathbb{N}_+}$  是  $\mathbb{F}$ -适应的 (即  $\forall t \in \mathbb{N}_+, X_t$  是  $\mathcal{F}_t$  可测的), 若  $\{X_t\}_{t \in \mathbb{N}_+}$  还满足下列条件:

$$(1) \mathbb{E}[X_t | \mathcal{F}_{t-1}] = X_{t-1} \text{ a.s. } \forall t \in \{2, 3, \dots\}$$

(2)  $X_t$  是可积的. ( $\iff \mathbb{E}[|X_t|] < \infty$ )

则称  $\{X_t\}_{t \in \mathbb{N}_+}$  是一个鞅. 若 (1) 中的 “=” 换成 “ $\leq$ ”, 则称  $\{X_t\}_{t \in \mathbb{N}_+}$  是一个上鞅. 反之, 若 (1) 中的 “=” 换成 “ $\geq$ ”, 则称  $\{X_t\}_{t \in \mathbb{N}_+}$  是一个下鞅.

**例 1.21.** 设  $\{S_t\}_{t \in \mathbb{N}_+}$  为一个定义为  $S_t := \sum_{i=1}^t (X_i - \mathbb{E}[X_i])$  的序列, 其中  $X_1, X_2, \dots$  均为独立的定义在  $(\Omega, \mathcal{F}, \mathbb{P})$  上的随机变量, 且  $\mathbb{E}[|X_t|] < \infty, \forall t \in \mathbb{N}_+$ . 则  $\{S_t\}_{t \in \mathbb{N}_+}$  是一个鞅.

证明. 设  $Y_i := X_i - \mathbb{E}[X_i], \forall i \in \mathbb{N}_+$ . 设  $\mathbb{F}^Y$  是  $\{Y_t\}_{t \in \mathbb{N}_+}$  的自然  $\sigma$  域流. 定义为:

$$\mathbb{F}^Y := \{\mathcal{F}_t^Y\}_{t \in \mathbb{N}_+}, \text{ 其中 } \mathcal{F}_t^Y := \sigma \left\{ \bigcup_{j=1}^t \sigma(Y_j) \right\}, \text{ 可知 } \forall j \in [t], Y_j \text{ 均为 } \mathcal{F}_t^Y \text{-可测的.}$$

因为  $S_t = \sum_{i=1}^t Y_i$ , 所以  $S_t$  是  $\mathcal{F}_t^Y$  可测的, 即  $\{S_t\}_{t \in \mathbb{N}_+}$  是  $\mathbb{F}^Y$ -适应的. 首先验证  $\mathbb{E}[S_t | \mathcal{F}_{t-1}] = S_{t-1}$  a.s.  $\forall t \in \{2, 3, \dots\}$ .

$$\begin{aligned} \mathbb{E}[S_t | \mathcal{F}_{t-1}] &= \mathbb{E}[S_{t-1} + Y_t | \mathcal{F}_{t-1}] \\ &= \mathbb{E}[S_{t-1} | \mathcal{F}_{t-1}] + \mathbb{E}[Y_t | \mathcal{F}_{t-1}] \\ &= S_{t-1} + \mathbb{E}[Y_t | \mathcal{F}_{t-1}] \\ &= S_{t-1} + \mathbb{E}[Y_t] \text{ a.s.} = S_{t-1} \text{ a.s. } \forall t \in \{2, 3, \dots\} \end{aligned}$$

然后验证  $\mathbb{E}[|S_t|] < \infty, \forall t \in \mathbb{N}_+$ .

$$\mathbb{E}[|S_t|] = \mathbb{E} \left[ \left| \sum_{i=1}^t Y_i \right| \right] \leq \mathbb{E} \left[ \sum_{i=1}^t |Y_i| \right] = \sum_{i=1}^t \mathbb{E}[|Y_i|] < \infty$$

故  $\{S_t\}_{t \in \mathbb{N}_+}$  是一个鞅. □

**定义 1.22. (可料性).** 给定一个过滤概率空间  $(\Omega, \mathcal{F}, \{\mathcal{F}_n\}_{n \in \mathbb{N}}, \mathbb{P})$ , 则随机过程  $\{X_n\}_{n \in \mathbb{N}}$  是可料的, 当  $\forall n \in \mathbb{N}, X_{n+1}$  均是  $\mathcal{F}_n$ -可测的.

**定理 1.23. (杜布分解).** 设  $(\Omega, \mathcal{F}, \mathbb{P})$  是一个概率空间, 指标集  $I = \{0, \dots, N\}$ , 其中  $N \in \mathbb{N}$  或  $I = \mathbb{N}$  (即指标集既可以有限, 又可以无限),  $\mathbb{F} := \{\mathcal{F}_n\}_{n \in I}$  为一个定义在  $\mathcal{F}$  上的  $\sigma$  域流, 且  $X = \{X_n\}_{n \in I}$  为一个  $\mathbb{F}$ -适应的随机过程, 满足  $\mathbb{E}[|X_n|] < \infty, \forall n \in I$ . 则存在一个鞅  $M = \{M_n\}_{n \in I}$  和一个可积可料过程  $A = \{A_n\}_{n \in I}$ , 其中  $A_0 = 0$ , 使得  $X_n = M_n + A_n$  对所有  $n \in I$  成立. 该分解是几乎处处唯一的.

证明. 先证存在性. 定义随机过程  $A$  和  $M$  为:

$$\forall n \in I \quad A_n = \sum_{k=1}^n (\mathbb{E}[X_k | \mathcal{F}_{k-1}] - X_{k-1})$$

$$M_n = X_0 + \sum_{k=1}^n (X_k - \mathbb{E}[X_k | \mathcal{F}_{k-1}])$$

定义  $n = 0$  时  $\sum_{k=1}^n := 0$ , 显然  $X_n = A_n + M_n (\forall n \in I)$ . 根据  $\{A_n\}_{n \in I}$  的定义, 显然可知  $\{A_n\}_{n \in I}$  是可料过程. 根据  $\{M_n\}_{n \in I}$  的定义, 显然  $\{M_n\}_{n \in I}$  适应  $\mathbb{F}$ , 且  $\mathbb{E}[|A_n|] < \infty, \mathbb{E}[|M_n|] < \infty$ . 这里以验证  $\mathbb{E}[|M_n|] < \infty$  为例, 验证  $\mathbb{E}[|A_n|] < \infty$  是类似的.

$$\begin{aligned} \mathbb{E}[|M_n|] &\leq \mathbb{E} \left[ |X_0| + \sum_{k=1}^n (|X_k| + \mathbb{E}[|X_k| | \mathcal{F}_{k-1}]) \right] \\ &= \mathbb{E}[|X_0|] + \sum_{k=1}^n (\mathbb{E}[|X_k|] + \mathbb{E}[\mathbb{E}[|X_k| | \mathcal{F}_{k-1}]]) \\ &= \mathbb{E}[|X_0|] + \sum_{k=1}^n (\mathbb{E}[|X_k|] + \mathbb{E}[|X_k|]) \\ &< \infty \end{aligned}$$

$\forall n \in I \setminus \{0\}$ ,

$$\begin{aligned} M_n - M_{n-1} &= X_0 + \sum_{k=1}^n (X_k - \mathbb{E}[X_k | \mathcal{F}_{k-1}]) - X_0 - \sum_{k=1}^{n-1} (X_k - \mathbb{E}[X_k | \mathcal{F}_{k-1}]) \\ &= X_n - \mathbb{E}[X_n | \mathcal{F}_{n-1}] \end{aligned}$$

从而  $M_n = X_n - \mathbb{E}[X_n | \mathcal{F}_{n-1}] + M_{n-1}$ , 求条件期望有

$$\begin{aligned} \mathbb{E}[M_n | \mathcal{F}_{n-1}] &= \mathbb{E}[X_n | \mathcal{F}_{n-1}] - \mathbb{E}[\mathbb{E}[X_n | \mathcal{F}_{n-1}] | \mathcal{F}_{n-1}] + \mathbb{E}[M_{n-1} | \mathcal{F}_{n-1}] \\ &= M_{n-1} \end{aligned}$$

因此  $\{M_n\}_{n \in I}$  是一个鞅.  $A$  和  $M$  是对  $X$  的杜布分解. 然后验证唯一性, 设  $X = M' + A'$  为另一个杜布分解. 则有

$$0 = X - X = M + A - M' - A' = (M - M') + (A - A')$$

从而有  $M - M' = A' - A$ , 设  $Y = M' - M$ . 因为  $M, M'$  均是适应  $\mathbb{F}$  的鞅, 可知  $Y$  也是适应  $\mathbb{F}$  的鞅. 即  $\forall n \in I \setminus \{0\}$ ,

$$\mathbb{E}[Y_n | \mathcal{F}_{n-1}] = Y_{n-1} \quad \text{a.s.}$$

因为  $Y = A' - A$ , 所以  $Y$  是可料过程. 即  $\forall n \in I \setminus \{0\}$ ,

$$\mathbb{E}[Y_n | \mathcal{F}_{n-1}] = Y_n \quad \text{a.s.}$$

结合上面两式, 得  $\forall n \in I \setminus \{0\}, Y_{n-1} = Y_n \quad \text{a.s.}$ . 根据杜布分解定义  $A'_0 = A_0 = 0$ , 所以  $Y_0 = A'_0 - A_0 = 0 \Rightarrow Y_0 = Y_1 = Y_2 = \dots = 0 \quad \text{a.s.}$ , 因此有  $M = M' \quad \text{a.s.}, A = A' \quad \text{a.s.}$   $\square$

**推论 1.24.** 一个实值随机过程  $X$  是一个下鞅当且仅当  $X$  存在杜布分解  $X = M + A$ , 其中  $M$  是一个下鞅,  $A$  是一个几乎处处增长的可料过程. 与此相对应地,  $X$  是一个上鞅, 当且仅当  $A$  是一个几乎处处下降的可料过程.

证明. 若  $X$  是下鞅, 则

$$\mathbb{E}[X_t | \mathcal{F}_{t-1}] \geq X_{t-1} \quad \text{a.s.} \quad \forall t \in I \setminus \{0\}$$

根据上个证明中的构造, 可知:  $A_n = \sum_{t=1}^n (\mathbb{E}[X_t | \mathcal{F}_{t-1}] - X_{t-1})$  a.s.  $\forall n \in I \setminus \{0\}$ ,  
可知  $A_n$  是一个几乎处处增长的随机过程.  $\square$

**定理 1.25. (吾妻不等式).** 设  $\{X_t\}_{t \in \mathbb{N}}$  为一个相对于  $\sigma$  域流  $\mathbb{F} := \{\mathcal{F}_t\}_{t \in \mathbb{N}}$  的上鞅, 假设存在  $\mathbb{F}$ -可料过程  $\{A_t\}_{t \in \mathbb{N}}, \{B_t\}_{t \in \mathbb{N}}$  和常数列  $\{c_t\}_{t \in \mathbb{N}}$  满足  $0 < c_t < \infty, \forall t \in \mathbb{N}$  使得  $A_t \leq X_t - X_{t-1} \leq B_t$  及  $B_t - A_t \leq c_t$  几乎处处成立. 则  $\forall \varepsilon \geq 0$  有:

$$\mathbb{P}(X_n - X_0 \geq \varepsilon) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{t=1}^n c_t^2}\right)$$

证明. 根据杜布分解 (推论(1.24)),  $X$  可分解为  $X_t = Y_t + Z_t, \forall t \in \mathbb{N}$ , 其中  $\{Y_t, \mathcal{F}_t\}$  是鞅,  $\{Z_t, \mathcal{F}_t\}$  是几乎处处下降的可料过程. 则有:

$$X_n - X_0 = (Y_n + Z_n) - (Y_0 + Z_0) = (Y_n - Y_0) + (Z_n - Z_0)$$

因为  $\{Z_t\}$  几乎处处下降, 所以  $Z_n - Z_0 \leq 0$  a.s.  $\Rightarrow Y_n - Y_0 \geq X_n - X_0$ , 可知

$$X_n - X_0 \geq \varepsilon \Rightarrow Y_n - Y_0 \geq \varepsilon$$

则  $\{\omega \in \Omega : X_n(\omega) - X_0(\omega) \geq \varepsilon\} \subset \{\omega \in \Omega : Y_n(\omega) - Y_0(\omega) \geq \varepsilon\}$ , 即

$$\mathbb{P}(X_n - X_0 \geq \varepsilon) \leq \mathbb{P}(Y_n - Y_0 \geq \varepsilon)$$

故只需证明:

$$\mathbb{P}(Y_n - Y_0 \geq \varepsilon) \leq \exp\left(-\frac{2\varepsilon^2}{\sum_{t=1}^n c_t^2}\right)$$

对  $Y_n - Y_0$  应用 Chernoff 界

$$\begin{aligned} \mathbb{P}(Y_n - Y_0 \geq \varepsilon) &\leq \min_{\lambda > 0} e^{-\lambda\varepsilon} \mathbb{E}[e^{\lambda(Y_n - Y_0)}] \\ &= \min_{\lambda > 0} e^{-\lambda\varepsilon} \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^n (Y_t - Y_{t-1})\right)\right] \\ &= \min_{\lambda > 0} e^{-\lambda\varepsilon} \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n-1} (Y_t - Y_{t-1})\right) \exp(\lambda(Y_n - Y_{n-1}))\right] \\ &= \min_{\lambda > 0} e^{-\lambda\varepsilon} \mathbb{E}\left[\mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n-1} (Y_t - Y_{t-1})\right) \exp(\lambda(Y_n - Y_{n-1})) \mid \mathcal{F}_{n-1}\right]\right] \quad (\clubsuit) \end{aligned}$$

由于  $\exp\left(\lambda \sum_{t=1}^{n-1} (Y_t - Y_{t-1})\right)$  是对  $\mathcal{F}_{n-1}$  可测的, 所以

$$\mathbb{P}(Y_n - Y_0 \geq \varepsilon) \leq \min_{\lambda > 0} e^{-\lambda\varepsilon} \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n-1} (Y_t - Y_{t-1})\right) \mathbb{E}[\exp(\lambda(Y_n - Y_{n-1})) \mid \mathcal{F}_{n-1}]\right]$$

根据已知条件,  $\forall t \in \mathbb{N}_+, A_t \leq X_t - X_{t-1} \leq B_t$  a.s.

根据杜布分解,  $\forall t \in \mathbb{N}_+, A_t \leq Y_t + Z_t - Y_{t-1} - Z_{t-1} \leq B_t$

$$A_t + Z_{t-1} - Z_t \leq Y_t - Y_{t-1} \leq B_t + Z_{t-1} - Z_t \quad \text{a.s.}$$

在  $Y_t - Y_{t-1}$  上套用条件 Hoeffding 引理:

$$\begin{aligned} & \mathbb{E}[\exp(\lambda(Y_t - Y_{t-1})) \mid \mathcal{F}_{t-1}] \\ & \leq \lambda \mathbb{E}[Y_t - Y_{t-1} \mid \mathcal{F}_{t-1}] + \exp\left(\frac{\lambda^2(B_t + Z_{t-1} - Z_t - A_t - Z_{t-1} + Z_t)^2}{8}\right) \\ & = \exp\left(\frac{\lambda^2(B_t - A_t)^2}{8}\right) \\ & \leq \exp\left(\frac{\lambda^2 c_t^2}{8}\right) \quad (\text{因为 } B_t - A_t \leq c_t \quad \text{a.s.} \quad \forall t \in \mathbb{N}_+) \end{aligned}$$

前面已知

$$\mathbb{P}(Y_n - Y_0 \geq \varepsilon) \leq \min_{\lambda > 0} e^{-\lambda\varepsilon} \exp\left(\frac{\lambda^2 c_t^2}{8}\right) \mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n-1} (Y_t - Y_{t-1})\right)\right]$$

对  $\mathbb{E}\left[\exp\left(\lambda \sum_{t=1}^{n-1} (Y_t - Y_{t-1})\right)\right]$  重复上述 (♣) 式中的操作  $n-1$  遍得到 (即不断对  $\mathcal{F}_{n-t}$  求条件期望, 并把其中条件期望的式子用上界来替代)

$$\begin{aligned} \mathbb{P}(Y_n - Y_0 \geq \varepsilon) & \leq \min_{\lambda > 0} e^{-\lambda\varepsilon} \prod_{t=1}^n \exp\left(\frac{\lambda^2 c_t^2}{8}\right) \\ & = \min_{\lambda > 0} e^{-\lambda\varepsilon} \exp\left(\sum_{t=1}^n \frac{\lambda^2 c_t^2}{8}\right) \\ & = \min_{\lambda > 0} \exp\left(\frac{\lambda^2 \sum_{t=1}^n c_t^2}{8} - \lambda\varepsilon\right) \end{aligned}$$

因为对数函数严格单增

$$\arg \min_{\lambda \in \mathbb{R}} \exp\left(\frac{\lambda^2 \sum_{t=1}^n c_t^2}{8} - \lambda\varepsilon\right) = \arg \min_{\lambda \in \mathbb{R}} \frac{\lambda^2 \sum_{t=1}^n c_t^2}{8} - \lambda\varepsilon = \lambda^*$$

求导, 得:  $\frac{\sum_{t=1}^n c_t^2}{4} \lambda - \varepsilon$ . 置为 0, 求  $\lambda^*$ :

$$\frac{\sum_{t=1}^n c_t^2}{4} \lambda^* - \varepsilon = 0$$

即  $\lambda^* = \frac{4\varepsilon}{\sum_{t=1}^n c_t^2}$ , 将  $\lambda^*$  代回界中:

$$\begin{aligned}\mathbb{P}(Y_n - Y_0 \geq \varepsilon) &\leq \min_{\lambda > 0} \exp \left( \frac{\lambda^2 \sum_{t=1}^n c_t^2}{8} - \lambda \varepsilon \right) \\ &= \exp \left( \frac{(\lambda^*)^2 \sum_{t=1}^n c_t^2}{8} - \lambda^* \varepsilon \right) \\ &= \exp \left( \frac{16\varepsilon^2}{(\sum_{t=1}^n c_t^2)^2} \cdot \frac{\sum_{t=1}^n c_t^2}{8} - \frac{4\varepsilon}{\sum_{t=1}^n c_t^2} \varepsilon \right) \\ &= \exp \left( -\frac{2\varepsilon^2}{\sum_{t=1}^n c_t^2} \right)\end{aligned}$$

□

**推论 1.26.** 在与定理(1.25)相同的条件下, 若  $\{X_t\}_{t \in \mathbb{N}}$  为一个下鞅, 则:

$$\mathbb{P}(Y_n - Y_0 \geq -\varepsilon) \leq \exp \left( -\frac{2\varepsilon^2}{\sum_{t=1}^n c_t^2} \right)$$

若  $\{X_t\}_{t \in \mathbb{N}}$  为一个鞅, 则:

$$\mathbb{P}(|X_n - X_0| \geq \varepsilon) \leq 2 \exp \left( -\frac{2\varepsilon^2}{\sum_{t=1}^n c_t^2} \right)$$

证明. 因对下鞅取负号后即得到上鞅, 故第一个式子得证. 组合针对上鞅, 下鞅的吾妻不等式, 可得第二个式子. □

## 2 最小最大定理及其证明

最小最大定理是一个非常基础的定理, 在机器学习和强化学习理论分析中被广泛使用. 它本质上做的事情只有一个, 它提供了先求最小值再求最大值和先求最大值再求最小值这两个式子相等的条件. 它具体的应用都有哪些呢? 首先我们在解决优化问题时经常会使用最小最大定理, 其次我们在对算法决策鲁棒性进行分析时也会用到最小最大定理. 此外, 最小最大定理在博弈论中也占据了重要地位, 事实上, 最小最大定理的提出和证明标志了博弈论这个新学科的诞生. 博弈论之父冯诺依曼曾说过

据我所知, 在最小最大定理被证明之前, 不存在任何一个值得发表的博弈论工作.

可见其重要性非同一般. 因此最小最大定理也被称作博弈论基本定理. 接下来我们通过一个大家都很熟悉的结论来引出最小最大定理.

**定理 2.1. (最大最小不等式).** 对任意函数  $f : Z \times W \rightarrow \mathbb{R}$ , 有:

$$\max_{z \in Z} \min_{w \in W} f(z, w) \leq \min_{w \in W} \max_{z \in Z} f(z, w)$$

证明. 定义  $g(z) := \min_{w \in W} f(z, w)$ , 显然  $\forall w \in W, \forall z \in Z$  有  $g(z) \leq f(z, w)$ . 从而

$$\forall w \in W, \max_z g(z) \leq \max_z f(z, w)$$

$$\Rightarrow \max_z g(z) \leq \min_{w \in W} \max_z f(z, w)$$

代入  $g(z)$  的定义有

$$\max_{z \in Z} \min_{w \in W} f(z, w) \leq \min_{w \in W} \max_{z \in Z} f(z, w)$$

□

最大-最小不等式 (Max-Min Inequality) 应用范围非常广, 因为其在任意条件下都成立, 称为**弱对偶性**. 与其相对应的是**强对偶性**, 强对偶性的条件更严格, 结论更强, 最小最大定理就是介绍强对偶性需要满足哪些条件的定理, 最初由冯诺依曼在 1928 年提出.

**定理 2.2.** (冯诺依曼最小最大定理, von Neumann's Minimax Theorem). 设  $X \subset \mathbb{R}, Y \subset \mathbb{R}$  为紧致凸集, 若  $f : X \times Y \rightarrow \mathbb{R}$  为一个连续函数满足下列性质:

- (1)  $f(\cdot, y) : X \rightarrow \mathbb{R}$  对任意固定的  $y$  都是凹函数.
- (2)  $f(x, \cdot) : Y \rightarrow \mathbb{R}$  对任意固定的  $x$  都是凸函数.

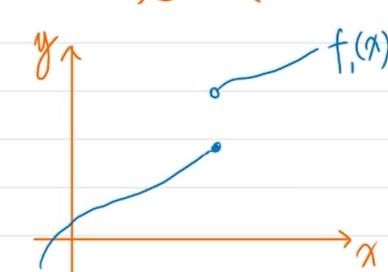
则有:  $\max_{x \in X} \min_{y \in Y} f(x, y) = \min_{y \in Y} \max_{x \in X} f(x, y).$

事实上, 我们在机器学习的公式推导和定理证明中经常需要交换  $\min$  和  $\max$  的顺序, 所以经常用到最小最大定理. 通常将取到最小最大值的  $x$  和  $y$  称为函数的**鞍点**. 我们这里不直接证明冯诺依曼的最小最大定理, 而去证明它的一个推广, 即 Sion 最小最大定理 (Sion's Minimax Theorem), Sion 最小最大定理同样广泛使用, 其条件比冯诺依曼最小最大定理更弱, 但得出的结论却比冯诺依曼的最小最大定理更强. 在介绍 Sion 最小最大定理之前需要介绍一些基本概念.

**定义 2.3. (上/下半连续).** 说一个定义在拓扑空间  $X$  上的函数  $f : X \rightarrow [-\infty, \infty]$  是

- (1) 下半连续的, 当对任意  $c \in \mathbb{R}$ , 有  $\{x \in X : f(x) \leq c\}$  是闭集.
- (2) 上半连续的, 当对任意  $c \in \mathbb{R}$ , 有  $\{x \in X : f(x) \geq c\}$  是闭集.

下半连续函数:



上半连续函数:



**引理 2.4.** 设  $f : X \rightarrow [-\infty, +\infty]$  为一个定义在拓扑空间  $X$  上的函数, 则:

- (1)  $f$  是下半连续的当且仅当  $x_\alpha \rightarrow x \Rightarrow \liminf_{\alpha} f(x_\alpha) \geq f(x)$ .
- (2)  $f$  是上半连续的当且仅当  $x_\alpha \rightarrow x \Rightarrow \limsup_{\alpha} f(x_\alpha) \leq f(x)$ .

**定义 2.5. (拟凸/拟凹).** 设  $f : C \rightarrow \mathbb{R}$  为定义在向量空间的凸子集  $C$  上的实函数, 则说  $f$  是:

- (1) 拟凸的, 若  $f(\alpha x + (1 - \alpha)y) \leq \max\{f(x), f(y)\}, \forall x, y \in C, 0 \leq \alpha \leq 1$ .
- (2) 拟凹的, 若  $-f$  是拟凸的.

**定理 2.6. (Sion 最小最大定理).** 设  $X$  为一个拓扑向量空间的紧致凸子集,  $Y$  是拓扑向量空间的凸集. 若  $f$  是定义在  $X \times Y$  上的实值函数. 满足:

- (1)  $\forall y \in Y, f(\cdot, y)$  为定义在  $X$  上的下半连续拟凸函数.
- (2)  $\forall x \in X, f(x, \cdot)$  为定义在  $Y$  上的上半连续拟凹函数.

则有  $\min_{x \in X} \sup_{y \in Y} f(x, y) = \sup_{y \in Y} \min_{x \in X} f(x, y)$ .

我们的证明参考自小宫英敏 (Hidetoshi Komia) 提出的证明方法. 感兴趣的读者可以阅读他的论文

*elementary proof for Sion's minimax theorem*

**引理 2.7.** 在与 Sion 最小最大定理相同的假设下, 对于任意  $y_1, y_2 \in Y$  和任意  $\alpha \in \mathbb{R}$  满足  $\alpha < \min_{x \in X} \max(f(x, y_1), f(x, y_2))$ , 则存在  $y_0 \in Y$  使得  $\alpha < \min_{x \in X} f(x, y_0)$ .

证明. 反证法, 假设对于任意  $y_0 \in Y$  都有  $\alpha \geq \min_{x \in X} f(x, y_0)$ . 设  $\beta$  满足

$$\alpha < \beta < \min_{x \in X} \max(f(x, y_1), f(x, y_2))$$

令  $[y_1, y_2]$  表示  $y_1, y_2$  之间的线段. 对于任意  $z \in [y_1, y_2]$ , 定义集合

$$C_z = \{x \in X : f(x, z) \leq \alpha\} \quad C'_z = \{x \in X : f(x, z) \leq \beta\}$$

显然  $C_z \subset C'_z$ ,  $C_z \neq \emptyset, C'_z \neq \emptyset$ . 设  $A = C'_{y_1}, B = C'_{y_2}$ , 则  $A \neq \emptyset, B \neq \emptyset$ . 根据  $f(\cdot, y)$  对任意  $y \in Y$  均是下半连续的, 可知  $C_z, C'_z, A, B$  均为闭集. 设  $x_0 \in A = C'_{y_1}$ , 则

$$\begin{aligned} f(x_0, y_1) &\leq \beta < \min_{x \in X} \max(f(x, y_1), f(x, y_2)) \\ &\leq \max(f(x_0, y_1), f(x_0, y_2)) = f(x_0, y_2) \end{aligned}$$

即  $\beta < f(x_0, y_2) \Rightarrow x_0 \notin B \Rightarrow A \cap B = \emptyset$ . 因为  $f(x, \cdot)$  是拟凹函数, 根据拟凹性定义, 有:

$$f(x, z) \geq \min(f(x, y_1), f(x, y_2)) \quad \forall x \in X, z \in [y_1, y_2]$$

设  $x_0 \in C'_z$ , 则  $\min(f(x_0, y_1), f(x_0, y_2)) \leq f(x_0, z) \leq \beta$ , 从而有  $f(x_0, y_1) \leq \beta$  或者  $f(x_0, y_2) \leq \beta$ , 则  $x_0 \in A$  或者  $x_0 \in B$ . 从而有  $C'_z \subset A \cup B$ . 根据  $f(\cdot, z)$  的拟凸性, 可知  $C'_z$  是凸集, 故也是连通集. 假设  $C'_z \cap A \neq \emptyset, C'_z \cap B \neq \emptyset$ , 则  $A' := C'_z \cap A$  为闭集,  $B' = C'_z \cap B$  也为闭集. 由于  $C'_z = A' \cup B'$ ,  $A' \cap B' = \emptyset$ , 所以  $A', B'$  既开又闭, 且既非全集也非空集, 则  $C'_z$  不连通, 矛盾. 则  $C_z \subset C'_z \subset A$  或  $C_z \subset C'_z \subset B$ . 定义

$$I = \{z \in [y_1, y_2] : C_z \subset A\} \quad J = \{z \in [y_1, y_2] : C_z \subset B\}$$

则  $I \cap J = \emptyset, I \cup J = [y_1, y_2]$ . 假设  $I = \emptyset \Rightarrow \forall z \in [y_1, y_2], C_z \not\subset A = C'_{y_1}$ , 令  $z = y_1$  即得矛盾, 所以  $I \neq \emptyset$ , 同理可得  $J \neq \emptyset$ . 设  $\{z_n\}$  为  $I$  中的一个序列满足  $\lim z_n = z \in [y_1, y_2]$ , 设  $x$  为  $C_z$  中任意一点. 则有  $f(x, z) < \beta$ . 根据  $f(x, \cdot)$  的上半连续性,  $\limsup f(x, z_n) < \beta$ .

$$\beta - \limsup f(x, z_n) = \varepsilon > 0 \Rightarrow \forall x \in \mathbb{R}, \text{若 } |x - \limsup f(x, z_n)| < \varepsilon, \text{ 则 } x < \beta$$

则  $\{f(x, z_n)\}$  存在一个收敛到  $\limsup f(x, z_n)$  的子列, 设其为  $f(x, z_{n_i})$ , 即

$$\forall \varepsilon > 0, \exists N \in \mathbb{N} \text{ 使得 } j \geq N \Rightarrow |f(x, z_{n_j}) - \limsup f(x, z_n)| < \varepsilon$$

则有  $f(x, z_{n_j}) < \beta$ . 换言之  $\exists m$  使得  $f(x, z_m) < \beta$ , 即  $x \in C'_{z_m}$ . 因  $\{z_n\}$  定义在  $I$  中, 故  $z_m \in I$ . 根据  $I$  定义,  $C_{z_m} \subset A$ , 又因为  $\forall z \in [y_1, y_2]$  有  $C_z \subset C'_z$ , 故  $C'_{z_m} \subset A$ . 因此  $x \in A$ . 因此  $z \in I$ ,  $I$  是  $[y_1, y_2]$  中的闭集. 利用相似方法可以证明  $J$  也是闭集. 可得  $[y_1, y_2]$  不连通, 和  $[y_1, y_2]$  定义相矛盾.  $\square$

**引理 2.8.** 在与 Sion 最小最大定理相同的假设下, 对于任意有限个  $y_1, y_2, \dots, y_n$  及任意  $\alpha \in \mathbb{R}$  满足  $\alpha < \min_{x \in X} \max_{1 \leq i \leq n} f(x, y_i)$ , 存在  $y_0 \in Y$  满足  $\alpha < \min_{x \in X} f(x, y_0)$ .

证明. 利用数学归纳法对  $n$  进行归纳.

第 1 步: 假设  $n = 1$ , 取  $y_0 = y_1$  即可得证.

第 2 步: 假设  $n = k - 1$  时性质成立. 设  $X' := \{x \in X : f(x, y_k) \leq \alpha\}$  可知  $X'$  是闭集, 凸集, 紧致集合.

假设  $X' = \emptyset \Rightarrow \forall x \in X, f(x, y_k) > \alpha \Rightarrow \min_{x \in X} f(x, y_k) > \alpha$ , 取  $y_0 = y_k$  即可得证.

假设  $X' \neq \emptyset$ ,

$$\alpha < \min_{x \in X} \max_{1 \leq i \leq k} f(x, y_i) \leq \min_{x \in X'} \max_{1 \leq i \leq k} f(x, y_i)$$

假设  $i = k$  时  $f(x, y_i)$  取到最大值, 会产生矛盾. 则

$$\alpha < \min_{x \in X} \max_{1 \leq i \leq k} f(x, y_i) \leq \min_{x \in X'} \max_{1 \leq i \leq n} f(x, y_i) = \min_{x \in X'} \max_{1 \leq i \leq k-1} f(x, y_i)$$

将  $f$  限制在  $X' \times Y$  上, 然后在其上应用归纳假设, 可知存在  $y'_0 \in Y$ , 使得  $\alpha < \min_{x \in X} f(x, y'_0)$ , 则

$$\alpha < \min_{x \in X'} \max(f(x, y'_0), f(x, y_k))$$

可以验证

$$\alpha < \min_{x \in X} \max(f(x, y'_0), f(x, y_k))$$

因为若  $x^* \in X'$ , 替换  $X'$  为  $X$  后依然成立. 若  $x^* \notin X'$ , 则根据  $X'$  定义,  $\alpha < f(x^*, y_k)$ , 则

$$\alpha < \max(f(x^*, y'_0), f(x^*, y_k)) = \min_{x \in X} \max(f(x, y'_0), f(x, y_k))$$

应用引理(2.7), 即可得证存在  $y_0 \in Y$  满足  $\alpha < \min_{x \in X} f(x, y_0)$ . □

下面开始证明 Sion 最小最大定理

证明. 显然  $\sup_{y \in Y} \min_{x \in X} f(x, y) \leq \min_{x \in X} \sup_{y \in Y} f(x, y)$ . 因此只需证  $\geq$ . 设  $\alpha$  是任意满足  $\alpha < \min_{x \in X} \sup_{y \in Y} f(x, y)$  的实数. 设  $X_y := \{x \in X : f(x, y) \leq \alpha\}, \forall y \in Y$ . 可知  $X_y$  是

紧支闭集, 则  $\bigcap_{y \in Y} X_y = \emptyset$ . 反证, 假设  $\bigcap_{y \in Y} X_y \neq \emptyset$ , 则  $\exists x \in X$  使得  $\forall y \in Y$ ,

$$f(x, y) \leq \alpha < \min_{x \in X} \sup_{y \in Y} f(x, y)$$

从而有

$$\exists x \text{ s.t. } \sup_{y \in Y} f(x, y) < \min_{x \in X} \sup_{y \in Y} f(x, y)$$

令  $g(x) = \sup_{y \in Y} f(x, y)$ , 则

$$\exists x \text{ s.t. } g(x) < \min_{x \in X} g(x)$$

矛盾. 设  $X$  为全集, 则  $\forall y \in Y, X_y^c = X \setminus X_y$ .

$$\left( \bigcap_{y \in Y} X_y \right)^c = \emptyset^c \Rightarrow \bigcup_{y \in Y} X_y^c = X$$

则  $\{X_y^c\}_{y \in Y}$  为对  $X$  的一个开覆盖, 则存在有限子覆盖, 即存在  $y_1, y_2, \dots, y_n \in Y$  使得  $\bigcup_{i=1}^n X_{y_i}^c = X$ . 则  $\bigcap_{i=1}^n X_{y_i} = \emptyset$ . 即  $\forall x \in X, \exists i \in [n], \text{s.t. } f(x, y_i) > \alpha$ . 也就是  $\forall x \in X, \max_{1 \leq i \leq n} f(x, y_i) > \alpha$ . 则有

$$\min_{x \in X} \max_{1 \leq i \leq n} f(x, y_i) > \alpha$$

应用引理(2.8), 可知存在  $y_0 \in Y$  使得  $\alpha < \min_{x \in X} f(x, y_0) \leq \sup_{y \in Y} \min_{x \in X} f(x, y)$ , 故可知

$$\min_{x \in X} \sup_{y \in Y} f(x, y) \leq \sup_{y \in Y} \min_{x \in X} f(x, y)$$

□

## 3 Bandit 问题

### 3.1 Bandit 简介与遗憾分解引理

在通常的强化学习课程中,往往介绍的第一类算法就是解决 Bandit 问题的一些方法.Bandit 问题在中文的文献中通常被翻译成赌博机问题或者老虎机问题,博彩机问题等.由于不同的中文文献对 Bandit 术语翻译并不统一,我们讲义中就直接用英语 Bandit 来称呼这种问题.之所以很多强化学习课程一上来会讲 Bandit 问题,是因为 Bandit 问题比很多强化学习问题简单.由于强化学习的环境是具有状态转移的,因此强化学习的智能体在决策的时候,需要考虑长期的累计回报.而 Bandit 问题中的环境是没有状态转移的,因此可以将其看成强化学习问题的特例.不过实际上,Bandit 问题虽然简单,但其应用非常多.最有名的应用就是 AlphaGo,它是在 2016 年以 4 比 1 的总比分战胜了曾经的世界冠军李世石的围棋 AI.在 AlphaGo 中使用了强化学习和 MCTS 这两种方法,而 MCTS 这种思想就借鉴了 UCB 这种针对 Bandit 问题的算法.除了 AlphaGo,现实生活中 Bandit 问题也被应用于很多领域.比如在医疗的领域,Bandit 问题的算法通常被应用于临床测试以及大脑和行为方面的建模.在金融中,Bandit 问题的相关算法被应用于投资组合的选择.在电商的相关领域,Bandit 问题的相关算法被应用于零售价格方面的决策.此外,Bandit 问题的一些算法还被应用于推荐系统,社交网络,信息检索,对话系统,异常检测以及通信等领域.除了现实中的应用以外,Bandit 算法还被用来和其他算法相结合,譬如用于算法的选择,超参数的优化,特征选择,主动学习,聚类以及强化学习当中.关于 Bandit 问题更多应用的相关细节,读者可以参考这篇论文 [A Survey on Practical Applications of Multi-Armed and Contextual Bandits](#)

除了 Bandit 本身之外,与 Bandit 相关的方向也十分具有研究与了解的价值.比如完全信息在线学习,它与 Bandit 的区别仅仅是从环境中得到了不同的反馈信号.由于目前还没有说 Bandit 问题的定义,这里不再细说.为什么完全信息在线学习很重要?因为很多强化学习算法的理论分析中都借鉴或引用了完全信息在线学习的相关理论.比如在模仿学习的算法 DAgger 中借鉴了针对在线学习 FTL(Follow-the-leader) 算法的相关理论.此外在线学习的理论框架还和博弈论有一定的联系,像针对扑克牌这种不完美信息扩展式博弈的强化学习算法 CFR,也

就是反事实遗憾最小化算法, 它的理论也是基于在线学习遗憾最小化的相关理论. 在线学习还有许许多多的其他分类, 比如当在线学习中的动作集是凸集, 损失函数是凸函数的时候, 这个问题就可以被分类为在线凸优化问题. 还有一类在线学习叫做 Partial Monitoring, 这类问题中可以更加灵活的定义环境提供给智能体的反馈. 这一讲中, 我们专注于 Bandit 问题, 因此除了 Bandit 问题以外的在线学习问题, 我们不做详细的展开. 读者若对在线学习的相关理论感兴趣, 可参考这本书 [Prediction, Learning, and Games by Nicolò Cesa-Bianchi](#)

这并不是针对 Bandit 问题的书, 针对 Bandit 问题的书可以参考这本 [Lattimore, Tor, and Csaba Szepesvári. Bandit algorithms. Cambridge University Press, 2020.](#)

下面我们开始介绍 Bandit 问题. Bandit 的意思是匪徒, 我们这里的意思是老虎机, 机器上有一个拉杆, 每次去拉这个拉杆的时候, 机器都有一定的概率吐出一定的奖励来. 奖励在现实中可以是钱或者游戏币. 这种被称作老虎机的机器也就是所谓的 Bandit, 有人会问既然 Bandit 的本意是匪徒, 那么为什么用这个单词来称呼老虎机? 这里是一个比喻, 俗话说逢赌必输, 一个人在老虎机之前一直玩下去是注定会输光所有钱的. 这就好像这个人遭到了匪徒的打劫一样, 钱财都被洗劫一空, 因此我们就理解了 Bandit 为什么是老虎机的意思. 正常的老虎机都是只有一个拉杆的, 但是在 Bandit 问题中我们往往假设老虎机有多个拉杆, 并且拉动不同的拉杆, 机器所吐出的奖励的分布是有区别的. 举个例子, 比如一台机器有两个拉杆, 其中一个拉杆每次拉动都会吐出一块钱, 另一个拉杆每次拉动的时候则会以 0.01 的概率吐出 100 块钱, 其余 0.99 的概率则不吐出钱. 就意味着选择不同的拉杆 (臂), 获得的奖励的期望和方差都不相同. 对于我们刚刚举的例子, 这两个臂的奖励的期望是一样的, 区别在于这两个臂所对应的奖励的方差. 对于风险偏好型的决策者来说, 往往更适合选择第二个臂; 对于风险厌恶型决策者来说, 往往更适合选择第一个臂; 对于风险中性决策者来说, 选择这两个臂并没有什么区别. 即风险中性的决策者是只关心期望, 不关心方差的. 在我们这讲中的所有算法, 都是把智能体是风险中性的决策者作为假设的. 我们的算法也只关心奖励的期望, 不关心奖励的方差.

在 Bandit 问题的语境下, 我们把一个拉杆叫做一个动作, 把机器上拉杆的集合叫做这个问题的动作集, 或者动作空间. 事实上, Bandit 问题的动作空间里面包含的元素并不仅限于有限个, 也就是我们可以定义一个具有无穷个拉杆的老虎机,

即这个 Bandit 问题的动作空间里有无穷个动作.

在 Bandit 问题中, 赌徒对老虎机上每个拉杆对应的奖励的分布是一无所知的. 它首先需要做的是尝试不同的拉杆, 看看在拉动不同的拉杆下, 机器所吐出的奖励各自有多少, 不过由于无论他拉动哪一个拉杆, 机器所吐出的奖励都是一个随机的数, 因此只拉动一遍这些拉杆, 远远不能完全确定这些拉杆所对应的奖励分布. 因此即使所有拉杆都被拉了一遍, 赌徒也有必要去拉一些一开始看上去返回的奖励并不是很多的拉杆. 因为只拉了一次返回的奖励不多, 很可能只是运气不好, 说明不了什么问题. 另一方面, 如果某个拉杆经过许许多多次尝试后, 其结果都是拉动它比拉动其他拉杆得到的奖励更多, 赌徒就可以相当的肯定这个拉杆比其他的拉杆更好, 因此就应该在今后多拉这个拉杆.Bandit 问题中这两种情况分别对应“探索”与“利用”这两个概念. 所谓的探索就是赌徒在对各个拉杆都不熟悉的情况下所做的尝试, 所谓的利用就是赌徒利用过去的历史经验去拉动那个他认为可为他带来最大奖励的拉杆, 从而谋求利益的最大化. 赌徒的目标就是最大化到赌场赢的钱, 在现实中, 任何人都只能赌有限轮, 但是在 Bandit 问题的设定中, 我们允许赌徒去赌无限轮.

上面是我们对 Bandit 问题的形象化描述, 接下来我们将使用更加严格的语言去描述 Bandit 问题.Bandit 问题可以看作智能体与环境之间的序贯博弈, 这个博弈既可以进行有限轮, 又可以进行无穷轮. 当进行有限轮的时候, 我们可以设它进行了  $N$  轮, 即总轮数, 在文献中被称作Horizon, 由于没有找到合适的中文翻译, 接下来的讲解中将继续使用 Horizon 这个单词. 在每一轮中智能体会先选择一个动作, 也就是对应前面的例子中, 智能体会先选择一个拉杆. 智能体选好动作以后, 环境就会给智能体返回一个奖励, 然后这个博弈就会不断进行下去. 智能体在每一轮的决策所参考的是前面几轮的决策和对应的奖励. 更严格的说, 第  $t$  轮智能体的决策参考的是第一轮到第  $t - 1$  轮智能体选择的动作和对应的奖励. 我们把第一轮到第  $t - 1$  轮智能体选择的动作和对应的奖励称作一条历史, 智能体在决策的时候, 是通过这条历史来决策. 它相当于一个映射, 把第一轮到第  $t - 1$  轮的历史映射为第  $t$  轮智能体选择的动作, 我们把这个映射称为第  $t$  轮时智能体的策略, 以上就是 Bandit 问题基本的介绍. 由于 Bandit 问题有很多分类, 我们后面将会严格并且详细的定义不同类型的 Bandit 问题.

当一个 Bandit 问题是有限动作集时, 我们称它为多臂 Bandit 问题 (Multi-Armed Bandit); 当动作的个数是  $K$  时, 我们就可以称为  $K$ -Armed Bandit.

为什么要研究 Bandit 问题? 因为 Bandit 问题比很多强化学习问题简单, 研究 Bandit 问题有助于启发我们去研究更复杂的问题. 2018 年图灵奖获得者, 人工智能三位先驱之一约书亚 · 本吉奥 (Yoshua Bengio) 曾说过

我们总是过快的放弃了玩具问题,  
而去过度专注于那些需要数周来运行的非常困难的基准测试.

这句话实际上就强调了玩具问题的重要性. Bandit 问题的定义非常简单, 我们就可以把它当成玩具问题. 此外 Bandit 问题虽然是强化学习中简单的问题, 但是研究起来还是很具有挑战性的. 首先由于环境对智能体是未知的, 在 Bandit 问题中智能体就会面临探索与利用的困境. 探索指的就是通过不同的动作来获取更多关于环境的信息, 所谓的利用就是根据已有的信息来做出最优的决策. 举一个现实生活的例子, 假如你搬到一个新的小区, 周围有 10 家饭馆, 过去你只尝试过其中的 5 家, 所谓探索指的就是你去一家没去过的饭馆, 所谓的利用就好比你去已经去过的 5 家中你认为最好的一家. 对于你这个决策者来说, 只去探索和只去利用都是不可取的. 只去探索相当于你每次都随机选一家饭馆, 假如你已经知道某家饭馆很好吃, 某些饭馆很难吃, 那还有什么理由去吃那些很难吃的饭馆呢? 而如果你本身探索的很少, 你只去过 10 家饭馆中的 5 家, 只去利用的话, 也许你心中最好吃的那家并不是真实中最好吃的那家, 也许最好吃的那家饭馆出现在你没去过的那些饭馆中. 因此作为一个决策者, 无论在哪个时刻都会面临探索与利用的困境. 目光长远的策略往往包含短期的牺牲, 因为探索这件事从长期来看是有益的, 毕竟探索之后你可以获得更多的信息, 从而在今后做出更好的决策. 但是从短期来看, 探索的价值往往没有选择已知的饭馆中最好吃的一家, 也就是利用的价值高.

除了探索与利用的困境, Bandit 问题第二个难点是奖励是随机的量. 即使知道奖励的分布, 我们也需要一个效用函数来评估不同分布所对应的效用. 效用函数分成三类, 也就是前面所说的, 风险中性的, 风险厌恶的, 风险偏好的. 后面我们讲的算法中都假设智能体是风险中性的.

另一个 Bandit 问题的难点是, 智能体在博弈的时候并不知道后面还剩多少轮. 对这个难点, 不同算法各自都有区别. 有些算法在不知道总共多少轮的情况下, 也可以获得较好的性能保证. 而另一些算法则需要在一开始就知道总共要玩多少轮. 这些具体区别我们将会在后面详细介绍. 接下来我们就开始介绍最简单的一类 Bandit, 随机 Bandit 问题.

**定义 3.1. 随机 Bandit 问题(Stochastic Bandit).** 一个随机 Bandit 问题由一组分布  $\nu = \{P_a \mid a \in \mathcal{A}\}$  定义, 其中  $\mathcal{A}$  为智能体可选择的动作的集合. 智能体与环境依次交互  $n$  轮, 在第  $t \in \{1, \dots, n\}$  轮交互中, 智能体选择动作  $A_t \in \mathcal{A}$  并交给环境, 环境之后从分布  $P_{A_t}$  中采样奖励  $X_t \in \mathbb{R}$ , 并将  $X_t$  的值告知智能体. 智能体与环境的交互所产生的结果序列  $A_1, X_1, A_2, X_2, \dots, A_n, X_n$  (其中总轮数  $n$  既可以是有限的, 又可以是无限的) 的概率分布应满足下列假设:

1. 奖励  $X_t$  在给定  $A_1, X_1, \dots, A_{t-1}, X_{t-1}$  时的条件分布为  $P_{A_t}$ .
2. 动作  $A_t$  在给定  $A_1, X_1, \dots, A_{t-1}, X_{t-1}$  时的条件分布为  $\pi_t(\cdot \mid A_1, X_1, \dots, A_{t-1}, X_{t-1})$ , 其中  $\pi_1, \pi_2, \dots$  为一个刻画智能体的概率核序列.

### 注记

- (1) 这里策略  $\pi_t$  的定义与强化学习中策略的定义是有区别的, 强化学习是基于马尔可夫决策过程的, 其满足马尔可夫性, 因此其中的策略并不需要依赖整个历史, 只需依赖当前时刻的状态, 而 Bandit 问题中的策略需要依赖整个历史的. Bandit 问题中智能体与环境交互产生的轨迹是不满足马尔可夫性的.
- (2) 同时满足这两条假设的概率空间是存在的. 当博弈是有限轮的时候, 可以直接构造出这个随机过程背后的概率空间; 当这个交互是无限轮的时候, 它背后概率空间存在性由定理 Ionescu-Tulcea 定理保证.
- (3) 完全信息在线学习, Bandit 问题, Partial Monitoring 问题的区别在于环境告诉智能体的信息是有区别的. Bandit 问题中, 环境只会告诉智能体他所选择的动作对应的奖励, 不会告诉智能体其他动作对应的奖励. 完全信息在线学习中, 环境会告诉智能体所有动作对应的奖励. 而在 Partial Monitoring 问题中, 问题的设计者可以指定环境告诉智能体哪些信息.
- (4) 随机 Bandit 问题中假设奖励是有界的.

下面介绍 Bandit 问题中非常重要的概念 **遗憾 (regret)**.

**定义 3.2. 遗憾(regret).** 设  $\nu = \{P_a \mid a \in \mathcal{A}\}$  为一个随机 Bandit, 定义

$$\mu_a(\nu) := \int_{-\infty}^{\infty} x dP_a(x), \quad \mu^*(\nu) := \max_{a \in \mathcal{A}} \mu_a(\nu)$$

假设对于任意  $a \in \mathcal{A}$ ,  $\mu_a(\nu)$  均存在并且有限, 且  $\arg \max_{a \in \mathcal{A}} \mu_a(\nu)$  均非空, 则策略  $\pi$  在 Bandit 实例  $\nu$  下的遗憾为

$$R_n(\pi, \nu) = n\mu^*(\nu) - \mathbb{E} \left[ \sum_{t=1}^n X_t \right]$$

其中的期望是对  $\pi$  和  $\nu$  交互的结果所服从的概率分布求的.

### 注记

(1) 遗憾的本质是一个期望, 一个算法就算遗憾很低, 实际运行时也可能表现很差, 因为实际运行时获得的奖励是个随机变量, 在运气非常不好的情况下, 算法依然可能表现的很糟糕. 这就像即使一个人既聪明又努力, 也未必事业有成. 我们不能夸大运气的作用, 也不能否定运气的影响.

(2) 在 Bandit 问题中, 我们通常会让智能体去最小化遗憾, 而不是最大化奖励. 虽然从逻辑上说最小化遗憾和最大化奖励是等价的, 但从遗憾的增长速度去评估算法的好坏比用累计奖励的增长速度更加合理, 一个遗憾始终为 0 的策略一定是最优策略, 而一个遗憾增长比较慢的算法, 比如以对数速度, 往往也是比较令人满意的算法. 而一个遗憾呈线性增长的算法就不尽如人意. 因为一个智能体即使每次都选择最坏的动作, 遗憾也是呈线性增长的. 若我们只讨论奖励就无法做这种评估, 因为无论最好的策略还是最差的策略, 它的累计期望奖励都是线性增长的.(这里假设奖励是有界的) 这就是我们为什么要定义遗憾.

(3) 可以验证一个随机选动作的算法的遗憾也是线性增长的. 一个随机选动作的算法相当于只做探索不做利用. 一个算法在没有做够充足的探索之前就只去利用, 比如只选择次优动作, 算法的遗憾也是线性增长的.

想要一个算法的遗憾的增长比线性增长慢, 我们就需要合理的权衡探索与利用. 如何界定遗憾的增长比线性增长慢, 这里需要介绍次线性的概念.

**定义 3.3. (次线性).** 函数  $f : \mathbb{Z}^+ \rightarrow \mathbb{R}$  被称作是次线性, 若  $\lim_{n \rightarrow \infty} \frac{f(n)}{n} = 0$ .

想设计一个遗憾是次线性增长的算法并不容易, 事实上很多看似表现不错的算法, 甚至在现实中有许多应用的算法, 它们的遗憾都是线性增长的. 比如  $\varepsilon$ -greedy 算法, 当  $\varepsilon$  是常数的时候, 尽管算法依然会表现出不错的性能, 但从长期来看, 遗憾的增长是线性的. 当我们把  $\varepsilon$  以特定的方式进行衰减的时候, 就可以实现次线性的遗憾增长.

接下来介绍一个非常重要的引理, 遗憾分解引理, 它在后面的证明中会反复用到. 它将遗憾这个量分解为对每个动作求和的形式, 由于我们在理论分析中, 经常需要对每个动作分别讨论, 因此这种对不同动作求和的形式在使用的时候比原始遗憾定义更方便.

**引理 3.4. 遗憾分解引理(regret decomposition lemma).** 对任意动作空间  $\mathcal{A}$  为有限或可数集的随机 Bandit 问题  $\nu$ , 以及任意的策略  $\pi$ , 均满足

$$R_n = \sum_{a \in \mathcal{A}} \Delta_a \mathbb{E}[T_a(n)]$$

其中  $\Delta_a := \mu^*(\nu) - \mu_a(\nu)$  称作动作  $a$  的次优间隙,  $T_a(t) := \sum_{s=1}^t \mathbb{I}_{\{A_s=a\}}$  为第  $t$  轮博弈结束后动作  $a$  被选择的次数.

证明. 根据遗憾定义,

$$\begin{aligned} R_n &= n\mu^* - \mathbb{E}\left[\sum_{t=1}^n X_t\right] \\ &= \sum_{t=1}^n \mu^* - \sum_{t=1}^n \mathbb{E}[X_t] \\ &= \sum_{t=1}^n (\mu^* - \mathbb{E}[X_t]) = \sum_{t=1}^n \mathbb{E}[\mu^* - X_t] \end{aligned}$$

注意到  $\sum_{a \in \mathcal{A}} \mathbb{I}_{\{A_t=a\}} = 1, \forall t \in \{1, \dots, n\}$ . 因为每个时刻只能选择一个动作, 求和中只有一项为 1, 其他项为 0. 从而有

$$R_n = \sum_{t=1}^n \mathbb{E} \left[ (\mu^* - X_t) \sum_{a \in \mathcal{A}} \mathbb{I}_{\{A_t=a\}} \right]$$

我们想要把对动作的求和号移到最外面, 我们分三步来做. 首先因为  $\mu^*$  是存在有限的,  $X_t$  是有限的, 且它们与  $a$  无关. 因此有

$$R_n = \sum_{t=1}^n \mathbb{E} \left[ \sum_{a \in \mathcal{A}} (\mu^* - X_t) \mathbb{I}_{\{A_t=a\}} \right]$$

要交换期望和求和, 因为期望是对概率测度积分, 而求和可以看做对计数测度进行积分, 积分的交换可以应用 Fubini 定理, 只需验证

$$\sum_{a \in \mathcal{A}} \mathbb{E} [ |(\mu^* - X_t) \mathbb{I}_{\{A_t=a\}} | ] < \infty$$

而由于  $(\mu^* - X_t) \mathbb{I}_{\{A_t=a\}}$  为有界随机变量, 且求和项只有一项不为 0, 因此上式成立. 从而由 Fubini 定理

$$R_n = \sum_{t=1}^n \sum_{a \in \mathcal{A}} \mathbb{E} [(\mu^* - X_t) \mathbb{I}_{\{A_t=a\}}]$$

再由对计数测度积分的线性性, 有

$$R_n = \sum_{a \in \mathcal{A}} \sum_{t=1}^n \mathbb{E} [(\mu^* - X_t) \mathbb{I}_{\{A_t=a\}}]$$

再由条件期望的性质

$$R_n = \sum_{a \in \mathcal{A}} \sum_{t=1}^n \mathbb{E} [\mathbb{E} [(\mu^* - X_t) \mathbb{I}_{\{A_t=a\}} | A_t]]$$

定义  $f(x) = \begin{cases} 1, & x = a \\ 0, & x \neq a \end{cases}$ , 则  $\mathbb{I}_{\{A_t=a\}} = f \circ A_t$ , 因为显然  $f$  是可测的, 所以

$\mathbb{I}_{\{A_t=a\}}$  是  $\sigma(A_t)$ -可测的. 从而其可以提到条件期望外面.

$$\begin{aligned}
 R_n &= \sum_{a \in \mathcal{A}} \sum_{t=1}^n \mathbb{E} [\mathbb{I}_{\{A_t=a\}} \mathbb{E} [(\mu^* - X_t) \mid A_t]] \\
 &= \sum_{a \in \mathcal{A}} \sum_{t=1}^n \mathbb{E} [\mathbb{I}_{\{A_t=a\}} (\mu^* - \mathbb{E} [X_t \mid A_t])] \\
 &= \sum_{a \in \mathcal{A}} \sum_{t=1}^n \mathbb{E} [\mathbb{I}_{\{A_t=a\}} (\mu^* - \mathbb{E} [X_t \mid a])] \\
 &= \sum_{a \in \mathcal{A}} \sum_{t=1}^n \mathbb{E} [\mathbb{I}_{\{A_t=a\}} (\mu^* - \mu_a)] \\
 &= \sum_{a \in \mathcal{A}} \sum_{t=1}^n \mathbb{E} [\mathbb{I}_{\{A_t=a\}} \Delta_a] \\
 &= \sum_{a \in \mathcal{A}} \sum_{t=1}^n \mathbb{E} [\mathbb{I}_{\{A_t=a\}}] \Delta_a \\
 &= \sum_{a \in \mathcal{A}} \mathbb{E} \left[ \sum_{t=1}^n \mathbb{I}_{\{A_t=a\}} \right] \Delta_a \\
 &= \sum_{a \in \mathcal{A}} \mathbb{E} [T_a(n)] \Delta_a
 \end{aligned}$$

□

## 3.2 Explore-Then-Commit 算法

本节我们介绍针对随机 Bandit 的第一个算法:Explore-Then-Commit(ETC)算法,ETC 算法在一些文献中也被称作 explore-first 或 exploration first 算法,这个算法的假设有

- (1) 针对随机 Bandit 问题.
- (2) 有限动作集 (即多臂 Bandit).
- (3) 每个动作所对应的奖励分布均是 1-次高斯的.

因为在 Bandit 问题中我们对奖励做正系数的缩放不会影响最优解,也不会影响两个任意解之间比较遗憾大小得到的结果,因此可以对奖励的次高斯参数做一

个归一化, 即第 (3) 条假设不失一般性的认为每个动作所对应的奖励分布次高斯参数都为 1.

顾名思义, ETC 算法分成两个阶段. 第一个阶段是探索阶段, 第二个阶段是利用阶段. 在探索阶段算法会对每一个动作尝试  $m$  次, 这里的  $m$  是算法唯一的超参数. 在利用阶段算法每次选择的动作都是一样的, 即探索阶段中平均奖励最大的那个动作.

伪代码: 设  $k$  个动作,  $N$  轮.

### 算法 1.(Explore-Then-Commit)

输入  $m$  (超参数, 含义为每个动作的探索轮数)

For  $t \in [n]$  do

$$A_t = \begin{cases} (t - 1 \bmod k) + 1 & \text{if } t \leq mk(\text{探索阶段}) \\ \arg \max_{i \in [k]} \hat{\mu}_i(mk) & \text{if } t > mk(\text{利用阶段}) \end{cases}$$

(若  $\arg \max$  结果为多个元素, 则任选一个). 其中

$$\hat{\mu}_i(t) := \frac{1}{T_i(t)} \sum_{s=1}^t \mathbb{I}_{\{A_s=i\}} X_s, \quad T_i(t) := \sum_{s=1}^t \mathbb{I}_{\{A_s=i\}}$$

执行  $A_t$  记录  $X_t$ .

End For.

**定理 3.5.** 当 ETC 算法与任意 1-次高斯 Bandit 交互时, 有:

$$R_n \leq \underbrace{m \sum_{i=1}^k \Delta_i}_{\text{探索阶段的遗憾}} + \underbrace{(n - mk) \sum_{i=1}^k \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right)}_{\text{利用阶段的遗憾上界}}$$

证明. 不失一般性地假设第一个动作最优, 即  $\mu_1 = \mu^* = \max_{i \in [k]} \mu_i$ . 根据遗憾分解引理,

$$(3.6) \quad R_n = \sum_{i=1}^k \Delta_i \mathbb{E}[T_i(n)]$$

由于  $\Delta_i$  是确定的, 我们只需分析  $\mathbb{E}[T_i(n)]$  的上界. 易知

$$(3.7) \quad \mathbb{E}[T_i(n)] = m + (n - mk)\mathbb{P}\left(\hat{\mu}_i(mk) \geq \max_{j \neq i} \hat{\mu}_j(mk)\right)$$

我们只需分析右侧概率的上界

$$\begin{aligned} \mathbb{P}\left(\hat{\mu}_i(mk) \geq \max_{j \neq i} \hat{\mu}_j(mk)\right) &\leq \mathbb{P}(\hat{\mu}_i(mk) \geq \hat{\mu}_1(mk)) \quad (\text{依据测度的单调性}) \\ &= \mathbb{P}(\hat{\mu}_i(mk) - \hat{\mu}_1(mk) - (\mu_i - \mu_1) \geq \Delta_i) \end{aligned}$$

根据  $\hat{\mu}_i$  的定义:

$$\begin{aligned} \hat{\mu}_i(mk) &= \frac{1}{T_i(mk)} \sum_{s=1}^{mk} \mathbb{I}_{\{A_s=i\}} X_s \\ &= \frac{1}{m} \sum_{s=1}^m X_{k(s-1)+i} \end{aligned}$$

因此有

$$\begin{aligned} &\mathbb{P}(\hat{\mu}_i(mk) - \hat{\mu}_1(mk) - (\mu_i - \mu_1) \geq \Delta_i) \\ &= \mathbb{P}\left(\frac{1}{m} \sum_{s=1}^m X_{k(s-1)+i} - \frac{1}{m} \sum_{s=1}^m X_{k(s-1)+1} - (\mu_i - \mu_1) \geq \Delta_i\right) \\ &= \mathbb{P}\left(\frac{1}{m} \sum_{s=1}^m (X_{k(s-1)+i} - X_{k(s-1)+1}) - \frac{1}{m} \sum_{s=1}^m (\mu_i - \mu_1) \geq \Delta_i\right) \\ &= \mathbb{P}\left(\sum_{s=1}^m (X_{k(s-1)+i} - X_{k(s-1)+1} - (\mu_i - \mu_1)) \geq m\Delta_i\right) \\ &= \mathbb{P}\left(\sum_{s=1}^m (X_{k(s-1)+i} - X_{k(s-1)+1} - (\mathbb{E}[X_{k(s-1)+i}] - \mathbb{E}[X_{k(s-1)+1}])) \geq m\Delta_i\right) \\ &= \mathbb{P}\left(\sum_{s=1}^m (X_{k(s-1)+i} - X_{k(s-1)+1} - (\mathbb{E}[X_{k(s-1)+i}] - X_{k(s-1)+1})) \geq m\Delta_i\right) \end{aligned}$$

根据假设  $\forall i, X_i$  均是 1-次高斯的, 所以根据次高斯性的相关性质  $X_{k(s-1)+i} - X_{k(s-1)+1}$  是  $\sqrt{1^2 + 1^2} = \sqrt{2}$ -次高斯的. 由 Hoeffding 界可得

$$\begin{aligned} &\mathbb{P}\left(\sum_{s=1}^m (X_{k(s-1)+i} - X_{k(s-1)+1} - (\mathbb{E}[X_{k(s-1)+i}] - X_{k(s-1)+1})) \geq m\Delta_i\right) \\ &\leq \exp\left(-\frac{(m\Delta_i)^2}{2 \sum_{i=1}^m \sqrt{2}^2}\right) = \exp\left(-\frac{m\Delta_i^2}{4}\right) \end{aligned}$$

因此有

$$(3.8) \quad \mathbb{P}\left(\hat{\mu}_i(mk) \geq \max_{j \neq i} \hat{\mu}_j(mk)\right) \leq \exp\left(-\frac{m\Delta_i^2}{4}\right)$$

结合(3.6),(3.7),(3.8)式得

$$\begin{aligned} R_n &= \sum_{i=1}^k \Delta_i \mathbb{E}[T_i(n)] \quad (\text{根据(3.6)式}) \\ &\leq \sum_{i=1}^k \Delta_i \left( m + (n - mk) \mathbb{P}\left(\hat{\mu}_i(mk) \geq \max_{j \neq i} \hat{\mu}_j(mk)\right) \right) \quad (\text{代入(3.7)式}) \\ &\leq \sum_{i=1}^k \Delta_i \left( m + (n - mk) \exp\left(-\frac{m\Delta_i^2}{4}\right) \right) \quad (\text{代入(3.8)式}) \\ &= m \sum_{i=1}^k \Delta_i + (n - mk) \sum_{i=1}^k \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right) \end{aligned}$$

□

### 注记

(1) 这个上界刻画了探索与利用之间的困境. 当  $m$  越大, 第一项越大, 第二项越小, 直觉上的解释是探索越多, 就越容易在利用阶段选到最优动作, 这样虽然探索阶段遗憾增加了, 但利用阶段遗憾降低了; 反过来若  $m$  非常小, 则虽然探索的遗憾小了, 但很容易选到不太好的次优动作, 并且一直执行下去, 因此利用阶段的遗憾就增加了.

(2) 当  $m$  是常数时, 算法的遗憾线性增长.

**性质 3.9.** 在 ETC 算法中, 若动作数  $k = 2$ , 且算法可提前得知每个动作  $i$  的次优间隙  $\Delta_i$  及总轮数  $n$ , 则通过适当地选取  $m$ , 算法可以实现次线性遗憾增长.

证明. 根据定理(3.5):

$$R_n \leq m \sum_{i=1}^k \Delta_i + (n - mk) \sum_{i=1}^k \Delta_i \exp\left(-\frac{m\Delta_i^2}{4}\right)$$

不失一般性假设  $\Delta_1 = 0$ , 将  $\Delta_2$  改写为  $\Delta$ . 则有

$$\begin{aligned} R_n &\leq m\Delta + (n - 2m)\Delta \exp\left(-\frac{m\Delta^2}{4}\right) \\ &\leq m\Delta + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right) \end{aligned}$$

显然, 不等号右侧相对于  $m$  是凸函数. 故选取使之最小化的  $m$ , 只需对  $m$  求导, 再将导函数置为 0.

$$\begin{aligned} &\Delta + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right) \cdot \left(-\frac{\Delta^2}{4}\right) \\ &= \Delta - \frac{n\Delta^3}{4} \exp\left(-\frac{m\Delta^2}{4}\right) \end{aligned}$$

将其置为零:

$$\Delta - \frac{n\Delta^3}{4} \exp\left(-\frac{m^*\Delta^2}{4}\right) = 0$$

即  $\frac{n\Delta^3}{4} \exp\left(-\frac{m^*\Delta^2}{4}\right) = \Delta$ , 假设  $\Delta \neq 0$ , 则

$$\frac{n\Delta^2}{4} \exp\left(-\frac{m^*\Delta^2}{4}\right) = 1$$

解得

$$m^* = -\frac{4}{\Delta^2} \ln\left(\frac{4}{n\Delta^2}\right) = \frac{4}{\Delta^2} \ln\left(\frac{n\Delta^2}{4}\right)$$

由于  $m$  只可取正整数, 故设

$$m = \max\left\{1, \left\lceil \frac{4}{\Delta^2} \ln\left(\frac{n\Delta^2}{4}\right) \right\rceil\right\}$$

因为有

$$\begin{aligned} \max\left\{1, \frac{4}{\Delta^2} \ln\left(\frac{n\Delta^2}{4}\right)\right\} &\leq m \leq \max\left\{1, \frac{4}{\Delta^2} \ln\left(\frac{n\Delta^2}{4}\right) + 1\right\} \\ &= \max\left\{0, \frac{4}{\Delta^2} \ln\left(\frac{n\Delta^2}{4}\right)\right\} + 1 \end{aligned}$$

将  $m$  的范围代入遗憾界:

$$\begin{aligned}
 R_n &\leq m\Delta + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right) \\
 &\leq \Delta \left( \max \left\{ 0, \frac{4}{\Delta^2} \ln \left( \frac{n\Delta^2}{4} \right) \right\} + 1 \right) + n\Delta \exp\left(-\frac{m\Delta^2}{4}\right) \\
 &\leq \Delta \max \left\{ 0, \frac{4}{\Delta^2} \ln \left( \frac{n\Delta^2}{4} \right) \right\} + \Delta + n\Delta \exp\left(-\frac{\max \left\{ 1, \frac{4}{\Delta^2} \ln \left( \frac{n\Delta^2}{4} \right) \right\} \Delta^2}{4}\right) \\
 &= \max \left\{ 0, \frac{4}{\Delta} \ln \left( \frac{n\Delta^2}{4} \right) \right\} + \Delta + \min \left\{ n\Delta \exp\left(-\frac{\Delta^2}{4}\right), n\Delta \exp\left(-\frac{\frac{4}{\Delta^2} \ln \left( \frac{n\Delta^2}{4} \right) \Delta^2}{4}\right) \right\} \\
 &\leq \max \left\{ 0, \frac{4}{\Delta} \ln \left( \frac{n\Delta^2}{4} \right) \right\} + \Delta + \left\{ n\Delta, \frac{4}{\Delta} \right\} \\
 &= \Delta + \frac{4}{\Delta} \left( \max \left\{ 0, \ln \left( \frac{n\Delta^2}{4} \right) \right\} + \min \left\{ \frac{n\Delta^2}{4}, 1 \right\} \right)
 \end{aligned}$$

因为  $\ln x \leq \sqrt{x}, \forall x > 0$ , 所以

$$\max \left\{ 0, \ln \left( \frac{n\Delta^2}{4} \right) \right\} \leq \max \left\{ 0, \sqrt{\frac{n\Delta^2}{4}} \right\} \leq \sqrt{\frac{n\Delta^2}{4}}$$

若  $\frac{n\Delta^2}{4} \geq 1$ , 则  $\sqrt{\frac{n\Delta^2}{4}} \geq 1$ , 有

$$\min \left\{ \frac{n\Delta^2}{4}, 1 \right\} \leq \sqrt{\frac{n\Delta^2}{4}}$$

若  $\frac{n\Delta^2}{4} < 1$ , 则  $\frac{n\Delta^2}{4} \leq \sqrt{\frac{n\Delta^2}{4}}$ , 也有

$$\min \left\{ \frac{n\Delta^2}{4}, 1 \right\} \leq \sqrt{\frac{n\Delta^2}{4}}$$

综上有

$$\begin{aligned}
 R_n &\leq \Delta + \frac{4}{\Delta} \left( \sqrt{\frac{n\Delta^2}{4}} + \sqrt{\frac{n\Delta^2}{4}} \right) \\
 &= \Delta + 4\sqrt{n}
 \end{aligned}$$

所以算法可以实现次线性遗憾增长.

□

## 注记

- 命题的缺点:(1) 只支持两个动作的环境.  
(2) 需提前知道总轮数  $n$ .(非 Anytime)  
(3) 使用 doubling trick 可以将算法改进为 Anytime 的算法.

### 3.3 UCB 算法: 简介, 流程与公式推导

UCB 全称为 Upper Confidence Bound, 有多个中文翻译: 上置信界/上限置信区间/置信区间上界等. UCB 算法是被广泛应用于游戏 AI 和推荐系统的算法. 其具体应用举例如下:

- AlphaGo 围棋 AI 中采用了 UCT 算法, 而 UCT 使用了 UCB 的公式.
- 即时策略 (RTS) 游戏 AI 中, 使用 UCB 选策略.(存在致命缺点)
- 推荐系统

我们首先回顾 ETC 算法的缺点:

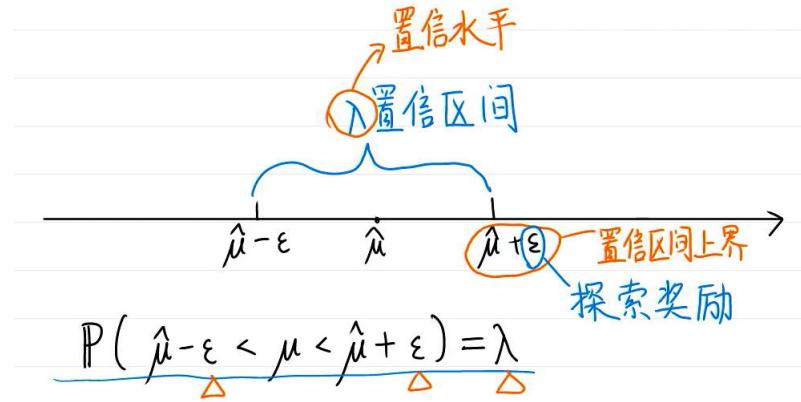
1. 若想获得次线性遗憾增长, 则  $m$  的设定需要依赖  $n$ (博弈的总轮数) 及动作的次优间隙.
2. 若想获得次线性遗憾增长, 则动作数  $k$  只能为 2.
3. ETC 对每个动作均探索相同次数, 即使对一个尝试几次后几乎肯定很差的动作也会尝试  $m$  次, 从而造成浪费.

UCB 会在每轮中对每个动作进行评分 (计算 Index), 之后选评分最高的动作. 故设计合理的评分计算公式十分重要.

计算每个动作评分的公式的设计目标:

- (1) 若一个动作探索的次数比别的动作少, 则它更该被选中.(评分应更高)
- (2) 若一个动作以往探索所取得的奖励高于别的动作, 则它更该被选中.

下面介绍置信区间的概念. 首先我们手上有获得的某个动作奖励的历史数据, 根据它可以估计出动作奖励分布的均值  $\hat{\mu}$ , 虽然真实均值  $\mu$  是未知的, 但它大概率出现在估计量  $\hat{\mu}$  附近. 参考如下示意图:



样本容量越大，置信区间越窄；样本容量越小，置信区间越宽。这个规律从直觉上不难理解，因为我们收集的奖励的样本越多，对奖励的均值的估计也就越准确，那么置信区间自然也会跟着缩窄。UCB 算法把置信区间上界作为动作的评分，那么这种设计是否能满足我们上面提到的设计目标呢？首先验证第 (1) 条，若一个动作探索的次数比别的动作少，则收集的关于那个动作的奖励数据少，即样本容量小，导致置信区间半径  $\varepsilon$  大，由于置信区间上界为  $\hat{\mu} + \varepsilon$ ，因此随着  $\varepsilon$  变大，置信区间上界也会随之变大，这样一来该动作的评分就变高了，也就更容易被选中了，这就精准的体现了第一条设计目标。再看第 (2) 条设计目标，由于  $\hat{\mu}$  是以往探索取得奖励的经验平均，它受以往探索取得的奖励影响而变化，若一个动作以往探索所取得的奖励高于别的动作，则  $\hat{\mu}$  就会很高，则置信区间上界  $\hat{\mu} + \varepsilon$  就会变高，则评分更高，也就更容易被选中，第 (2) 条设计目标也满足！也易知 UCB 算法可以改进 ETS 算法的缺点 3. UCB 的推导：

$$P(\hat{\mu} - \varepsilon < \mu < \hat{\mu} + \varepsilon) = \lambda$$

因为 UCB 算法只关心置信区间的上界，对其修改为  $P(\mu < \hat{\mu} + \varepsilon) = \lambda$ 。从而有

$$P(\mu - \hat{\mu} \geq \varepsilon) = 1 - \lambda =: \delta$$

适当放宽条件，得到保守置信区间

$$P(\mu - \hat{\mu} \geq \varepsilon) \leq \delta$$

依据 Hoeffding 界：

$$P(\mu - \hat{\mu} \geq \varepsilon) \leq \exp\left(-\frac{n\varepsilon^2}{2}\right) = \delta$$

解得

$$\varepsilon = \sqrt{\frac{2 \ln(\frac{1}{\delta})}{n}}$$

故  $t - 1$  时刻第  $i$  个动作的指标 (index) 可定义为:

$$\text{UCB}_i(t-1, \delta) := \begin{cases} \infty & \text{若 } T_i(t-1) = 0 \\ \hat{\mu}_i(t-1) + \sqrt{\frac{2 \ln(\frac{1}{\delta})}{T_i(t-1)}} & \text{其他情况} \end{cases}$$

### 算法 2(UCB1)

输入  $k$

for  $t \in 1, \dots, n$  do

选择  $A_t = \arg \max_{i \in [k]} \text{UCB}_i \left( t-1, \frac{1}{t-1} \right)$ .

记录  $X_t$ , 更新置信区间上界.

end for

## 3.4 UCB1 算法的理论分析

本节主要参考以下论文:

[Finite-time Analysis of Multiarmed Bandit Problem, Peter Auer 等人](#)

UCB1 算法理论分析的大体思路:

算法产生了遗憾  $\iff$  算法选择了次优动作

$\iff$  次优动作的指标高于其他所有动作

$\implies$  次优动作的指标高于最优动作

$\implies \begin{cases} \text{最优动作的指标过低} \\ \text{次优动作的指标过高} \end{cases}$

因为一个动作的指标由两部分组成, 一部分是奖励均值的经验估计  $\hat{\mu}$ , 一部分是探索奖励 (置信半径)  $\varepsilon$ , 而由于探索奖励是随着轮数增加而衰减的, 所以其只影响前有限轮的遗憾, 无长期影响.

次优动作的指标过高  $\begin{cases} \text{次优动作奖励均值的经验估计过高} \\ \text{次优动作的探索奖励过高 (只影响前有限轮的遗憾, 无长期影响)} \end{cases}$

**定理 3.10.** 对于任意  $K > 1$ , 在奖励分布为  $P_1, \dots, P_K$ , 支撑集为  $[0, 1]$  的  $K$  臂随机 Bandit 问题上运行 UCB1 策略, 则  $n$  轮后该策略所产生的遗憾最多为:

$$\left[ 8 \sum_{i:\mu_i < \mu^*} \left( \frac{\ln n}{\Delta_i} \right) \right] + \left( 1 + \frac{\pi^2}{3} \right) \left( \sum_{j=1}^K \Delta_j \right)$$

其中  $\mu_1, \dots, \mu_K$  为分布  $P_1, \dots, P_K$  的均值.

证明. 设  $X_{t,i}$  为第  $i$  个动作第  $t$  次被选择后返回的奖励, 故可知  $X_t = X_{T_{A_t}(t), A_t}$ . 设  $\hat{\mu}_{i,s} := \frac{1}{s} \sum_{u=1}^s X_{u,i}$  为基于前  $s$  个样本计算的  $i$  动作的奖励的平均值, 可知  $\hat{\mu}_i(t) = \hat{\mu}_{i,T_i(t)}$ . 不失一般性假设第 1 个动作最优, 即  $\mu_1 = \mu^* = \max_{i \in [K]} \mu_i$ . 设  $c_{t,s} := \sqrt{\frac{2 \ln t}{s}}$ . 可知若  $T_i(t-1) > 0$ , 则

$$\text{UCB}_i \left( t-1, \frac{1}{t-1} \right) = \hat{\mu}_{i,T_i(t-1)} + c_{t-1,T_i(t-1)}$$

设  $\ell$  为一个任意正整数, 假设  $n > K$ , 则有:

$$T_i(n) = \sum_{t=1}^n \mathbb{I}_{\{A_t=i\}} = \sum_{t=1}^K \mathbb{I}_{\{A_t=i\}} + \sum_{t=K+1}^n \mathbb{I}_{\{A_t=i\}}$$

因此 UCB1 算法前  $K$  轮必然每个动作分别访问 1 次, 所以有

$$\begin{aligned} T_i(n) &= 1 + \sum_{t=K+1}^n (\mathbb{I}_{\{A_t=i, T_i(t-1) \geq \ell\}} + \mathbb{I}_{\{A_t=i, T_i(t-1) < \ell\}}) \\ &= 1 + \sum_{t=K+1}^n \mathbb{I}_{\{A_t=i, T_i(t-1) \geq \ell\}} + \sum_{t=K+1}^n \mathbb{I}_{\{A_t=i, T_i(t-1) < \ell\}} \end{aligned}$$

注意到求和式  $\sum_{t=K+1}^n \mathbb{I}_{\{A_t=i, T_i(t-1) < \ell\}}$  中最多  $\ell - 1$  项不为 0, 反证法, 假设存在第  $\ell$  个不为 0 的项, 即  $\exists t \in \{K+1, \dots, n\}$  使得  $T_i(t) = \ell + 1$ , 则  $T_i(t-1) \geq \ell$ , 与

$T_i(t-1) < \ell$  矛盾. 因此  $\sum_{t=K+1}^n \mathbb{I}_{\{A_t=i, T_i(t-1)<\ell\}} \leq \ell - 1$ , 从而有

$$T_i(n) \leq \ell + \sum_{t=K+1}^n \mathbb{I}_{\{A_t=i, T_i(t-1)\geq\ell\}}$$

注意到

$$\begin{aligned} A_t = i &\iff \forall j \in [K], \text{UCB}_j\left(t-1, \frac{1}{t-1}\right) \leq \text{UCB}_i\left(t-1, \frac{1}{t-1}\right) \\ &\implies \text{UCB}_1\left(t-1, \frac{1}{t-1}\right) \leq \text{UCB}_i\left(t-1, \frac{1}{t-1}\right) \end{aligned}$$

从而有

$$\begin{aligned} T_i(n) &\leq \ell + \sum_{t=K+1}^n \mathbb{I}_{\{\text{UCB}_1\left(t-1, \frac{1}{t-1}\right) \leq \text{UCB}_i\left(t-1, \frac{1}{t-1}\right), T_i(t-1) \geq \ell\}} \\ &\leq \ell + \sum_{t=K+1}^n \mathbb{I}_{\{\hat{\mu}_{1,T_1(t-1)} + c_{t-1,T_1(t-1)} \leq \hat{\mu}_{i,T_i(t-1)} + c_{t-1,T_i(t-1)}, T_i(t-1) \geq \ell\}} \end{aligned}$$

若指示函数非零, 易知  $1 \leq T_1(t-1) \leq t-1, \ell \leq T_i(t-1) \leq t-1$ , 则有

$$\begin{aligned} T_i(n) &\leq \ell + \sum_{t=K+1}^n \mathbb{I} \left\{ \bigcup_{s \in \{1, \dots, t-1\}} \bigcup_{s_i \in \{\ell, \dots, t-1\}} \{\hat{\mu}_{1,s} + c_{t-1,s} \leq \hat{\mu}_{i,s_i} + c_{t-1,s_i}\} \right\} \\ &\leq \ell + \sum_{t=K}^{n-1} \mathbb{I} \left\{ \bigcup_{s \in \{1, \dots, t\}} \bigcup_{s_i \in \{\ell, \dots, t\}} \{\hat{\mu}_{1,s} + c_{t,s} \leq \hat{\mu}_{i,s_i} + c_{t,s_i}\} \right\} \\ &\leq \ell + \sum_{t=K}^{n-1} \sum_{s=1}^t \sum_{s_i=\ell}^t \mathbb{I}_{\{\hat{\mu}_{1,s} + c_{t,s} \leq \hat{\mu}_{i,s_i} + c_{t,s_i}\}} \end{aligned}$$

$\hat{\mu}_{1,s} + c_{t,s} \leq \hat{\mu}_{i,s_i} + c_{t,s_i}$  的意义为: 第  $t$  时刻, 最优动作被选过  $s$  次, 第  $i$  个动作被选过  $s_i$  次, 最优动作指标  $\leq$  动作  $i$  指标. 根据前面的思路分析, 发生这种现象共有 3 种情况: 最优动作的指标过高, 次优动作奖励均值的经验估计过高, 次优动作的探索奖励过高. 断言: 若  $\hat{\mu}_{1,s} + c_{t,s} \leq \hat{\mu}_{i,s_i} + c_{t,s_i}$  成立, 则下列 3 个不等式至少一个成立:

$$(3.11) \quad \hat{\mu}_{1,s} \leq \mu_1 - c_{t,s}$$

$$(3.12) \quad \hat{\mu}_{i,s_i} \geq \mu_i + c_{t,s_i}$$

$$(3.13) \quad \mu_1 < \mu_i + 2c_{t,s_i}$$

- (3.11)式的含义: 最优动作的指标过低 (最优动作的指标比真实奖励均值还要低, 严重低估!).
- (3.12)式的含义: 第  $i$  个动作对应的奖励期望的估计过大.
- (3.13)式的含义: 第  $i$  个动作对应的探索奖励过大, 大于第  $i$  个动作次优间隙的一半.

为验证(3.11),(3.12),(3.13)式至少成立一个, 采用反证法, 假设 3 个式子均不成立. 则

$$\hat{\mu}_{1,s} > \mu_1 - c_{t,s} \quad (\text{由于(3.11)式不成立})$$

即

$$(3.14) \quad \hat{\mu}_{1,s} + c_{t,s} > \mu_1$$

此外,

$$\hat{\mu}_{i,s_i} < \mu_i + c_{t,s_i} \quad (\text{由于(3.12)式不成立})$$

从而有

$$(3.15) \quad \hat{\mu}_{i,s_i} + c_{t,s_i} < \mu_i + 2c_{t,s_i}$$

由于  $\hat{\mu}_{1,s} + c_{t,s} \leq \hat{\mu}_{i,s_i} + c_{t,s_i}$ , 结合(3.14),(3.15)式可得:

$$\mu_1 < \hat{\mu}_{1,s} + c_{t,s} \leq \hat{\mu}_{i,s_i} + c_{t,s_i} < \mu_i + 2c_{t,s_i}$$

有  $\mu_1 < \mu_i + 2c_{t,s_i}$ , 由此可知(3.13)式成立, 与反证假设矛盾, 故待证命题成立. 对(3.11),(3.12)式所对应的事件应用 Hoeffding 界, 得:

$$\begin{aligned} \mathbb{P}(\hat{\mu}_{1,s} \leq \mu_1 - c_{t,s}) &= \mathbb{P}((-\hat{\mu}_{1,s}) - (-\mu_1) \geq c_{t,s}) \\ &\leq \exp\left(\frac{-2s^2 c_{t,s}^2}{s(0 - (-1))^2}\right) \\ &= \exp\left(-2s \cdot \frac{2 \ln t}{s}\right) = \exp(-4 \ln t) = t^{-4} \end{aligned}$$

$$\begin{aligned}\mathbb{P}(\hat{\mu}_{i,s_i} \geq \mu_i + c_{t,s_i}) &= \mathbb{P}(\hat{\mu}_{i,s_i} - \mu_i \geq c_{t,s_i}) \\ &\leq \exp\left(\frac{-2s_i^2 c_{t,s_i}^2}{s_i(1-0)^2}\right) = t^{-4}\end{aligned}$$

设  $\ell = \left\lceil \frac{8 \ln n}{\Delta_i^2} \right\rceil$  可使(3.13)式不成立, 验证过程如下: 使用反证法, 假设(3.13)式成立, 则  $\mu_1 - \mu_i - 2c_{t,s_i} < 0$ . 但是另一方面:

$$\mu_1 - \mu_i - 2c_{t,s_i} = \mu_1 - \mu_i - 2\sqrt{\frac{2 \ln t}{s_i}}$$

注意到  $s_i \geq \ell \geq \frac{8 \ln n}{\Delta_i^2}$ , 所以

$$\begin{aligned}\mu_1 - \mu_i - 2c_{t,s_i} &\geq \mu_1 - \mu_i - 2\sqrt{\frac{2 \ln t \Delta_i^2}{8 \ln n}} \\ &= \mu_1 - \mu_i - \sqrt{\frac{\ln t \Delta_i^2}{\ln n}} \\ &\geq \mu_1 - \mu_i - \sqrt{\frac{\ln n \Delta_i^2}{\ln n}} \quad (t \leq n-1 < n) \\ &= \mu_1 - \mu_i - \Delta_i = 0\end{aligned}$$

与  $\mu_1 - \mu_i - 2c_{t,s_i} < 0$  矛盾, 故(3.13)式不成立. 因此

$$\begin{aligned}\mathbb{E}[T_i(n)] &\leq \mathbb{E} \left[ \ell + \sum_{t=K}^{n-1} \sum_{s=1}^t \sum_{s_i=\ell}^t \mathbb{I}_{\{\hat{\mu}_{1,s} + c_{t,s} \leq \hat{\mu}_{i,s_i} + c_{t,s_i}\}} \right] \\ &= \ell + \sum_{t=K}^{n-1} \sum_{s=1}^t \sum_{s_i=\ell}^t \mathbb{P}\{\hat{\mu}_{1,s} + c_{t,s} \leq \hat{\mu}_{i,s_i} + c_{t,s_i}\} \\ &\leq \ell + \sum_{t=K}^{n-1} \sum_{s=1}^t \sum_{s_i=\ell}^t \mathbb{P}\{\hat{\mu}_{1,s} \leq \mu_1 - c_{t,s}\} \cup \{\hat{\mu}_{i,s_i} \geq \mu_i + c_{t,s_i}\} \\ &\leq \ell + \sum_{t=K}^{n-1} \sum_{s=1}^t \sum_{s_i=\ell}^t [\mathbb{P}\{\hat{\mu}_{1,s} \leq \mu_1 - c_{t,s}\} + \mathbb{P}\{\hat{\mu}_{i,s_i} \geq \mu_i + c_{t,s_i}\}] \\ &\leq \ell + \sum_{t=K}^{n-1} \sum_{s=1}^t \sum_{s_i=\ell}^t (t^{-4} + t^{-4}) \\ &= \left\lceil \frac{8 \ln n}{\Delta_i^2} \right\rceil + \sum_{t=K}^{n-1} \sum_{s=1}^t \sum_{s_i=\ell}^t 2t^{-4}\end{aligned}$$

易知  $\left\lceil \frac{8 \ln n}{\Delta_i^2} \right\rceil \leq \frac{8 \ln n}{\Delta_i^2} + 1$ ,  $n > K, K > 1$ , 所以  $n \geq 3$ , 则  $\frac{8 \ln n}{\Delta_i^2} > 0$ , 所以  $\ell \geq 1$ .  
则

$$\begin{aligned}\mathbb{E}[T_i(n)] &\leq \frac{8 \ln n}{\Delta_i^2} + 1 + \sum_{t=K}^{n-1} \sum_{s=1}^t \sum_{s_i=1}^t 2t^{-4} \\ &= \frac{8 \ln n}{\Delta_i^2} + 1 + \sum_{t=K}^{n-1} t \cdot t \cdot 2t^{-4} \\ &\leq \frac{8 \ln n}{\Delta_i^2} + 1 + \sum_{t=1}^{\infty} 2t^{-2}\end{aligned}$$

由于  $\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}$ , 所以

$$\mathbb{E}[T_i(n)] \leq \frac{8 \ln n}{\Delta_i^2} + 1 + \frac{\pi^2}{3}$$

将上式代入遗憾分解引理, 得:

$$\begin{aligned}R_n &= \sum_{i=1}^K \mathbb{E}[T_i(n)] \Delta_i \\ &= \sum_{i=2}^K \mathbb{E}[T_i(n)] \Delta_i \\ &\leq 8 \sum_{i=2}^K \frac{\ln n}{\Delta_i} + \left(1 + \frac{\pi^2}{3}\right) \left(\sum_{i=2}^K \Delta_i\right)\end{aligned}$$

□

### 注记

该定理给出了 UCB1 算法一个次线性的遗憾界, 但遗憾界依赖于次优间隙的倒数, 这点非常不好. 因为只要我们这个问题中有一个动作次优间隙可以任意小的话, 则该定理给出的界就会任意大, 任意大的上界没有任何意义, 这就是该定理一个关键的不足之处. 但我们稍加修改, 可以证明出一个不依赖于次优间隙的倒数的遗憾界.

下面从直觉角度分析为什么次优间隙很小的话, 我们刚证明的界会特别大. 问题出在定理(3.10) 的推导中被我们忽略了(3.13)式上. 我们之所以忽略(3.13)式, 是由于在(3.13)式有可能成立的情况下, 次优动作被选中的概率就变得不好分析了, 因此为了排除这种情况, 也就是为了使得(3.13)式永远都不成立, 就要求动作至少被选择  $\ell$  次, 而我们的定理(3.10) 近似的认为算法会先给每个次优动作选择  $\ell$  次, 直到使得(3.13)式不成立, 之后再按照正常的逻辑按部就班地选动作, 这个近似非常的不准确, 因为现实中对于某个次优动作来说, 即使(3.13)式是成立的, 算法也不一定会选择那个动作, 现实中(3.13)式的成立只会增加那个动作被选中的概率, 而绝不是说一定要选它. 尽管这个近似不准确, 但由于我们推的是遗憾的上界, 也只能说这个近似使得上界变松了, 但它却是完全合法的. 这种近似通常不会带来很大的问题, 因为当动作的次优间隙不是特别小的时候,  $\ell$  的值就在一个可以接受的范围内. 但是当次优间隙非常小的时候, 由于  $\ell$  的解析式中分母是次优间隙的平方, 因此  $\ell$  就会非常大. 这在直觉上也很容易理解, 因为要想让(3.13)式不成立, 就一定要让探索奖励小于次优间隙的一半. 假如次优间隙本身就很小, 探索奖励就必须动作被选择相当多的次数后才能衰减到小于次优间隙的一半, 这个相当多的次数就是这里的  $\ell$ , 此时  $\ell$  会非常大, 可能甚至比总博弈轮数  $N$  都要大. 这样定理(3.10)假设每个动作会先选择  $\ell$  次就非常不符合实际, 因此此时定理(3.10)算出的上界就非常松了. 因此一个看似合理的改进方案就是单独处理次优间隙小的动作, 通过这样的处理, 我们就得到了一个不依赖于次优间隙倒数的遗憾界.

**定理 3.16.** 在和定理(3.10)相同的假设下, UCB1 的遗憾满足:

$$R_n \leq 4\sqrt{2nk \log n} + \sum_{i=1}^K \left( \Delta_i + \frac{\pi^2 \Delta_i}{3} \right)$$

且该界是次线性的.

证明. 为了衡量次优间隙的大小, 设阈值  $\Delta > 0$ . 根据定理(3.10)的证明过程, 对任意次优动作  $i$ , 都有:

$$\mathbb{E}[T_i(n)] \leq \frac{8 \ln n}{\Delta_i^2} + 1 + \frac{\pi^2}{3}$$

根据遗憾分解引理:

$$\begin{aligned} R_n &= \sum_{i=1}^K \Delta_i \mathbb{E}[T_i(n)] = \sum_{i:\Delta_i < \Delta} \Delta_i \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i \geq \Delta} \Delta_i \mathbb{E}[T_i(n)] \\ &< \Delta \sum_{i:\Delta_i < \Delta} \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i \geq \Delta} \left( \frac{8 \ln n}{\Delta_i} + \Delta_i + \frac{\pi^2 \Delta_i}{3} \right) \end{aligned}$$

注意到

$$\sum_{i:\Delta_i < \Delta} \mathbb{E}[T_i(n)] = \mathbb{E} \left[ \sum_{i:\Delta_i < \Delta} T_i(n) \right] \leq \mathbb{E} \left[ \sum_{i=1}^K T_i(n) \right] = \mathbb{E}[n] = n$$

从而有

$$\begin{aligned} R_n &\leq n\Delta + \sum_{i:\Delta_i \geq \Delta} \frac{8 \ln n}{\Delta_i} + \sum_{i=1}^K \left( \Delta_i + \frac{\pi^2 \Delta_i}{3} \right) \\ &\leq n\Delta + \frac{8K \ln n}{\Delta} + \sum_{i=1}^K \left( \Delta_i + \frac{\pi^2 \Delta_i}{3} \right) \end{aligned}$$

令  $\Delta = \sqrt{\frac{8K \ln n}{n}}$ , 有

$$\begin{aligned} R_n &\leq n \sqrt{\frac{8K \ln n}{n}} + 8K \ln n \cdot \sqrt{\frac{n}{8K \ln n}} + \sum_{i=1}^K \left( \Delta_i + \frac{\pi^2 \Delta_i}{3} \right) \\ &= 4\sqrt{2nK \ln n} + \sum_{i=1}^K \left( \Delta_i + \frac{\pi^2 \Delta_i}{3} \right) \end{aligned}$$

最后验证该界是次线性的:

$$\begin{aligned} &\lim_{n \rightarrow \infty} \frac{4\sqrt{2nK \ln n} + \sum_{i=1}^K \left( \Delta_i + \frac{\pi^2 \Delta_i}{3} \right)}{n} \\ &= \lim_{n \rightarrow \infty} \frac{4\sqrt{2nK \ln n}}{n} + \lim_{n \rightarrow \infty} \frac{\sum_{i=1}^K \left( \Delta_i + \frac{\pi^2 \Delta_i}{3} \right)}{n} \end{aligned}$$

由于  $\sum_{i=1}^K \left( \Delta_i + \frac{\pi^2 \Delta_i}{3} \right)$  为常数, 且根据洛必达法则,  $\lim_{n \rightarrow \infty} \frac{\ln n}{n} = \lim_{n \rightarrow \infty} \frac{1}{n} = 0$ , 所以

$$\lim_{n \rightarrow \infty} \frac{4\sqrt{2nK \ln n} + \sum_{i=1}^K \left( \Delta_i + \frac{\pi^2 \Delta_i}{3} \right)}{n} = 4\sqrt{2K \cdot 0} + 0 = 0$$

故该界是次线性的. □

## 3.5 贝叶斯定理简介及测度论角度的解释

在概率论的发展过程中，人们曾经以四种不同的角度来理解概率：古典概率，试验概率，主观概率，公理化概率。

### 1. 古典概率

- 假设：随机现象是由有限个互不相交的等概率基本事件构成。
- 对应的现实问题：抛硬币，掷骰子等。
- 计算公式：假设实验结果由  $X$  中等概率基本事件构成，事件  $Y$  包含其中  $Z$  种，则  $Y$  发生的概率为  $\frac{Z}{X}$ 。

### 2. 试验概率

- 关键概念：总体，个体，样本。

### 3. 主观概率

- 描述主观信念的概率。主观概率的意义在于拓宽了概率论的应用范围，使得我们不仅可以用它来讨论客观事物，还可以用它来描述主观信念。这种思想对贝叶斯统计来说是十分重要的。因为在贝叶斯统计中，我们认为统计模型的参数是服从概率分布的，即主观概率。

### 4. 公理化概率

Kolmogorov 概率公理：

- 事件的概率是非负实数，即： $\forall E \in \mathcal{F}, \mathbb{P}(E) \in \mathbb{R}$  且  $\mathbb{P}(E) \geq 0$ ，其中  $\mathcal{F}$  为事件空间。
- 至少发生一个基本事件的概率为 1，即  $\mathbb{P}(\Omega) = 1$ 。
- ( $\sigma$  可加性) 任意可数不相交的集合序列（或称作互斥事件序列） $E_1, E_2, \dots$  满足：

$$\mathbb{P}\left(\bigcup_{i=1}^{\infty} E_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(E_i)$$

我们在机器学习和强化学习中会同时涉及到频率派和贝叶斯派的方法，它们各有千秋。由于可以给假设赋予概率，贝叶斯方法可以解决的问题更多。随着概率图模

型相关算法的不断发展, 贝叶斯统计在机器学习中的地位越来越高, 但贝叶斯统计的一大缺点是计算上更加困难. 贝叶斯统计要使用贝叶斯公式进行后验概率的推断, 计算过程比频率派的参数估计过程麻烦得多, 因此概率分布的表示, 计算和采样算法的时空复杂度就都成为了痛点. 这些痛点随着维数变得越来越高, 模型变得越来越复杂时会进一步加剧.

**定理 3.17. (针对随机事件的贝叶斯定理).** 设  $A_1, A_2, \dots$  为 (可能无穷个) 不相交的事件序列使得  $\bigcup_{i=1}^n A_i = \Omega$  且  $\mathbb{P}(A_i) > 0, \forall i$  成立. 设  $B$  为另一个事件使得  $\mathbb{P}(B) > 0$ , 则有:

$$\mathbb{P}(A_i | B) = \frac{\mathbb{P}(B | A_i)\mathbb{P}(A_i)}{\sum_{j=1}^n \mathbb{P}(B | A_j)\mathbb{P}(A_j)}$$

证明. 根据条件概率的定义,  $\mathbb{P}(A_i | B) = \frac{\mathbb{P}(A_i \cap B)}{\mathbb{P}(B)}$ , 其中  $\mathbb{P}(A_i \cap B) = \mathbb{P}(B | A_i)\mathbb{P}(A_i)$ . 因此

$$\mathbb{P}(A_i | B) = \frac{\mathbb{P}(B | A_i)\mathbb{P}(A_i)}{\mathbb{P}(B)}$$

由于  $\bigcup_{j=1}^n A_j = \Omega$ , 可知:

$$B = B \cap \Omega = B \cap \bigcup_{j=1}^n A_j = \bigcup_{j=1}^n B \cap A_j \quad (\text{交对并的分配律})$$

因此  $A_j$  彼此不相交, 所以  $B \cap A_j$  彼此也不相交. 因为  $\mathbb{P}$  是概率测度, 依据  $\sigma$  可加性, 可知:

$$\mathbb{P}(B) = \mathbb{P}\left(\bigcup_{j=1}^n B \cap A_j\right) = \sum_{j=1}^n \mathbb{P}(B \cap A_j) = \sum_{j=1}^n \mathbb{P}(B | A_j)\mathbb{P}(A_j)$$

□

证明上述定理在贝叶斯统计中存在局限性的例子: 假设观测随机变量  $X$  的值为  $x'$ , 则上述定理中  $B$  事件为  $\{X = x'\}$ . 但当  $X$  是连续型随机变量时,  $\mathbb{P}(B) =$

$\mathbb{P}(X = x') = 0$ , 违反了上述定理的前提假设. 故上述定理不适用于连续型随机变量.

**定理 3.18. (使用PMF或PDF的贝叶斯定理).** 若  $f_Y(y) \neq 0$ , 则

$$\underbrace{f_{X|Y}(x | y)}_{\text{后验概率}} = \frac{\overbrace{f_{Y|X}(y | x)}^{\text{似然函数}} \overbrace{f_X(x)}^{\text{先验概率}}}{\underbrace{f_Y(y)}_{\text{证据}}}$$

其中:  $f_X$  为随机变量  $X$  的 PMF 或 PDF;  $f_Y$  为随机变量  $Y$  的 PMF 或 PDF;  $f_{X|Y}$  为  $X$  给定  $Y$  的条件 PMF 或 PDF;  $f_{Y|X}$  为  $Y$  给定  $X$  的条件 PMF 或 PDF. 且若  $X$  为连续性随机变量, 则

$$f_Y(y) = \int_{-\infty}^{\infty} f_{Y|X}(y | x) f_X(x) dx$$

若  $X$  为离散型随机变量, 则

$$f_Y(y) = \sum_x f_{Y|X}(y | x) f_X(x)$$

因为存在分布既不是连续分布也不是离散分布, 我们需要测度论版本的贝叶斯定理. 下面先介绍一些相关概念:

**定义 3.19. (统计模型).** 设随机实验的观测结果为  $n$  个 i.i.d. 随机变量  $X_1, X_2, \dots, X_n$ , 它们取值范围是同一个可测空间  $E$ , 它们共同的分布为  $\mathbb{P}$ , 则与该随机实验对应的统计模型定义为一个序对:  $(E, \{P_\theta | \theta \in \Omega\})$ . 其中:  $E$  为样本空间, 指标集  $\{P_\theta | \theta \in \Omega\}$  为  $E$  上的概率测度族, 称作分布族.  $\Omega$  为任意集合, 称作参数空间, 若  $\exists d, \Omega \subseteq \mathbb{R}^d$ , 则模型被称作参数化模型.

贝叶斯设定下的概率空间, 随机变量的定义:

设  $(S, \mathcal{A}, \mathbb{P})$  为概率空间,  $(\mathcal{X}, \mathcal{H})$  和  $(\Omega, \mathcal{G})$  为可测空间,  $X : S \rightarrow \mathcal{X}$  和  $\Theta : S \rightarrow \Omega$  为可测映射. 若要在贝叶斯设定下定义参数化分布族, 需定义条件概率分布  $\mathbb{P}(X \in A | \Theta = \theta)$ , 其中  $A \in \mathcal{H}, \theta \in \Omega$ .

针对随机事件的条件概率定义式:

$$\mathbb{P}(A | B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}, \quad A, B \text{ 均为集合}, \mathbb{P}(B) \neq 0$$

若  $\Theta$  为连续型随机变量, 则  $\forall \theta \in \Omega, \mathbb{P}(B) = \mathbb{P}\{\Theta = \theta\} = 0$ , 违反  $\mathbb{P}(B) \neq 0$  的假设. 故无法使用针对随机事件的条件概率来定义贝叶斯设定下的参数化分布族  $\mathbb{P}(X \in \cdot | \Theta \in \cdot)$ .

注意下列性质:  $\mathbb{E}[\mathbb{I}_A] = \mathbb{P}(A)$ , 其中  $A$  为任意事件. 给定随机变量的条件期望的定义:

$$\mathbb{E}[f | \Theta] := \mathbb{E}[f | \sigma(\Theta)]$$

其中  $\sigma(\Theta) := \{\Theta^{-1}(A) | A \in \mathcal{G}\}$ ,  $f$  为可测函数. 问题: 可否通过  $\mathbb{E}[\mathbb{I}_{\{X \in A\}} | \Theta]$  定义  $\mathbb{P}(X \in A | \Theta = \theta)$ ?

不行! 因为固定  $A$  之后,  $\mathbb{E}[\mathbb{I}_{\{X \in A\}} | \Theta]$  依照条件期望定义是  $L_1(S)$  中的元素, 即  $S$  上定义的可测函数的等价类, 而我们希望定义的  $\mathbb{P}(X \in A | \Theta = \theta)$  是关于  $A, \theta$  的二元函数, 固定  $A$  后为关于  $\theta$  的一元函数, 两者在固定  $A$  后定义域不同.

解决方法: 借助 Doob-Dynkin 引理!

**定义 3.20. (Borel同构).** 称两个可测空间  $S$  和  $T$  是 Borel 同构的, 当存在双射  $f : S \rightarrow T$  使得  $f$  和  $f^{-1}$  同时可测.

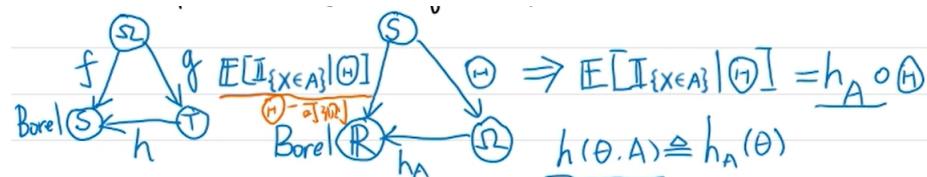
**定义 3.21. (Borel空间).** 称空间  $S$  为 Borel 空间 (或标准Borel空间,nice空间) 当它 Borel 同构于  $[0, 1]$  闭区间中的 Borel 子集.

### 注记

所有装备 Borel  $\sigma$  代数的波兰空间均是 Borel 空间.

**引理 3.22. (Doob-Dynkin).** 给定两个定义在  $\Omega$  上的可测函数  $f$  和  $g$  分别映射到  $(S, \mathcal{F})$  和  $(T, \mathcal{G})$ , 其中  $(S, \mathcal{F})$  为 Borel 空间, 则  $f$  是  $g$ -可测的当且仅当存在可测映射  $h : T \rightarrow S$  使得  $f = h \circ g$ .

由于  $\mathbb{E}[\mathbb{I}_{\{X \in A\}} | \Theta]$  给定  $A$  之后为一个从  $S$  到  $\mathbb{R}$  的可测映射,  $\Theta$  为一个  $S$  到  $\Omega$  的可测映射,  $\mathbb{R}$  是一个波兰空间. 故依据 Doob-Dynkin 引理, 存在  $\mathcal{G}$ -可测映射  $h_A(\theta) : \Omega \rightarrow \mathbb{R}$  使得  $\mathbb{E}[\mathbb{I}_{\{X \in A\}} | \Theta] = h_A \circ \Theta$ . 由于  $h_A(\theta)$  的值依赖于  $A$  和  $\theta$ , 故将其记作  $h(\theta, A)$ .



问题: 可否使用  $h(\theta, A)$  作为  $\mathbb{P}(X \in A | \Theta = \theta)$  的定义?

其中存在两个技术问题:

- (i)  $h(\theta, A)$  有多个不同版本
- (ii) 需保证  $\forall \theta \in \Omega, h(\theta, \cdot)$  是一个概率测度.

是否存在一个  $h$  的版本, 使得 (ii) 中的条件成立?

不一定存在, 若存在满足 (ii) 的版本, 则称该版本为  $X$  给定  $\Theta$  的正则条件分布.

**定义 3.23. (概率核).** 设  $(\mathcal{X}, \mathcal{F})$  和  $(\mathcal{Y}, \mathcal{G})$  为可测空间. 从  $(\mathcal{X}, \mathcal{F})$  到  $(\mathcal{Y}, \mathcal{G})$  的概率核 (也称作马尔可夫核, 随机核) 为一个二元函数  $K : \mathcal{X} \times \mathcal{G} \rightarrow [0, 1]$  使得:

- (i)  $\forall x \in \mathcal{X}, K(x, \cdot)$  是一个概率测度.
- (ii)  $\forall A \in \mathcal{G}, K(\cdot, A)$  是一个  $\mathcal{F}$ -可测映射.

有时也可将  $K(x, A)$  记作  $K(A | x)$ .

在马尔可夫过程中也会使用概率核这个概念. 其实对于强化学习来说, 策略和环境的状态转移函数都是概率核.

**定义 3.24. (正则条件分布).** 设  $(S, \mathcal{A}, \mathbb{P})$  为概率空间,  $(\mathcal{X}, \mathcal{H})$  和  $(\Omega, \mathcal{G})$  为可测空间. 设  $X : S \rightarrow \mathcal{X}$  和  $\Theta : S \rightarrow \Omega$  为可测映射, 设  $K : \Omega \times \mathcal{H} \rightarrow [0, 1]$  是  $\mathbb{P}(X \in \cdot | \Theta \in \cdot)$  的一个版本, 则  $K$  被称作正则的当  $K$  是一个概率核.

**定理 3.25.** (正则条件分布的存在性和唯一性的成立条件). 设  $X$  和  $Y$  是同一个概率空间  $(\Omega, \mathcal{F}, \mathbb{P})$  上定义的随机元素, 取值范围分别为  $\mathcal{X}$  和  $\mathcal{Y}$ , 假设  $\mathcal{X}$  为 Borel 空间, 则存在概率核  $K(\cdot | \cdot)$  使得它是  $\mathbb{P}(X | Y)$  的正则版本 (即  $K$  是  $X$  给定  $Y$  的正则条件分布). 此外, 若  $K_1$  与  $K_2$  同时满足该条件, 则对于几乎所有  $y$  有  $K_1(\cdot | y) = K_2(\cdot | y)$  成立.

**定义 3.26.** (贝叶斯设定下的参数化分布族). 设  $(S, \mathcal{A}, \mathbb{P})$  为概率空间, 设  $(\mathcal{X}, \mathcal{B})$  和  $(\Omega, \tau)$  为 Borel 空间, 设  $X : S \rightarrow \mathcal{X}$  和  $\Theta : S \rightarrow \Omega$  为可测映射. 则称  $\Theta$  为参数, 称  $\Omega$  为参数空间,  $X$  给定  $\Theta$  的正则条件分布称作  $X$  的参数化分布族, 记作:

$$\mathcal{P}_0 = \{P_\theta \mid \forall A \in \mathcal{B}, P_\theta(A) = \mathbb{P}(X \in A \mid \Theta = \theta), \theta \in \Omega\}$$

$\Theta$  的先验分布是  $\Theta$  诱导的  $(\Omega, \tau)$  上的概率测度, 记作  $\mu_\Theta$ .

$X$  的边缘分布是  $X$  诱导的  $(\mathcal{X}, \mathcal{B})$  上的概率测度, 记作  $\mu_X$ .

后验分布  $\mu_{\Theta|X} : \mathcal{X} \times \tau \rightarrow [0, 1]$  是从  $(\mathcal{X}, \mathcal{B})$  到  $(\Omega, \tau)$  的概率核. 记作  $(x, A) \mapsto \mu_{\Theta|X}(A \mid x)$ , 表示  $\Theta$  给定  $X$  的正则条件分布.

设  $\nu$  为一个  $(\mathcal{X}, \mathcal{B})$  上的测度使得  $P_\theta \ll \nu$  成立, 假设  $P_\theta$  和  $\nu$  均是  $\sigma$  有限的. 设函数  $f_{X|\Theta}(x \mid \theta) := \frac{dP_\theta}{d\nu}(x)$  固定  $x$  后得到的一元函数称作似然函数, 记作  $L(\theta)$ .

**定理 3.27.** (测度论语言描述的贝叶斯定理). 假设  $X$  有一个参数空间为  $\Omega$  的参数化分布族  $\mathcal{P}_0$ . 假设存在  $(X, \mathcal{B})$  上的一个测度  $\nu$  使得  $P_\theta \ll \nu, \forall \theta \in \Omega$ , 设  $f_{X|\Theta}(x \mid \theta)$  为  $X$  给定  $\Theta = \theta$  的条件密度 (相对于测度  $\nu$ ), 设  $\mu_\Theta$  为  $\Theta$  的先验分布. 设  $\mu_{\Theta|X}(\cdot \mid x)$  为  $\Theta$  给定  $X = x$  的条件分布. 则  $\mu_{\Theta|X} \ll \mu_\Theta, \mu_X$ -a.s., 且其 Radon-Nikodym 导数为:

$$\frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta \mid x) = \frac{f_{X|\Theta}(x \mid \theta)}{\int_\Omega f_{X|\Theta}(x, t) d\mu_\Theta(t)}$$

上式对那些使分母既不为 0 也不为  $\infty$  的  $x$  成立. 而使分母为 0 或  $\infty$  的  $x$  构成的集合的边缘概率为 0, 故对于那些  $x$  的取值, 后验分布可以任意定义.

证明. 证明主要参考

*Theory of statistics, Mark J. Schervish*

先证关于分母的命题, 设:

$$C_0 = \left\{ x \left| \int_{\Omega} f_{X|\Theta}(x | t) d\mu_{\Theta}(t) = 0 \right. \right\}, C_{\infty} = \left\{ x \left| \int_{\Omega} f_{X|\Theta}(x | t) d\mu_{\Theta}(t) = \infty \right. \right\}$$

$X$  的边缘分布  $\mu_X$  的表达式为:

$$\mu_X(A) = \int_A \int_{\Omega} f_{X|\Theta}(x | \theta) d\mu_{\Theta}(\theta) d\nu(x)$$

验证上式:  $\forall B \in \tau, A \in \mathcal{B}$  有

$$(3.28) \quad \mathbb{P}(\Theta \in B, X \in A) = \mathbb{E}[\mathbb{I}(\{\Theta \in B\} \cap \{X \in A\})] = \mathbb{E}[\mathbb{I}\{\Theta \in B\} \cdot \mathbb{I}\{X \in A\}]$$

依据条件期望定义

$$\begin{aligned} \mathbb{P}(\Theta \in B, X \in A) &= \mathbb{E}[\mathbb{I}\{\Theta \in B\} \mathbb{E}[\mathbb{I}\{X \in A | \Theta\}]] \\ &= \int_{s: \Theta(s) \in B} \mathbb{E}[\mathbb{I}\{X \in A | \Theta\}](s) d\mathbb{P}(s) \\ &= \int_B \mathbb{E}[\mathbb{I}\{X \in A | \Theta = \theta\}] d\mu_{\Theta}(\theta) \quad (\text{换元}) \\ &= \int_B P_{\theta}(A) d\mu_{\Theta}(\theta) \quad (\text{因为} \mathbb{E}[\mathbb{I}\{X \in A | \Theta = \theta\}] = P_{\theta}(A), \mu_{\Theta} \text{ a.s.}) \\ &= \int_B \int_A \frac{dP_{\theta}}{d\nu}(x) d\nu(x) d\mu_{\Theta}(\theta) \quad (\text{Radon - Nikodym定理}) \\ &= \int_B \int_A f_{X|\Theta}(x | \theta) d\nu(x) d\mu_{\Theta}(\theta) \quad (\text{代入} f_{X|\Theta} \text{的定义}) \end{aligned}$$

不失一般性假设  $f_{X|\Theta}$  是  $\mathcal{B} \otimes \tau$ -可测的, 应用 Tonelli 定理可得:

$$(3.29) \quad \mathbb{P}(\Theta \in B, X \in A) = \int_A \int_B f_{X|\Theta}(x | \theta) d\mu_{\Theta}(\theta) d\nu(x)$$

故

$$\mu_X(A) = \mathbb{P}(\Theta \in \Omega, X \in A) = \int_A \int_{\Omega} f_{X|\Theta}(x | \theta) d\mu_{\Theta}(\theta) d\nu(x)$$

依上式可知

$$(3.30) \quad \mu_X \ll \nu, \quad \text{且} \frac{d\mu_X}{d\nu} = \int_{\Omega} f_{X|\Theta}(x | \theta) d\mu_{\Theta}(\theta)$$

则

$$\begin{aligned}\mu_X(C_0) &= \int_{C_0} \int_{\Omega} f_{X|\Theta}(x | \theta) d\mu_{\Theta}(\theta) d\nu(x) = 0 \\ \mu_X(C_{\infty}) &= \int_{C_{\infty}} \int_{\Omega} f_{X|\Theta}(x | \theta) d\mu_{\Theta}(\theta) d\nu(x) \\ &= \int_{C_{\infty}} \infty d\nu(x) = \begin{cases} 0 & \text{若 } \nu(C_{\infty}) = 0 \\ \infty & \text{其他情况} \end{cases}\end{aligned}$$

由于  $\mu_X$  是概率测度,  $\mu_X(C_{\infty})$  不可能为  $\infty$ , 故  $\mu_X(C_{\infty}) = 0$ , 即关于分母的命题得证. 接下来证明分母既不为 0 也不为  $\infty$  的情况, 依据 (3.28) 式,  $\forall B \in \tau, A \in \mathcal{B}$  有:

$$\begin{aligned}\mathbb{P}(\Theta \in B, X \in A) &= \mathbb{E}[\mathbb{I}\{\Theta \in B\} \cdot \mathbb{I}\{X \in A\}] \\ &= \mathbb{E}[\mathbb{E}[\mathbb{I}\{\Theta \in B | X\}] \mathbb{I}\{X \in A\}] \quad (\text{条件期望定义}) \\ &= \int_{s:X(s) \in A} \mathbb{E}[\mathbb{I}\{\Theta \in B\} | X](s) d\mathbb{P}(s) \\ &= \int_A \mathbb{E}[\mathbb{I}\{\Theta \in B\} | X = x](x) d\mu_X(x) \quad (\text{换元}) \\ &= \int_A \mu_{\Theta|X}(B | x) d\mu_X(x) \quad (\text{代入 } \mu_{\Theta|X} \text{ 定义}) \\ &= \int_A \mu_{\Theta|X}(B | x) \frac{d\mu_X}{d\nu}(x) d\nu(x) \\ &= \int_A \mu_{\Theta|X}(B | x) \int_{\Omega} f_{X|\Theta}(x | \theta) d\mu_{\Theta}(\theta) d\nu(x) \quad (\text{代入 (3.30) 式})\end{aligned}$$

结合 (3.29) 式, 得:

$$\begin{aligned}\int_A \int_B f_{X|\Theta}(x | \theta) d\mu_{\Theta}(\theta) d\nu(x) &= \int_A \mu_{\Theta|X}(B | x) \int_{\Omega} f_{X|\Theta}(x | \theta) d\mu_{\Theta}(\theta) d\nu(x) \\ &\quad \text{对 } \forall B \in \tau, A \in \mathcal{B} \text{ 成立}\end{aligned}$$

$$\iff \int_B f_{X|\Theta}(x | \theta) d\mu_{\Theta}(\theta) = \mu_{\Theta|X}(B | x) \int_{\Omega} f_{X|\Theta}(x | \theta) d\mu_{\Theta}(\theta) \quad \text{a.s., } \forall B \in \tau$$

由于前面已经假设贝叶斯公式的分母不为 0, 因此:

$$(3.31) \quad \mu_{\Theta|X}(B | x) = \frac{\int_B f_{X|\Theta}(x | \theta) d\mu_{\Theta}(\theta)}{\int_{\Omega} f_{X|\Theta}(x | \theta) d\mu_{\Theta}(\theta)} \quad \text{a.s., } \forall B \in \tau$$

可知  $\mu_{\Theta|X} \ll \mu_\Theta$ , 且

$$\frac{d\mu_{\Theta|X}}{d\mu_\Theta}(\theta | x) = \frac{f_{X|\Theta}(x | \theta)}{\int_\Omega f_{X|\Theta}(x | \theta) d\mu_\Theta(\theta)} \quad \text{a.s.}$$

□

从定理(3.27)推定理(3.18):

证明. 设测度  $\nu'$  使得  $\mu_\Theta \ll \nu'$  且  $\mu_{\Theta|X} \ll \nu'$ . 设存在  $f_\Theta = \frac{d\mu_\Theta}{d\nu'}$ ,  $f_{\Theta|X} = \frac{d\mu_{\Theta|X}}{d\nu'}$ . 根据公式 (3.31):

$$\mu_{\Theta|X}(B | x) = \frac{\int_B f_{X|\Theta}(x | \theta) d\mu_\Theta(\theta)}{\int_\Omega f_{X|\Theta}(x | \theta) d\mu_\Theta(\theta)} \quad \text{a.s., } \forall B \in \tau$$

由 Radon-Nikodym 导数

$$\begin{aligned} \int_B f_{\Theta|X}(\theta | x) d\nu'(\theta) &= \frac{\int_B f_{X|\Theta}(x | \theta) d\mu_\Theta(\theta)}{\int_\Omega f_{X|\Theta}(x | \theta) d\mu_\Theta(\theta)} \\ &= \frac{\int_B f_{X|\Theta}(x | \theta) f_\Theta(\theta) d\nu'(\theta)}{\int_\Omega f_{X|\Theta}(x | \theta) f_\Theta(\theta) d\nu'(\theta)} \quad \text{a.s., } \forall B \in \tau \text{ (更换测度)} \\ \implies f_{\Theta|X}(\theta | x) d\nu'(\theta) &= \frac{f_{X|\Theta}(x | \theta) f_\Theta(\theta)}{\int_\Omega f_{X|\Theta}(x | \theta) f_\Theta(\theta) d\nu'(\theta)} \quad \text{a.s.} \end{aligned}$$

最后, 根据  $X$  和  $\Theta$  各自是离散型还是连续型的随机变量来指定  $\nu$  和  $\nu'$  为对应的计数测度或 Lebesgue 测度. □

### 3.6 共轭先验 (Conjugate Priors)

先验分布应该如何选择? 这个问题问不同的人会给出不同的答案. 有的统计学家认为应该采访领域专家来获得先验分布, 有的统计学家认为应该设置一个不包含主观偏见的先验分布, 即所谓的无信息先验. 这里不细讲这些不同的方法, 因为这些方法多少带有统计学家的个人偏好, 我们主要关注计算方面. 如何选择先验分布才能使得贝叶斯统计的过程更加容易计算是一个非常现实的问题, 我们将从这个问题出发, 从而引出贝叶斯统计非常重要的主题, 即共轭先验.

为什么说贝叶斯统计是难以计算的? 一个最直接的困难是我们通常难以保证后验分布具有解析解. 后验分布具有解析解, 指的是后验分布的密度函数具有解析解. 当然即使后验分布没有解析解, 在现实应用中我们也可以尝试对后验分布进行近似的推断, 或者对后验分布进行采样. 这里我们不关注对后验分布进行近似或采样, 只去关注后验分布是否有解析解的问题.

针对连续型随机变量的贝叶斯公式:  $p(\theta | x) = \frac{p(x | \theta)p(\theta)}{\int_{\Omega} p(x | \theta)p(\theta) d\theta}$  a.s. 显然分母是个积分的形式, 故后验分布不一定具有解析解. 由于分母的被积函数是模型和先验分布相乘的形式, 所以积分是否具有解析解取决于模型和先验分布的选取. 是不是只要选择了使积分具有解析解的模型和先验分布就可以了呢? 事实上, 这样也是不够的. 因为贝叶斯统计是个不断收集实验数据并更新参数的迭代过程, 之所以说是迭代过程, 是因为数据是不断收集的, 而不是一次收集的. 每当收集一定的实验数据之后, 我们都会更新参数的分布, 也就是利用贝叶斯公式来计算后验分布. 这样一轮一轮的迭代更新, 除了第一轮以外, 每轮更新的过程都会把上一轮得到的后验分布设为新一轮迭代的先验分布, 然后再去计算新的先验分布. 所以我们不仅要保证模型和先验分布代入贝叶斯公式可以产生具有解析解的后验分布, 还要保证在下一轮迭代中把上一轮产生的后验分布作为新的先验分布之后得出新的后验分布依然具有解析解. 想要实现这一点, 一个直接的想法是让先验分布和后验分布同属于一个参数化分布族. 也就是让先验分布和后验分布表达式的形式相同. 这不仅保证了每一轮迭代中后验分布都是具有解析解的, 还保证了每一轮迭代中可以使用完全相同的公式来计算后验分布, 从而方便了计算以及编程实现.

**定义 3.32. (共轭先验).** 称一个分布族为模型  $X \sim f_{X|\Theta}(x | \theta)$ ,  $\theta \in \Omega$  的共轭先验. 若只要先验分布  $f_{\Theta}(\theta)$  是从该分布族中选取的, 最终得到的后验分布  $f_{\Theta|X}(\theta | x)$  就也属于该分布族.

**例 3.33.** 所有概率分布构成的分布族是一个共轭先验.

这个例子是显然成立的, 但在现实生活中没什么用. 因为全体概率分布太大了, 它里面也包含了所有难以表示, 难以计算的分布. 而在贝叶斯统计中, 我们想要的并不是共轭先验本身, 而是方便表示和计算的共轭先验. 我们采用共轭先验的目的就是让先验分布和后验分布都方便表示和计算.

**定义 3.34. (伯努利分布).** 伯努利分布是以  $\theta$  的概率取 1, 以  $1 - \theta$  的概率取 0 的离散型随机变量的分布, 其 PMF 为:

$$p(x; \theta) = \theta^x(1 - \theta)^{1-x}, \quad \text{其中 } x \in \{0, 1\}$$

**定义 3.35. (beta 分布).** beta 分布是由定义在  $[0, 1]$  上的连续型概率分布构成的分布族, 具有两个参数  $\alpha$  和  $\beta$ . 其 PDF 为:

$$p(x; \alpha, \beta) = \frac{1}{B(\alpha, \beta)} x^{\alpha-1} (1-x)^{\beta-1}$$

其中  $B$  为 beta 函数,  $B(\alpha, \beta)$  是一个使得 PDF 积分等于 1 的归一化常数.

在实际应用中,beta 分布通常是作为先验分布来使用的. 当模型是伯努利分布或二项分布的时候,beta 分布都是这个模型的共轭先验. 既然 beta 分布是用来充当先验分布的, 它就是模型参数的分布. 但 beta 分布本身也是一个参数化分布族, 它也有自己的参数  $\alpha$  和  $\beta$ . 那么我们怎么区分模型的参数和先验分布的参数? 在贝叶斯统计中, 我们通常称先验分布或后验分布的参数为超参数. 如果我们把 beta 分布当成先验分布, 那么 beta 分布的两个参数  $\alpha$  和  $\beta$  就可以称作超参数, 而模型的参数依然称为参数, 这样就可以在术语上把模型的参数和先验分布的参数加以区分, 从而避免了混淆带来的困惑.

**定理 3.36. (beta – 伯努利共轭).** 若  $X \sim \text{Bernoulli}(\Theta)$  且  $\Theta \sim \text{Beta}(\alpha, \beta)$ , 则观测到  $X = x$  的后验分布可以选作  $\text{Beta}(x + \alpha, \beta - x + 1)$ .

证明. 设  $p(\theta)$  为  $\Theta$  的分布的 PDF,  $p(x | \theta)$  为  $X$  的 PMF, 后验分布的 PDF 为:

$$p(\theta | x) = \frac{p(x | \theta)p(\theta)}{\int_{\Omega} p(x | \theta)p(\theta) d\theta} \quad \text{a.s.}$$

由于分母不依赖于  $\theta$ , 所以

$$\begin{aligned} p(\theta | x) &\propto p(x | \theta)p(\theta) \\ &= \theta^x(1 - \theta)^{1-x} \cdot \frac{1}{B(\alpha, \beta)} \theta^{\alpha-1} (1-\theta)^{\beta-1} \\ &\propto \theta^{x+\alpha-1} (1-\theta)^{(\beta-x+1)-1} \end{aligned}$$

注意到  $\frac{1}{B(x+\alpha, \beta-x+1)}\theta^{x+\alpha-1}(1-\theta)^{(\beta-x+1)-1}$  刚好是 Beta( $x+\alpha, \beta-x+1$ ) 的 PDF, 可知后验分布  $p(\theta|x)$  与 Beta( $x+\alpha, \beta-x+1$ ) 的 PDF 之间相差一个常系数. 但因 PDF 在定义域内积分等于 1, 故后验概率分布可选择为参数为  $x+\alpha$  和  $\beta-x+1$  的 Beta 分布.  $\square$

**定义 3.37. (高斯分布).** 高斯分布(或正态分布)是一个具有两个参数由连续型分布所构成的参数化分布族, 其参数为均值  $\mu$  和方差  $\sigma^2$ . 高斯分布的 PDF 为

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

**定义 3.38. (多元高斯分布).** 称一个随机向量服从  $k$  元高斯分布, 当它的  $k$  个分量的任意线性组合均服从高斯分布.

**定理 3.39. (高斯 - 高斯共轭).** 若  $X \sim \mathcal{N}(\mu_s, \sigma_s^2), \mu_s \sim \mathcal{N}(\mu_p, \sigma_p^2)$ , 其中  $\sigma_s^2$  已知. 则观测到  $X=x$  的后验分布可以选作:

$$\mathcal{N}\left(\frac{\sigma_p^2}{\sigma_s^2 + \sigma_p^2}x + \frac{\sigma_s^2}{\sigma_s^2 + \sigma_p^2}\mu_p, \left(\frac{1}{\sigma_p^2} + \frac{1}{\sigma_s^2}\right)^{-1}\right)$$

**引理 3.40.** 若  $(X_1, X_2)$  服从二元高斯分布, 则给定  $X_2 = x_2$  时  $X_1$  的分布可选作均值为:

$$\mathbb{E}[X_1] + \frac{\text{Cov}(X_1, X_2)}{\text{Var}[X_2]}(x_2 - \mathbb{E}[X_2])$$

方差为:

$$\text{Var}[X_1] - \frac{\text{Cov}^2(X_1, X_2)}{\text{Var}[X_2]}$$

的高斯分布.

定理(3.39)的证明:

证明.  $X$  和  $\mu_s$  可分解为

$$X = \mu_s + \sigma_s \varepsilon, \quad \text{其中 } \varepsilon \sim \mathcal{N}(0, 1)$$

$$\mu_s = \mu_p + \sigma_p \delta, \quad \text{其中 } \delta \sim \mathcal{N}(0, 1)$$

由此可得  $\mathbb{E}[X] = \mathbb{E}[\mu_s] + \mathbb{E}[\sigma_s \varepsilon] = \mathbb{E}[\mu_p] + \mathbb{E}[\sigma_p \varepsilon] = \mu_p$

$$\begin{aligned} \text{Var}[X] &= \text{Var}[\mu_s + \sigma_s \varepsilon] = \text{Var}[\mu_s] + \text{Var}[\sigma_s \varepsilon] \\ &= \text{Var}[\mu_p] + \text{Var}[\sigma_p \delta] + \sigma_s^2 = \sigma_p^2 + \sigma_s^2 \end{aligned}$$

$$\begin{aligned} \text{Cov}(X, \mu_s) &= \mathbb{E}[(X - \mathbb{E}[X])(\mu_s - \mathbb{E}[\mu_s])] \\ &= \mathbb{E}[(X - \mu_p)](\mu_s - \mu_p) \\ &= \mathbb{E}[(\mu_s + \sigma_s \varepsilon - \mu_p)(\mu_s - \mu_p)] \\ &= \mathbb{E}[(\mu_s - \mu_p)(\mu_s - \mu_p) + \sigma_s \varepsilon (\mu_s - \mu_p)] \\ &= \mathbb{E}[(\mu_s - \mu_p)^2] + \mathbb{E}[\sigma_s \varepsilon (\mu_s - \mu_p)] \\ &= \mathbb{E}[(\mu_s - \mu_p)^2] + \mathbb{E}[\sigma_s \varepsilon] \cdot \mathbb{E}[\mu_s - \mu_p] \\ &= \text{Var}[\mu_s] = \sigma_p^2 \end{aligned}$$

接下来验证  $(\mu_s, X)$  服从二元高斯分布, 设  $a, b \in \mathbb{R}$  为任意实数, 则有

$$\begin{aligned} aX + b\mu_s &= a\mu_s + a\sigma_s \varepsilon + b\mu_s \\ &= (a+b)\mu_s + a\sigma_s \varepsilon \\ &= (a+b)\mu_p + (a+b)\sigma_p \delta + a\sigma_s \varepsilon \end{aligned}$$

由于  $\delta$  和  $\varepsilon$  服从正态分布且相互独立, 其余量均为常数, 可知整个式子服从正态分布, 即  $aX + b\mu_s$  服从正态分布. 根据多元高斯分布定义, 可知  $(\mu_s, X)$  服从二元高斯分布. 根据引理(3.40)可知  $\mu_s$  给定  $X = x$  的分布可选作高斯分布. 该分布均值为

$$\begin{aligned} \mathbb{E}[\mu_s] + \frac{\text{Cov}(\mu_s, X)}{\text{Var}[X]}(x - \mathbb{E}[X]) &= \mu_p + \frac{\sigma_p^2}{\sigma_s^2 + \sigma_p^2}(x - \mu_p) \\ &= \mu_p + \frac{\sigma_p^2}{\sigma_s^2 + \sigma_p^2}x - \frac{\sigma_p^2 \mu_p}{\sigma_s^2 + \sigma_p^2} \\ &= \frac{\sigma_s^2 \mu_p + \sigma_p^2 \mu_p}{\sigma_s^2 + \sigma_p^2} + \frac{\sigma_p^2}{\sigma_s^2 + \sigma_p^2}x - \frac{\sigma_p^2 \mu_p}{\sigma_s^2 + \sigma_p^2} \\ &= \frac{\sigma_s^2}{\sigma_s^2 + \sigma_p^2} \mu_p + \frac{\sigma_p^2}{\sigma_s^2 + \sigma_p^2} x \end{aligned}$$

方差为:

$$\begin{aligned}\text{Var}[\mu_s] &= \frac{\text{Cov}(\mu_s, X)}{\text{Var}[X]} \\ &= \sigma_p^2 - \frac{\sigma_p^4}{\sigma_p^2 + \sigma_s^2} = \frac{\sigma_p^4 + \sigma_p^2 \sigma_s^2 - \sigma_p^4}{\sigma_p^2 + \sigma_s^2} \\ &= \frac{1}{\frac{1}{\sigma_p^2} + \frac{1}{\sigma_s^2}} = \left( \frac{1}{\sigma_p^2} + \frac{1}{\sigma_s^2} \right)^{-1}\end{aligned}$$

□

本节中我们简单介绍了共轭先验的定义和作用, 并且介绍了 3 个共轭先验的例子, 其中后两个是比较常用的例子. 当然对于研究机器学习来说, 只知道这 3 个例子是远远不够的. 此外, 对于后两个例子还有许多推广, 比如可以把 beta-伯努利共轭中的伯努利分布替换成二项分布. beta 分布本身可以看作 Dirichlet 分布的一个特例, 而二项分布可以看作多项分布的一个特例, 因此如果进一步推广 beta-二项共轭的话, 还可以得出 Dirichlet-多项共轭. 对于刚提到的高斯-高斯共轭也是可以推广的, 刚讲的例子中它所给定的只是单个  $X$  的观测值, 还可以进一步推广到多个观测值. 感兴趣的读者可以找资料进一步了解, 在维基百科上有一个共轭先验的表格, 上面包含了各种常见的共轭先验, 但没有必要死记硬背, 只要了解常见共轭先验, 使用的时候现去查表即可.

### 3.7 贝叶斯 Bandit(Bayesian Bandits)

Bandit 问题的根本性难点在于环境是未知的, 因为假如环境已知, 那么寻找最优策略就是求解一个优化问题. 但如果环境是未知的, 我们就不能求出环境的最优策略. 因为一个环境的最优策略对于另一个环境不一定是最优的, 甚至对另一个环境来说是最差的. 所以通常在 Bandit 问题中我们会通过与环境交互来收集信息, 借助这些信息来决策. 然而由于这些信息是随机试验得到的样本, 用这些信息我们依然无法确定真实的环境是什么, 我们需要权衡很多因素来决策. 这就涉及了统计决策论这门学科, 它是博弈论和统计学相互借鉴而产生的一门学科, 它和强化学习的理论有很多交集. 这里我们只去介绍统计决策论中非常有限的几个概念, 感兴趣的读者可以阅读其他书籍了解, 比如

*Statistical Decision Theory and Bayesian Analysis, James O. Berger*

决策论中的概念	Bandit 领域中类似的概念
自然的状态 (state of nature)	环境
决策法则 (decision rule)	策略
风险	遗憾

问题: 选择策略时可以权衡的因素有哪些?

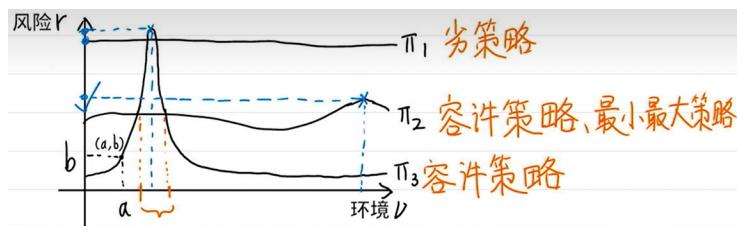
1. 容许性
2. 最小最大原则
3. 贝叶斯原则
4.  $\Gamma$ -最小最大原则

在统计决策论中, 策略分为两类: 容许策略和劣策略. 对于某个策略  $A$  来说, 若存在另一个策略  $B$  使得在所有环境下,  $A$  的风险都比  $B$  的风险大, 则策略  $A$  为劣策略. 作为决策者, 我们是完全没有理由选择劣策略的. 因为劣策略在任何环境下都不是最优的. 当一个策略不是劣策略时, 它就是一个容许策略 (admissible policy). 与劣策略不同, 容许策略至少在某些环境下是优于其他所有策略的. 容许策略通常不止一个, 因此除了容许性还需要考虑其他因素来选出最合适的策略. 然而容许策略之间很难明确的比较谁好谁坏, 就像如果一个学生每科成绩都比另一个学生高, 那么我们可以明确的说他的学习成绩比另一个学生好. 可是如果学生只是偏不同的学科, 没有哪个学生每个学科都碾压另一个学生, 我们在比较两个学生谁的成绩更好时就会左右为难. 那么我们如何在不同的容许策略之间做选择呢?

一个比较常见的想法是我们不希望策略太依赖于运气因素, 我们希望策略可以稳健 (robustness) 一些. 我们可以把策略在最坏情况下的风险当成稳健性的指标, 也就是当我选择了对策略最不利的那个环境下, 如果策略表现的依然不错, 那么我们就认为该策略是稳健的, 这种定义是符合我们的常识的. 假如一个人去投资, 一个稳健的投资通常是比较保守的投资, 换言之在最坏的情况下也不会赔的太多的投资. 我们把稳健的策略叫做最小最大策略 (minimax policy), 最小最大策略适用于我们对环境一无所知的情况, 它蕴含着做最坏的打算的思想, 对真实环境的预期是较为悲观的. 当我们选择最小最大策略时, 通常说我们采用了最小最大原则 (minimax principle) 来选取策略.

但很多情况下最小最大策略并不是最好的选择, 比如下图中的  $\pi_3$  在绝大多数情况下都优于  $\pi_2$ , 仅仅因为少数情况下不如  $\pi_2$  而没有选它, 这在很多时候并不合理, 假如我们真实环境出现在那个小区间的概率很低的话, 我们没有任何理由去选

择  $\pi_2$  而不是  $\pi_3$ . 另一方面, 我们也不能肯定的说  $\pi_3$  一定比  $\pi_2$  更好, 因为真实环境如果出现在  $\pi_3$  不如  $\pi_2$  的那个区间的概率非常大的话, 我们依然应该选择  $\pi_2$  而不是  $\pi_3$ . 总而言之应该把环境的分布考虑进去, 因为只有得知环境的分布, 我们才能知道真实分布落在这个区间的概率, 才能在  $\pi_2$  和  $\pi_3$  之间做出更好的选择. 那么我们该如何把环境的分布考虑进来呢? 这就引出了本节重要的主题: 贝叶斯原则.



图中  $\pi_3$  上的点  $(a, b)$  的含义:  $\pi_3$  与  $a$  环境交互的风险为  $b$ . 寻找最小最大策略, 只需比较不同曲线最高点的大小即可.

假设我们对环境一无所知, 我们针对环境最坏的情况做准备. 但在现实很多情况下, 我们并非对环境一无所知, 而是知晓了先验分布. 这样采取最小最大策略就是不恰当的. 我们希望利用掌握的先验分布来帮助我们做决策, 我们依然给不同策略排序, 只不过排序标准不再是坏情况下的风险, 而是改为环境服从先验分布时策略风险的期望, 我们通常称这个期望为贝叶斯风险, 把最大化贝叶斯风险的策略叫做贝叶斯最优策略, 注意这里的贝叶斯最优策略是 Bandit 语境下的习惯叫法, 在统计决策论语境中习惯把这个概念叫做贝叶斯决策法则. 当我们选取贝叶斯最优策略时, 就可以说我们采取贝叶斯原则来选取策略.

贝叶斯原则在我们可以精准的给出先验分布时是无可挑剔的, 但是现实往往介于决策者对于环境一无所知和决策者能精确提供环境先验分布这两者之间. 具体来说, 决策者可能只掌握先验分布的一部分信息, 这些信息可以给先验分布划定一个范围, 也就是一个分布族, 但却无法确定具体分布是分布族中哪一个分布. 此时最小最大原则和贝叶斯原则就都不在适用, 因为它们要么没有利用已有的先验信息, 要么需要我们提供明确的先验分布. 因此为了做出合理的决策, 我们需要引入新的概念, 叫做  $\Gamma$ -最小最大策略. 该策略是在可能的先验分布族中对决策者最不利的那个分布下风险最低的策略. 使用  $\Gamma$ -最小最大策略的目的是在于利用先验知识的前提下, 在面对不确定性时尽量做到稳健.  $\Gamma$ -最小最大策略的理念对于最大熵强化学习理论非常重要. 因为最大熵强化学习的理论中十分重要的一环是试图

合理化在决策问题中应用最大熵原则的做法. 而合理化最大熵原则的方法之一就是证明使用最大熵原则在一些决策的场景下是等价于  $\Gamma$ -最小最大原则的, 具体内容会在后面讲解最大熵强化学习时展开.

**定义 3.41. ( $K$ 臂贝叶斯 Bandit 环境).**  $K$  臂贝叶斯 Bandit 环境定义为四元组  $(\mathcal{E}, \mathcal{G}, Q, P)$ , 其中  $(\mathcal{E}, \mathcal{G})$  为一个可测空间, 先验分布  $Q$  为一个定义在  $(\mathcal{E}, \mathcal{G})$  上的概率测度,  $P$  为一个从  $\mathcal{E} \times [K]$  到  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  的概率核. 用  $P_{\nu_i}$  表示在 Bandit 环境  $\nu$  下第  $i$  个动作所对应的奖励分布. 贝叶斯 Bandit 环境与策略  $\pi = \{\pi_t\}_{t=1}^n$  交互产生的一系列随机变量  $\nu \in \mathcal{E}, \{A_t\}_{t=1}^n$  以及  $\{X_t\}_{t=1}^n$ , 其中  $A_t \in [K]$  且  $X_t \in \mathbb{R}$ , 满足下列三个条件:

1.  $\mathbb{P}(\nu \in \cdot) = Q(\cdot)$
2. 动作  $A_t$  给定  $\nu, A_1, X_1, \dots, A_{t-1}, X_{t-1}$  的条件分布为  $\pi_t(\cdot | A_1, X_1, \dots, A_{t-1}, X_{t-1})$  a.s.
3. 奖励  $X_t$  给定  $\nu, A_1, X_1, \dots, A_t$  的条件分布为  $P_{\nu A_t}$  a.s.

为了和贝叶斯 Bandit 区分, 我们把之前几节提到的 Bandit 称为频率派 Bandit, 前几节定义的遗憾为频率派遗憾. 贝叶斯 Bandit 与频率派 Bandit 的区别在于存在一个先验分布, 一开始的时候会先从先验分布中采样出一个频率派 Bandit 环境来和策略进行交互. 贝叶斯 Bandit 和环境交互所产生的随机变量的联合分布以及携带这些随机变量的概率空间的存在性也是由 Ionescu Tulcea 定理保证.

**例 3.42. ( $K$  臂贝叶斯伯努利 Bandit).** 一个  $K$  臂贝叶斯伯努利 Bandit 环境可通过设  $\mathcal{E} = [0, 1]^K$ ,  $\mathcal{G} = \mathcal{B}(\mathcal{E})$ ,  $P_{\nu_i} = \text{Bernoulli}(\nu_i)$  来定义. 该环境的一个自然的先验分布为 Beta( $\alpha, \beta$ ) 分布的乘积:

$$Q(A) = \int_A \prod_{i=1}^K q_i(\nu_i) d\nu$$

其中,  $q_i(\nu) = \nu^{\alpha-1}(1-\nu)^{\beta-1} \frac{\Gamma(a+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ .

回顾  $K$  臂 Bandit 环境  $\nu$  与策略  $\pi$  交互  $n$  轮的频率派遗憾为:

$$R_n(\pi, \nu) = n\mu^* - \mathbb{E} \left[ \sum_{t=1}^n X_t \right]$$

其中  $\mu^* = \max_{i \in [K]} \mu_i$ ,  $\mu_i$  为  $P_{\nu_i}$  的均值.

**定义 3.43. (贝叶斯遗憾).** 给定一个贝叶斯 Bandit 环境  $(\mathcal{E}, \mathcal{G}, Q, P)$  和一个策略  $\pi$ , 贝叶斯遗憾定义为:

$$\text{BR}_n(\pi, Q) := \int_{\mathcal{E}} R_n(\pi, \nu) dQ(\nu)$$

**定义 3.44. (贝叶斯最优遗憾).** 给定一个贝叶斯 Bandit 环境  $(\mathcal{E}, \mathcal{G}, Q, P)$ , 则策略  $\pi$  在第  $n$  轮博弈时的贝叶斯最优遗憾定义为:

$$\text{BR}_n^* := \inf_{\pi} \text{BR}_n(\pi, Q)$$

**定义 3.45. (贝叶斯最优策略).** 贝叶斯最优策略定义为最小化贝叶斯遗憾的策略

$$\pi^* \in \arg \min_{\pi} \text{BR}_n(\pi, Q)$$

### 注记

- (1) 最小化遗憾的策略和最大化平均累计奖励的策略是相同的, 这点和频率派遗憾相同.
- (2) 贝叶斯最优策略不一定存在, 但贝叶斯最优遗憾一定存在, 因为非负数的下确界一定存在, 而  $\arg \min_{\pi} \text{BR}_n(\pi, Q)$  可能为空集. 尽管贝叶斯最优策略不一定存在, 但现实生活中选取足够好的次优策略就可以了. 次优策略的贝叶斯遗憾可以任意逼近最优策略的贝叶斯遗憾.

**性质 3.46.** 对于任意  $\varepsilon > 0$ , 存在策略  $\pi$  使得  $\text{BR}_n(\pi, Q) \leq \text{BR}_n^*(Q) + \varepsilon$ .

证明. 反证法. 假设存在  $\varepsilon > 0$ , 使得对于任意策略  $\pi$  都有  $\text{BR}_n(\pi, Q) > \text{BR}_n^*(Q) + \varepsilon$ . 设常数  $\varepsilon_0 > 0$  为一个使  $\text{BR}_n(\pi, Q) > \text{BR}_n^*(Q) + \varepsilon_0, \forall \pi$  成立的数, 可知  $\text{BR}_n^*(Q) + \varepsilon_0$  为  $\text{BR}_n(\pi, Q)$  的下界. 由于  $\varepsilon_0 > 0$ , 可知  $\text{BR}_n^*(Q) + \varepsilon_0 > \text{BR}_n^*(Q)$ . 故  $\text{BR}_n^*(Q)$  不是  $\text{BR}_n(\pi, Q)$  的最大下界, 与  $\text{BR}_n^*(Q)$  的定义相矛盾. 故原命题成立.  $\square$

先验分布反应了我们的主观信念, 但主观信念有可能是不真实的. 比如我们对环境是一无所知的, 甚至对环境持有错误信息, 因此随着和环境交互过程中我们收集到了更多数据, 我们就需要去更新我们的信念. 换言之, 我们就需要去算出后验分布, 把后验分布当作新的主观信念.

**定理 3.47.** (*K*臂贝叶斯Bandit的后验分布). 设  $(\mathcal{E}, \mathcal{G}, Q, P)$  为一个 *K* 臂贝叶斯 Bandit 环境. 假设  $(\mathcal{E}, \mathcal{G})$  为一个 Borel 空间. 假设存在  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$  上的  $\sigma$ -有限测度  $\lambda$  使得  $P_{\nu i} \ll \lambda, \forall i \in [K]$  和  $\nu \in \mathcal{E}$ . 则  $t$  轮交互之后的后验分布为:

$$Q(B \mid a_1, x_1, \dots, a_t, x_t) = \frac{\int_B \prod_{s=1}^t p_{\nu a_s}(x_s) dQ(\nu)}{\int_{\mathcal{E}} \prod_{s=1}^t p_{\nu a_s}(x_s) dQ(\nu)}$$

其中  $p_{\nu a}$  为奖励分布  $P_{\nu a}$  相对于  $\lambda$  的密度函数. 当分母取 0 或正无穷时后验分布可任取, 这两种情况发生的概率为 0.

**引理 3.48.** 在与定理(3.47)相同的假设下, 令  $\rho$  为计数测度. 当给定环境  $\nu$  和策略  $\pi$  时, 历史  $A_1, X_1, \dots, A_n, X_n$  的分布律  $\mathbb{P}_{\nu \pi}$  相对于  $(\rho \times \lambda)^n$  的 Radon-Nikodym 导数为:

$$(3.49) \quad p_{\nu \pi}(a_1, x_1, \dots, a_n, x_n) = \prod_{t=1}^n \pi_t(a_t \mid a_1, x_1, \dots, a_{t-1}, x_{t-1}) p_{\nu a_t}(x_t)$$

定理(3.47)的证明:

证明. 设  $\mathbb{P}_{\nu \pi}$  为历史  $A_1, X_1, \dots, A_n, X_n$  给定环境  $\nu$  和策略  $\pi$  的分布. 依据引理(3.48)可知  $\mathbb{P}_{\nu \pi} \ll (\theta \times \lambda)^n$ . 设  $\rho_{\nu \pi} = \frac{d\mathbb{P}_{\nu \pi}}{d(\theta \times \lambda)^n}$ . 根据贝叶斯定理:

$$\frac{dQ(\cdot \mid a_1, x_1, \dots, a_t, x_t)}{dQ} = \frac{p_{\nu \pi}(a_1, x_1, \dots, a_t, x_t)}{\int_{\mathcal{E}} p_{\nu \pi}(a_1, x_1, \dots, a_t, x_t) dQ(\nu)}$$

$$\begin{aligned}
& Q(B \mid a_1, x_1, \dots, a_t, x_t) \\
&= \frac{\int_B p_{\nu\pi}(a_1, x_1, \dots, a_t, x_t) dQ(\nu)}{\int_{\mathcal{E}} p_{\nu\pi}(a_1, x_1, \dots, a_t, x_t) dQ(\nu)} \quad (\text{Radon - Nikodym}) \\
&= \frac{\int_B \prod_{s=1}^t \pi_s(a_s \mid a_1, x_1, \dots, a_{s-1}, x_{s-1}) p_{\nu a_s}(x_s) dQ(\nu)}{\int_{\mathcal{E}} \prod_{s=1}^t \pi_s(a_s \mid a_1, x_1, \dots, a_{s-1}, x_{s-1}) p_{\nu a_s}(x_s) dQ(\nu)} \quad (\text{代入(3.49)式}) \\
&= \frac{\int_B (\prod_{s=1}^t \pi_s(a_s \mid a_1, x_1, \dots, a_{s-1}, x_{s-1})) (\prod_{s=1}^t p_{\nu a_s}(x_s)) dQ(\nu)}{\int_{\mathcal{E}} (\prod_{s=1}^t \pi_s(a_s \mid a_1, x_1, \dots, a_{s-1}, x_{s-1})) (\prod_{s=1}^t p_{\nu a_s}(x_s)) dQ(\nu)} \\
&= \frac{\prod_{s=1}^t \pi_s(a_s \mid a_1, x_1, \dots, a_{s-1}, x_{s-1}) \int_B \prod_{s=1}^t p_{\nu a_s}(x_s) dQ(\nu)}{\prod_{s=1}^t \pi_s(a_s \mid a_1, x_1, \dots, a_{s-1}, x_{s-1}) \int_{\mathcal{E}} \prod_{s=1}^t p_{\nu a_s}(x_s) dQ(\nu)} \\
&= \frac{\int_B \prod_{s=1}^t p_{\nu a_s}(x_s) dQ(\nu)}{\int_{\mathcal{E}} \prod_{s=1}^t p_{\nu a_s}(x_s) dQ(\nu)}
\end{aligned}$$

□

说明了贝叶斯 Bandit 的后验分布不依赖于策略.

**例 3.50. (贝叶斯伯努利Bandit的后验分布).** 例(3.42)中的  $K$  臂贝叶斯伯努利 Bandit 的后验分布相对于 Lebesgue 测度的密度函数为:

$$q(\nu \mid h_t) \propto \prod_{i=1}^K \nu^{\alpha+s_i(h_t)-1} (1-\nu)^{\beta+t_i(h_t)-s_i(h_t)-1}$$

其中  $h_t = a_1, x_1, \dots, a_t, x_t$ ;  $s_i(h_t) = \sum_{u=1}^t x_u \mathbb{I}\{a_u = i\}$ ;  $t_i(h_t) = \sum_{u=1}^t \mathbb{I}\{a_u = i\}$ . 上式反映出后验分布的密度函数与先验分布具有相同的形式, 即例(3.42)中定义的先验分布是共轭先验.

证明. 设  $\lambda$  为 Lebesgue 测度. 根据定理(3.47),  $K$  臂贝叶斯 Bandit 的后验分布为:

$$\begin{aligned}
Q(B \mid h_t) &= \frac{\int_B \prod_{s=1}^t p_{\nu a_s}(x_s) dQ(\nu)}{\int_{\mathcal{E}} \prod_{s=1}^t p_{\nu a_s}(x_s) dQ(\nu)} \\
&= \frac{\int_B (\prod_{s=1}^t p_{\nu a_s}(x_s)) \frac{dQ(\nu)}{d\lambda} d\lambda(\nu)}{\int_{\mathcal{E}} \prod_{s=1}^t p_{\nu a_s}(x_s) dQ(\nu)} \quad (\text{更换测度})
\end{aligned}$$

根据 Radon-Nikodym 定理, 可知:

$$(3.51) \quad q(\nu | h_t) = \frac{(\prod_{s=1}^t p_{\nu a_s}(x_s)) \frac{dQ(\nu)}{d\lambda}}{\int_{\mathcal{E}} \prod_{s=1}^t p_{\nu a_s}(x_s) dQ(\nu)} \propto \left( \prod_{s=1}^t p_{\nu a_s}(x_s) \right) \frac{dQ(\nu)}{d\lambda}$$

因为在例(3.42)中先验分布定义为:

$$Q(A) = \int_A \prod_{i=1}^K q_i(\nu_i) d\nu$$

其中  $q_i(\nu) = \nu^{\alpha-1}(1-\nu)^{\beta-1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}$ . 根据 Radon-Nikodym 定理:

$$(3.52) \quad \frac{dQ(\nu)}{d\lambda} = \prod_{i=1}^K q_i(\nu_i) = \prod_{i=1}^K \nu_i^{\alpha-1}(1-\nu_i)^{\beta-1} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \propto \prod_{i=1}^K \nu_i^{\alpha-1}(1-\nu_i)^{\beta-1}$$

此外, 由于  $P_{\nu i} = \text{Bernoulli}(\nu_i)$ , 可知:

$$(3.53) \quad p_{\nu a_s}(x_s) = \nu_{a_s}^{x_s} (1 - \nu_{a_s})^{1-x_s} = \prod_{i=1}^K \nu_i^{\mathbb{I}\{a_s=i\}x_s} (1 - \nu_i)^{\mathbb{I}\{a_s=i\}(1-x_s)}$$

将(3.52),(3.53)式代入(3.51)式, 得:

$$\begin{aligned} q(\nu | h_t) &\propto \left( \prod_{s=1}^t \prod_{i=1}^K \nu_i^{\mathbb{I}\{a_s=i\}x_s} (1 - \nu_i)^{\mathbb{I}\{a_s=i\}(1-x_s)} \right) \left( \prod_{i=1}^K \nu_i^{\alpha-1}(1-\nu_i)^{\beta-1} \right) \\ &= \left( \prod_{i=1}^K \prod_{s=1}^t \nu_i^{\mathbb{I}\{a_s=i\}x_s} (1 - \nu_i)^{\mathbb{I}\{a_s=i\}(1-x_s)} \right) \left( \prod_{i=1}^K \nu_i^{\alpha-1}(1-\nu_i)^{\beta-1} \right) \\ &= \left( \prod_{i=1}^K \left( \prod_{s=1}^t \nu_i^{\mathbb{I}\{a_s=i\}x_s} \right) \left( \prod_{s=1}^t (1 - \nu_i)^{\mathbb{I}\{a_s=i\}(1-x_s)} \right) \right) \left( \prod_{i=1}^K \nu_i^{\alpha-1}(1-\nu_i)^{\beta-1} \right) \\ &= \left( \prod_{i=1}^K \nu_i^{\sum_{s=1}^t \mathbb{I}\{a_s=i\}x_s} (1 - \nu_i)^{\sum_{s=1}^t \mathbb{I}\{a_s=i\}(1-x_s)} \right) \left( \prod_{i=1}^K \nu_i^{\alpha-1}(1-\nu_i)^{\beta-1} \right) \\ &= \prod_{i=1}^K \nu_i^{\sum_{s=1}^t \mathbb{I}\{a_s=i\}x_s} (1 - \nu_i)^{\sum_{s=1}^t \mathbb{I}\{a_s=i\}(1-x_s)} \nu_i^{\alpha-1}(1-\nu_i)^{\beta-1} \\ &= \prod_{i=1}^K \nu_i^{\alpha-1 + \sum_{s=1}^t \mathbb{I}\{a_s=i\}x_s} (1 - \nu_i)^{\beta-1 + \sum_{s=1}^t \mathbb{I}\{a_s=i\}(1-x_s)} \\ &= \prod_{i=1}^K \nu_i^{\alpha-1 + \sum_{s=1}^t \mathbb{I}\{a_s=i\}x_s} (1 - \nu_i)^{\beta-1 + \sum_{s=1}^t \mathbb{I}\{a_s=i\} - \sum_{s=1}^t \mathbb{I}\{a_s=i\}x_s} \end{aligned}$$

设  $s_i(h_t) = \sum_{u=1}^t x_u \mathbb{I}\{a_u = i\}$  以及  $t_i(h_t) = \sum_{u=1}^t \mathbb{I}\{a_u = i\}$ , 则有:

$$q(\nu \mid h_t) \propto \prod_{i=1}^K \nu^{\alpha+s_i(h_t)-1} (1-\nu)^{\beta+t_i(h_t)-s_i(h_t)-1}$$

□

### 3.8 汤普森采样 (Thompson Sampling)

目前已有的求解贝叶斯最优策略的方法往往都存在计算复杂度过高的问题, 而汤普森采样由于计算复杂度非常低, 所以常被用于替代那些算法. 尽管汤普森采样策略并不一定是贝叶斯最优策略, 但它通常也足够好, 即贝叶斯遗憾是次线性增长的, 这个理论保证将在下一节推导. 总的来说, 汤普森采样可以解决两类问题:

- (1) 给定已知的贝叶斯 Bandit, 最小化贝叶斯遗憾.
- (2) 给定未知的频率派 Bandit, 最小化频率派遗憾.

第二类问题比第一类问题更具有现实意义, 但本质上可以转化为第一类问题. 因为想要解决第二类问题, 我们可以人为指定一个先验分布, 该先验分布刻画的是关于未知的频率派 Bandit 的先验知识. 有了这个先验分布之后, 我们就可以把装备了这个先验分布的贝叶斯 Bandit 当作第一类问题的输入, 从而得到一个使得贝叶斯遗憾次线性增长的汤普森采样策略, 而巧合的是该策略的频率派遗憾通常也是次线性增长的, 这样我们就可以利用该策略来解决第二类问题, 也就是给定未知的频率派 Bandit, 最小化频率派遗憾的问题. 这也是为什么很多文献中把汤普森采样算法和 UCB 之类的算法放在一起比较, 因为它们都是可以解决第二类问题的算法.

在历史中, 汤普森采样提出的非常早, 它早在 1933 年就由汤普森提出. 与之相比, UCB 算法主要思想 1985 年才在 Lai 和 Robbins 合作的论文中提出. 不过最初的汤普森采样算法和今天的版本并不完全相同, 最初的版本针对的只是一个特定的问题, 也就是临床试验问题, 这个问题等同于只有两个动作的 Bandit 问题. 我们现在学习的汤普森采样是后来被推广到多臂 Bandit 问题上的版本. 尽管汤普森采样出现的比 UCB 更早, 但它的性能并不比 UCB 差. 根据下面这篇文章的研究

*An Empirical Evaluation of Thompson Sampling*

汤普森采样在该文中测试的几乎所有环境都比 UCB 具有更好的性能, 并且在其中一部分环境下显著优于 UCB. 汤普森采样的思想还可以推广到强化学习的问题上, 感兴趣的读者可以了解 Posterior Sampling for Reinforcement Learning 算法.

下面介绍汤普森采样的大概思想, 其实在日常生活中我们做的很多决策的思想都和汤普森采样是类似的. 举个之前提到过的例子, 假如你搬家到了一个陌生的地点, 而在那周围有很多家饭馆, 那么你该如何通过尝试来快速找到口味最好的那家饭馆呢? 首先你会对饭馆有些第一印象, 第一印象可能来自饭馆的名称, 地理位置, 店铺的装修风格等等. 然后你会根据你的第一印象去大致猜测每家饭馆的口味如何, 并且根据猜测去挑选口味最好的饭馆, 然后去试吃. 之后再根据试吃的结果来调整对这些饭馆的印象, 然后再去猜测哪些最好吃, 再去试吃, 再根据新的试吃的结果去调整对这些饭馆的印象. 如此循环往复, 就能快速对每家饭馆的口味有个准确了解. 这个过程就和汤普森采样的思想类似. 尽管这个例子很贴近日常生活, 但并不是很具体, 因为饭菜的口味如何以及自己对这些口味的主观印象等概念都是没有严格定义的, 接下来讲个更加具体的例子.

**例 3.54.** 在网站上刊登某件商品的广告, 共 10 个方案可选, 每个方案的点击率各不相同且均未知. 需要用算法通过尝试来找到点击率最高的方案并最大化累计点击次数.

当然现实的情况会更加复杂, 比如网站需要多个打广告的商品以及多个广告位. 此外现实中的网站还会记录每个用户的偏好, 以便针对性的推送广告. 这里我们并不考虑这些复杂因素. 显然这个问题可以归类到之前的第二类问题: 给定未知的频率派 Bandit, 最小化频率派遗憾.

决策者: 服务器

动作空间: 10 个动作, 每个动作对应于把 10 个方案中的其中一个显示在网页上.

奖励: 若用户未点击则为 0, 若用户点击则为 1, 服从伯努利分布.

汤普森采样解决上述问题的方案:

1. 为算法设定每个方案的先验分布.(等同于设定了频率派 Bandit 环境的先验分布)
2. 从每个广告方案的先验分布中分别采一个样本.(等同于从环境的先验分布中采一个样本)

3. 根据上一步的样本选出点击率最高的广告方案并呈现给访客.
4. 根据用户是否点击广告来更新后验分布, 然后开始下一轮的采样, 执行动作, 更新参数的循环.

而具体计算后验分布的方法取决于我们选择的先验分布. 对于当前这个例子, 我们把先验分布设置成 beta 分布, 因此只需要每个广告显示的次数和被点击次数, 就可以轻易算出后验分布的参数  $\alpha$  和  $\beta$ . 其中  $\alpha$  设置成被点击的次数加 1, 而  $\beta$  设置成展示了广告却未被点击的次数加 1.(这样设置的前提假设是每个动作的先验分布都设成了 Beta(1,1)). 这些细节到后面会展开, 总的来说通过不断循环上面的流程, 算法就可以找出效果最好的广告, 同时最大化累计点击次数, 这样就介绍完了汤普森采样的流程.

这个例子中的奖励是延迟的, 网站的访客不可能在网站加载出来的一瞬间就点击广告, 而是延迟一段时间去点击. 这段时间访客需要完成观看广告, 理解广告内容, 移动鼠标等一系列操作. 这样可能在上个访客还没来得及点击广告就迎来下一个访客. 用 Bandit 语言来说, 就是可能在上一轮奖励还未知的前提下, 就提前开始了下一轮的博弈. 遇到这种情况是否会影响汤普森采样算法的执行呢? 答案是并不会. 因为在实际应用中, 汤普森采样并不会再每一轮决策完都必须更新参数. 也就是说我们完全可以在不更新参数的前提下执行很多次决策, 然后等数据收集到一定数量时, 再统一更新参数. 并且从经验上看, 汤普森采样对于延迟奖励具有很好的鲁棒性, 而很多其他算法比如 UCB 在这方面就差的多, 因此这也是汤普森采样相比 UCB 的一大优点.

汤普森采样和 UCB 的区别对比:

1. 汤普森采样比 UCB 出现的更早 (汤普森采样于 1933 年提出, UCB 于 1985 年提出)
2. 从经验上看汤普森采样比 UCB 性能更优.
- \* 3. **汤普森采样是一个随机化策略, 而 UCB 是确定性策略.**由此导致了:
  - (a) 汤普森采样存在因随机性而导致的方差, 而 UCB 不存在此问题.
  - (b) 随机性使汤普森采样更稳定.
  - (c) 汤普森采样对延迟奖励的鲁棒性比 UCB 更高.
  - (d) 汤普森采样可推广到对抗性环境.
4. 汤普森采样可通过设定先验分布来引入环境的先验知识.

**定义 3.55. (汤普森采样策略).** 设  $(\mathcal{E}, \mathcal{B}(\mathcal{E}), Q, P)$  为一个贝叶斯 Bandit 环境. 设  $\mu_i(\nu) = \int_{\mathbb{R}} x dP_{\nu_i}(x)$  为环境  $\nu \in \mathcal{E}$  的第  $i$  个动作的平均奖励. 汤普森采样策略定义为  $\pi = \{\pi_t\}_{t=1}^{\infty}$ . 其中

$$\pi_t(\{a\} \mid A_1, X_1, \dots, A_{t-1}, X_{t-1}) = Q(B_a \mid A_1, X_1, \dots, A_{t-1}, X_{t-1})$$

其中  $B_a = \{\nu \in \mathcal{E} : a = \arg \max_b \mu_b(\nu)\} \in \mathcal{B}(\mathcal{E})$ . 其中的  $\arg \max$  可采取任意但一致的方式处理存在多个最大值的情况.

$\pi_t(\{a\} \mid A_1, X_1, \dots, A_{t-1}, X_{t-1}) = Q(B_a \mid A_1, X_1, \dots, A_{t-1}, X_{t-1})$  用通俗的语言描述:  $t$  时刻选择动作  $a$  的概率等于给定历史交互数据之后  $a$  是最优动作这个事件  $B_a$  在后验分布中的概率. 一个动作在后验分布中是最优动作概率越大, 则它被选中的几率就越大.

**性质 3.56.** 若  $A_t = \arg \max_a \mu_a(\nu_t)$ , 其中  $\nu_t \sim Q(\cdot \mid A_1, X_1, \dots, A_{t-1}, X_{t-1})$ . 则

$$A_t \sim \pi_t(\cdot \mid A_1, X_1, \dots, A_{t-1}, X_{t-1})$$

证明.

$$\begin{aligned} \mathbb{P}(\{A_t = a\}) &= \mathbb{P}(\arg \max_a \mu_a(\nu_t) = a) \\ &= Q(\arg \max_a \mu_a(\nu_t) = a \mid A_1, X_1, \dots, A_{t-1}, X_{t-1}) \\ &= \pi_t(\{a\} \mid A_1, X_1, \dots, A_{t-1}, X_{t-1}) \end{aligned}$$

所以  $A_t \sim \pi_t(\cdot \mid A_1, X_1, \dots, A_{t-1}, X_{t-1})$ .  $\square$

### 算法 3(汤普森采样)

输入贝叶斯 Bandit 环境  $(\mathcal{E}, \mathcal{B}(\mathcal{E}), Q, P)$ .

for  $t \in 1, \dots, n$  do

    采样  $\nu_t \sim Q(\cdot \mid A_1, X_1, \dots, A_{t-1}, X_{t-1})$ .

    选择动作  $A_t = \arg \max_{i \in [K]} \mu_i(\nu_t)$ , 观测  $X_t$ .

end for

**定义 3.57. (使用Beta先验的汤普森采样).** 使用 Beta 先验的汤普森采样定义为以  $\alpha = \beta = 1$  为超参数的  $K$  臂贝叶斯伯努利 Bandit 作为输入的汤普森采样算法.

回顾  $K$  臂贝叶斯伯努利 Bandit. 先验分布  $Q$  定义

$$Q(A) = \int_A \prod_{i=1}^K q(\nu_i; \alpha, \beta) d\nu = \int_A \prod_{i=1}^K q(\nu_i; 1, 1) d\nu$$

其中  $q(\cdot; 1, 1)$  为 Beta( $1, 1$ ) 的 PDF. Beta( $1, 1$ ) 恰好就是  $[0, 1]$  上的均匀分布. 注意到密度函数是对每个动作的密度函数求乘积的形式, 这意味着每个动作奖励分布的参数彼此之间是独立的. 我们称这种形式的先验为乘积先验, 它的好处是我们不需要采样所有动作的联合先验分布, 取而代之我们可以分别采样每个动作的先验分布, 分别采样每个动作比采样联合分布实现起来要容易的多, 并且运行效率也高的多. 因此采用乘积先验可以帮助我们降低代码实现的难度, 提升代码的运行效率.

回顾 Beta 分布的共轭先验性质:

设  $X \sim \text{Bernoulli}(\Theta)$ ,  $\Theta \sim \text{Beta}(\alpha, \beta)$ , 则  $\mathbb{P}_{\Theta|X=x}(x + \alpha, \beta - x + 1)$ . 上述结论可重写为:

$$\mathbb{P}_{\Theta|X=x} = \begin{cases} \text{Beta}(\alpha + 1, \beta), & \text{若 } x = 1 \\ \text{Beta}(\alpha, \beta + 1), & \text{若 } x = 0 \end{cases}$$

上式可推广到多个观测的情况. 假设有  $K$  个观测值  $D = [x_1, x_2, \dots, x_K]$ , 设  $S$  为其中 1 的个数,  $F$  为其中 0 的个数 ( $K = S + F$ ). 则后验分布为:

$$\mathbb{P}_{\Theta|D} = \text{Beta}(\alpha + S, \beta + F)$$

总结上述讨论:

1. 由于每个动作的奖励分布的参数彼此独立, 可分别采样每个动作的先验/后验分布.
2. 每个动作的先验分布均为 Beta( $1, 1$ ).
3. 对于每个动作  $i$ , 设  $S_i$  表示观测到奖励为 1 的次数,  $F_i$  表示观测到奖励为 0 的次数, 则动作  $i$  的后验分布为  $\text{Beta}(1 + S_i, 1 + F_i)$ .

**算法 4(使用 Beta 先验的汤普森采样)**

初始化  $S_i \leftarrow 0$  和  $F_i \leftarrow 0, \forall i \in [K]$

for  $t \in 1, \dots, n$  do

    采样  $\theta_i(t) \sim \text{Beta}(1 + S_i, 1 + F_i), \forall i \in [K]$

    选择  $A_t = \arg \max_{i \in [K]} \theta_i(t)$ , 观测奖励  $X_t$

    若  $X_t = 1$  则  $S_{A_t} \leftarrow S_{A_t} + 1$ , 否则  $F_{A_t} \leftarrow F_{A_t} + 1$

end for

使用 Beta 先验的汤普森采样可以被推广到奖励取值为  $[0, 1]$  闭区间的版本.  
这个版本最早由下面这篇文章提出, 感兴趣读者可查阅

*Analysis of Thompson Sampling for the Multi-armed Bandit*

作者:Shipra Agrawal 和 Navin Goyal

这篇文章的推广思路非常简单, 其主要思路是把非伯努利分布的奖励转化为伯努利分布的奖励. 假设在  $t$  时刻观测到的奖励  $X_t \in [0, 1]$ , 想把  $X_t$  转化为伯努利随机变量, 设  $Y_t \sim \text{Bernoulli}(X_t)$ , 用  $Y_t$  取代  $X_t$  成为新的奖励, 即可套用之前的算法.

**算法 5(奖励值取  $X_t \in [0, 1]$  的汤普森采样)**

初始化  $S_i \leftarrow 0$  和  $F_i \leftarrow 0, \forall i \in [K]$

for  $t \in 1, \dots, n$  do

    采样  $\theta_i(t) \sim \text{Beta}(1 + S_i, 1 + F_i), \forall i \in [K]$

    选择  $A_t = \arg \max_{i \in [K]} \theta_i(t)$ , 观测奖励  $X_t$

    采样  $Y_t \sim \text{Bernoulli}(X_t)$

    若  $Y_t = 1$ , 则  $S_{A_t} \leftarrow S_{A_t} + 1$ , 否则  $F_{A_t} \leftarrow F_{A_t} + 1$

end for

### 3.9 汤普森采样的贝叶斯遗憾分析

本节主要参考 *Bandit Algorithms* 这本书. 当然这个理论分析并不是完全这本书的原创, 而是基于下面这篇论文

*Learning to Optimize Via Posterior Sampling*

作者:Daniel Russo, Benjamin Van Roy

不过尽管这篇文章给出的贝叶斯遗憾是次线性的, 但并不是已知的界里面最紧的一个, 因为下面这篇论文证明出一个更紧的贝叶斯遗憾界:

*Prior-free and Prior-dependent Regret Bounds for Thompson Sampling*

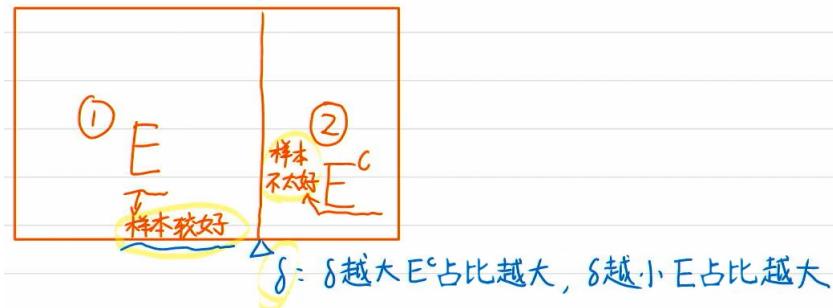
作者:Sébastien Bubeck, Che-Yu Liu

**定理 3.58.** (汤普森采样的贝叶斯遗憾界). 假设  $(\mathcal{E}, \mathcal{B}(\mathcal{E}), Q, P)$  为一个贝叶斯 Bandit 环境满足对于任意  $\nu \in \mathcal{E}$  和  $i \in [K]$ ,  $P_{\nu i}$  均为均值在  $[0, 1]$  内的 1-次高斯分布. 则汤普森采样策略  $\pi$  满足:

$$\text{BR}_n(\pi, Q) = O\left(\sqrt{nK \log(n)}\right)$$

注:  $O(g(n)) := \{f(n) : \exists c > 0, n_0 > 0 \text{ 使得 } \forall n \geq n_0 \text{ 有 } 0 \leq f(n) \leq cg(n) \text{ 成立}\}$ .

推导 Bandit 算法遗憾界的一个常见思路就是分而治之, 具体来说就是先对算法执行的过程分情况处理, 然后再整合出我们想要的结论. 这点对于汤普森采样来说也并不例外, 我们如果直接从贝叶斯遗憾定义出发去推次线性上界是非常困难的, 因此我们需要分两种情况来处理.



样本较好的情况很容易推出次线性增长的理论保证, 而样本不太好的情况很难推出次线性增长. 但我们可以换种思路, 压根不去推第二种情况的次线性遗憾界, 而是去衰减  $\delta$  这个阈值, 因为  $\delta$  越小, 第二种情况的占比就越小. 因此我们只要以一定的方式去衰减第二种情况的阈值, 就可以使第二种情况的概率本身呈次线性增长.

证明. 设  $\mu_i(\nu)$  为环境  $\nu$  的第  $i$  个动作的平均奖励简写作  $\mu_i$ . 设  $A^* = \arg \max_{i \in [K]} \mu_i$  为最优动作.(这里需要注意  $\mu_i, A^*$  均为随机变量, 在前面的 ETC, UCB1 的理论分析中这两个量都为常数, 之所以有这个区别, 是因为这两个量都是依赖于环境  $\nu$  的, 在前面的算法中环境  $\nu$  并不是一个随机变量, 在这里  $\nu$  是服从先验分布  $Q$  的随机变量, 其根源是我们这里采用的是贝叶斯设定, 而 ETC, UCB1 算法采用频率派设定. 既然  $\nu$  是随机变量, 后面依赖于  $\nu$  的所有量都是随机变量, 这点需要格外注意.) 定义事件  $E$  为:

$$E = \left\{ |\hat{\mu}_i(t-1) - \mu_i| < \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}}, \forall t \in [n], \forall i \in [K] \right\}$$

其中  $\hat{\mu}_i(t-1)$  为  $t-1$  轮博弈结束后动作  $i$  的奖励的经验估计,  $T_i(t-1)$  为第  $i$  个动作在第  $t-1$  轮博弈结束后动作  $i$  被选中的次数,  $0 < \delta \leq 1$  为一个实数. 规定  $\hat{\mu}_i(t-1) = 0$  若  $T_i(t-1) = 0$ . 则有  $\mathbb{P}(E^c) \leq 2nK\delta$  成立. 其验证如下:

$$\begin{aligned} E^c &= \left\{ |\hat{\mu}_i(t-1) - \mu_i| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}}, \exists t \in [n], i \in [K] \right\} \\ &= \bigcup_{i=1}^K \bigcup_{t=1}^n \left\{ |\hat{\mu}_i(t-1) - \mu_i| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \right\} \end{aligned}$$

因此,

$$\mathbb{P}(E^c) = \mathbb{P} \left( \bigcup_{i=1}^K \bigcup_{t=1}^n \left\{ |\hat{\mu}_i(t-1) - \mu_i| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \right\} \right)$$

由概率的次可加性

$$(3.59) \quad \mathbb{P}(E^c) \leq \sum_{i=1}^K \sum_{t=1}^n \mathbb{P} \left( |\hat{\mu}_i(t-1) - \mu_i| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \right)$$

可以通过集中不等式来计算其上界, 在此之前先把分母的  $\max$  去掉, 否则用完集中不等式后的式子会比较复杂, 不利于化简. 通过全概率公式分情况讨论把  $\max$

去掉:

$$\begin{aligned}
& \mathbb{P} \left( |\hat{\mu}_i(t-1) - \mu_i| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \right) \\
&= \mathbb{P}(T_i(t-1) < 1) \mathbb{P} \left( |\hat{\mu}_i(t-1) - \mu_i| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \middle| T_i(t-1) < 1 \right) \\
&\quad + \mathbb{P}(T_i(t-1) \geq 1) \mathbb{P} \left( |\hat{\mu}_i(t-1) - \mu_i| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \middle| T_i(t-1) \geq 1 \right) \\
&= \mathbb{P}(T_i(t-1) = 0) \underbrace{\mathbb{P} \left( \mu_i \geq \sqrt{2 \ln \left( \frac{1}{\delta} \right)} \middle| T_i(t-1) = 0 \right)}_{(a)} \\
&\quad + \underbrace{\mathbb{P}(T_i(t-1) \neq 0) \mathbb{P} \left( |\hat{\mu}_i(t-1) - \mu_i| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{T_i(t-1)}} \middle| T_i(t-1) \neq 0 \right)}_{(b)}
\end{aligned}$$

若 (a)  $\leq 2\delta$  和 (b)  $\leq 2\delta$  同时成立, 则上式  $\leq 2\delta$ . 下面先证: (a)  $\leq 2\delta$ :

$$\begin{aligned}
(a) &= \mathbb{P} \left( \mu_i \geq \sqrt{2 \ln \left( \frac{1}{\delta} \right)} \middle| T_i(t-1) = 0 \right) \\
&= \mathbb{Q}_1 \left( \mu_i \geq \sqrt{2 \ln \left( \frac{1}{\delta} \right)} \right) \quad (\text{其中 } \mathbb{Q}_1(\cdot) := \mathbb{P}(\cdot \mid T_i(t-1) = 0)) \\
&\leq \mathbb{Q}_1 \left( \mu_i^2 \geq 2 \ln \left( \frac{1}{\delta} \right) \right) \quad \left( \text{因为 } \mu_i \geq \sqrt{2 \ln \left( \frac{1}{\delta} \right)} \Rightarrow \mu_i^2 \geq 2 \ln \left( \frac{1}{\delta} \right) \right) \\
&= \mathbb{Q}_1 \left( \frac{\mu_i^2}{2} \geq \ln \left( \frac{1}{\delta} \right) \right) = \mathbb{Q}_1 \left( \exp \left( \frac{\mu_i^2}{2} \right) \geq \frac{1}{\delta} \right) \\
&\leq \frac{\mathbb{E}_{\mathbb{Q}_1} \left[ \exp \left( \frac{\mu_i^2}{2} \right) \right]}{\frac{1}{\delta}} \quad \left( \text{因为 } \exp \left( \frac{\mu_i^2}{2} \right) > 0, \text{ 根据Markov不等式} \right) \\
&\leq \exp \left( \frac{1}{2} \right) \delta \quad (\text{因为 } \mu_i \in [0, 1]) \\
&\approx 1.649\delta < 2\delta
\end{aligned}$$

然后证明 (b)  $\leq 2\delta$ , 固定环境  $\nu$ , 此时  $\mu_i$  为常数 (因为贝叶斯遗憾是假设环境服从

先验分布时的平均遗憾, 只要证明对任意确定的环境  $\nu$  的取值, 某个遗憾的上界都成立, 则对服从先验分布的  $\nu$ , 这个上界也是成立的. 即任意取值都成立的上界对期望就也成立), 则有

$$\begin{aligned} \text{(b)} &= \mathbb{P} \left( |\hat{\mu}_i(t-1) - \mu_i| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{T_i(t-1)}} \mid T_i(t-1) \neq 0 \right) \\ &= \mathbb{Q}_2 \left( |\hat{\mu}_i(t-1) - \mu_i| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{T_i(t-1)}} \right) \quad (\text{定义 } \mathbb{Q}_2(\cdot) := \mathbb{P}(\cdot \mid T_i(t-1) \neq 0)) \end{aligned}$$

由于  $T_i(t-1) \neq 0$ , 根据定义:

$$\hat{\mu}_i(t-1) = \frac{\sum_{\tau=1}^{t-1} \mathbb{I}\{A_\tau = i\} X_\tau}{T_i(t-1)}$$

注意到分子求和项只有  $T_i(t-1)$  项不等于 0, 设  $S_{ij}(t-1)$  为第  $j$  个使得  $A_\tau = i$  的  $\tau$ , 则上式可改写为:

$$\hat{\mu}_i(t-1) = \frac{\sum_{j=1}^{T_i(t-1)} X_{S_{ij}(t-1)}}{T_i(t-1)}$$

将上式代入 (b) 的表达式,

$$\begin{aligned} \text{(b)} &= \mathbb{Q}_2 \left( \left| \frac{\sum_{j=1}^{T_i(t-1)} X_{S_{ij}(t-1)}}{T_i(t-1)} - \mu_i \right| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{T_i(t-1)}} \right) \\ &= \mathbb{Q}_2 \left( \left| \frac{\sum_{j=1}^{T_i(t-1)} X_{S_{ij}(t-1)}}{T_i(t-1)} - \frac{\sum_{j=1}^{T_i(t-1)} \mathbb{E}[X_{S_{ij}(t-1)}]}{T_i(t-1)} \right| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{T_i(t-1)}} \right) \\ &= \mathbb{Q}_2 \left( \frac{1}{T_i(t-1)} \left| \sum_{j=1}^{T_i(t-1)} (X_{S_{ij}(t-1)} - \mathbb{E}[X_{S_{ij}(t-1)}]) \right| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{T_i(t-1)}} \right) \end{aligned}$$

在本章第(3.3)节中推导了用  $\delta$  表示  $\varepsilon$  的 Hoeffding 界:

$$\mathbb{P} \left( \frac{1}{n} \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{n}} \right) \leq \delta$$

注: 上述该 Hoeffding 界要求对于任意的  $i \in [n]$ ,  $X_i$  的次高斯参数均为 1.

其双侧尾概率的版本为:

$$\mathbb{P} \left( \frac{1}{n} \left| \sum_{i=1}^n (X_i - \mathbb{E}[X_i]) \right| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{n}} \right) \leq 2\delta$$

应用上面的 Hoeffding 界可以得到  $(b) \leq 2\delta$ . 至此已经验证 (a) 和 (b) 均不大于  $2\delta$ , 故可知

$$\mathbb{P} \left( |\hat{\mu}_i(t-1) - \mu_i| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \right) \leq 2\delta$$

将上式代入(3.59)式:

$$\begin{aligned} \mathbb{P}(E^c) &\leq \sum_{i=1}^K \sum_{t=1}^n \mathbb{P} \left( |\hat{\mu}_i(t-1) - \mu_i| \geq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \right) \\ &\leq \sum_{i=1}^K \sum_{t=1}^n 2\delta = 2nK\delta \end{aligned}$$

这个式子的含义为  $\delta$  越小, 则样本不太好的情况发生的概率就越低. 后期我们可以通过衰减  $\delta$  就可以衰减第二种情况发生概率. 接下来我们推导  $E$  事件发生时算法的次线性贝叶斯遗憾界. 这里的技巧是对贝叶斯遗憾进行适当的分解, 具体来说, 贝叶斯遗憾定义是对频率派遗憾求期望. 频率派遗憾定义为每一轮最优动作的平均奖励到策略选择的动作的平均奖励之间的距离之和. 由于每一轮博弈对遗憾的贡献都是两个平均奖励的数值距离, 因此我们可以取一个所谓的中间值, 从而把每一轮博弈对遗憾的贡献分解为两个数值到中间值各自的距离之和. 这种分解之所以对我们的理论分析有帮助, 是因为我们通过巧妙选取中间值, 可以使得在  $E$  事件发生的情况下, 分解之后的第一项永远是负的, 而第二项的上界是次线性的. 这就是接下来推导的思路, 先定义中间值,  $\forall t \in [n], i \in [K]$  定义随机变量  $U_t(i)$  为:

$$U_t(i) = \text{clip}_{[0,1]} \left( \hat{\mu}_i(t-1) + \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \right)$$

其中  $\text{clip}_{[0,1]}(x) = \max(0, \min(1, x))$ . 设  $\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t)$ , 可知  $U_t(i)$  是  $\mathcal{F}_{t-1}$  可测的. 接下来将证明下式几乎处处成立:

$$(3.60) \quad \mathbb{P}(A^* = \cdot \mid \mathcal{F}_{t-1}) = \mathbb{P}(A_t = \cdot \mid \mathcal{F}_{t-1})$$

验证上式:

$$\begin{aligned}
 \forall i \in [k] \quad \mathbb{P}(A^* = i \mid \mathcal{F}_{t-1}) &= \mathbb{P}(\arg \max_{j \in [K]} \mu_j(\nu) = i \mid \mathcal{F}_{t-1}) \\
 &= Q(\{\nu' \in \mathcal{E} : \arg \max_{j \in [K]} \mu_j(\nu') = i\} \mid \mathcal{F}_{t-1}) \\
 &= \mathbb{P}(\arg \max_{j \in [K]} \mu_j(\nu_t) = i \mid \mathcal{F}_{t-1}) \\
 &= \mathbb{P}(A_t = i \mid \mathcal{F}_{t-1})
 \end{aligned}$$

基于(3.60)可验证  $\mathbb{E}[U_t(A^*)] = \mathbb{E}[U_t(A_t)]$ , 步骤如下:

$$\begin{aligned}
 &\mathbb{E}[U_t(A^*)] - \mathbb{E}[U_t(A_t)] \\
 &= \mathbb{E}[\mathbb{E}[U_t(A^*) \mid \mathcal{F}_{t-1}]] - \mathbb{E}[\mathbb{E}[U_t(A_t) \mid \mathcal{F}_{t-1}]] \\
 &= \mathbb{E}[\mathbb{E}[U_t(A^*) \mid \mathcal{F}_{t-1}] - \mathbb{E}[U_t(A_t) \mid \mathcal{F}_{t-1}]] \\
 &= \mathbb{E}\left[\mathbb{E}\left[\sum_{a \in [K]} \mathbb{I}\{A^* = a\} U_t(a) \mid \mathcal{F}_{t-1}\right] - \mathbb{E}\left[\sum_{a \in [K]} \mathbb{I}\{A_t = a\} U_t(a) \mid \mathcal{F}_{t-1}\right]\right] \\
 &= \mathbb{E}\left[\sum_{a \in [K]} U_t(a) \mathbb{E}[\mathbb{I}\{A^* = a\} \mid \mathcal{F}_{t-1}] - \sum_{a \in [K]} U_t(a) \mathbb{E}[\mathbb{I}\{A_t = a\} \mid \mathcal{F}_{t-1}]\right] \\
 &= \mathbb{E}\left[\sum_{a \in [K]} U_t(a) (\mathbb{P}(A^* = a \mid \mathcal{F}_{t-1}) - \mathbb{P}(A_t = a \mid \mathcal{F}_{t-1}))\right] \\
 &= 0
 \end{aligned}$$

所以

$$(3.61) \quad \mathbb{E}[U_t(A^*)] = \mathbb{E}[U_t(A_t)]$$

根据贝叶斯遗憾的定义,

$$\begin{aligned}
 \text{BR}_n &= \mathbb{E}\left[\sum_{t=1}^n (\mu_{A^*} - \mu_{A_t})\right] \\
 &= \mathbb{E}\left[\sum_{t=1}^n (\mu_{A^*} - \mathbb{E}[U_t(A^*)] + \mathbb{E}[U_t(A_t)] - \mu_{A_t})\right] \\
 &= \mathbb{E}\left[\sum_{t=1}^n (\mu_{A^*} - U_t(A^*) + U_t(A_t) - \mu_{A_t})\right]
 \end{aligned}$$

因此

$$\text{BR}_n = \mathbb{E} \left[ \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) + \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \right]$$

取条件期望有

$$\begin{aligned} \text{BR}_n &= \mathbb{E} \left[ \mathbb{E} \left[ \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) + \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \middle| \mathbb{I}\{E\} \right] \right] \\ &= \mathbb{P}(\mathbb{I}\{E\} = 1) \cdot \mathbb{E} \left[ \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) + \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \middle| \mathbb{I}\{E\} = 1 \right] \\ &\quad + \mathbb{P}(\mathbb{I}\{E\} = 0) \cdot \mathbb{E} \left[ \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) + \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \middle| E \right] \\ &= \underbrace{\mathbb{P}(E) \cdot \mathbb{E} \left[ \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) + \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \middle| \mathbb{I}\{E\} = 1 \right]}_{(1)} \\ &\quad + \underbrace{\mathbb{P}(E^c) \cdot \mathbb{E} \left[ \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) + \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \middle| E^c \right]}_{(2)} \end{aligned}$$

先推 (2) 的上界, 因已经验证过  $\mathbb{P}(E^c) \leq 2nK\delta$ , 故只需推出

$\mathbb{E} \left[ \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) + \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \middle| E^c \right]$  的上界. 由于  $\forall i \in [K], \mu_i \in [0, 1]$  以及  $\forall i \in [K], t \in [n], U_t(i) \in [0, 1]$ , 因此可知

$$-1 \leq mu_i - U_t(i) \leq 1, \forall i \in [K], t \in [n]$$

故

$$\mathbb{E} \left[ \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) + \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \middle| E^c \right] \leq \mathbb{E} \left[ \sum_{t=1}^n 1 + \sum_{t=1}^n 1 \middle| E^c \right] = 2n$$

因此  $(2) \leq 2nK\delta \cdot 2n = 4n^2K\delta$ . 接下来推导 (1) 的上界,

$$\begin{aligned} (1) &= \mathbb{P}(E) \cdot \mathbb{E} \left[ \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) + \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \middle| \mathbb{I}\{E\} = 1 \right] \\ &= \mathbb{E} \left[ \mathbb{I}\{E\} \left( \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) + \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \right) \right] \end{aligned}$$

因此

$$(1) = \mathbb{E} \left[ \mathbb{I}\{E\} \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) + \mathbb{I}\{E\} \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \right]$$

首先证明  $\mathbb{I}\{E\} \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) \leq 0$ . 若  $E$  不发生, 则

$$\mathbb{I}\{E\} \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) = 0$$

若  $E$  发生, 则根据  $E$  的定义可知  $\forall t \in [n]$  和  $\forall i \in [K]$  有:

$$\begin{aligned} |\hat{\mu}_i(t-1) - \mu_i| &< \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \\ \Rightarrow \mu_i - \hat{\mu}_i(t-1) &< \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \end{aligned}$$

则有

$$\mu_i < \hat{\mu}_i(t-1) + \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}}$$

从而

$$\text{clip}_{[0,1]}(\mu_i) \leq \text{clip}_{[0,1]} \left( \hat{\mu}_i(t-1) + \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \right)$$

即  $\mu_i \leq U_t(i)$ . 因此  $E$  发生时

$$\mathbb{I}\{E\} \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) = \sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) \leq 0$$

接下来推导  $\mathbb{I}\{E\} \sum_{t=1}^n (U_t(A_t) - \mu_{A_t})$  的上界. 若  $E$  不发生则该式值为 0, 假设  $E$  发生有:

$$\begin{aligned} &\mathbb{I}\{E\} \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \\ &= \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) = \sum_{t=1}^n \sum_{i=1}^K \mathbb{I}(A_t = i)(U_t(i) - \mu_i) \\ &= \sum_{i=1}^K \sum_{t=1}^n \mathbb{I}(A_t = i) \left( \text{clip}_{[0,1]} \left( \hat{\mu}_i(t-1) + \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \right) - \mu_i \right) \end{aligned}$$

由于 clip 中的数为非负的, 本质上就是和 1 取小, 把 clip 丢掉放大有

$$\mathbb{E}\{E\} \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \leq \sum_{i=1}^K \sum_{t=1}^n \mathbb{I}(A_t = i) \left( \hat{\mu}_i(t-1) + \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} - \mu_i \right)$$

由于  $E$  发生时有  $\hat{\mu}_i(t-1) - \mu_i \leq \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}}$ , 所以

$$\begin{aligned} & \mathbb{E}\{E\} \sum_{t=1}^n (U_t(A_t) - \mu_{A_t}) \\ & \leq \sum_{i=1}^K \sum_{t=1}^n \mathbb{I}(A_t = i) \left( \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} + \sqrt{\frac{2 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \right) \\ & = \sum_{i=1}^K \sum_{t=1}^n \mathbb{I}(A_t = i) \sqrt{\frac{8 \ln(\frac{1}{\delta})}{\max(1, T_i(t-1))}} \\ & = \sum_{i=1}^K \sum_{t=1}^n \mathbb{I}(A_t = i) \sqrt{\frac{8 \ln(\frac{1}{\delta})}{\max(1, \sum_{t'=1}^{t-1} \mathbb{I}\{A_{t'} = i\})}} \\ & = \sum_{i=1}^K \sum_{s=0}^{T_i(n)-1} \sqrt{\frac{8 \ln(\frac{1}{\delta})}{\max(1, s)}} \\ & \leq \sum_{i=1}^K \int_0^{T_i(n)} \sqrt{\frac{8 \ln(\frac{1}{\delta})}{s}} ds \quad (*) \\ & = \sum_{i=1}^K \sqrt{32 T_i(n) \ln\left(\frac{1}{\delta}\right)} \\ & = K \sum_{i=1}^K \frac{1}{K} \sqrt{32 T_i(n) \ln\left(\frac{1}{\delta}\right)} \\ & \leq K \sqrt{\sum_{i=1}^K \frac{1}{K} 32 T_i(n) \ln\left(\frac{1}{\delta}\right)} \quad (\text{Jensen不等式}) \\ & = K \sqrt{32 \ln\left(\frac{1}{\delta}\right) \frac{1}{K} \sum_{i=1}^K T_i(n)} \\ & = K \sqrt{32 \ln\left(\frac{1}{\delta}\right) \frac{1}{K} n} = \sqrt{32 n K \ln\left(\frac{1}{\delta}\right)} \end{aligned}$$

其中 (\*) 式验证较为复杂, 放在最后验证. 将上面的结论代入 (1):

$$\begin{aligned}
 (1) &= \mathbb{E} \left[ \underbrace{\mathbb{I}\{E\} \sum_{t=1}^n (\mu_{A^*} - U_t(A^*))}_{\leq 0} + \underbrace{\mathbb{I}\{E\} \sum_{t=1}^n (U_t(A_t) - \mu_{A_t})}_{\leq \sqrt{32nK \ln(\frac{1}{\delta})}} \right] \\
 &\leq \sqrt{32nK \ln\left(\frac{1}{\delta}\right)}
 \end{aligned}$$

令  $\delta = n^{-2}$ , 将 (1) 和 (2) 的上界代入贝叶斯遗憾:

$$\begin{aligned}
 \text{BR}_n &= \mathbb{P}(E) \cdot \mathbb{E} \left[ \underbrace{\sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) + \sum_{t=1}^n (U_t(A_t) - \mu_{A_t})}_{(1) \leq \sqrt{32nK \ln(\frac{1}{\delta})} = 8\sqrt{nK \ln n}} \middle| \mathbb{I}\{E\} = 1 \right] \\
 &\quad + \mathbb{P}(E^c) \cdot \mathbb{E} \left[ \underbrace{\sum_{t=1}^n (\mu_{A^*} - U_t(A^*)) + \sum_{t=1}^n (U_t(A_t) - \mu_{A_t})}_{(2) \leq 4n^2 K \delta = 4K} \middle| E^c \right] \\
 &= 8\sqrt{nK \ln n} + 4K \\
 &= O\left(\sqrt{nK \log n}\right)
 \end{aligned}$$

验证 (\*) 步骤:

$$\sum_{i=1}^K \sum_{s=0}^{T_i(n)-1} \sqrt{\frac{8 \ln(\frac{1}{\delta})}{\max(1, s)}} \leq \sum_{i=1}^K \int_0^{T_i(n)} \sqrt{\frac{8 \ln(\frac{1}{\delta})}{s}} ds$$

若证上式, 只需证

$$\forall n \in \mathbb{N} \quad \sum_{s=0}^{n-1} \sqrt{\frac{8 \ln(\frac{1}{\delta})}{\max(1, s)}} \leq \int_0^n \sqrt{\frac{8 \ln(\frac{1}{\delta})}{s}} ds$$

当  $n = 1$  时上式显然成立, 故假设  $n \geq 2$ , 需证明:

$$\forall n \in \{2, 3, \dots\} \quad \sum_{s=0}^{n-1} \sqrt{\frac{8 \ln(\frac{1}{\delta})}{\max(1, s)}} \leq \int_0^n \sqrt{\frac{8 \ln(\frac{1}{\delta})}{s}} ds$$

$$\begin{aligned}
&\iff \sum_{s=0}^{n-1} \sqrt{\frac{1}{\max(1, s)}} \leq \int_0^n \sqrt{\frac{1}{s}} ds \\
&\iff \sum_{s=0}^{n-1} \sqrt{\frac{1}{\max(1, s)}} \leq 2\sqrt{n} \\
&\iff 1 + \sum_{s=1}^{n-1} \sqrt{\frac{1}{\max(1, s)}} \leq 2\sqrt{n} \\
&\iff 1 + \sum_{s=1}^{n-1} \sqrt{\frac{1}{s}} - 2\sqrt{n} \leq 0 \\
&\iff 1 + \sum_{s=2}^n \sqrt{\frac{1}{s-1}} - 2(\sqrt{1} + \sum_{s=2}^n (\sqrt{s} - \sqrt{s-1})) \leq 0 \\
&\iff -1 + \sum_{s=2}^n \left( \sqrt{\frac{1}{s-1}} - 2(\sqrt{s} - \sqrt{s-1}) \right) \leq 0
\end{aligned}$$

为了验证上式, 首先证明

$$\begin{aligned}
&\sqrt{\frac{1}{s-1}} - 2(\sqrt{s} - \sqrt{s-1}) \geq 0 \\
&\iff \sqrt{\frac{1}{s-1}} \geq 2(\sqrt{s} - \sqrt{s-1}) \\
&\iff 1 \geq 2(\sqrt{s}\sqrt{s-1} - (s-1)) \\
&\iff 1 \geq 2\sqrt{s^2-s} - (2s-2) \\
&\iff 2s-1 \geq 2\sqrt{s^2-s} \\
&\iff 4s^2 - 4s + 1 \geq 4s^2 - 4s \iff 1 \geq 0
\end{aligned}$$

故  $\sqrt{\frac{1}{s-1}} - 2(\sqrt{s} - \sqrt{s-1}) \geq 0$  成立. 因此:

$$\sum_{s=2}^n \sqrt{\frac{1}{s-1}} - 2(\sqrt{s} - \sqrt{s-1}) \leq \sum_{s=2}^{\infty} \sqrt{\frac{1}{s-1}} - 2(\sqrt{s} - \sqrt{s-1})$$

由 Wolfram Alpha 可算出级数  $\sum_{s=2}^{\infty} \sqrt{\frac{1}{s-1}} - 2(\sqrt{s} - \sqrt{s-1}) = \zeta\left(\frac{1}{2}\right) + 2 \approx 0.54 \leq 1$ . 代入前面的不等式即可完成证明.  $\square$

### 3.10 汤普森采样的频率派遗憾分析

我们首先从贝叶斯遗憾和频率派遗憾的区别说起:

1. 贝叶斯遗憾次线性增长的理论保证不依赖于先验分布的选取.
2. 与贝叶斯遗憾的情况不同. 并非所有先验分布都可以使频率派遗憾次线性增长.(例:狄拉克先验 (Dirac Prior)会使频率派遗憾线性增长.) 原因:
  - 每次采样狄拉克先验都会得到相同的环境样本, 从而选择相同动作.
  - 当先验分布是狄拉克先验时, 后验分布永远等于先验分布.
3. 常见的使用Beta 先验和高斯先验的汤普森采样都具有频率派遗憾次线性增长的理论保证.

不同的先验对应的频率派遗憾界并不相等, 其推导过程也并不相同. 由于时间关系, 本节仅推导采用 Beta 先验的汤普森采样算法的频率派遗憾界, 参考的论文:

*Further Optimal Regret Bounds for Thompson Sampling*

作者:Shipra Agrawal 和 Navin Goyal

**定理 3.62.** (采用Beta先验的汤普森采样的频率派遗憾界). 给定一个  $K$  臂伯努利 Bandit 环境  $\nu \in \mathcal{E}_B^K$ , 假设  $\nu$  的第一个动作是唯一最优动作, 则使用 Beta 先验的汤普森采样算法与  $\nu$  交互的遗憾满足

$$R(n) = O\left(\sqrt{nK \log n}\right)$$

注记

假设  $\nu$  的第一个动作是唯一最优动作不失一般性, 若存在多个最优动作, 则遗憾一定比我们证明的遗憾界更小.

**定理 3.63. (针对伯努利随机变量的Hoeffding界).** 设  $X_1, X_2, \dots, X_n$  为  $n$  个独立的伯努利随机变量满足  $\mathbb{E}[X_i] = p_i$ . 设  $X = \frac{1}{n} \sum_{i=1}^n X_i, \mu = \mathbb{E}[X] = \frac{1}{n} \sum_{i=1}^n p_i$ , 则对于任意  $0 < \varepsilon < 1 - \mu$  有

$$\mathbb{P}(X - \mu \geq \varepsilon) \leq \exp(-nd(\mu + \varepsilon, \mu))$$

且对任意  $0 < \varepsilon < \mu$  有

$$\mathbb{P}(X - \mu \leq -\varepsilon) \leq \exp(-nd(\mu - \varepsilon, \mu))$$

其中  $d(x, y) := x \ln \frac{x}{y} + (1-x) \ln \frac{1-x}{1-y}$  为参数分别为  $x$  和  $y$  的两个伯努利分布之间的 KL 散度.

定理(3.62)的证明:

证明.  $\forall i \in [K] \setminus \{1\}$ , 设  $x_i$  和  $y_i$  为满足  $\mu_i < x_i < y_i < \mu_1$  的两个实数. 显然, 因假设第一个动作为唯一的最优动作, 即  $\mu_i < \mu_1, \forall i \in [K] \setminus \{1\}$ , 故  $x_i$  和  $y_i$  存在.  $\forall i \in [K]$  定义  $T_i(t) = \sum_{s=1}^t \mathbb{I}\{A_s = i\}$ , 以及

$$\hat{\mu}_i(t) = \frac{\sum_{s=1}^t \mathbb{I}\{A_s = i\} X_s}{T_i(t) + 1}$$

为了防止分母为 0, 所以分母取为  $T_i(t) + 1$ .  $\forall i \in [K] \setminus \{1\}$ , 定义事件

$$E_i^\mu(t) := \{\hat{\mu}_i(t-1) \leq x_i\} \text{ 和 } E_i^\theta(t) := \{\theta_i(t) \leq y_i\}$$

这里的  $\hat{\mu}_i(t-1)$  是从第 1 时刻到第  $t-1$  时刻第  $i$  个动作的奖励的经验平均, 当然由于分母加了 1, 所以这里的  $\hat{\mu}_i(t-1)$  并不是真正意义上的经验平均, 而是一个低估.  $\theta_i(t)$  表示第  $t$  时刻从第  $i$  个动作的后验分布中采样出的奖励期望, 无论是  $\hat{\mu}_i(t-1)$  还是  $\theta_i(t)$  都是对第  $i$  个动作真实奖励期望的一个近似, 前者是经验估计, 后者是从后验分布中采样出的一个样本, 而  $x_i$  和  $y_i$  则是第  $i$  个动作真实奖励期望的上界. 因此直观上看, 这两个事件的意义是第  $i$  个动作在第  $t$  时刻所对应的这两个近似值不要高估的太多. 具体来说不要高估到超过  $x_i$  和  $y_i$  这两个阈值. 其实

这里定义的事件类似于上一节我们定义的  $E$  事件, 也就是样本比较好所对应的事件, 只不过在上节中我们只定义了一个事件, 这一节中我们对每个次优动作都定义了两个事件, 共  $2(K - 1)$  个事件.

接下来我们利用刚刚定义的事件来分解遗憾, 我们并不直接分解遗憾, 而是分解次优动作平均被选取的次数. 因为根据遗憾分解引理, 把每个动作平均选择次数乘以次优间隙, 再对所有动作求和就得到了最终的遗憾, 而最优动作对遗憾没有贡献, 所以只要分解次优动作平均被选取的次数. 对  $\forall i \in [K] \setminus \{1\}$  有:

$$\begin{aligned}\mathbb{E}[T_i(n)] &= \mathbb{E}\left[\sum_{t=1}^n \mathbb{I}\{A_t = i\}\right] = \sum_{t=1}^n \mathbb{E}[\mathbb{I}\{A_t = i\}] = \sum_{t=1}^n \mathbb{P}(A_t = i) \\ &= \sum_{t=1}^n \mathbb{P}(A_t = i, E_i^\mu(t)) + \sum_{t=1}^n \mathbb{P}(A_t = i, \overline{E_i^\mu(t)})\end{aligned}$$

其中  $\overline{E_i^\mu(t)}$  表示事件  $E_i^\mu(t)$  的补集, 继续对上式第一项按事件  $E_i^\theta(t)$  进行拆分, 有  
(3.64)

$$\mathbb{E}[T_i(n)] = \underbrace{\sum_{t=1}^n \mathbb{P}(A_t = i, E_i^\theta(t), E_i^\mu(t))}_{(1)} + \underbrace{\sum_{t=1}^n \mathbb{P}(A_t = i, \overline{E_i^\theta(t)}, E_i^\mu(t))}_{(2)} + \underbrace{\sum_{t=1}^n \mathbb{P}(A_t = i, \overline{E_i^\mu(t)})}_{(3)}$$

设  $\mathcal{F} = \{\mathcal{F}_t\}_{t=1}^{n-1}$ , 其中  $\mathcal{F}_t = \sigma(A_1, X_1, \dots, A_t, X_t)$ . 由于  $\mathbb{E}[\mathbb{P}(E | Y)] = \mathbb{P}(E)$ , 则  
(1) 可改写为

$$(3.65) \quad (1) = \sum_{t=1}^n \mathbb{E}[\mathbb{P}(A_t = i, E_i^\theta(t), E_i^\mu(t) | \mathcal{F}_{t-1})]$$

定义  $p_{i,t} = \mathbb{P}(\theta_1(t) > y_i | \mathcal{F}_{t-1})$ . 则  $\forall t \in [n]$  和  $\forall i \in [K] \setminus \{1\}$  有:

$$(3.66) \quad \mathbb{P}(A_t = i, E_i^\mu(t), E_i^\theta(t) | \mathcal{F}_{t-1}) \leq \frac{1 - p_{i,t}}{p_{i,t}} \mathbb{P}(A_t = 1, E_i^\mu(t), E_i^\theta(t) | \mathcal{F}_{t-1})$$

因  $E_i^\mu(t) \in \mathcal{F}_{t-1}$ , 故上式中  $E_i^\mu(t)$  是否发生完全由  $\mathcal{F}_{t-1}$  中的信息决定. 当  $\mathcal{F}_{t-1}$  使得  $E_i^\mu(t)$  不发生时, (3.66) 式不等号两侧均为 0, 不等式平凡成立. 故假设  $\mathcal{F}_{t-1}$  可以使  $E_i^\mu(t)$  发生. 只需证:

$$\begin{aligned}\mathbb{P}(A_t = i, E_i^\theta(t) | \mathcal{F}_{t-1}) &\leq \frac{1 - p_{i,t}}{p_{i,t}} \mathbb{P}(A_t = 1, E_i^\theta(t) | \mathcal{F}_{t-1}) \\ \iff \mathbb{P}(A_t = i | E_i^\theta(t), \mathcal{F}_{t-1}) \mathbb{P}(E_i^\theta(t) | \mathcal{F}_{t-1}) &\leq \frac{1 - p_{i,t}}{p_{i,t}} \mathbb{P}(A_t = 1 | E_i^\theta(t), \mathcal{F}_{t-1}) \mathbb{P}(E_i^\theta(t) | \mathcal{F}_{t-1})\end{aligned}$$

若  $\mathbb{P}(E_i^\theta(t) \mid \mathcal{F}_{t-1}) = 0$ , 则上式平凡成立. 故假设  $\mathbb{P}(E_i^\theta(t) \mid \mathcal{F}_{t-1}) \neq 0$ , 只需证:

$$(3.67) \quad \mathbb{P}(A_t = i \mid E_i^\theta(t), \mathcal{F}_{t-1}) \leq \frac{1 - p_{i,t}}{p_{i,t}} \mathbb{P}(A_t = 1 \mid E_i^\theta(t), \mathcal{F}_{t-1})$$

首先推导不等号左侧的上界

$$\begin{aligned} & \mathbb{P}(A_t = i \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ &= \mathbb{P}(\arg \max_{j \in [K]} \theta_j(t) = i \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ &= \mathbb{P}(\theta_j(t) \leq \theta_i(t), \forall j \in [K] \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ &\leq \mathbb{P}(\theta_1(t) \leq y_i, \theta_j(t) \leq \theta_i(t), \forall j \in [K] \setminus \{1\} \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ &= \mathbb{P}(\theta_1(t) \leq y_i \mid \theta_j(t) \leq \theta_i(t), \forall j \in [K] \setminus \{1\}, E_i^\theta(t), \mathcal{F}_{t-1}) \\ &\quad \cdot \mathbb{P}(\theta_j(t) \leq \theta_i(t), \forall j \in [K] \setminus \{1\} \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ &= \mathbb{P}(\theta_1(t) \leq y_i \mid \mathcal{F}_{t-1}) \cdot \mathbb{P}(\theta_j(t) \leq \theta_i(t), \forall j \in [K] \setminus \{1\} \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ &= (1 - p_{i,t}) \cdot \mathbb{P}(\theta_j(t) \leq \theta_i(t), \forall j \in [K] \setminus \{1\} \mid E_i^\theta(t), \mathcal{F}_{t-1}) \end{aligned}$$

将上式中的  $\theta_j(t) \leq \theta_i(t), \forall j \in [K] \setminus \{1\}$  事件记为  $M_i(t)$ . 根据定义  $\mu_i < y_i < \mu_1$  且  $\mu_i \in [0, 1], \forall i \in [K]$ , 故  $y_i \neq 1$ . 因  $\theta_1(t)$  服从 Beta 分布, 所以当  $y_i \neq 1$  时  $\mathbb{P}(\theta_1(t) > y_i \mid \mathcal{F}_{t-1}) = p_{i,t} \neq 0$ . 因此:

$$\begin{aligned} \mathbb{P}(A_t = i \mid E_i^\theta(t), \mathcal{F}_{t-1}) &\leq (1 - p_{i,t}) \cdot \mathbb{P}(M_i(t) \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ &= \frac{1 - p_{i,t}}{p_{i,t}} \cdot p_{i,t} \cdot \mathbb{P}(M_i(t) \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ &= \frac{1 - p_{i,t}}{p_{i,t}} \cdot \mathbb{P}(\theta_1(t) > y_i \mid \mathcal{F}_{t-1}) \cdot \mathbb{P}(M_i(t) \mid E_i^\theta(t), \mathcal{F}_{t-1}) \end{aligned}$$

注意到  $\theta_1(t), \theta_2(t), \dots, \theta_k(t)$  给定  $\mathcal{F}_{t-1}$  时条件独立.  $M_i(t)$  和  $E_i^\theta(t)$  由  $\theta_2(t), \dots, \theta_k(t)$  确定. 所以

$$\mathbb{P}(A_t = i \mid E_i^\theta(t), \mathcal{F}_{t-1}) \leq \frac{1 - p_{i,t}}{p_{i,t}} \cdot \mathbb{P}(\theta_1(t) > y_i \mid M_i(t), E_i^\theta(t), \mathcal{F}_{t-1}) \cdot \mathbb{P}(M_i(t) \mid E_i^\theta(t), \mathcal{F}_{t-1})$$

根据定义有  $E_i^\theta(t) = \{\theta_i(t) \leq y_i\}, M_i(t) = \{\theta_j(t) \leq \theta_i(t), \forall j \in [K] \setminus \{1\}\}$ , 故当  $E_i^\theta(t)$  和  $M_i(t)$  同时发生时有:

$$\theta_j(t) \leq \theta_i(t) \leq y_i \quad \forall j \in [K] \setminus \{i\}$$

因此

$$\theta_1(t) > y_i \Rightarrow \theta_1(t) > \theta_j(t), \forall j \in [K] \setminus \{1\} \Rightarrow A_t = 1$$

故

$$\begin{aligned} \mathbb{P}(A_t = i \mid E_i^\theta(t), \mathcal{F}_{t-1}) &\leq \frac{1 - p_{i,t}}{p_{i,t}} \cdot \mathbb{P}(\theta_1(t) > y_i \mid M_i(t), E_i^\theta(t), \mathcal{F}_{t-1}) \cdot \mathbb{P}(M_i(t) \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ &\leq \frac{1 - p_{i,t}}{p_{i,t}} \cdot \mathbb{P}(A_t = 1 \mid M_i(t), E_i^\theta(t), \mathcal{F}_{t-1}) \cdot \mathbb{P}(M_i(t) \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ &= \frac{1 - p_{i,t}}{p_{i,t}} \cdot \mathbb{P}(A_t = 1, M_i(t) \mid E_i^\theta(t), \mathcal{F}_{t-1}) \\ &\leq \frac{1 - p_{i,t}}{p_{i,t}} \cdot \mathbb{P}(A_t = 1 \mid E_i^\theta(t), \mathcal{F}_{t-1}) \end{aligned}$$

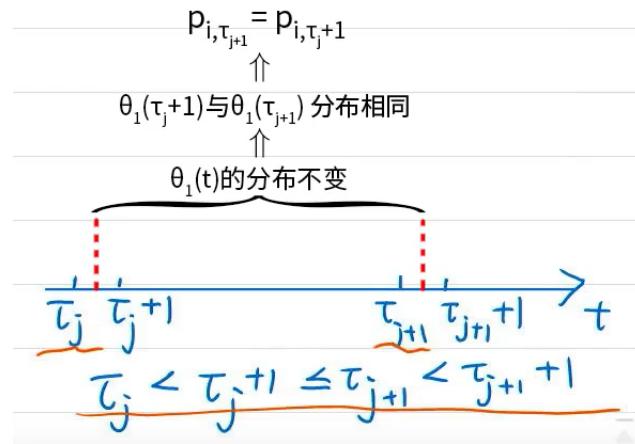
因此(3.67)式得证. 由于(3.67)式成立仅当(3.66)式成立, 因此(3.66)式得证. 将(3.66)式代入(3.65)式

$$\begin{aligned} (1) &= \sum_{t=1}^n \mathbb{E}[\mathbb{P}(A_t = i, E_i^\theta(t), E_i^\mu(t) \mid \mathcal{F}_{t-1})] \\ &\leq \sum_{t=1}^n \mathbb{E}\left[\frac{1 - p_{i,t}}{p_{i,t}} \mathbb{P}(A_t = 1, E_i^\mu(t), E_i^\theta(t) \mid \mathcal{F}_{t-1})\right] \\ &= \sum_{t=1}^n \mathbb{E}\left[\frac{1 - p_{i,t}}{p_{i,t}} \mathbb{E}[\mathbb{I}\{A_t = 1, E_i^\mu(t), E_i^\theta(t)\} \mid \mathcal{F}_{t-1}]\right] \\ &= \sum_{t=1}^n \mathbb{E}\left[\frac{1 - p_{i,t}}{p_{i,t}} \mathbb{I}\{A_t = 1, E_i^\mu(t), E_i^\theta(t)\}\right] \\ &\leq \sum_{t=1}^n \mathbb{E}\left[\frac{1 - p_{i,t}}{p_{i,t}} \mathbb{I}\{A_t = 1\}\right] \\ &= \mathbb{E}\left[\sum_{t=1}^n \frac{1 - p_{i,t}}{p_{i,t}} \mathbb{I}\{A_t = 1\}\right] \end{aligned}$$

由于上式求和中只有  $T(n)$  项不等于 0, 这些项对应着第 1 个动作被选中的时刻, 故可重新对求和进行整理, 只保留不为 0 的项. 令  $\tau_j$  表示第  $j$  次选择第 1 个动作

的时刻, 其中  $j \geq 1$ . 令  $\tau_0 = 0$ . 则有:

$$\begin{aligned}
 (1) &\leq \mathbb{E} \left[ \sum_{t=1}^n \frac{1 - p_{i,t}}{p_{i,t}} \mathbb{I}\{A_t = 1\} \right] \\
 &= \mathbb{E} \left[ \sum_{j=1}^{T_1(n)} \frac{1 - p_{i,\tau_j}}{p_{i,\tau_j}} \right] \\
 &= \mathbb{E} \left[ \sum_{j=0}^{T_1(n)-1} \frac{1 - p_{i,\tau_j+1}}{p_{i,\tau_j+1}} \right] \\
 &= \mathbb{E} \left[ \sum_{j=0}^{T_1(n)-1} \frac{1 - p_{i,\tau_j+1}}{p_{i,\tau_{j+1}}} \right] \\
 &= \mathbb{E} \left[ \sum_{j=0}^{T_1(n)-1} \left( \frac{1}{p_{i,\tau_{j+1}}} - 1 \right) \right] \\
 &\leq \mathbb{E} \left[ \sum_{j=0}^{n-1} \left( \frac{1}{p_{i,\tau_{j+1}}} - 1 \right) \right] \\
 &\leq \sum_{j=0}^{n-1} \mathbb{E} \left[ \frac{1}{p_{i,\tau_{j+1}}} - 1 \right]
 \end{aligned}$$



注意到  $p_{i,t} = \mathbb{P}(\theta_1(t) > y_i \mid \mathcal{F}_{t-1})$ , 只与  $\theta_1(t)$  的分布有关, 所以  $p_{i,\tau_j+1} = p_{i,\tau_{j+1}}$ .

由原论文中的引理 2(由于其证明篇幅较长, 此处略去, 感兴趣读者可阅读原论文), 有

$$(1) \leq \sum_{j=0}^{n-1} \mathbb{E} \left[ \frac{1}{p_{i,\tau_{j+1}}} - 1 \right] \\ (3.68) \quad \leq \frac{24}{\Delta_i'^2} + \sum_{j=1}^{n-1} O \left( e^{\frac{-\Delta_i'^2 j}{2}} + \frac{1}{(j+1)\Delta_i'^2} e^{-d(y_i, \mu_1)j} + \frac{1}{e^{\frac{\Delta_i'^2 j}{4}} - 1} \right)$$

其中  $\Delta_i' = \mu_1 - y_i, d(x, y) = x \ln \frac{x}{y} + (1-x) \ln \frac{1-x}{1-y}$  为两个参数分别为  $x$  和  $y$  的 Bernoulli 分布之间的 KL 散度. 接下来推导 (2) 式上界. 对于每个动作  $i$ , 定义  $L_i(t) = \frac{\ln t}{d(x_i, y_i)}$ . 设  $\tau$  为使  $T_i(t) \leq L_i(n)$  成立的最大的  $t$ . 可知  $\forall t \leq \tau, T_i(t) \leq L_i(n)$  均成立. 则有:

$$(2) = \sum_{t=1}^n \mathbb{P}(A_t = i, \overline{E_i^\theta(t)}, E_i^\mu(t)) \\ = \mathbb{E} \left[ \sum_{t=1}^n \mathbb{P}(A_t = i, \overline{E_i^\theta(t)}, E_i^\mu(t)) \right] \\ = \mathbb{E} \left[ \sum_{t=1}^{\tau} \mathbb{P}(A_t = i, \overline{E_i^\theta(t)}, E_i^\mu(t)) + \sum_{t=\tau+1}^n \mathbb{P}(A_t = i, \overline{E_i^\theta(t)}, E_i^\mu(t)) \right] \\ \leq \mathbb{E} \left[ \sum_{t=1}^{\tau} \mathbb{P}(A_t = i) + \sum_{t=\tau+1}^n \mathbb{P}(A_t = i, \overline{E_i^\theta(t)}, E_i^\mu(t)) \right] \\ = \mathbb{E} \left[ \sum_{t=1}^{\tau} \mathbb{P}(A_t = i) + \sum_{t=\tau+1}^n \mathbb{P}(A_t = i, \overline{E_i^\theta(t)} \mid E_i^\mu(t)) \mathbb{P}(E_i^\mu(t)) \right]$$

因此有

$$(3.69) \quad (2) \leq \mathbb{E} \left[ \sum_{t=1}^{\tau} \mathbb{P}(A_t = i) + \sum_{t=\tau+1}^n \underbrace{\mathbb{P}(A_t = i, \overline{E_i^\theta(t)} \mid E_i^\mu(t))}_{(4)} \right]$$

接下来将证明  $t > \tau$  成立时  $(4) \leq \frac{1}{n}$ . 由  $t > \tau$  可知  $T_i(t) > L_i(n)$ . 则有:

$$(4) = \mathbb{P}(A_t = i, \overline{E_i^\theta(t)} \mid E_i^\mu(t)) \\ = \mathbb{P}(A_t = i, \theta_i(t) > y_i \mid \hat{\mu}_i(t-1) \leq x_i) \\ \leq \mathbb{P}(\theta_i(t) > y_i \mid \hat{\mu}_i(t-1) \leq x_i)$$

其中

$$\begin{aligned}
 \theta_i(t) &\sim \text{Beta}(S_i(t) + 1, F_i(t) + 1) \\
 &= \text{Beta}\left(\sum_{s=1}^{t-1} \mathbb{I}\{A_s = i\} X_s + 1, \sum_{s=1}^{t-1} \mathbb{I}\{A_s = i\} (1 - X_s) + 1\right) \\
 &= \text{Beta}(\hat{\mu}_i(t-1)(T_i(t-1) + 1) + 1, T_i(t-1) - \hat{\mu}_i(t-1)(T_i(t-1) + 1) + 1) \\
 &= \text{Beta}(\hat{\mu}_i(t-1)(T_i(t-1) + 1) + 1, (1 - \hat{\mu}_i(t-1))(T_i(t-1) + 1))
 \end{aligned}$$

定义  $\mathbb{P}(\text{Beta}(\alpha, \beta) > y_i)$  表示一个服从  $\text{Beta}(\alpha, \beta)$  的随机变量取值大于  $y_i$  的概率. 则有:

$$\begin{aligned}
 (4) &\leq \mathbb{P}(\text{Beta}(\hat{\mu}_i(t-1)(T_i(t-1) + 1) + 1, (1 - \hat{\mu}_i(t-1))(T_i(t-1) + 1)) > y_i \mid \hat{\mu}_i(t-1) \leq x_i) \\
 &\leq \mathbb{P}(\text{Beta}(x_i(T_i(t-1) + 1) + 1, (1 - x_i)(T_i(t-1) + 1)) > y_i) \\
 &= 1 - \mathbb{P}(\text{Beta}(x_i(T_i(t-1) + 1) + 1, (1 - x_i)(T_i(t-1) + 1)) \leq y_i)
 \end{aligned}$$

设  $F_{\alpha, \beta}^{\text{Beta}}(\cdot)$  为  $\text{Beta}(\alpha, \beta)$  的 CDF 函数, 设  $F_{n, p}^B$  为二项分布的 CDF 函数, 则:

$$(4) \leq 1 - F_{x_i(T_i(t-1)+1)+1, (1-x_i)(T_i(t-1)+1)}^{\text{Beta}}(y_i)$$

代入 Beta 分布与二项分布之间的关系式:  $F_{\alpha, \beta}^{\text{Beta}}(y) = 1 - F_{\alpha+\beta-1, y}^B(\alpha - 1)$ ,

$$(4) \leq 1 - \left(1 - F_{x_i(T_i(t-1)+1)+1+(1-x_i)(T_i(t-1)+1)-1, y_i}^B(x_i(T_i(t-1) + 1) + 1 - 1)\right)$$

即有

$$\begin{aligned}
 (4) &\leq F_{x_i(T_i(t-1)+1)+(1-x_i)(T_i(t-1)+1), y_i}^B(x_i(T_i(t-1) + 1)) \\
 &= F_{T_i(t-1)+1, y_i}^B(x_i(T_i(t-1) + 1))
 \end{aligned}$$

定义  $\mathbb{P}(\text{Binomial}(n, p) > y)$  表示服从  $\text{Binomial}(n, p)$  的随机变量取值大于  $y$  的概率. 则:

$$\begin{aligned}
 (4) &\leq \mathbb{P}(\text{Binomial}(T_i(t-1) + 1, y_i) \leq x_i(T_i(t-1) + 1)) \\
 &= \mathbb{P}\left(\frac{\text{Binomial}(T_i(t-1) + 1, y_i)}{T_i(t-1) + 1} \leq x_i\right) \\
 &= \mathbb{P}\left(\frac{\text{Binomial}(T_i(t-1) + 1, y_i)}{T_i(t-1) + 1} - y_i \leq -(y_i - x_i)\right) \\
 &\leq \exp\{-(T_i(t-1) + 1)d(y_i - (y_i - x_i), y_i)\} \\
 &\quad (\text{应用针对伯努利随机变量的Hoeffding界})
 \end{aligned}$$

因此

$$\begin{aligned}
 (4) &\leq \exp\{-(T_i(t-1)+1)d(x_i, y_i)\} \\
 &\leq \exp\{-L_i(n)d(x_i, y_i)\} \quad (\text{由于 } T_i(t-1)+1 \geq T_i(t) > L_i(n)) \\
 (3.70) \quad &= \exp\left\{-\frac{\ln n}{d(x_i, y_i)}d(x_i, y_i)\right\} \\
 &= \frac{1}{n}
 \end{aligned}$$

将(3.70)式代入(3.69)式:

$$\begin{aligned}
 (2) &\leq \mathbb{E}\left[\sum_{t=1}^{\tau} \mathbb{P}(A_t = i) + \sum_{t=\tau+1}^n \frac{1}{n}\right] \\
 &= \mathbb{E}\left[\sum_{t=1}^{\tau} \mathbb{P}(A_t = i) + (n - \tau)\frac{1}{n}\right] \\
 &= \mathbb{E}\left[\sum_{t=1}^{\tau} \mathbb{P}(A_t = i)\right] + \mathbb{E}\left[(n - \tau)\frac{1}{n}\right] \\
 &\leq \mathbb{E}\left[\sum_{t=1}^{\tau} \mathbb{E}[\mathbb{I}\{A_t = i\}]\right] + \mathbb{E}\left[n\frac{1}{n}\right] \\
 &= \mathbb{E}\left[\sum_{t=1}^{\tau} \mathbb{I}\{A_t = i\}\right] + 1 \\
 &= \mathbb{E}[T_i(\tau)] + 1 \leq \mathbb{E}[L_i(n)] + 1 \\
 &= \frac{\ln n}{d(x_i, y_i)} + 1
 \end{aligned}$$

即 (2) 式的上界推导结束,

$$(3.71) \quad (2) \leq \frac{\ln n}{d(x_i, y_i)} + 1$$

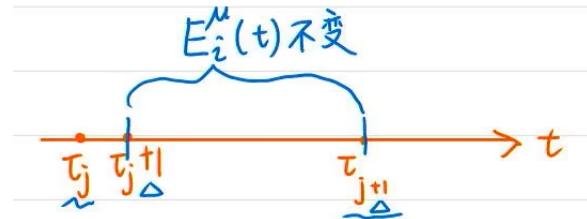
接下来推导 (3) 式的上界. 设  $\tau_j$  为第  $j$  次选择动作  $i$  的时刻, 定义  $\tau_0 = 0$ . 则有:

$$\begin{aligned}
 (3) &= \sum_{t=1}^n \mathbb{P}(A_t = i, \overline{E_i^\mu(t)}) = \sum_{t=1}^n \mathbb{E}\left[\mathbb{I}\{A_t = i, \overline{E_i^\mu(t)}\}\right] \\
 &= \mathbb{E}\left[\sum_{t=1}^n \mathbb{I}\{A_t = i, \overline{E_i^\mu(t)}\}\right] \\
 &= \mathbb{E}\left[\sum_{t=1}^n \mathbb{I}\{A_t = i\} \cdot \mathbb{I}\{\overline{E_i^\mu(t)}\}\right]
 \end{aligned}$$

变换求和指标有

$$(3) = \mathbb{E} \left[ \sum_{j=1}^{T_i(n)} \mathbb{I}\{\overline{E_i^\mu(\tau_j)}\} \right]$$

$$= \mathbb{E} \left[ \sum_{j=0}^{T_i(n)-1} \mathbb{I}\{\overline{E_i^\mu(\tau_{j+1})}\} \right]$$



所以  $E_i^\mu(\tau_{j+1}) = E_i^\mu(\tau_j + 1)$ . 则

$$(3) = \mathbb{E} \left[ \sum_{j=0}^{T_i(n)-1} \mathbb{I}\{\overline{E_i^\mu(\tau_j + 1)}\} \right]$$

$$\leq \mathbb{E} \left[ \sum_{j=0}^{n-1} \mathbb{I}\{\overline{E_i^\mu(\tau_j + 1)}\} \right]$$

$$= \mathbb{E} \left[ \mathbb{I}\{\overline{E_i^\mu(\tau_0 + 1)}\} + \sum_{j=1}^{n-1} \mathbb{I}\{\overline{E_i^\mu(\tau_j + 1)}\} \right]$$

$$= 1 + \mathbb{E} \left[ \sum_{j=1}^{n-1} \mathbb{I}\{\overline{E_i^\mu(\tau_j + 1)}\} \right]$$

$$= 1 + \sum_{j=1}^{n-1} \mathbb{P}(\overline{E_i^\mu(\tau_j + 1)})$$

$$= 1 + \sum_{j=1}^{n-1} \mathbb{P}(\hat{\mu}_i(\tau_j + 1 - 1) > x_i)$$

$$= 1 + \sum_{j=1}^{n-1} \mathbb{P} \left( \frac{\sum_{t=1}^{\tau_j} \mathbb{I}\{A_t = i\} X_t}{T_i(\tau_j) + 1} > x_i \right)$$

由  $\tau_j$  的定义知  $T_i(\tau_j) = j$ , 从而有

$$\begin{aligned}
 (3) &= 1 + \sum_{j=1}^{n-1} \mathbb{P} \left( \frac{\sum_{t=1}^{\tau_j} \mathbb{I}\{A_t = i\} X_t}{j+1} > x_i \right) \\
 &\leq 1 + \sum_{j=1}^{n-1} \mathbb{P} \left( \frac{\sum_{t=1}^{\tau_j} \mathbb{I}\{A_t = i\} X_t}{j} > x_i \right) \\
 &= 1 + \sum_{j=1}^{n-1} \mathbb{P} \left( \frac{\sum_{t=1}^{\tau_j} \mathbb{I}\{A_t = i\} X_t}{j} = \mu_i > x_i - \mu_i \right) \\
 &\leq 1 + \sum_{j=1}^{n-1} \exp\{-jd(\mu_i + (x_i - \mu_i), \mu_i)\} \quad (\text{应用针对伯努利随机变量的Hoeffding界}) \\
 &\leq 1 + \sum_{j=1}^{\infty} \exp\{-jd(x_i, \mu_i)\} \\
 &= 1 + \sum_{j=1}^{\infty} \left( \frac{1}{\exp\{d(x_i, \mu_i)\}} \right)^j
 \end{aligned}$$

由于  $x_i > \mu_i \Rightarrow d(x_i, \mu_i) > 0 \Rightarrow \exp\{d(x_i, \mu_i)\} > 1$ , 所以  $\sum_{j=1}^{\infty} \left( \frac{1}{\exp\{d(x_i, \mu_i)\}} \right)^j$  收敛. 所以有

$$\begin{aligned}
 (3) &= 1 + \sum_{j=0}^{\infty} \left( \frac{1}{\exp\{d(x_i, \mu_i)\}} \right)^j - 1 \\
 &= 1 + \frac{1}{1 - \frac{1}{\exp\{d(x_i, \mu_i)\}}} - 1 \\
 &= 1 + \frac{\exp\{d(x_i, \mu_i)\}}{\exp\{d(x_i, \mu_i)\} - 1} - 1 \\
 &= 1 + \frac{1}{\exp\{d(x_i, \mu_i)\} - 1}
 \end{aligned}$$

由不等式  $e^x - 1 > x, \forall x \neq 0$  有

$$(3.72) \quad (3) \leq 1 + \frac{1}{d(x_i, \mu_i)}$$

将(3.68),(3.71),(3.72)式代入(3.64)式

(3.73)

$$\mathbb{E}[T_i(n)] \leq \frac{24}{\Delta_i'^2} + \sum_{j=1}^{n-1} O \left( e^{-\frac{\Delta_i'^2 j}{2}} + \frac{1}{(j+1)\Delta_i'^2} e^{-d(y_i, \mu_i)j} + \frac{1}{e^{\frac{\Delta_i'^2 j}{4}} - 1} \right) + \frac{\ln n}{d(x_i, y_i)} + 1 + \frac{1}{d(x_i, \mu_i)}$$

为进一步化简(3.73)式需先推导  $d(x, y)$  的下界. 依据 Pinsker 不等式, KL 散度与 TV 距离 (total variation distance) 有如下关系:

$$D_{\text{KL}}(P||Q) \geq 2(D_{\text{TV}}(P||Q))^2$$

其中  $P, Q$  为定义在同一个可测空间的两个分布. 接下来从直觉角度分析 TV 距离的意义. 假设  $P, Q$  为概率分布且  $P \neq Q$ , 则存在事件  $A$  使得  $P(A) \neq Q(A)$ .  $|P(A) - Q(A)|$  可以衡量  $P(A)$  和  $Q(A)$  的距离, 由于使得  $P(A) \neq Q(A)$  的事件  $A$  并不唯一, 因此  $|P(A) - Q(A)|$  无法衡量  $P$  和  $Q$  的距离. 例如  $A$  取  $A_1$  和  $A_2$  都满足  $P(A) \neq Q(A)$ , 但  $|P(A_1) - Q(A_1)| \neq |P(A_2) - Q(A_2)|$ . 所以选取事件  $A$  使得  $|P(A) - Q(A)|$  最大, 并将其值作为 TV 距离定义, 即

$$D_{\text{TV}}(P||Q) = \max_A |P(A) - Q(A)|^2$$

这里最大值并不一定存在, 因此正式定义中求的是上确界

$$D_{\text{TV}}(P||Q) = \sup_A |P(A) - Q(A)|^2$$

TV 距离在机器学习中的理论分析中一大应用就是作为 KL 散度的下界. 之所以把 KL 散度转化为 TV 距离, 就是因为 TV 距离计算公式简单, 因此在分析理论保证时通常借助 TV 距离. 但是 TV 距离相比于 KL 散度也有一个很明显的缺点, 就是不太容易通过样本去估计, 与之相比, KL 散度的定义本身就是一个期望, 很容易通过样本去估计 KL 散度. 正是因为 TV 距离存在这一缺点, 因此它还是主要出现在理论分析中, 而不是出现在算法中. 接下来我们就利用 Pinsker 不等式具体推导  $d(x, y)$  的下界.

设  $f^B(\cdot; \theta)$  为参数为  $\theta$  的伯努利分布的 PMF 函数. 则:

$$\begin{aligned} d(x, y) &= D_{\text{KL}}(\text{Bernoulli}_i(x)||\text{Bernoulli}_i(y)) \\ &\geq 2(D_{\text{TV}}(\text{Bernoulli}_i(x)||\text{Bernoulli}_i(y))^2) \end{aligned}$$

$$\begin{aligned} d(x, y) &\geq 2 \left( \frac{1}{2} |f^B(0; x) - f^B(0; y)| + \frac{1}{2} |f^B(1; x) - f^B(1; y)| \right)^2 \\ &= \frac{1}{2} (|f^B(0; x) - f^B(0; y)| + |f^B(1; x) - f^B(1; y)|)^2 \\ &= \frac{1}{2} (|1 - x - (1 - y)| + |x - y|)^2 = \frac{1}{2} (2|x - y|)^2 = 2(x - y)^2 \end{aligned}$$

其中使用了公式  $D_{\text{TV}}(P||Q) = \frac{1}{2} \sum_{x \in E} |f^P(x) - f^Q(x)|$ . 选取  $x_i = \mu_i + \frac{\Delta_i}{3}$ ,  $y_i = \mu_1 - \frac{\Delta_i}{3}$ . 由此可知  $d(x_i, y_i) \geq 2(x_i - \mu_i)^2 = \frac{2\Delta_i^2}{9} \Rightarrow \frac{1}{d(x_i, \mu_i)} \leq \frac{9}{2\Delta_i^2}$  以及  $d(x_i, y_i) \geq 2(x_i - y_i)^2 = 2\left(\mu_i - \mu_1 + \frac{2}{3}\Delta_i\right)^2 = 2\left(-\frac{1}{3}\Delta_i\right)^2 = \frac{2}{9}\Delta_i^2 \Rightarrow \frac{\ln n}{d(x_i, y_i)} \leq \frac{9 \ln n}{2\Delta_i^2}$  以及  $\Delta_i'^2 = (\mu_1 - y_i)^2 = \frac{\Delta_i^2}{9}$ . 将这些代入(3.73)式

$$\begin{aligned} E[T_i(n)] &\leq \frac{216}{\Delta_i^2} + \frac{9 \ln n}{2\Delta_i^2} + 2 + \frac{9}{2\Delta_i^2} + \sum_{j=1}^{n-1} O\left(e^{-\frac{\Delta_i'^2 j}{2}} + \frac{1}{(j+1)\Delta_i'^2} + \frac{4}{\Delta_i'^2 j}\right) \\ &\leq O\left(\frac{216}{\Delta_i^2} + \frac{9 \ln n}{2\Delta_i^2} + 2 + \frac{9}{2\Delta_i^2}\right) + \sum_{j=1}^{n-1} O\left(e^{-\frac{\Delta_i'^2 j}{2}} + \frac{1}{(j+1)\Delta_i'^2} + \frac{4}{\Delta_i'^2 j}\right) \\ &= O\left(\frac{\ln n}{\Delta_i^2}\right) + O\left(\sum_{j=1}^{n-1} e^{-\frac{\Delta_i'^2 j}{2}} + \sum_{j=1}^{n-1} \frac{1}{(j+1)\Delta_i'^2} + \sum_{j=1}^{n-1} \frac{4}{\Delta_i'^2 j}\right) \end{aligned}$$

因为  $\sum_{j=1}^{n-1} e^{-\frac{\Delta_i'^2 j}{2}} \leq \sum_{j=1}^{\infty} e^{-\frac{\Delta_i'^2 j}{2}}$  是几何级数, 易知其收敛到常数  $C$ . 而

$$\sum_{j=1}^{n-1} \frac{1}{(j+1)\Delta_i'^2} \leq \int_0^{n-1} \frac{1}{(j+1)\Delta_i'^2} dj = \frac{\ln n}{\Delta_i'^2}$$

$$\sum_{j=1}^{n-1} \frac{4}{\Delta_i'^2 j} = \frac{4}{\Delta_i'^2} + \sum_{j=2}^{n-1} \frac{4}{\Delta_i'^2 j} \leq \frac{4}{\Delta_i'^2} + \int_1^{n-1} \frac{4}{\Delta_i'^2 j} dj = \frac{4}{\Delta_i'^2} + \frac{4 \ln(n-1)}{\Delta_i'^2}$$

因此

$$\begin{aligned} E[T_i(n)] &\leq O\left(\frac{\ln n}{\Delta_i^2}\right) + O\left(\frac{\ln n}{\Delta_i'^2} + \frac{4}{\Delta_i'^2} + \frac{4 \ln(n-1)}{\Delta_i'^2}\right) \\ &\leq O\left(\frac{\ln n}{\Delta_i^2}\right) + O\left(\frac{5 \ln n}{\Delta_i'^2}\right) \\ &= O\left(\frac{\ln n}{\Delta_i^2}\right) + O\left(\frac{45 \ln n}{\Delta_i^2}\right) \\ &= O\left(\frac{\ln n}{\Delta_i^2}\right) \end{aligned}$$

定义  $I_1 = \left\{ i \in [K] \mid \Delta_i \geq \sqrt{\frac{K \ln n}{n}} \right\}$ ,  $I_2 = \left\{ i \in [K] \mid \Delta_i < \sqrt{\frac{K \ln n}{n}} \right\}$ .

显然  $I_1 \cap I_2 = \emptyset$  且  $I_1 \cup I_2 = [K]$ . 根据遗憾分解引理,

$$R_n = \sum_{i \in [K]} \Delta_i \mathbb{E}[T_i(n)] = \sum_{i \in I_1} \Delta_i \mathbb{E}[T_i(n)] + \sum_{i \in I_2} \Delta_i \mathbb{E}[T_i(n)]$$

其中

$$\begin{aligned} \sum_{i \in I_1} \Delta_i \mathbb{E}[T_i(n)] &= \sum_{i \in I_1} \Delta_i O\left(\frac{\ln n}{\Delta_i^2}\right) = \sum_{i \in I_1} O\left(\frac{\ln n}{\Delta_i}\right) \leq \sum_{i \in I_1} O\left(\frac{\sqrt{n} \ln n}{\sqrt{K} \ln n}\right) \\ &= \sum_{i \in I_1} O\left(\sqrt{\frac{n \ln n}{K}}\right) \leq K O\left(\sqrt{\frac{n \ln n}{K}}\right) = O\left(\sqrt{n K \ln n}\right) \end{aligned}$$

$$\begin{aligned} \sum_{i \in I_2} \Delta_i \mathbb{E}[T_i(n)] &= \mathbb{E} \left[ \sum_{i \in I_2} \Delta_i T_i(n) \right] \leq \mathbb{E} \left[ \sum_{i \in I_2} \sqrt{\frac{K \ln n}{n}} T_i(n) \right] \\ &= \mathbb{E} \left[ \sqrt{\frac{K \ln n}{n}} \sum_{i \in I_2} T_i(n) \right] \leq \mathbb{E} \left[ \sqrt{\frac{K \ln n}{n}} \cdot n \right] = O\left(\sqrt{n K \ln n}\right) \end{aligned}$$

故

$$R_n = O\left(\sqrt{n K \ln n}\right) + O\left(\sqrt{n K \ln n}\right) = O\left(\sqrt{n K \ln n}\right)$$

□