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 $\varepsilon_{SG}^k(1)$ 上是一致的。总是选择第一个动作的决策在任何上都是不一致的，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

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

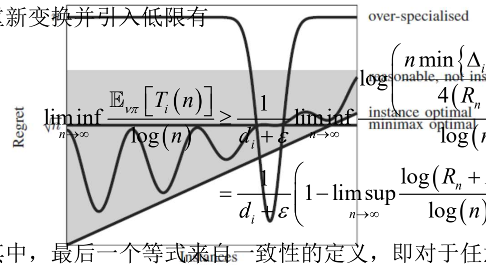

图 16.1 在 x 轴上，实例按照难度度量进行排序，y 轴上显示的是遗憾（在某种程度上）。在前一章中，我们证明了没有任何决策可以完全低于水平“极大极小最优”线。本章的结果表明，如果某个决策的遗憾在任何时候都低于“实例最优”线，那么对于其他实例，该策略的遗憾必须高于阴影区域。例如，“过度指定”决策。

定理 16.2。设 $\varepsilon=\mathcal{M}_1\times\cdots\times\mathcal{M}_k$ 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，$\nu'=(P'_j)$ 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 $d_{\inf}$ 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，

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

重新变换并引入低限有

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

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 $R_n+R'_n\le C_p n^p$，这意味着

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

由于 $p>0$ 是任意的，取 ε 趋于零的极限，再由引理 4.5 即可得证。

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

$$\liminf_{n\to\infty}\frac{R_n}{\log(n)}=\liminf_{n\to\infty}\sum_{i:\Delta_i>0}\frac{\Delta_i\mathbb{E}[T_i(n)]}{\log(n)}\ge\sum_{i:\Delta_i>0}\Delta_i\liminf_{n\to\infty}\frac{\mathbb{E}[T_i(n)]}{\log(n)}\ge\sum_{i:\Delta_i>0}\frac{\Delta_i}{d_i+\varepsilon}\to\sum_{i:\Delta_i>0}\frac{\Delta_i}{d_{\inf}(P_i,\mu^*,\mathcal{M}_i)}$$

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，则可以在常... 上调用渐近最优决策。例如，第 8 章

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

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，

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

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。仅仅假设就足以表明，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

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

$_n(\pi,\nu')\le Cn^p$，有

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

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 $_j(\nu')=\mu_i+\Delta_i(1+\varepsilon)$。再基于引理 16.3 及 $\lambda=\Delta_i(1+\varepsilon)$，$D(P_i,P_i')\le d_i+\varepsilon$，$_n+R_n'\le C_p n^p$ 有，

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

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。

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。