多臂老虎机（论文研读）

2024-2025 学年春夏学期计算经济学讨论班

郑涵文

zhwen@zju.edu.cn

2025 年 4 月 18 日

Introduction

在前面的多臂老虎机问题中，我们研究了有限个老虎臂的情况．现在我们将视角转向无限，即我们的“老虎臂”有无穷个．

问题模型

我们考虑这样一个问题模型．一个卖家有非常多的、一样的商品，现在有依次有 n 个买家前来购买，卖家需要对每个买家进行报价 b．每个买家都有自己的估值 v，如果 $b < v$ 那么买家会支付 b 以买下物品，否则不进行购买．我们从卖家的视角研究问题，即怎样能够使得收益最大化？卖家需要根据前面的买卖结果找出买家估值的规律，并不断完善自己的报价以最接近买家的可能估值，从而获得到最大的利益．自然，买家的估值也是要遵循一定的规律的，在论文中研究了三种情况．

估值模型

1. 完全相同：所有买家具有相同估值 $p \in [0, 1]$．
2. 随机：所有买家估值独立同分布，服从于 $[0, 1]$ 上的某个概率分布．
3. 最差情况（对抗）：买家没有预设估值．买家知道卖家的策略并可以在卖家报价之前任意选择估值．

我们这里依然使用遗憾值去衡量卖家策略的优劣，通常遗憾值是指最优固定策略收益与实际策略收益的差值．

估值相同模型

定理

在所有买家估值相同的情况下，存在一个定价策略使得遗憾值为 $O(\log\log n)$，并且没有定价策略使得遗憾值为 $o(\log\log n)$.

上界

一个显然的结论就是在估值相同模型下，最优固定策略的损失为 0（直接定价为估值）. 我们首先研究上界，即构造一个策略使得遗憾值为 $O(\log\log n)$.

策略

策略的核心思路是维护一个可行区间 $[a,b]$，其初始值为 $[0,1]$. 与此同时维护一个精度常数 ϵ，初始为 $\frac{1}{2}$. 在算法的单个阶段中，卖家依次报价 $a, a+\epsilon, \ldots$ 直到某个报价被拒绝. 假设 $a+k\epsilon$ 是最后一个被接受的报价，那说明可行区间可以缩小到 $[a+k\epsilon, a+(k+1)\epsilon]$. 同时，让精度常数变为 ϵ^2，随后进行下个阶段. 当可行区间长度小于 $\frac{1}{n}$ 时，停止上述操作，转而以 a 的价格出售商品给剩下的所有买家.
我们可以证明上述策略达到了 $O(\log\log n)$ 的遗憾.

上界

我们首先研究上界，即构造一个策略使得遗憾值为 $O(\log\log n)$.

证明

首先算法总共花费的阶段数是 $O(\log\log n)$ 的. 因为区间长度从 $\frac{1}{2}$ 到 $\frac{1}{n}$，每次区间长度都进行平方.

我们考虑遗憾值的来源:

1. 物品以 $q < p$ 的价格卖出去了，遗憾值为 $p - q$;
2. 物品的报价被拒绝了，遗憾值为 p.

在每个阶段中，最多有一个物品被拒绝，因此遗憾值为 $p < 1$，总共有 $O(\log\log n)$ 个阶段，因此产生遗憾值 $O(\log\log n)$.

除了第一个阶段和最后一个阶段，可行区间的长度 $b - a = \sqrt{\epsilon}$，并且在这个阶段中每个子区间的长度为 ϵ，因此最多进行 $\frac{\sqrt{\epsilon}}{\epsilon} = \frac{1}{\epsilon}$ 次报价并被接受，遗憾值小于 $(b-a)\frac{\epsilon}{=}1$，总共有 $O(\log\log n)$ 个阶段，因此产生遗憾值 $O(\log\log n)$.

对于第一个阶段显然遗憾值小于 1，对于最后一个阶段，可行区间长度小于 $\frac{1}{n}$ 并且最多 n 个报价，遗憾值也小于 1.

综上证明了遗憾值为 $O(\log\log n)$.

下界

接下来考虑下界，即证明不存在策略使得遗憾值为 $o(\log\log n)$．具体而言，定理的表述如下

定理

对于任意的随机定价策略 S，并且所有买家估值 p 服从 $[0,1]$ 的均匀分布且相等，那么期望遗憾值为 $\Omega(\log\log n)$.

下界

事实上这等价于证明上述定理对任意确定性定价策略成立，因为随机策略是确定性策略的概率混合．我们仍然使用可行区间这一说法，对于可行区间 $[a, b]$ 显然理性的卖家不会报出偏离此区间的价格．所以，我们可以断言，所有策略都可以写成这样的形式：卖家在可行区间内提供一串逐渐递增的报价，直到某个报价被拒绝；此时更新可行区间，重复上述过程进行递增报价．由于连续性，理论上这个过程能够无限进行下去．

更具体地，我们把策略写成这样的形式（容易证明任意策略都能被写成这样）：把报价用阶段划分（从阶段 0 开始）：阶段 k 在阶段 $k-1$ 结束后立即开始，并在该阶段已经接受了 $2^{2^k}-1$ 个买家（从而限制阶段的长度）或该阶段第一次被拒绝时，结束阶段 k.

下界

Claim

记 $\mathbb{I}_k$ 表示阶段 k 开始时的可能的当前可行区间 $[a, b]$ 构成的集合．该集合大小最多是 2^{2^k}．

证明

对 k 进行归纳．当 $k = 0$ 的时候区间为 $[0, 1]$ 显然成立．假设对 k 成立，并假设在阶段 k 开始时的可行区间为 $[a_k, b_k]$，由于阶段 k 最多可以接受 $2^{2^k} - 1$ 个买家，也就是说，区间可以切 $2^{2^k} - 1$ 刀，得到 2^{2^k} 个子区间，每个子区间都是可能的下一阶段的可行区间．同时，根据我们的归纳，阶段 k 开始时的可行区间有 2^{2^k} 个，所以阶段 $k + 1$ 开始时的可行区间有 $2^{2^{k+1}}$ 个．

下界

Claim

记 $|I|$ 表示区间 I 的长度，我们断言，$|I_k| \geq \frac{1}{4} \cdot 2^{-2^k}$ 以至少 $\frac{3}{4}$ 的概率成立.

证明

我们考虑计算 $1/|I_k|$ 的期望. 事实上，在确定策略的情况下，只有估值 p 是随机变量. 因此我们在关于 p 求期望，有 $E(1/|I_k|) = \sum_{I \in \mathbb{I}_k} \Pr(p \in I)(1/|I|) = \sum |I|/|I| = \sum_{I \in \mathbb{I}_k} 1 \leq 2^{2^k}$.

使用马尔可夫不等式，有

$$\Pr(|I_k| < \frac{1}{4} \cdot 2^{-2^k}) = \Pr(1/|I_k| > 4 \cdot 2^{2^k}) < 1/4$$

下界

Claim

阶段 k 的期望遗憾至少是 $\frac{1}{64}$.

证明

用 E_k 表示如下事件：$p \geq \frac{1}{4}$ 并且 $|I_k| \geq \frac{1}{4} \cdot 2^{-2^k}$. 这是两个事件的交，并且每个事件都有至少 $\frac{3}{4}$ 的概率成立，因此 E_k 至少有 $\frac{1}{2}$ 概率成立. 因此我们只需要证明在 E_k 成立的情况下，阶段 k 的期望遗憾至少是 $\frac{1}{32}$. 以下讨论均假设 $p \geq \frac{1}{4}$ 并且 $|I_k| \geq \frac{1}{4} \cdot 2^{-2^k}$.

记 m 为 I_k 的中点. 方便起见，记 $j = 2^{2^k} - 1$，$x_1 \leq x_2 \leq \cdots \leq x_j$ 表示阶段 k 提供的若干递增报价（如果没有被拒绝的话）.

下界

证明（续）

讨论两种情况：

1. $x_j \geq m$. 由于 p 落在 $I_k = [a, b]$ 上，因此 $p < m$ 的概率为 $\frac{1}{2}$. 并且在这种情况下，肯定会出现拒绝报价的结果，它导致了大小为 p 的遗憾，根据我们的假设有 $p > \frac{1}{4}$，因此这个情况的遗憾值至少是 $\frac{1}{8} > \frac{1}{32}$.
2. $x_j < m$. 同理 $p > m$ 的概率为 $\frac{1}{2}$，在这种情况下，不会出现拒绝报价. 并且，可以计算条件期望 $E(p - m | p > m) = |I_k|/4 \geq 2^{-2^k}/16$. 由于所有报价都被接受，造成的遗憾值至少是 $(2^{2^k} - 1)(2^{-2^k})/16$，总遗憾值至少是 $(2^{2^k} - 1)(2^{-2^k})/32 \approx \frac{1}{32}$（事实上它小于 $1/32$ 但是由于我们最开始的放缩把 $9/16$ 放到了 $1/2$ 所以结论肯定是对的）.

所以我们证明了每个阶段的期望遗憾值是 $\Omega(1)$ 的，并且总阶段数是 $\Omega(\log \log n)$ 的，因此总的期望遗憾值是 $\Omega(\log \log n)$ 的.

随机估值模型

接下来我们研究更复杂的情况．现在每个买家估值不一样了，而是独立同分布的．估值的分布遵循需求曲线

$$D(x) = \Pr(v \geq x)$$

在知道需求曲线的情况下，很容易想到怎么获得期望下的最大利润．因为 $D(x)$ 实际上就是卖家报价 x 时买家接受的概率，那么期望收益为 $xD(x)$．则 $x^* = \arg\max_{x\in[0,1]} xD(x)$ 使得卖家期望收益最大化．我们记这个策略为 S^*，它的期望收益为 $\rho(S^*)$．容易发现对于任意的在线定价策略 S，都有

$$\rho(S) \leq \rho(S^*) \leq \rho(S^{\text{opt}})$$

这里的 S^{opt} 可能会引起疑惑．但事实上它就是所有固定定价策略中最优的，而 S^* 本身就是一个固定定价策略，它自然不优于 S^{opt}，它甚至不太可能等于 S^{opt}，因为它只是期望情况下的最优．

接下来我们将给出一个 $\rho(S^*) - \rho(S)$ 的下界和一个 $\rho(S^{opt}) - \rho(S)$ 的上界．

决策树

确定性定价策略可以通过一系列有根平面二叉树 $T_1, T_2, \ldots$ 来指定，其中第 n 棵树指定了卖家在与 n 个买家交互时所采用的决策树．(因此，T_n 是一棵深度为 n 的完全二叉树．) 我们将用 a 表示这样一个决策树中的一个通用内部节点，用 ℓ 表示一个通用叶节点．关系 $a \prec b$ 表示 b 是 a 的后代；这里 b 可以是叶节点或另一个内部节点．如果 e 是树 T 的一条边，我们还将使用 $a \prec e$ (分别 $e \prec a$) 来表示 e 在 T 中位于 a 的下方 (上方)，即 e 的至少一个端点是 a 的后代 (祖先)．以 a 为根的左子树记为 $T_l(a)$，右子树记为 $T_r(a)$．注意，$T_l(a)$ ($T_r(a)$) 包括从 a 到其左 (右) 子节点的边．

树的内部节点被标记为数字 $x_a \in [0,1]$，表示卖家在节点 a 处提供的价格，以及随机变量 $v_a \in [0,1]$，表示卖家在该节点与之交互的买家的估值．买家的选择由随机变量表示：

$$\chi_a = \begin{cases} 1 & \text{如果} v_a \geq x_a, \\ 0 & \text{如果} v_a < x_a. \end{cases}$$

换句话说，如果买家接受提供的价格，则 χ_a 为 1，否则为 0．

下界

在随机估值模型下，买家估值的自由度很大，可以是任意的概率分布，这显然不利于研究，所以这里对其概率分布（需求曲线）进行一些限制，并研究在这种情况下的结果．
需求曲线 D 均来自集合 $\mathcal{D}$．$\mathcal{D}$ 表示一个单参数需求曲线族 $\{D_t : 0.3 \le t \le 0.4\}$，定义如下．
设：

$$\tilde{D}_t(x) = \max\left\{1 - 2x, \frac{2t}{7} - \frac{t}{2}x, 1 - \frac{2}{x}\right\}.$$

换句话说，$\tilde{D}_t$ 的图由三条线段组成：中间的线段在点 $(t, 1/(7t))$ 处与曲线 $xy = 1/7$ 相切，而左右线段属于位于该曲线下方且与 t 无关的直线．现在我们通过对 $\tilde{D}_t$ 进行平滑处理来获得 D_t．具体来说，设 $b(x)$ 是一个非负的、偶的 C^∞ 函数，其支撑在区间 $[-0.01, 0.01]$ 上，并满足：

$$\int_{-0.01}^{0.01} b(x)\, dx = 1.$$

通过将 $\tilde{D}_t$ 与 b 进行卷积来定义 D_t，即：

$$D_t(x) = \int_{-\infty}^{\infty} \tilde{D}_t(y) b(x - y)\, dy.$$

下界

我们将通过指定 t 在 $[0.3, 0.4]$ 上均匀分布，为 $D = \{D_t : 0.3 \le t \le 0.4\}$ 赋予一个概率测度. 设 $x_t^* = \arg\max_{x\in[0,1]} xD_t(x)$. 可以计算得出 $x_t^* = t$.（如果用 $\tilde{D}_t$ 代替 D_t，这将是显而易见的. 现在，除非 x 在 $\tilde{D}_t'$ 不连续的两个点附近 0.01 范围内，否则 $D_t(x) = \tilde{D}_t(x)$，而这两个点远离使 $x\tilde{D}_t(x)$ 最大化的点，因此 $xD_t(x)$ 也在 $x = t$ 处达到最大值.）除了使我们能够证明以下引理中指定的性质外，D 的具体构造细节并不重要.

引理 A

存在常数 $\alpha, \beta > 0$ 和 $\gamma < \infty$ 使得对于所有 $D = D_{t_0} \in D$ 和 $x \in [0, 1]$:

1. $\frac{dt}{dt}(x_t^*)\big|_{t=t_0} > \alpha$;
2. $x^* D(x^*) - xD(x) > \beta(x^* - x)^2$;
3. $|\dot{D}(x)/D(x)| < \gamma|x^* - x|$ 且 $|\dot{D}(x)/(1 - D(x))| < \gamma|x^* - x|$;
4. $|D^{(k)}(x)/D(x)| < \gamma$ 且 $|D^{(k)}(x)/(1 - D(x))| < \gamma$，其中 $k = 2, 3, 4$.

这里 x^* 表示 $x_{t_0}^*$，$D^{(k)}(x)$ 表示 $D_t(x)$ 在 $t = t_0$ 处的 k 阶导数，$\dot{D}(x)$ 表示 D 的一阶导数.

D 的具体细节我们并不需要清楚，只需要知道它具有上述性质即可.

下界

对后悔值下界的证明基于以下直觉．如果对需求曲线存在不确定性，那么没有一个单一价格能在所有需求曲线上实现低期望后悔值．上面展示的需求曲线族由单个参数 t 参数化，我们将看到，如果对 t 的不确定性在 ε 的量级，那么每个买家的后悔值为 $\Theta(\varepsilon^2)$．(这一陈述将在下面的引理 3.7 中被精确化．) 因此，为了避免在最后的 $\Theta(n)$ 个买家上累积 $\Theta(\sqrt{n})$ 的后悔值，定价策略必须确保在与最初的 (n) 个买家的交互中将不确定性降低到 $O(n^{-1/4})$．然而——这是证明的关键——我们将展示，提供远离 x^* 的价格比提供接近 x^* 的价格具有更多的信息量，因此降低 t 的不确定性是有成本的．特别是，将不确定性降低到 $O(n^{-1/4})$ 在期望后悔值方面要付出 $\Theta(\sqrt{n})$ 的代价．

为了精确化这些想法，我们将引入“知识”的概念，它量化了卖家根据过去交易获得的信息来区分实际需求曲线与附近曲线的能力，以及“条件后悔”的概念，其期望值是定价策略期望后悔值的下界．我们将证明条件后悔与知识的比率有下界，因此策略无法在不累积 $\Theta(\sqrt{n})$ 后悔值的情况下累积 $\Theta(\sqrt{n})$ 知识．最后，我们将证明，当期望知识小于 $\sqrt{n}$ 的一个小的常数倍时，对真实需求曲线存在如此多的不确定性，以至于期望后悔值以高概率为 $\Theta(\sqrt{n})$ (在需求曲线的概率测度上)．

下界

在以下定义中，log 表示自然对数函数．T 表示一棵有限的平面二叉树，按照第 3.1 节中解释的定价策略进行标记．当 f 是定义在 T 的叶节点上的函数时，我们将使用符号 $E_D f$ 表示 f 在叶节点上的概率分布 p_D 下的期望值，即：

$$E_D f = \sum_{\ell \in T} p_D(\ell) f(\ell).$$

对于给定的需求曲线 $D = D_{t_0}$，我们定义叶节点 $\ell \in T$ 的**无穷小相对熵**为：

$$IRE_D(\ell) = \frac{d}{dt}(-\log p_{D_t}(\ell))\Big|_{t=t_0},$$

并定义 ℓ 的**知识**为无穷小相对熵的平方：

$$K_D(\ell) = (IRE_D(\ell))^2.$$

下界

$IRE_D(\ell)$ 的一个重要特性是它可以表示为从根节点到 ℓ 的路径上的边的项的和. 对于边 $e = (a, b) \in T$, 设:

$$ire_D(e) = \begin{cases} \frac{d}{dt}(\log D(x_a)) & \text{如果} e \in T_r(a), \\ \frac{d}{dt}(\log(1 - D(x_a))) & \text{如果} e \in T_l(a) \end{cases}$$

即:

$$ire_D(e) = \begin{cases} \frac{\dot{D}(x_a)}{D(x_a)} & \text{如果} e \in T_r(a), \\ -\frac{\dot{D}(x_a)}{1-D(x_a)} & \text{如果} e \in T_l(a). \end{cases}$$

那么有

$$IRED(\ell) = \sum_{e \prec \ell} ire_D(e).$$

下界

设

$$r_D(x) = x^* D(x^*) - xD(x),$$

其中 $x^* = \arg\max_{x\in[0,1]}\{xD(x)\}$. 注意，如果两个不同的卖家分别以价格 x^* 和 x 向一个估值分布符合 D 的买家提供商品，那么 $r_D(x)$ 是他们期望收入的差异. 现在定义

$$R_D(\ell) = \sum_{a\prec\ell} r_D(x_a).$$

虽然 $R_D(\ell)$ 并不等于卖家在结果 ℓ 条件下的实际后悔值，但它是一个有用的不变量，因为 $E_D R_D(\ell)$ 等于策略 S 相对于 S^* 的实际期望后悔值.
我们有如下引理:

引理

设 S 是一个具有决策树 T 的策略，S^* 是一个固定价格策略，向每个买家提供价格 x^*. 如果买家的估值是来自 D 指定的分布的独立随机样本，那么 S^* 的期望收入超过 S 的期望收入的值恰好等于 $E_D R_D(\ell)$.

下界

在阐述以下引理时，我们将引入常数 $c_1, c_2, \ldots$. 当我们引入这些常数时，我们隐含地断言存在一个仅依赖于需求曲线族 $\mathcal{D}$ 的常数 $0 < c_i < \infty$，并且该常数满足相应引理中指定的性质. 我们首先通过一系列引理来证明 $E_D K_D$ 被 $E_D R_D$ 的一个常数倍数限制了上界. 假设 $\mathcal{D}$ 是固定的，因此 x^* 也是固定的，并设 $h_a = x_a - x^*$.

引理 & 推论

1. $E_D R_D(\ell) \geq c_1 \sum_{a \in T} p_D(a) h_a^2$.
2. $E_D K_D(\ell) \leq c_2 \sum_{a \in T} p_D(a) h_a^2$.

根据上述两个引理我们能得到推论

$$E_D K_D(\ell) \leq c_3 E_D R_D(\ell).$$

下界

上述引理可以用于证明下面的重要引理：

引理

1. 对于所有足够大的 n，如果 $E_D R_D(\ell) < \sqrt{n}$，则存在一个叶节点集合 S，使得 $p_D(S) \geq 1/2$，并且对于所有 $\ell \in S$ 和 $t \in [t_0, t_0 + n^{-1/4}]$，有 $p_{D_t}(\ell) > c_4 p_D(\ell)$.
2. 对于所有 M 和足够大的 n，如果 $E_D R_D < c(M)\sqrt{n}$，那么对于所有 $t \in [t_0 + (1/M)n^{-1/4}, t_0 + n^{-1/4}]$，有 $E_{D_t} R_{D_t} > c(M)\sqrt{n}$.

下界

最终，我们可以根据前面的引理证明最后的结论．

定理

设 S 是任何随机的非均匀策略，$R_D(S,n)$ 表示 S 在 n 个买家群体上的期望事前后悔值，这些买家的估值是根据需求曲线 D 指定的概率分布的独立随机样本．则：

$$\Pr_{D\leftarrow\mathcal{D}}\left(\limsup_{n\to\infty}\frac{R_D(S,n)}{\sqrt{n}}>0\right)=1.$$

换句话说，如果 D 是从 $\mathcal{D}$ 中随机抽取的，那么几乎可以肯定 $R_D(S,n)$ 不是 $o(\sqrt{n})$．

后续，我们能够把 $\mathcal{D}$ 扩展到更一般的曲线 D，并且仍然能够得到类似的结论．至此，我们就证明了遗憾至少是 $\sqrt{n}$ 级别的．接下来我们研究上界．

上界

这里的上界是基于多臂老虎机问题中的算法得到的．我们考虑一个离散的策略，卖家的报价集合只有 $\{1/K, 2/K, ..., 1-1/K, 1\}$，其中 K 是卖家可以自己随意选择的（事实上 $K = \Theta((n/\log n)^{1/4})$ 是最优的）．这事实上就是一个多臂老虎机，卖家选择某个报价能够得到一定的收益，卖家需要找出哪个报价能够获得最大的收益．

我们把 $\{1/K, 2/K, ..., 1-1/K, 1\}$ 这 K 个行动依次编号为 $1, ..., K$，定义 μ_i 为行动 i 的期望收益，并记 $\mu^* = \max\{\mu_1, ..., \mu_K\}$，$\Delta_i = \mu^* - \mu_i$．

有了以上的定义，我们可以得到以下的定理：

定理

存在一种名为 ucb1 的算法，对于所有 $K > 1$，如果在具有任意奖励分布 $P_1, \dots, P_K$ 的 K 个动作集上运行 ucb1，且这些奖励分布的支持集在 $[0, 1]$ 内，那么在任何次数 n 的运行后，其期望后悔值最多为：

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

上界

接下来我们将证明若干引理，以最终确定 $S^*, S^{opt}, UCB1$ 等策略之间的收益关系，最终证明 UCB1 策略具有较好的遗憾上界.

引理 3.11

存在常数 C_1 和 C_2，使得对于所有 $x \in [0,1]$，

$$C_1(x^* - x)^2 < f(x^*) - f(x) < C_2(x^* - x)^2.$$

证明

$f''(x^*)$ 的存在性和严格负性保证了存在常数 A_1、A_2 和 $\varepsilon > 0$，使得对于所有 $x \in (x^* - \varepsilon, x^* + \varepsilon)$ 有 $A_1(x^* - x)^2 < f(x^*) - f(x) < A_2(x^* - x)^2$.
集合 $X = \{x \in [0,1] : |x^* - x| \geq \varepsilon\}$ 的紧致性，以及 $f(x^*) - f(x)$ 对所有 $x \in X$ 严格为正的事实，保证了存在常数 B_1 和 B_2，使得对于所有 $x \in X$，

$$B_1(x^* - x)^2 < f(x^*) - f(x) < B_2(x^* - x)^2.$$

现在取 $C_1 = \min\{A_1, B_1\}$ 和 $C_2 = \max\{A_2, B_2\}$，即可得到该引理.

上界

推论 3.12

对于所有 i，有 $\Delta_i \geq C_1(x^* - i/K)^2$. 如果 $\tilde{\Delta}_0 \leq \tilde{\Delta}_1 \leq \ldots \leq \tilde{\Delta}_{K-1}$ 是集合 $\{\Delta_1, \ldots, \Delta_K\}$ 按升序排列的元素，那么 $\tilde{\Delta}_j \geq C_1(j/(2K))^2$.

证明

不等式 $\Delta_i \geq C_1(x^* - i/K)^2$ 是使用引理 3.11 中的公式对 Δ_i 和 μ_i 的重新表述. 对 $\tilde{\Delta}_j$ 的下界估计是基于观察到集合 $\{1/K, 2/K, \ldots, 1\}$ 中最多有 j 个元素位于 x^* 的 $j/(2K)$ 范围内.

上界

推论 3.13

$\mu^* > x^* D(x^*) - C_2/K^2$.

证明

在集合 $\{1/K, 2/K, \ldots, 1\}$ 中至少有一个数位于 x^* 的 $1/K$ 范围内；现在应用引理 3.11 中关于 $f(x^*) - f(x)$ 的上界.

将所有这些结果综合起来，我们得到了以下上界.

定理 3.14

假设函数 $f(x) = xD(x)$ 在 $x^* \in (0, 1)$ 处有一个唯一的全局最大值，并且 $f'(x^*)$ 存在且严格为负，那么选择 $K = \lceil (n/\log n)^{1/4} \rceil$ 的 ucb1 策略实现的期望后悔值为 $O(\sqrt{n \log n})$.

上界

证明

我们要证明 UCB1 和 opt 策略的关系，这里选择考虑一些中间策略，间接地得到我们想要的关系．考虑以下四种策略：

1. ucb1
2. S_{opt}，最优固定价格策略．
3. S^*，向每个买家提供 x^* 的固定价格策略．
4. S_K^*，向每个买家提供 i^*/K 的固定价格策略，其中 i^*/K 是集合 $\{1/K, 2/K, \ldots, 1\}$ 中最接近 x^* 的元素．

我们使用 $\rho(\cdot)$ 表示策略获得的期望收入．我们将证明 $\rho(S_K^*) - \rho(\text{ucb1})$、$\rho(S^*) - \rho(S_K^*)$ 和 $\rho(S_{\text{opt}}) - \rho(S^*)$ 的 $O(\sqrt{n \log n})$ 上界，那么定理将得证．

上界

证明（续）

首先，我们使用定理 3.10 证明 $\rho(S_K^*) - \rho(\text{ucb1}) = O(\sqrt{n \log n})$．根据推论 3.12，

$$\sum_{i:\mu_i<\mu^*} \left(\frac{1}{\Delta_i}\right) < \sum_{i=1}^{K} \frac{1}{C_1} \left(\frac{2K}{i}\right)^2 < \frac{4K^2}{C_1} \sum_{i=1}^{\infty} \frac{1}{i^2} = \frac{4K^2}{C_1} \cdot \frac{\pi^2}{6} = O((n/\log n)^{1/2}).$$

将这些估计代入定理 3.10 的后悔值上界，我们发现 ucb1 相对于 S_K^* 的后悔值为 $O(\sqrt{n \log n})$.
接下来，我们限制 $\rho(S^*) - \rho(S_K^*)$ 的差异．S^* 和 S_K^* 的期望收入分别为 $nx^* D(x^*)$ 和 $n\mu^*$．应用推论 3.13，S_K^* 相对于 S^* 的后悔值上限为

$$\frac{C_2 n}{K^2} \leq \frac{C_2 n}{(n/\log n)^{1/2}} = O(\sqrt{n \log n}).$$

上界

证明（续）

最后，我们必须限制 $\rho(S_{\mathsf{opt}}) - \rho(S^*)$．对于任何 $x \in [0,1]$，设 $\rho(x)$ 表示以固定价格 x 提供时获得的收入，且 $x_{\mathsf{opt}} = \arg\max_{x\in[0,1]} \rho(x)$．我们首先观察到，对于所有 $x < x_{\mathsf{opt}}$，

$$\rho(x) \geq \rho(x_{\mathsf{opt}}) - n(x_{\mathsf{opt}} - x).$$

这是因为接受价格 x_{opt} 的每个买家也会接受价格 x，而设置较低价格所损失的收入是 $x_{\mathsf{opt}} - x$.

现在有

$$\Pr(\rho(x) - \rho(x^*) > \lambda) \geq \Pr(\rho(x_{\mathsf{opt}}) - \rho(x^*) > 2\lambda \text{ and } x_{\mathsf{opt}} - x < \lambda/n)$$

$$= \frac{\lambda}{n} \Pr(\rho(x_{\mathsf{opt}}) - \rho(x^*) > 2\lambda),$$

不等式成立是因为左边的事件包含了右边的事件．因此，对固定 x 的概率界限可以转化为对 $\Pr(\rho(x_{\mathsf{opt}}) - \rho(x^*) > \lambda)$ 的界限．

上界

证明（续）

对于固定 x，所讨论的概率是 n 个独立同分布的随机变量之和超过 λ 的概率，每个随机变量的取值范围在 $[-1,1]$ 且期望为负．这适用切诺夫-霍夫丁界，我们可以得到：

$$\Pr(\rho(x)-\rho(x^*)>\lambda)<e^{-\lambda^2/2n},$$

因此有 $\Pr(\rho(x_{\mathsf{opt}})-\rho(x^*)>2\lambda)<\min\left\{1,\frac{n}{\lambda}e^{-\lambda^2/2n}\right\}$．最后，

$$E(\rho(x_{\mathsf{opt}})-\rho(x^*))<\int_0^\infty \Pr(\rho(x_{\mathsf{opt}})-\rho(x^*)>y)\,dy$$

$$<\int_0^\infty \min\left\{1,\frac{2n}{y}e^{-y^2/2n}\right\}dy$$

$$<\int_0^{\sqrt{4n\log n}} dy+\frac{2n}{\sqrt{4n\log n}}\int_{\sqrt{4n\log n}}^\infty e^{-y^2/2n}dy=O(\sqrt{n\log n})$$

Exp3

在最坏情况估值模型中，我们假设买家的估值是由一个对手选择的，该对手知道 n 和定价策略，但不知道算法的随机选择．我们知道多臂老虎机中也有类似的对抗模型，并且 Exp3 算法是一个很经典的对抗策略算法，其实现的后悔值为 $O(\sqrt{nK\log K})$，其中 K 是可能动作的数量，并且他们展示了后悔值的下界为 $\Omega(\sqrt{nK})$．这里我们将参考 Exp3 算法，对最差情况的上下界进行分析．

上界

定理

如果在 n 步中运行 Exp3 算法，且每一步每个动作的奖励属于 $[0,1]$，那么 Exp3 相对于最佳固定动作的期望后悔值最多为：

$$2(\sqrt{e}-1)\sqrt{nK\log K}.$$

因此，设 S^K_{opt} 表示从集合 X 中选择最佳报价 i^*/K 的固定价格策略，S_{opt} 表示从 $[0,1]$ 中选择最佳报价 x^* 的固定价格策略，我们有以下不等式：

$$\rho(S^K_{\text{opt}})-\rho(S)<\frac{2(\sqrt{e}-1)}{\sqrt{nK\log K}},$$

$$\rho(S^*)-\rho(S^K_{\text{opt}})<\frac{n}{K},$$

第二个不等式源于 S^K_{opt} 不会比向每个买家提供 $\frac{1}{K}\lceil Kx^*\rceil$ 的策略更差.

如果我们选择 $K=\lceil(n/\log n)^{1/3}\rceil$，那么 $\sqrt{nK\log K}$ 和 n/K 都是 $O(n^{2/3}(\log n)^{1/3})$.

因此，我们将 Exp3 的后悔值表示为两个项的和，每个项都是 $O(n^{2/3}(\log n)^{1/3})$，从而确立了上界.

下界

在多臂老虎机问题中，作者在研究下界的时候定义了一个随机收益模型（取决于 n 而不是算法），使得任何算法在来自这个分布的随机样本上的期望后悔值为 $\Omega(\sqrt{nK})$．其思想是随机均匀地选择 K 个动作中的一个，并将其指定为“好”动作．对于所有其他动作，每轮的收益是 $\{0,1\}$ 的均匀随机样本，但对于“好”动作，收益是有偏的样本，以 $1/2+\varepsilon$ 的概率为 1，其中 $\varepsilon=\Theta(\sqrt{K/n})$．一个“好”动作的策略将实现期望收益 $(1/2+\varepsilon)n=1/2+\Theta(\sqrt{nK})$．由于信息论的原因，可以证明没有策略能够足够快速和可靠地学习“好”动作，以在期望中玩它超过 $n/K+\Theta(\varepsilon\sqrt{n^3/K})$ 次，从而得出后悔值的下界．在我们这个在线定价拍卖问题中可以沿用类似的思路，构造买家估值的随机分布（取决于 n 而不是算法），使得任何算法在来自这个分布的随机样本上的期望后悔值为 $\Omega(n^{2/3})$．其思想大致与上述相同：随机选择 $[0,1]$ 的一个长度为 $1/K$ 的子区间作为“好价格”区间，并选择买家估值的分布，使得在“好价格”区间外的每个报价的期望收入是一个与报价无关的常数，而在“好价格”区间内的期望收入比这个常数高 ε．

下界

我们最终能得到如下定理：

定理

对于任何给定的 n，存在一个有限的分布族 $\mathcal{P} = \{p_j^n\}_{j=1}^K$ 在 $[0,1]$ 上的概率分布，使得如果从 $\mathcal{P}$ 中均匀随机选择 p_j^n，然后根据 p_j^n 独立随机采样买家的估值，那么没有定价策略能够实现期望后悔值 $o(n^{2/3})$，其中期望是同时对 D 的随机选择和随机采样的估值取的.