16. 实例依赖性下界

在上一章中，我们证明了在[0, 1]中具有次优性间隙的亚高斯老虎机的极大极小遗憾的下界。这样的界限可以作为衡量决策鲁棒性的有用指标，但通常过于保守。本章致力于理解实例依赖性下界，它试图捕捉决策在特定老虎机实例上的最佳性能。

由于遗憾是一个多目标准则，算法设计者可能会尝试设计在某种实例上表现良好的算法。一个极端的例子是为所有 t 选择 $A_t = 1$ 的决策，当第一个臂为最优时，该决策将遭受零遗憾，否则将遭受线性遗憾。这是一个苛刻的权衡，仅在少数情况下将遗憾从对数减少到零的代价是其他情况下的线性遗憾。令人惊讶的是，这就是老虎机游戏的本质。可以为每个实例指定一个难度度量，这样在某些实例上相对于此度量执行得太好的决策会为其他实例付出高昂的代价。情况如图 16.1 所示

在有限的时间内，情况有点混乱，但如果将这些想法推向极限，那么对于许多类别的老虎机来说，可以定义依赖实例的最优性的精确概念。

个人注 1：次优差距(sub-optimality gap)是指该策略与最优策略之间的性能差距。

个人注 2：本章主要论述了适用于任何非结构化类的随机老虎机的通用下界，论证了有限时间实例依赖性下界，同时提供了 $\mathcal{M}$ 对应的 $d_{\inf}(P,\mu^*,\mathcal{M})$ 的明确公式。

16.1 渐近界

我们需要准确定义合理决策的含义。如果只关心渐近性，那么一个相当保守的定义就足够了。

定义 16.1。如果对于所有 $\nu \in \varepsilon$ 和 $p>0$，在一类老虎机 ε 上称决策 π 是一致的，则有

$$\lim_{n\to\infty} \frac{R_n(\pi,\nu)}{n^p} = 0. \tag{1.1}$$

ε 上的一致决策类用 $\prod_{cons}(\varepsilon)$ 表示。

定理 7.1 表明 UCB 在 $\varepsilon_{\mathrm{SG}}^k(1)$ 上是一致的。总是选择第一个动作的决策在任何 ε 上都是不一致的，除非第一个臂对每个 $\nu \in \varepsilon$ 都是最优的。

一致性是一个渐进的概念。一个决策可以是一致的，但对所有 $t \le 10^{100}$ 都是 $A_t = 1$。因此，一致性假设不足以推导出非渐近下界。在第 16.2 节中，我们介绍了一个有限时间版本的一致性，它允许我们证明有限时间实例依赖性下界。

回想一下，如果 $\varepsilon = \mathcal{M}_1 \times \cdots \times \mathcal{M}_k$ 具有 $\mathcal{M}_1, \cdots, \mathcal{M}_k$ 分布集，则 ε 类随机老虎机是非结构化的。本章的主要定理是一个通用的下界，适用于任何非结构化类的随机老虎机。在证明之后，我们将看到一些特定类的应用程序。设 $\mathcal{M}$ 是一组具有有限平均数的分布，设 $\mu : \mathcal{M} \to \mathbb{R}$ 是将 $P \in \mathcal{M}$ 映射到其平均值的函数。对 $\mu^* \in \mathbb{R}$ 和 $P \in \mathcal{M}$，有 $\mu(P) < \mu^*$ 并定义

$$d_{\inf}(P,\mu^*,\mathcal{M}) = \inf_{P' \in \mathcal{M}} \{ D(P,P') : \mu(P') > \mu^* \}.$$

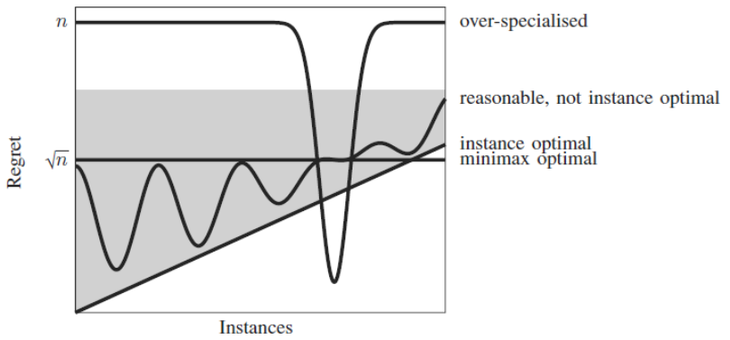

图 16.1 在 x 轴上，实例按照难度的度量进行排序，y 轴上显示的是遗憾（在某种程度上）。在前一章中，我们证明了没有任何决策可以完全低于水平“极大极小最优”线。本章的结果表明，如果某个决策的遗憾在任何时候都低于“实例最优”线，那么对于其他实例，该策略的遗憾必须高于阴影区域。例如，“过度指定”决策。

定理 16.2。令 $\varepsilon = \mathcal{M}_1 \times \cdots \times \mathcal{M}_k$ 和 $\pi \in \prod_{cons}(\varepsilon)$ 是 ε 上的一致决策。对所有 $\nu = (P_i)_{i=1}^k \in \varepsilon$，有

$$\liminf_{n\to\infty} \frac{R_n}{\log(n)} \ge c^*(\nu,\varepsilon) = \sum_{i:\Delta_i>0} \frac{\Delta_i}{d_{\inf}(P_i,\mu^*,\mathcal{M}_i)}, \tag{1.2}$$

式中，Δ_i 是 ν 中第 i 个臂的次优差距，μ^* 是最佳臂的平均值。

证明 设 μ_i 为 ν 中第 i 个臂的平均值，$d_i = d_{\inf}(P_i,\mu^*,\mathcal{M}_i)$。引理 4.5 给出的结果表明，对于任何次优臂 i，它都满足

$$\liminf_{n\to\infty} \frac{\mathbb{E}_{\nu\pi}[T_i(n)]}{\log(n)} \ge \frac{1}{d_i}.$$

固定一个次优臂 i，令 $\varepsilon > 0$ 为任意值，$\nu' = (P_j')_{j=1}^k \in \varepsilon$ 是一个老虎机，有 $P_j' = P_j$ $(j \ne i)$ 和 $P_i' \in \mathcal{M}_i$，后者根据 d_i 的定义存在 $D(P_i,P_i') \le d_i + \varepsilon$ 和 $\mu(P_i') > \mu^*$。令 $\mu' \in \mathbb{R}^k$ 是 μ' 分布均值的向量。根据引理 15.1，我们有 $D(\mathbb{P}_{\nu\pi},\mathbb{P}_{\nu'\pi}) \le \mathbb{E}_{\nu\pi}[T_i(n)](d_i+\varepsilon)$，根据定理 14.2，对于任意事件 A 有，

$$\mathbb{P}_{\nu\pi}(A) + \mathbb{P}_{\nu'\pi}(A^c) \ge \frac{1}{2}\exp\left(-D(\mathbb{P}_{\nu\pi},\mathbb{P}_{\nu'\pi})\right) \ge \frac{1}{2}\exp\left(-\mathbb{E}_{\nu\pi}[T_i(n)](d_i+\varepsilon)\right).$$

再选择 $A=\{T_i(n)>n/2\}$，并且令 $R_n=R_n(\pi,\nu)$ 和 $R'_n=R_n(\pi,\nu')$。接着，

$$\begin{aligned}R_n+R'_n &\geq \frac{n}{2}\Big(\mathbb{P}_{\nu\pi}(A)\Delta_i+\mathbb{P}_{\nu'\pi}(A^c)\big(\mu'_i-\mu^*\big)\Big)\\ &\geq \frac{n}{2}\min\{\Delta_i,\mu'_i-\mu^*\}\Big(\mathbb{P}_{\nu\pi}(A)+\mathbb{P}_{\nu'\pi}(A^c)\Big)\\ &\geq \frac{n}{4}\min\{\Delta_i,\mu'_i-\mu^*\}\exp\Big(-\mathbb{E}_{\nu\pi}\big[T_i(n)\big](d_i+\varepsilon)\Big).\end{aligned}$$

重新变换并引入低限有

$$\begin{aligned}\liminf_{n\to\infty}\frac{\mathbb{E}_{\nu\pi}\big[T_i(n)\big]}{\log(n)} &\geq \frac{1}{d_i+\varepsilon}\liminf_{n\to\infty}\frac{\log\left(\dfrac{n\min\{\Delta_i,\mu'_i-\mu^*\}}{4(R_n+R'_n)}\right)}{\log(n)}\\ &=\frac{1}{d_i+\varepsilon}\left(1-\limsup_{n\to\infty}\frac{\log(R_n+R'_n)}{\log(n)}\right)=\frac{1}{d_i+\varepsilon},\end{aligned}$$

其中，最后一个等式来自一致性的定义，即对于任意 $P>0$，存在一个常数 C_p，使得对于足够大的 n，有 $R_n+R'_n\leq C_p n^p$。这意味着，

$$\limsup_{n\to\infty}\frac{\log(R_n+R')_n}{\log(n)}\leq\limsup_{n\to\infty}\frac{p\log(n)+\log(C_p)}{\log(n)}=p,$$

由于 $p>0$ 是任意的，取 ε 趋于零的极限，再由引理 4.5 即 $R_n=\sum_{a\in\mathcal{A}}\Delta_a\mathbb{E}\big[T_a(n)\big]$ 可得，

$$\begin{aligned}\liminf_{n\to\infty}\frac{R_n}{\log(n)} &=\liminf_{n\to\infty}\sum_{i:\Delta_i>0}\frac{\Delta_i\mathbb{E}\big[T_i(n)\big]}{\log(n)}\\ &=\sum_{i:\Delta_i>0}\Delta_i\liminf_{n\to\infty}\frac{\mathbb{E}\big[T_i(n)\big]}{\log(n)}\\ &\geq\sum_{i:\Delta_i>0}\frac{\Delta_i}{d_i+\varepsilon}=\sum_{i:\Delta_i>0}\frac{\Delta_i}{d_{\inf}(P_i,\mu^*,\mathcal{M}_i)}=c^*(\nu,\varepsilon).\end{aligned}$$

得证。

表 16.1 提供了常见选择 $\mathcal{M}$ 对应的 $d_{\inf}(P,\mu^*,\mathcal{M})$ 的明确公式。这些量的计算都很简单（练习 16.1）。$c^*(\nu,\varepsilon)$ 的下界和定义是非常基本的量，因为对于大多数类 ε，都存在一个决策 π 使得，

$$\lim_{n\to\infty}\frac{R_n(\pi,\nu)}{\log(n)}=c^*(\nu,\varepsilon)\quad\text{for all }\nu\in\varepsilon \tag{1.3}$$

如果等式（16.3）成立，则可以在类 ε 上调用渐近最优决策。例如，第 8 章中的 UCB 和第 10 章中的 KL-UCB 分别对 $\varepsilon_{\mathcal{N}}^k(1)$ 和 $\varepsilon_{\mathcal{B}}^k$ 是渐近最优的。

表 16.1 当 p 的平均值小于 μ^* 时，不同参数族的 $d_{\inf}$ 表达式

$\mathcal{M}$	P	$d_{\inf}(P,\mu^*,\mathcal{M})$
$\{\mathcal{N}(\mu,\sigma^2):\mu\in\mathbb{R}\}$	$\mathcal{N}(\mu,\sigma^2)$	$\dfrac{(\mu-\mu^*)^2}{2\sigma^2}$
$\{\mathcal{N}(\mu,\sigma^2):\mu\in\mathbb{R},\sigma^2\in(0,\infty)\}$	$\mathcal{N}(\mu,\sigma^2)$	$\dfrac{1}{2}\log\left(1+\dfrac{(\mu-\mu^*)^2}{2\sigma^2}\right)$
$\{\mathcal{B}(\mu):\mu\in[0,1]\}$	$\mathcal{B}(\mu)$	$\mu\log\left(\dfrac{\mu}{\mu^*}\right)+(1-\mu)\log\left(\dfrac{1-\mu}{1-\mu^*}\right)$
$\{\mathcal{U}(a,b):a,b\in\mathbb{R}\}$	$\mathcal{U}(a,b)$	$\log\left(1+\dfrac{2\big((a+b)/2-\mu^*\big)^2}{b-a}\right)$

16.2 有限时间界限

通过对一致性进行有限时间模拟，可以证明有限时间实例依赖性下界。首先，通过将 Bretagnolle-Huber 不等式（定理 14.2）与散度分解引理（引理 15.1）关联得到一条引理。

引理 16.3。 设 $\nu=(P_i)$ 和 $\nu'=(P'_i)$ 为 k 臂随机老虎机，它们仅在动作 $i\in[k]$ 的奖励分布上有所不同。假设 i 在 ν 中次优，在 ν' 中唯一最优。令 $\lambda=\mu_i(\nu')-\mu_i(\nu)$。对任意决策 π 有，

$$\mathbb{E}_{\nu\pi}\big[T_i(n)\big]\geq\frac{\log\left(\dfrac{\min\{\lambda-\Delta_i(\nu),\Delta_i(\nu)\}}{4}\right)+\log(n)-\log\big(R_n(\nu)+R_n(\nu')\big)}{D(P_i,P'_i)}. \tag{1.4}$$

该引理适用于有限 n 和任意 ν，并可用于推导任何足够丰富的环境类 ε 的有限时间实例依赖性下界。下面的结果提供了高斯老虎机的有限时间实例依赖性下界，其中一致性的渐近概念被极大极小遗憾不是太大的假设所取代。仅此假设就足以表明，在任何情况下，任何接近极大极小最优的决策都不可能比 UCB 好得多。

定理 16.4。 设 $\nu\in\varepsilon_{\mathcal{N}}^k$ 为具有平均向量 $\mu\in\mathbb{R}^k$ 和次优差距 $\Delta\in[0,\infty)^k$ 的 k 臂高斯老虎机。令

$$\varepsilon(\nu)=\{\nu'\in\varepsilon_{\mathcal{N}}^k:\mu_i(\nu')\in[\mu_i,\mu_i+2\Delta_i]\}.$$

假设 $C>0$ 和 $p\in(0,1)$ 是常数，π 是一个决策使得对所有 n 和 $\nu'\in\varepsilon(\nu)$ 满足 $R_n(\pi,\nu')\le Cn^p$，有

$$R_n(\pi,\nu)\ge\frac{2}{(1+\varepsilon)^2}\sum_{i:\Delta_i>0}\left(\frac{(1-p)\log(n)+\log\left(\frac{\varepsilon\Delta_i}{8C}\right)}{\Delta_i}\right)^+ \tag{1.5}$$

证明 设 i 在 ν 中次优，选择 $\nu'\in\varepsilon(\nu)$，使得 $\mu_j(\nu')=\mu_j(\nu)\,(j\ne i)$ 和 $\mu_j(\nu')=\mu_i+\Delta_i(1+\varepsilon)$。再基于引理 16.3 及 $\lambda=\Delta_i(1+\varepsilon)$，$D(P_i,P_i')\le d_i+\varepsilon$，$R_n+R_n'\le C_p n^p$ 有，

$$\begin{aligned}
\mathbb{E}_{\nu\pi}[T_i(n)]&\ge\frac{\log\left(\frac{\min\{\lambda-\Delta_i(\nu),\Delta_i(\nu)\}}{4}\right)+\log(n)-\log\left(R_n(\nu)+R_n(\nu')\right)}{D(P_i,P_i')}\\
&\ge\frac{2}{\Delta_i^2(1+\varepsilon)^2}\left(\log\left(\frac{n}{2\left(R_n(\nu)+R_n(\nu')\right)}\right)+\log\left(\frac{\min\{\lambda-\Delta_i,\Delta_i\}}{4}\right)\right)\\
&\ge\frac{2}{\Delta_i^2(1+\varepsilon)^2}\left(\log\left(\frac{n}{2Cn^p}\right)+\log\left(\frac{\min\{\lambda-\Delta_i,\Delta_i\}}{4}\right)\right)\\
&=\frac{2}{\Delta_i^2(1+\varepsilon)^2}\left((1-p)\log(n)+\log\left(\frac{\varepsilon\Delta_i}{8C}\right)\right).
\end{aligned}$$

将其代入到基本遗憾分解恒等式（引理 4.5），即 $R_n=\sum_{a\in\mathcal{A}}\Delta_a\mathbb{E}[T_a(n)]$ 中可得

$$\begin{aligned}
R_n(\pi,\nu)&=\sum_{i:\Delta_i>0}\Delta_i\mathbb{E}[T_i(n)]\\
&\ge\sum_{i:\Delta_i>0}\Delta_i\frac{2}{\Delta_i^2(1+\varepsilon)^2}\left((1-p)\log(n)+\log\left(\frac{\varepsilon\Delta_i}{8C}\right)\right)\\
&=\frac{2}{(1+\varepsilon)^2}\sum_{i:\Delta_i>0}\left(\frac{\left((1-p)\log(n)+\log\left(\frac{\varepsilon\Delta_i}{8C}\right)\right)}{\Delta_i^2}\right)^+
\end{aligned}$$

得证。

当 $p=1/2$ 时，此下界中的前导项约为渐近界的一半。这种影响可能是真实的。所考虑的决策类别大于渐近下界，因此针对给定环境进行最佳调整的决策有可能获得较小的遗憾。

16.3 注释

1. 我们认为对于大多数类 ε，有一个满足等式（16.3）的决策。其形式源自下界，并通过对基础分布进行一些附加假设而得出。有关详细信息，请参见 Burnetas 和 Katehakis[1996]的文章，这也是定理 16.2 的原始来源。
2. 本章中的分析仅适用于非结构化类。如果没有这一假设，决策可能会利用其他手臂来了解一只手臂的回报，这大大减少了遗憾。结构化老虎机的下界更为微妙，将在后续章节中逐案讨论。
3. 表 16.1 中分析的类都是参数化的，这使得分析计算成为可能。在非参数情况下的分析相对较少，但我们知道三种例外情况供读者参考。第一类是具有有界支撑的分布：$\mathcal{M}=\{P:Supp(P)\subseteq[0,1]\}$，已经得到了准确的分析【Honda 和 Takemura，2010 年】。第二类是具有半有界支撑的分布，$\mathcal{M}=\{P:Supp(P)\subseteq(-\infty,1]\}$【Honda 和 Takemura，2015 年】。第三类是具有峰度有界的分布，$\mathcal{M}=\{P:\mathrm{Kurt}_{X\sim P}[X]\le\kappa\}$【Lattimore，2017 年】。

16.4 书目备注

基于一致性假设的渐近最优性首先出现在 Lai 和 Robbins【1985】的开创性论文中，后来被 Burnetas 和 Katehakis【1996】推广。就上界而言，目前存在单参数指数族渐近最优的决策【Cappé 等人，2013 年】。直到最近，还没有关于多参数类回报分布的渐近最优性的结果。对于均值和方差未知的高斯分布【Cowan 等人，2018 年】和均匀分布【Cowan 和 Katehakis，2015 年】，最近在这一问题上取得了一些进展。对于非参数类的回报分布，有许多与渐近最优策略相关的开放性问题。当回报分布为离散且有限支持时，Burnetas 和 Katehakis【1996】给出了一个渐近最优策略，尽管精确常数很难解释。一个相对完整的解决方案适用于具有有限支持的类【Honda 和 Takemura，2010 年】。对于半有界的情况，事情已经变得不明朗【Honda 和 Takemura，2015 年】。其中一位作者认为峰度有界的类非常有趣的，但这里的情况只有在常数因子下才能理解【Lattimore，2017 年】。Salomon 等人【2013 年】提出了定理 16.4 的渐近变体。几位作者提出了有限时间实例依赖性下界，包括 Kulkarni 和 Lugosi【2000】针对双臂，Garivier 等人【2019】和 Lattimore【2018】针对一般情况。如前所述， ETC 决策和基于消除的算法都无法实现渐近最优：如 Garivier 等人[2016b]所作研究，与最优渐近遗憾相比，这些算法（无论如何调整）在标准高斯老虎机问题上都必须产生两倍的额外乘性惩罚。

第一章 文献翻译&个人理解

15 极大极小下界

在短暂的讨论信息论偏移之后，让我们回到 k-armed 随机赌博机上。在下面的内容中，我们规定范围 $n>0$ 和动作次数 $k>1$。本章有两个组成部分。首先是针对固定策略和不同的赌博机，精确计算典型赌博机模型中措施之间的相对熵。在第二部分中，我们证明了一个极大极小下界，将第 13 章给出的直觉论点形式化。

个人注：本章主要讨论决策理论问题中极大极小风险的下界。这种界限对于评估决策规则的质量很有用。

15.1 赌博机之间的相对熵

下面的结果将被反复使用，练习题中提供了一些概括。

引理 15.1：散度分解

令 $\nu=(P_1,\cdots P_k)$ 为一个 k-armed 赌博机相关的奖励分布，令 $\nu'=(P_1',\cdots P_k')$ 为另一个 k-armed 赌博机相关的奖励分布。固定一些政策 π，同时令 $\mathbb{P}_\nu=\mathbb{P}_{\nu\pi}$ 和 $\mathbb{P}_{\nu'}=\mathbb{P}_{\nu'\pi}$ 为由 π 和 ν（π 和 ν'）的 n 轮相互联络所得到的正则赌博机模型（4.6 节）的概率测度。

因此有：

$$\mathrm{D}(\mathbb{P}_\nu,\mathbb{P}_{\nu'})=\sum_{i=1}^{k}\mathbb{E}_\nu[T_i(n)]\,\mathrm{D}(P_i,P_i')$$

证明：

假设对于所有 $i\in[k]$ 都有 $\mathrm{D}(P_i,P_i')<\infty$。由此可见 $P_i\ll P_i'$。定义 $\lambda=\sum_{i=1}^{k}P_i+P_i'$，$\lambda(A)=\sum_{i=1}^{k}(P_i(A)+P_i'(A))$ 为任意可测集 A 定义的度量，定理 14.1 表明，只要 $(\frac{\mathrm{d}\mathbb{P}_\nu}{\mathrm{d}\mathbb{P}_{\nu'}})<+\infty$，则有：

$$\mathrm{D}(\mathbb{P}_\nu,\mathbb{P}_{\nu'})=\sum_{i=1}^{k}\mathbb{E}_\nu[\log(\frac{\mathrm{d}\mathbb{P}_\nu}{\mathrm{d}\mathbb{P}_{\nu'}})]$$

回顾 ρ 是 $[k]$ 上的计数测度，我们发现 $\mathbb{P}_\nu$ 关于乘积测度 $(\rho\times\lambda)^n$ 的 Radon-Nikodym 导数在式（4.7）中给出：

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

除了 p_{a_t} 被 p'_{a_t} 取代外，$\mathbb{P}_\nu$ 的密度相同，则有：

$$\log\frac{\mathrm{d}\mathbb{P}_\nu}{\mathrm{d}\mathbb{P}_{\nu'}}(a_1,x_1,\ldots,a_n,x_n)=\sum_{t=1}^{n}\log\frac{p_{a_t}(x_t)}{p'_{a_t}(x_t)}$$

其中，我们使用了 Radon-Nikodym 衍生品的链规则和涉及策略的条款取消的事实。取双方的期望:

$$\mathbb{E}_\nu[\log(\frac{\mathrm{d}\mathbb{P}_\nu}{\mathrm{d}\mathbb{P}_{\nu'}}(A_1,X_1,\ldots,A_n,X_n)]=\sum_{t=1}^{n}\mathbb{E}_\nu[\log\frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)}]$$

以及

$$\mathbb{E}_\nu[\log\frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)}]=\mathbb{E}_\nu[\mathbb{E}_\nu[\log\frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)}\mid A_t]]=\mathbb{E}_\nu[\mathrm{D}(\mathrm{P}_{A_t},\mathrm{P}'_{A_t})]$$

其中，第二个等式在 $\mathbb{P}_\nu(\cdot\mid A_t)$ 条件下，X_t 的分布是 $dP_{A_t}=p_{A_t}d\lambda$，回插到前式中得：

$$\begin{aligned}\mathbb{E}_\nu[\log(\frac{\mathrm{d}\mathbb{P}_\nu}{\mathrm{d}\mathbb{P}_{\nu'}}(A_1,X_1,\ldots,A_n,X_n)]&=\sum_{t=1}^{n}\mathbb{E}_\nu[\log\frac{p_{A_t}(X_t)}{p'_{A_t}(X_t)}]\\&=\sum_{t=1}^{n}\mathbb{E}_\nu[\mathrm{D}(P_{A_t},P'_{A_t})]=\sum_{i=1}^{k}\mathbb{E}_\nu[\sum_{t=1}^{n}\mathbb{I}\{A_t=i\}\mathrm{D}(P_{A_t},P'_{A_t})]\\&=\sum_{i=1}^{k}\mathbb{E}_\nu[T_i(n)]\mathrm{D}(P_{A_t},P'_{A_t})\end{aligned}$$

当（15.1）式的右边是无穷大时，由我们前面的计算不难看出，左边也是无穷大的。

我们注意到，无论动作集是否离散，发散分解都成立。在更一般的形式下，有关行动的总和必须用适当的非负措施的积分来代替，该积分概括了预期的臂拔出数量。详情见练习 15.8。

个人注：相对熵（relative entropy）又称为 KL 散度（Kullback–Leibler divergence，简称 KLD），信息散度（information divergence），信息增益（information gain）。KL 散度是两个概率分布 P 和 Q 差别的非对称性的度量。 KL 散度是用来 度量使用基于 Q 的编码来编码来自 P 的样本平均所需的额外的位元数。 典型情况下，P 表示数据的真实分布，Q 表示数据的理论分布，模型分布，或 P 的近似分布。简单来说，相对熵用来衡量两个取值为正的函数或概率分布之间的差异

15.2 极大极小下界

记 $\varepsilon_N{}^k(1)$ 是具有单位方差的高斯赌博机类，可以通过它们的平均向量进行参数化 $\mu\in\mathbb{R}^k$，令 $\mu\in\mathbb{R}^k$，ν_μ 为第 i 个有奖励分布为 $N(\mu_i,1)$ 的高斯赌博机。

引理 15.2：

令 $k>1$ 且 $n\geq k-1$，因此对于任何政策 π，存在一个平均向量 $\mu\in[0,1]^k$，使得：

$$R_n(\pi,\nu_\mu)\geq\frac{1}{27}\sqrt{(k-1)n}$$

从而 $\nu_\mu\in\varepsilon_N{}^k(1)$，因此，当 $n\geq k-1$ 时，极大极小边界 $\varepsilon_N{}^k(1)$ 的极小边界当由上述等

式右边主导的边缘下界：

$$R^*_n(\varepsilon_N{}^k(1)) \geq \frac{1}{27}\sqrt{(k-1)n}$$

证明的思想如图 15.1 所示。

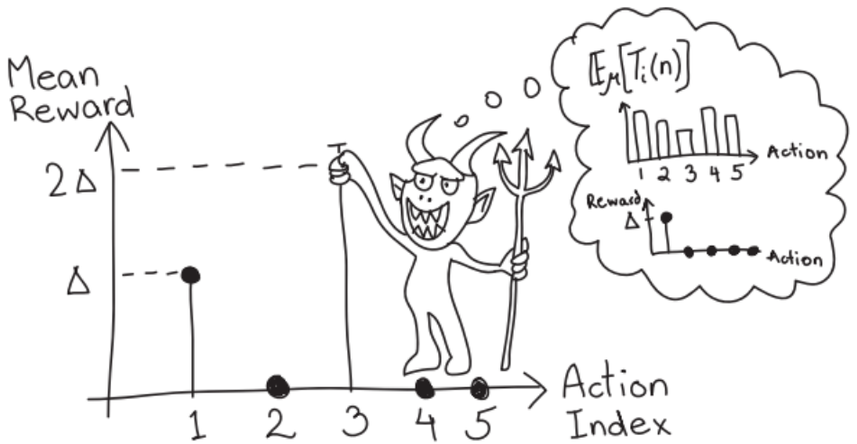

图 15.1 极大极小下界的概念。如果有一个政策和一个环境反对者选择另一个环境，使政策在至少一个环境中遭受巨大的偏差遗憾

证明：

固定策略 π，令 $\Delta \in [0,1/2]$ 是稍后选择的常数。根据第 13 章中的建议，我们从单位方差和平均向量 $\mu = (\Delta, 0, 0, \ldots, 0)$ 的高斯赌博机开始。这种环境下和 π 在正则空间上产生了分布 $\mathbb{P}_{\nu_\mu,\pi}$ 和赌博机模型 (H_n, F_n)。为了简洁起见，我们将使用 $\mathbb{P}_\mu$ 来代替 $\mathbb{P}_{\nu_\mu,\pi}$。同时，$\mathbb{P}_\mu$ 将用 $\mathbb{E}_\mu$ 表示。要选择第二个环境，令

$$i = \arg\min_{j>1} \mathbb{E}_\mu[T_i(n)]$$

从而，$\sum_{j=1}^k \mathbb{E}_\mu[T_i(n)] = n$，并且 $\mathbb{E}_\mu[T_i(n)] \leq n/(k-1)$。第二个高斯赌博机也是单位方差且均值为

$$\mu' = (\Delta, 0, 0, \ldots, 0, 2\Delta, 0, \ldots, 0)$$

其中 $\mu'_i = 2\Delta$。因此 $\mu_j = \mu'_j$，除了指数 i 和 ν_μ 是第一个臂，而在 ν_μ 中，臂 i 是最佳的。我们缩写 $\mathbb{P}_{\mu'} = \mathbb{P}_{\nu_{\mu'},\pi}$。将引理 4.5 进行一个简单的计算得到

$$R_n\left(\pi, \nu_\mu\right) \geq \mathbb{P}_\mu\left(T_1(n) \leq n/2\right)\frac{n\Delta}{2} \quad \text{and} \quad R_n\left(\pi, \nu_{\mu'}\right) > \mathbb{P}_{\mu'}\left(T_1(n) > n/2\right)\frac{n\Delta}{2}$$

然后，应用前一章中的 Bretagnolle–Huber 不等式（定理 14.2），

$$\begin{aligned} R_n\left(\pi, \nu_\mu\right) + R_n\left(\pi, \nu_{\mu'}\right) &> \frac{n\Delta}{2}\left(\mathbb{P}_\mu\left(T_1(n) \leq n/2\right) + \mathbb{P}_{\mu'}\left(T_1(n) > n/2\right)\right) \\ &\geq \frac{n\Delta}{4}\exp\left(-\mathrm{D}\left(\mathbb{P}_\mu, \mathbb{P}_{\mu'}\right)\right) \end{aligned}$$

它保持在上限 $\mathrm{D}\left(\mathbb{P}_\mu, \mathbb{P}_{\mu'}\right)$。为此，我们使用引理 15.1 和 μ, μ' 获得

$$\mathrm{D}\left(\mathbb{P}_\mu, \mathbb{P}_{\mu'}\right) = \mathbb{E}_\mu\left[T_i(n)\right]\mathrm{D}(\mathcal{N}(0,1), \mathcal{N}(2\Delta,1)) = \mathbb{E}_\mu\left[T_i(n)\right]\frac{(2\Delta)^2}{2} \leq \frac{2n\Delta^2}{k-1}$$

将此插入之前的显示，我们发现

$$R_n\left(\pi, \nu_\mu\right) + R_n\left(\pi, \nu_{\mu'}\right) \geq \frac{n\Delta}{4}\exp\left(-\frac{2n\Delta^2}{k-1}\right)$$

通过选择 $\Delta = \sqrt{(k-1)/4n} \leq 1/2$，其中不等式遵循定理陈述中的假设。最后一步是下限 $\exp(-1/2)$ 和 $2\max(a,b) \geq a+b$。

我们鼓励读者阅读练习 15.2 中概述的替代证明，其中采取了稍微不同的路径。

15.3 注记

1、我们使用高斯噪声模型，因为 KL 发散度在这种情况下很容易计算但我们实际使用的是 $\mathrm{D}\left(P_i, P_i'\right) = O\left(\left(\mu_i - \mu_i'\right)^2\right)$，当平均值之间的差距 $\Delta = \mu_i - \mu_i'$ 很小。虽然并非所有情况都是这样，但通常情况下确实如此。为什么呢？令 $\{P_\mu : \mu \in \mathbb{R}\}$ 是 Ω 上的一些参数分布族，并假设分布 p_μ 具有平均 μ。假设密度是两次可微的，并且所有的都是积分和导数可以交换（几乎总是这样），我们可以用一个关于 μ 的泰勒展开式来表示

$$
\begin{aligned}
\mathrm{D}\left(P_\mu, P_{\mu+\Delta}\right) &\approx \frac{\partial}{\partial\Delta}\mathrm{D}\left(P_\mu, P_{\mu+\Delta}\right)\Big|_{\Delta=0}\Delta + \frac{1}{2}\frac{\partial^2}{\partial\Delta^2}\mathrm{D}\left(P_\mu, P_{\mu+\Delta}\right)\Big|_{\Delta=0}\Delta^2 \\
&= \frac{\partial}{\partial\Delta}\int_\Omega \log\left(\frac{dP_\mu}{dP_{\mu+\Delta}}\right)dP_\mu\Big|_{\Delta=0}\Delta + \frac{1}{2}I(\mu)\Delta^2 \\
&= -\int_\Omega \frac{\partial}{\partial\Delta}\log\left(\frac{dP_{\mu+\Delta}}{dP_\mu}\right)\Big|_{\Delta=0} dP_\mu\Delta + \frac{1}{2}I(\mu)\Delta^2 \\
&= -\int_\Omega \frac{\partial}{\partial\Delta}\frac{dP_{\mu+\Delta}}{dP_\mu}\Big|_{\Delta=0} dP_\mu\Delta + \frac{1}{2}I(\mu)\Delta^2 \\
&= -\frac{\partial}{\partial\Delta}\int_\Omega \frac{dP_{\mu+\Delta}}{dP_\mu}dP_\mu\Big|_{\Delta=0}\Delta + \frac{1}{2}I(\mu)\Delta^2 \\
&= -\frac{\partial}{\partial\Delta}\int_\Omega dP_{\mu+\Delta}\Big|_{\Delta=0}\Delta + \frac{1}{2}I(\mu)\Delta^2 \\
&= \frac{1}{2}I(\mu)\Delta^2
\end{aligned}
$$

其中，第二行中引入的 $I_{(\mu)}$ 称为族的 Fisher 信息 $P(\mu)_\mu$ 在。注意，如果 λ 是 $\Delta small$ 的 $(P_{\mu+\Delta})$ 的常用主要度量，则 $dP_{\mu+\Delta} = p_{\mu+\Delta}d\lambda$ 我们可以写

$$
I(\mu) = -\int \frac{\partial^2}{\partial\Delta^2}\log p_{\mu+\Delta}\Big|_{\Delta=0} p_\mu d\lambda
$$

这是小学课文中通常给出的形式。这一切的结果是 $\mathrm{D}(P_\mu, P_{\mu+\Delta})$，因为 $\Delta small$ 实际上是 Δ 的二次方，通过提供的缩放 $I_{(\mu)}$，因此，最糟糕的遗憾总是 $O\sqrt{nk}$，提供所考虑的分配类别足够充分，也不太奇怪。

2、我们现在已经显示了一个下限 $O\sqrt{nk}$，虽然许多上限是 $O\sqrt{\log(n)}$。这并不矛盾，因为对数界限取决于次优间隙的倒数，次优间隙可能非常大。

3、我们的下限仅为 $n \geq k-1$。在练习 15.3 中，我们要求您理解当 $n < k-1$ 时，有这样一个赌博机

$$
R_n \geq \frac{n(2k-n-1)}{2k} > \frac{n}{2}
$$

4、用于证明定理 15.2 的方法可以看作是对统计中的 Le Cam 方法。回想一下，等式（15.2）对于任意 μ 和 μ'

$$
\inf_\pi \sup_\nu R_n(\pi, \nu) \geq \frac{n\Delta}{8}\exp\left(-\mathrm{D}\left(\mathbb{P}_\mu, \mathbb{P}_{\mu'}\right)\right)
$$

为了解释 Le Cam 的方法，我们需要一点符号。令 χ 为结果空间，$\mathcal{P}$ 为 χ 上的一系列措施，$\theta : \mathcal{P} \to \Theta$，其中 (Θ, d) 是一个度量空间。估计器是一个函数 $\hat{\theta} : \mathcal{X}^n \to \Theta$。Le-Cam 方法用于证明期望误差的极大极小下界估计量

$$
\inf_{P\in\mathcal{P}} \sup \mathbb{E}_{X_1,\ldots,X_n\sim P^n}\left[d\left(\hat{\theta}\left(X_1,\ldots,X_n\right), \theta(P)\right)\right]
$$

这个方法是用来选择 $P_0, P_1 \in \mathcal{P}$ 去最大化 $d\left(\theta\left(P_0\right), \theta\left(P_1\right)\right)\exp\left(-n\mathrm{D}\left(P_0, P_1\right)\right)$，在任意 $P_0, P_1 \in \mathcal{P}$ 的基础上

$$
Eq.(15.3) \geq \frac{\Delta}{8}\exp\left(-n\mathrm{D}\left(P_0, P_1\right)\right)
$$

式中，$\Delta = d\left(\theta\left(P_0\right), \theta\left(P_1\right)\right)$。与赌博机下限相比，有两个区别：（1）我们处理顺序设置；（2）选择 P_0 后，我们选择 P_1 的方式取决于关于算法。这提供了一个非常需要的额外提升，如果没有它，该方法将是无法捕捉 $\mathcal{P}$ 的特征如何反映在极大极小风险中（或遗憾，在我们的案例）

15.4 文献综述

我们知道的第一个关于下界的工作是 Vogel [1960]对双臂伯努利赌博机的非常精确的极大极小分析。Bubeck 等人[2013b]首次将 Bretagnolle–Huber 不等式（定理 14.2）用于赌博机。正如注记中所述，利用这个不等式证明下界，在统计学中被称为 Le Cam 方法[Le Cam, 1973]。定理 15.2 的证明采用了与 Gerchinovitz 和 Lattimore [2016]相同的思路，而练习题 15.2 中的另一种证明本质上是由 Auer 等人[1995]提出的，他们分析了奖励是伯努利的更困难的情况（见练习题 15.4）。Yu [1997]描述了一些替代 Le Cam 方法的被动、统计设定。这些备选方案可以（而且经常可以）适应序列设置。

第二章 关键证明过程公式推导

15.1 公式推导

缩写 $\hat{\theta} = \hat{\theta}\left(X_1,\ldots,X_n\right)$，同时令 $R(P) = \mathbb{E}_P[d(\hat{\theta}, P)]$。通过三角不等式得

$$
d\left(\hat{\theta}, P_0\right) + d\left(\hat{\theta}, P_1\right) \geq d\left(P_0, P_1\right) = \Delta
$$

令 $E = \left\{d\left(\hat{\theta}, P_0\right) \leq \Delta/2\right\}$ 在 E^c 上认为 $d\left(\hat{\theta}, P_0\right) \geq \Delta/2$，在 E 上认为 $d\left(\hat{\theta}, P_1\right) \geq \Delta - d\left(\hat{\theta}, P_0\right) \geq \Delta/2$

$$R(P_0)+R(P_1)\geq\frac{\Delta}{2}\left(P_0\left(E^c\right)+P_1(E)\right)\geq\frac{\Delta}{4}\exp\left(-\mathrm{D}\left(P_0,P_1\right)\right)$$

结果如下，因为$\max\{a,b\}\geq(a+b)/2$