东北财经大学
Dongbei University of Finance and Economics

多臂老虎机问题
Multi-Armed Bandit (MAB) Problem

唐加福 朱晗

东北财经大学

Road Map

秘书问题

多臂老虎机问题

多臂老虎机基本模型

多臂老虎机问题应用

多臂老虎机模型变种

多臂老虎机问题求解算法

秘书问题

- 秘书问题，又称未婚妻问题，在1949年由Merrill M. Flood 提出。
- 一个经理要为自己从N 个(能力强弱不同的)应聘者中录用一个秘书。
- 应聘者可能以**任意的（random）**顺序前来参加面试。
- 他在面试某个应聘者后必须立即决定是拒绝还是录用。
- 一旦录用某个应聘者,立即停止之后的面试；最后一个应聘者如果被面试则一定会被录用。
- 经理的目标是以最大可能性录用所有应聘者中能力最强的一个人。

秘书问题

- 这一类典型问题：决策者既希望多获得信息以提高未来的决策效果，也希望关注当前阶段的决策收益，因此需要对二者做出平衡。
- 针对这类问题有一个非常经典的策略框架：
- 将所有时刻(其中每个时刻依次面试一个秘书)分成探索(Exploration)和利用/守成（Exploitation）两个阶段。
- 在探索阶段一律拒绝应聘者，在之后的利用阶段一旦出现比之前所有人都要优秀的应聘者则录用，最后一位应聘者如果被面试一定会被聘用。
- 特别的，探索阶段占全阶段1/e比例能够达到最优效果。
- $e \approx 2.718; \frac{1}{e} = 36.8\%.$

秘书问题

- 实际情况中问题更为复杂，往往有多个面试周期，每个周期都会录用新的应聘者。
- 被面试的应聘者也往往会被分为几类，并根据之前已录用的各类应聘者的表现，针对性地(更多/更少)地招收特定类别的应聘者。
- 这对应著名的多摇臂赌博机（Multi-Armed Bandit）问题。

Multi-Armed Bandit(MAB) 多臂老虎机

MAB 基本模型

- 一家赌场，假设面前K台老虎机（arms）
- 老虎机本质上就是个运气游戏，我们假设每台老虎机每次吐出钱r_a 是不固定的，是一个随机变量，其均值为$u(a) := E[r_a]$。
- 例：以一定概率 p_i吐出一块钱，或者不吐钱（$1-p_i$ ）；则r_a=0或1，$u(a)$= p_i
- 注意： p_i可能各不相同，并且其值未知
- 假设你手上只有 T 枚代币（tokens），而每摇一次老虎机都需要花费一枚代币，也就是说你一共只能摇 T次.
- 目标：期望最大回报

MAB 模型应用

- 大部分需要在exploration和exploitation之间做tradeoff的问题都与MAB密切相关。
- 例如，需求函数不确定的库存控制问题。
- 一方面，产品需求函数不明确，报童需要explore更多需求与订货量之间的关系，进而learn需求量分布（demand distribution）；
- 另一方面，为了获得更多利润，报童需要利用已知信息，exploit现有需求与订货量关系。
- 类似问题：当销量与价格之间关系不明确时的产品定价问题。

MAB 模型应用

- 商品展示/搭配（assortment）问题：利用有限资源（如柜台或网站首页显著位置）展示哪种售卖商品或哪几种商品组合。
- 点菜/选餐馆问题：是选择自己熟悉的饭菜/饭店(exploit)， 还是选择新的饭餐/饭店（explore）
- 许多选择问题，本质上就是exploration与exploitation的tradeoff。
- 是选择已知的较为熟悉的，还是选择未知的。
- 从这个意义上说，MAB问题及其解法，可以说是“选择困难症”患者的福音。

MAB 分析

- **决策**是什么？即**每次摇哪台老虎机**
- **反馈**呢？即我们摇了某台特定的老虎机该回合可以**观察它吐了多少钱**。

- 重要的统计学/哲学问题：我们是**贝叶斯人**（Bayesian）还是**频率学家**（frequentist）？

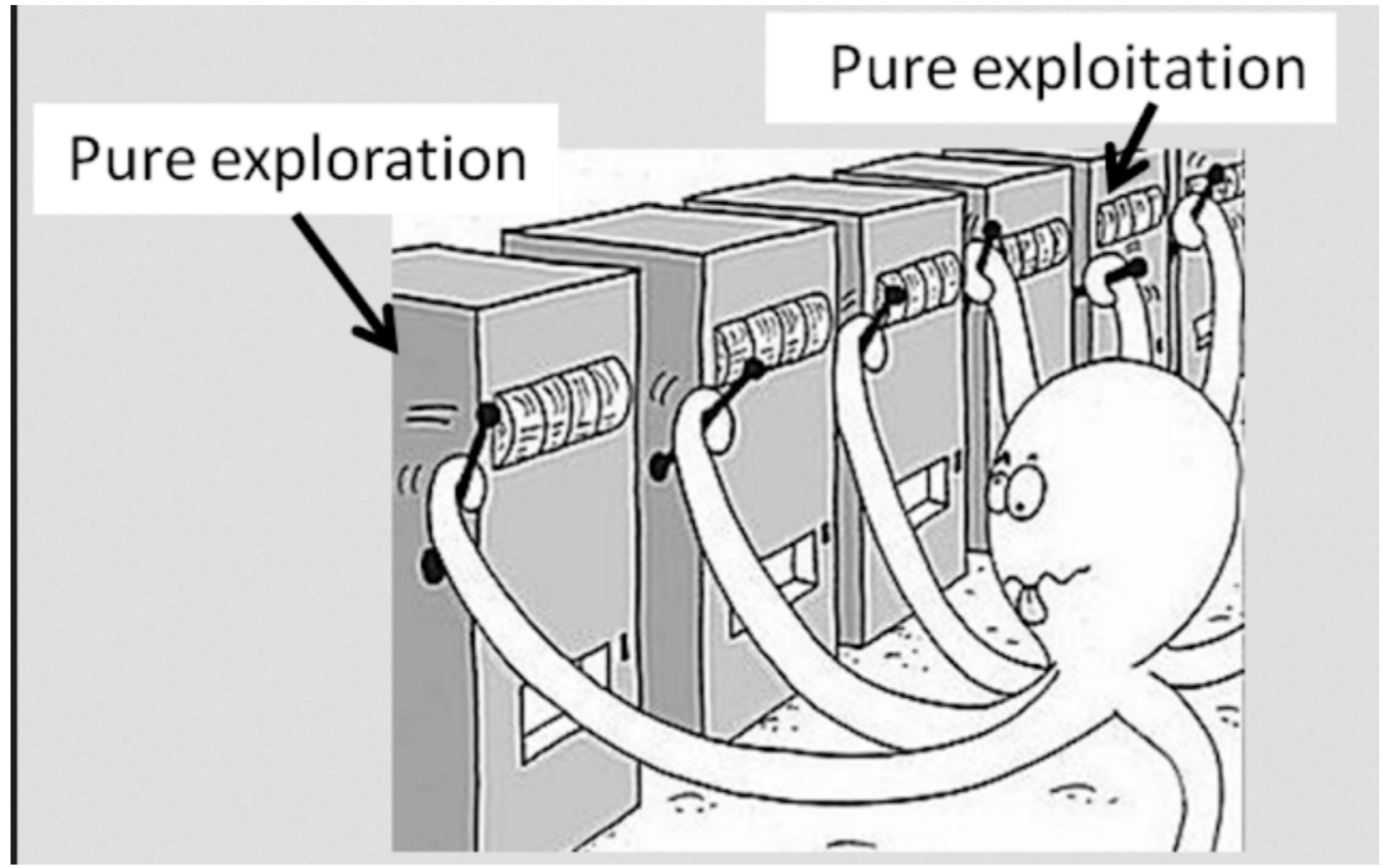
MAB 分析—频率学家

- 假设你一开始对这些机器的吐钱概率一无所知。你认为每个机器的 p_i是个确定的值。
- 那么，你的任务就是要在有限的时间内找到那些高p_i的机器，并尽可能多的去摇它们，以获得更多的回报。
- 这类问题的一大特点，即我们只有 T次摇机器的机会，如何去平衡这 T次中exploration（探索）和exploitation（挖掘）的次数。

MAB 分析—频率学家

- 探索（ exploration ）意味着广度: 一开始你至少每个机器都需要稍微摇几次（假设$T > K$）才能对每个机器吐钱概率有个大概感觉。
- 然后，你可能会缩小你的搜索范围，再几台机器里重点实验。
- 最后可能就专门摇一台你觉得最容易吐钱的机器了。（ exploitation ）

- 我们之后会看到这种办法也未必是最好的。
- 但本质上，所有的求解办法都是在exploration和exploitation之间做tradeoff。

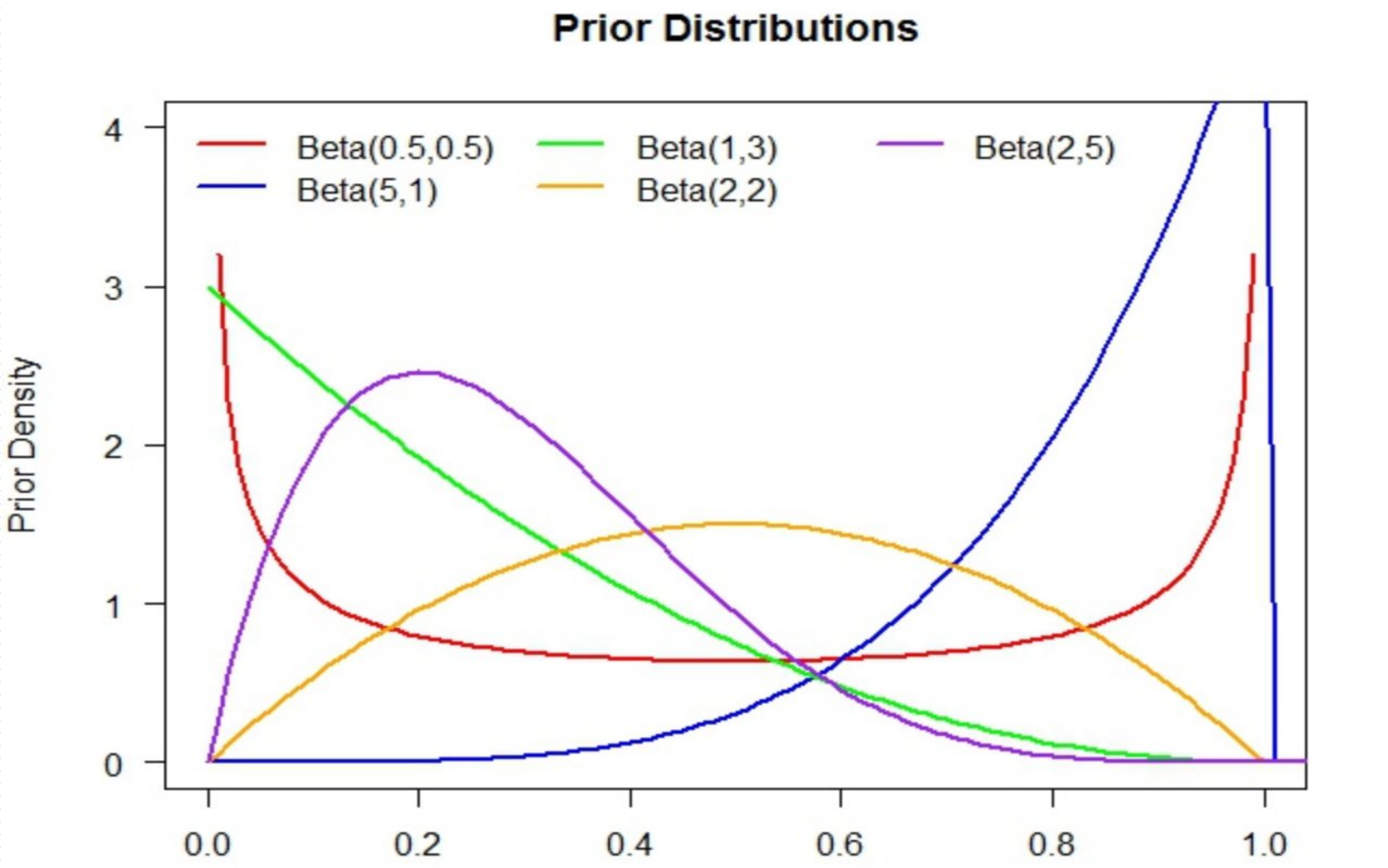


Multi-Armed Bandit - Explore Vs Exploit

MAB 分析—贝叶斯人

- 对贝叶斯人来说，我们在一进入赌场就对每台老虎机扔钱的概率 p_i就有一个先验分布（prior distribution）的假设了，比如我们可以用一个很常见的Beta分布。
- 如果我们认为大概率 p_i都应该是0.5，即对半开，而不太可能出现一些很极端的情况，我们就可以选择Beta(2,2)分布作为我们的先验分布。

MAB 分析—贝叶斯人


Prior Distributions
Beta(0.5,0.5)
Beta(1,3)
Beta(2,5)
Beta(5,1)
Beta(2,2)
Prior Density
0
1
2
3
4
0.0
0.2
0.4
0.6
0.8
1.0

MAB 分析—贝叶斯人

- 然后在我们真正摇了老虎机之后，根据相应的反馈，我们就可以调整 p_i们相应的后验分布（posterior distribution）。
- 如果某台机器摇了四五次一直吐不出钱，我们就应该将这台机器的吐钱概率的分布往左推，因为它的p_i大概率应该是小于0.5的。
- 你的任务便是要在有限的时间内找出p_i后验分布比较靠右的那些机器（因为他们更容易吐钱），并且尽可能多的去摇这些比较赚钱的机器。

MAB问题--模型变种

- 环境变化的MAB：如果一位玩家发现某个机器的 pi很高，一直摇之后赌场可能人为降低这台机器吐钱的概率。在这种情况下，MAB问题的环境就是随着时间/玩家的行为会发生变化。
- contextual MAB(cMAB)：几乎所有在线广告推送都可以看成是cMAB问题。在这类问题中，每个arm的回报会和当前时段出现的顾客的特征（也就是这里说的context）有关。

MAB问题--模型变种

- 如果每台老虎机每天摇的次数有上限，那我们就得到了一个Bandit with Knapsack问题，这类问题以传统组合优化里的背包问题命名。
- 还有很多变种，如Lipshitz bandit，我们不再有有限台机器，而有无限台（它们的reward function满足利普西茨连续性）等等。

MAB问题——算法

- 我们考虑一个概率空间，其中每个arm的收益都有一个相应的随机变量 r_a。算法每次选择一个arm a，就会观察到r_a的一个样本。我们记每个arm期望的回报是 $u(a) := E[r_a]$。那么“最好”的arm即是拥有最高期望回报的arm：

$$\mu^* := \max_{a \in \mathcal{A}} \mu(a) = \mu(a^*)$$

MAB问题——算法

- 我们评价任何一个bandit算法的好坏用的是regret。因为如果你事先知道哪个arm是最好的，你必然 T时间全部都会去选择这个最好的arm，这也是理论上最佳的算法。当然，我们的算法并不提前知道每个arm对应的 r_a的分布（需要learn），因此t时间的regret便定义为

$$R(t) \coloneqq \mu^* t - \sum_{s=1}^{t} \mu(a_s)$$

MAB问题——算法

$$R(t) \coloneqq \mu^* t - \sum_{s=1}^{t} \mu(a_s)$$

- 我们算法的目标便是让 $R(T)$尽可能地小,一般包含两个含义：
- (1) 以很大的概率（with high probability）小;
- (2) $\mathrm{E}[R(T)]$小。
- (1)一般也是蕴含(2)的，我们一般也叫期望的regret， $\mathrm{E}[R(T)]$，为pseudo regret。

MAB问题——Uniform Exploration算法

- 这边先考虑一个最简单的算法。即我们把所有的arm都先试个 N次，然后选择这 N次中得到reward feedback最好的那个arm，在余下的时间里就一直选择这个我们判断最好的arm。
- 这个uniform exploration的名字，是因为我们在开头对所有arm不加区分，都要试 N次而得名。
- 这个算法是一种贪心算法。
- 这个算法本质上是将explore和exploit这两个行为完全的分开了：它是**先纯粹地explore，再纯粹地exploit**。

MAB问题—— ϵ −greedy算法

- 但是纯贪心算法（uniform exploration）变现的并不好，下面我们介绍一种改进的贪心算法——ϵ −greedy 算法。
- 在每一轮 $t = 1 \ldots T$中，掷一个成功率为 ϵ_t的骰子。
- 如果中了，那么explore：以均匀概率选择每个arm;
- 如果没中，则exploit：选择目前为 $\overline{\mu}_t(a)$ 最高的arm（注意这里是 $\overline{\mu}_t(a)$，而不是$\mu_t(a)$，因为是动态变化的，而不像之前都是固定 N个样本的均值）。
- $\overline{\mu}_t(a) = \frac{1}{N_t(a)} \sum_{s=1}^{t} r_s \, \mathbf{1}[a_s = a]$
- $N_t(a) = \sum_{s=1}^{t} \mathbf{1}[a_s = a]$

MAB问题——UCB算法

- Random exploration: good or bad
- Solution 1: decrease ϵ in time;
- Solution 2: optimistic about options with high uncertainty and thus prefer actions for which we haven't had a confident value estimation yet (less-explored).
- We favor exploration of actions with a strong **potential** to have an optimal value.
- Potential: an upper confidence bound $U_t(a)$ of the reward value
- So the true value is below with bound $\mu(a) \leq \overline{\mu}_t(a) + U_t(a)$ with high probability
- $U_t(a)$ is a function of $N_t(a)$: $\downarrow$
- In UCB: $a_t^{\mathrm{UCB}} = arg\max_{a \in A} \overline{\mu}_t(a) + U_t(a)$

MAB问题——UCB算法

- 1.先把每个arm选择一次。
- 2.在每轮t，选择arm

$$A_t = arg\max_a(\bar{\mu}_{t(a)} + \sqrt{\frac{2\ln t}{N_a(t)}})$$

- $\bar{\mu}_{t(a)}$ 是arm a 到t 时刻为止所获得的的平均收益,
- $N_a(t)$是arm a 到t时刻为止所被选择的次数。

MAB问题——UCB算法

- (**Hoeffding's Inequality**) Let $X_1, \ldots, X_t$ be i.i.d. (independent and identically distributed) random variables and they are all bounded by the interval $[0, 1]$. The sample mean is $\overline{X}_t = \frac{1}{t}\sum_{s=1}^{t} X_s$. Then for $u > 0$, we have:

  $$\boldsymbol{Prob\left(E[X] > \overline{X}_t + u\right) \leq e^{-2tu^2}}$$

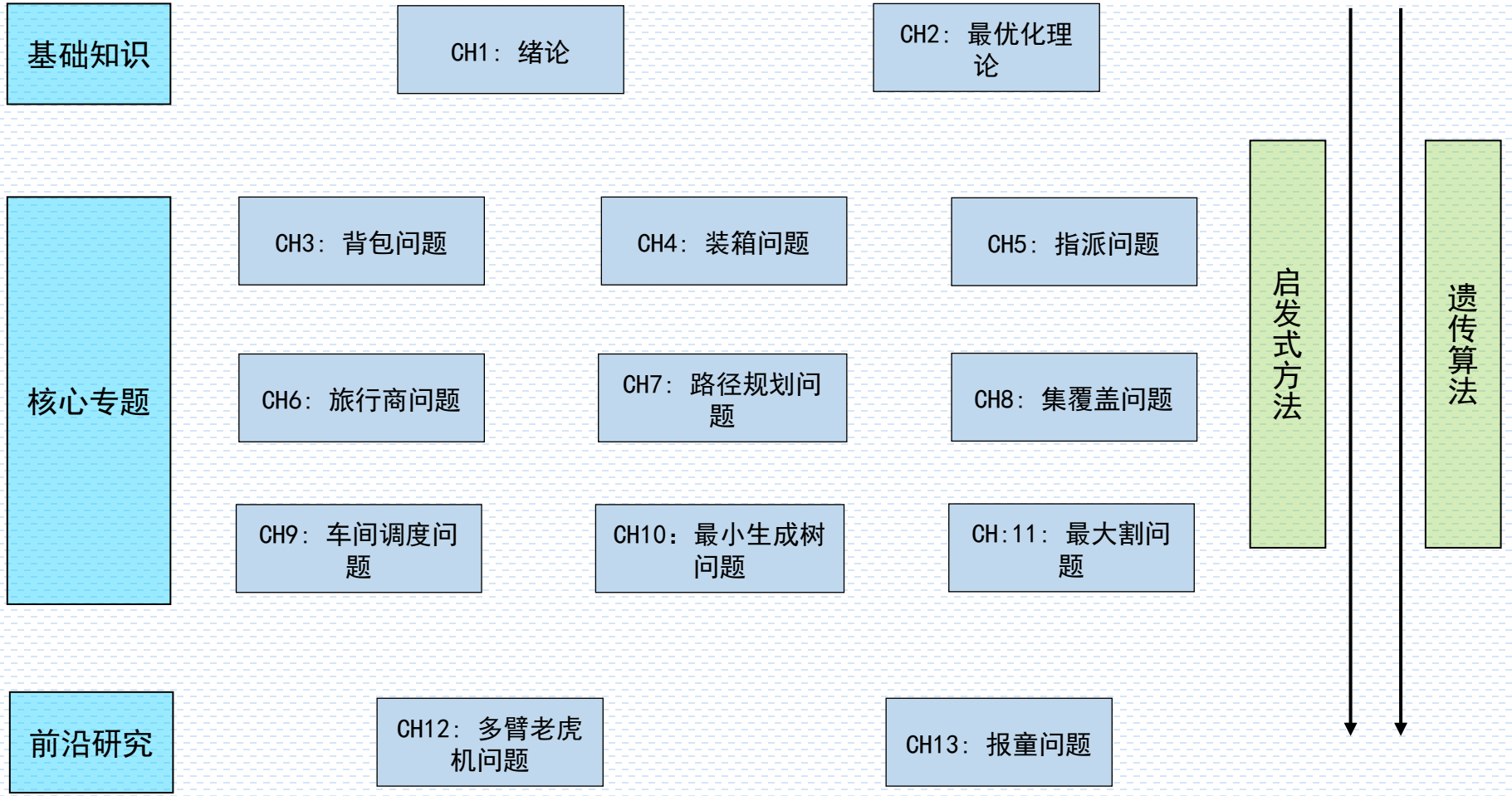
- Given one action a, let us consider
- $r_t(a)$ as the random variables
- $\mu(a)$ as the true mean
- $\overline{\mu}_t(a)$ as the sample mean
- u as the upper confidence bound (potential to perform well), $u = U_t(a)$
- Then, we have $Prob\left(\mu(a) > \overline{\mu}_t(a) + U_t(a)\right) \leq e^{-2N_t(a)U_t(a)^2}$

MAB问题——UCB算法

$$\boldsymbol{Prob}\left(\boldsymbol{\mu(a) > \bar{\mu}_t(a) + U_t(a)}\right) \leq \boldsymbol{e^{-2N_t(a)U_t(a)^2}}$$

- In UCB: $a_t^{UCB} \leq arg\max_{a \in A} \bar{\mu}_t(a) + U_t(a)$
- We want to pick a bound so that with high chances the true mean is blow the sample mean + the potential to perform well (upper confidence bound).
- Thus $e^{-2tU_t(a)^2}$ should be a small probability. Let $e^{-2N_t(a)U_t(a)^2} = \mathrm{p}$. Thus, $U_t(a) = \sqrt{\frac{-\log p}{2N_t(a)}}$
- Set $p = t^{-4}$. We get UCB1 algorithm: $U_t(a) = \sqrt{\frac{2\log t}{N_t(a)}}$ and $a_t = arg\max_{a \in A} \bar{\mu}_{t(a)} + \sqrt{\frac{2\ln \mathrm{t}}{N_t(a)}}$

典型运筹学问题与模型


基础知识
CH1：绪论
CH2：最优化理论
核心专题
CH3：背包问题
CH4：装箱问题
CH5：指派问题
CH6：旅行商问题
CH7：路径规划问题
CH8：集覆盖问题
CH9：车间调度问题
CH10：最小生成树问题
CH:11：最大割问题
前沿研究
CH12：多臂老虎机问题
CH13：报童问题
启发式方法
遗传算法

个人介绍—朱晗

- 工学学士 东北大学（2008-2012） 自动化专业 专业排名：1%
- 管理学博士 香港城市大学（2012-2017） 商学院 管理科学系
- 博士后研究员 加拿大麦吉尔大学（2017-2019） 管理学院运营管理系
- 助理教授 东北财经大学（2019-2021）管理科学与工程学院
- 教授 东北财经大学 （2021-至今）管理科学与工程学院
- 博士生导师、副院长、校学术委员会委员（2022-至今）东北财经大学
- 全校最年轻也是目前唯一的“90后”教授、博士生导师、处级干部
- 国家级人才——国家“优青”：全校唯一，东北地区近五年唯一的管理科学部优青
- 辽宁省向上向善好青年、辽宁省百千万人才工程、大连青年五四奖章、大连市高端人才
- **招收研究生**：运筹优化与博弈论方向 欢迎联系 hanzhu@dufe.edu.cn

岁月有着无声的力量，
陪伴才是最长情的告白。

祝大家 心怀梦想 眼沐春光 鹏程万里 一生向阳！