

2 有“专家建议”的预测

这一章介绍的“专家建议预测模型”构建了个体序列预测理论的基础，而我们将在整本书中进一步展开该理论。

“专家建议预测”基于以下用于序列决策的协议：决策者是一个预测者，其目标是预测一个未知的序列 y_1, y_2, \dots ，该序列来自一个结果空间 \mathcal{Y} 。预测者的预测 $\hat{p}_1, \hat{p}_2, \dots$ 属于一个决策空间 \mathcal{D} ，我们假设这个决策空间是一个向量空间的凸子集。在某些特殊情况下，我们取 $\mathcal{D} = \mathcal{Y}$ ，但通常 \mathcal{D} 可能与 \mathcal{Y} 不同。

预测者是按顺序计算其预测的，其预测表现与一组参考预测者（我们称之为“专家”）进行比较。更具体地说，在每个时刻 t ，预测者可以访问专家预测集 $\{f_{E,t} : E \in \mathcal{E}\}$ ，其中 $f_{E,t} \in \mathcal{D}$ ， \mathcal{E} 是专家的一个固定索引集。基于专家的预测，预测者计算自己对下一个结果 y_t 的预测 \hat{p}_t 。当 \hat{p}_t 计算完毕后，真实的结果 y_t 被揭示出来。

预测者和专家的预测通过一个非负损失函数 $\ell : \mathcal{D} \times \mathcal{Y} \rightarrow \mathbb{R}$ 进行评分。

这个预测协议可以自然地看作是预测者（“预测者”）和环境（“环境”）之间的重复博弈。预测者做出预测 \hat{p}_t ，而环境选择专家的建议 $\{f_{E,t} : E \in \mathcal{E}\}$ 并设定真实的结果 y_t 。

PREDICTION WITH EXPERT ADVICE

参数：决策空间 \mathcal{D} ，结果空间 \mathcal{Y} ，损失函数 ℓ ，专家索引集 \mathcal{E} 。

每一轮 $t = 1, 2, \dots$

- (1) 环境选择下一个结果 y_t 和专家建议 $\{f_{E,t} \in \mathcal{D} : E \in \mathcal{E}\}$ ，并将专家建议展示给预测者；
- (2) 预测者选择预测 $\hat{p}_t \in \mathcal{D}$ ；
- (3) 环境揭示下一个结果 $y_t \in \mathcal{Y}$ ；
- (4) 预测者承受损失 $\ell(\hat{p}_t, y_t)$ ，每位专家 E 承受损失 $\ell(f_{E,t}, y_t)$ 。

预测者的目标是尽可能减少相对于每位专家的累积遗憾（或简称为遗憾）。对于专家 E ，这个量通过以下总和定义：

$$R_{E,n} = \sum_{t=1}^n (\ell(\hat{p}_t, y_t) - \ell(f_{E,t}, y_t)) = \hat{L}_n - L_{E,n}$$

其中， $\hat{L}_n = \sum_{t=1}^n \ell(\hat{p}_t, y_t)$ 表示预测者的累积损失， $L_{E,n} = \sum_{t=1}^n \ell(f_{E,t}, y_t)$ 表示专家 E 的累积损失。因此， $R_{E,n}$ 是预测者在 n 轮预测后的总损失与专家 E 的总损失之间的差值。我们还可以定义相对于专家 E 在时刻 t 的瞬时遗憾 $r_{E,t} = \ell(\hat{p}_t, y_t) - \ell(f_{E,t}, y_t)$ ，因此 $R_{E,n} = \sum_{t=1}^n r_{E,t}$ 。你可以将 $r_{E,t}$ 理解为预测者在第 t 个结果 y_t 揭示之后，感受到的没有听从专家 E 建议的遗憾。

在本章的余下部分，我们假设专家的数量是有限的，即 $\mathcal{E} = \{1, 2, \dots, N\}$ ，并使用索引 $i = 1, \dots, N$ 来指代专家。预测者的目标是针对所有结果序列，尽量使遗憾最小化。例如，预测者可能希望实现每一轮的遗憾逐渐消失，也就是说，达到以下目标：

$$\max_{i=1, \dots, N} R_{i,n} = o(n) \quad \text{或等价于} \quad \frac{1}{n} \left(\hat{L}_n - \min_{i=1, \dots, N} L_{i,n} \right) \xrightarrow{n \rightarrow \infty} 0$$

这里的收敛是针对结果序列和专家建议的选择而言的。在下一节中，我们将展示，满足温和条件下，一个简单的预测者可以实现这一雄心勃勃的目标。

本章的结构如下：在 2.1 节中，我们引入了一个重要的加权平均预测者类，描述了基于势的预测者子类，并分析了两个重要的特殊情况：多项式加权平均预测者和指数加权平均预测者。指数加权平均预测者在我们的理论中占据核心地位，接下来的四节都与它的不同问题相关：2.2 节展示了某些最优性质，2.3 节讨论了动态调整势参数的问题，2.4 节探讨了当最佳专家的损失较小时如何获得更好的遗憾界限，2.5 节则研究了可微损失函数的特殊情况。从 2.6 节开始，我们发现重新缩放损失函数的优势。这个简单的技巧使我们能够推导出新的、更精确的性能界限。在 2.7 节中，我们介绍并分析了一个针对重新缩放损失的加权平均预测者，不同于之前的预测者，它不是基于势的概念。2.8 节中我们回到指数加权平均预测者，并基于损失函数的重新缩放推导出改进的遗憾界限。2.9 节和 2.10 节讨论了专家建议预测问题中的一些一般性问题，包括极小极大值的定义。最后，2.11 节讨论了一种引入折旧因子的遗憾变体。

2.1 加权平均预测

在这一框架中，一个自然的预测策略是基于计算专家预测的加权平均值。也就是说，预测者在时间 t 的预测由以下公式给出：

$$\hat{p}_t = \frac{\sum_{i=1}^N w_{i,t-1} f_{i,t}}{\sum_{j=1}^N w_{j,t-1}}$$

其中， $w_{1,t-1}, \dots, w_{N,t-1} \geq 0$ 是分配给专家的权重。注意， $\hat{p}_t \in \mathcal{D}$ ，因为它是专家建议 $f_{1,t}, \dots, f_{N,t} \in \mathcal{D}$ 的凸组合，并且根据我们的假设， \mathcal{D} 是凸的。由于我们的目标是最小化遗憾，因此根据时间 $t-1$ 的遗憾来选择权重是合理的。如果 $R_{i,t-1}$ 较大，那么我们给专家 i 分配较大的权重 $w_{i,t-1}$ ，反之亦然。由于 $R_{i,t-1} = \hat{L}_{t-1} - L_{i,t-1}$ ，这意味着对于那些累积损失 $L_{i,t-1}$ 较小的专家 i ，我们会给予更多的权重。因此，我们可以将权重看作是专家遗憾的一个任意递增函数。由于一些即将讨论的原因，我们将这个函数表示为一个非负、凸且递增的函数 $\phi: \mathbb{R} \rightarrow \mathbb{R}$ 的导数。我们用 ϕ' 表示这个导数。预测者使用 ϕ' 来确定分配给第 i 个专家的权重 $w_{i,t-1} = \phi'(R_{i,t-1})$ 。因此，加权平均预测者在时间 t 的预测 \hat{p}_t 定义为：

$$\hat{p}_t = \frac{\sum_{i=1}^N \phi'(R_{i,t-1}) f_{i,t}}{\sum_{j=1}^N \phi'(R_{j,t-1})}$$

（加权平均预测者）。

需要注意的是，这是一个合法的预测器，因为 \hat{p}_t 是基于专家在时间 t 的建议和到时间 $t-1$ 为止的累积遗憾计算出来的。

我们从一个简单的技术性观察开始对加权平均预测器进行分析。

引理 2.1 如果损失函数 ℓ 在它的第一个变量上是凸的，那么：

$$\sup_{y \in \mathcal{Y}} \sum_{i=1}^N r_{i,t} \phi'(R_{i,t-1}) \leq 0$$

证明. 使用 Jensen 不等式，对于所有 $y \in \mathcal{Y}$ ，

$$\ell(\hat{p}_t, y) = \ell\left(\frac{\sum_{i=1}^N \phi'(R_{i,t-1}) f_{i,t}}{\sum_{j=1}^N \phi'(R_{j,t-1})}, y\right) \leq \frac{\sum_{i=1}^N \phi'(R_{i,t-1}) \ell(f_{i,t}, y)}{\sum_{j=1}^N \phi'(R_{j,t-1})}$$

重新排列后，得到上述结论。

上述引理的简单观察使我们能够以一种有趣的方式解读加权平均预测器。为此，引入瞬时遗憾向量：

$$\mathbf{r}_t = (r_{1,t}, \dots, r_{N,t}) \in \mathbb{R}^N$$

相应的遗憾向量为 $\mathbf{R}_n = \sum_{t=1}^n \mathbf{r}_t$ 。同时，我们也引入了一个形式为

$$\Phi(\mathbf{u}) = \psi \left(\sum_{i=1}^N \phi(u_i) \right)$$

的势函数（Potential Function），其中 $\phi: \mathbb{R} \rightarrow \mathbb{R}$ 是一个非负、递增且二阶可微的函数，而 $\psi: \mathbb{R} \rightarrow \mathbb{R}$ 则是一个非负、严格递增、凹的且二阶可微的辅助函数。

利用势函数的概念，我们可以给出加权平均预测器的等价定义：

$$\hat{p}_t = \frac{\sum_{i=1}^N \nabla \Phi(\mathbf{R}_{t-1})_i f_{i,t}}{\sum_{j=1}^N \nabla \Phi(\mathbf{R}_{t-1})_j}$$

其中 $\nabla \Phi(\mathbf{R}_{t-1})_i = \partial \Phi(\mathbf{R}_{t-1}) / \partial R_{i,t-1}$ 。我们称上述定义的预测器是基于势函数 Φ 的。虽然加权平均预测器的定义与 ψ 的选择无关（在定义 \hat{p}_t 时 ψ' 被抵消了），但是本章主要结果的证明——定理 2.1——揭示了 ψ 在分析中起着重要作用。我们指出，证明定理 2.1 时不需要 ϕ 的凸性，这也是为什么在势函数的定义中没有提到凸性。然而，本书中所有基于势函数的并且具有消失的每轮遗憾的预测器，都是通过凸的 ϕ 构造的（参见习题 2.2）。

引理 2.1 的表述等价于：

$$\sup_{y_t \in \mathcal{Y}} \mathbf{r}_t \cdot \nabla \Phi(\mathbf{R}_{t-1}) \leq 0 \quad (\text{Blackwell 条件})$$

符号 $\mathbf{u} \cdot \mathbf{v}$ 表示两个向量的内积，定义为 $\mathbf{u} \cdot \mathbf{v} = u_1 v_1 + \dots + u_N v_N$ 。我们称上述不等式为 Blackwell 条件，因为它与著名的 Blackwell 可接近性定理证明中使用的一个关键性质非常相似。该定理及其与上述不等式的联系将在第 7.7 和 7.8 节中探讨。图 2.1 展示了满足 Blackwell 条件的预测实例。

Blackwell 条件表明，函数 Φ 的作用与动力系统势能有些类似：加权平均预测器通过让遗憾向量指向远离 Φ 梯度的方向（无论结果 y_t 是什么），趋向于保持点 \mathbf{R}_t 接近 Φ 的最小值。实际上，这一性质为分析提供了简单的思路，因为根据泰勒定理，势函数 Φ 的增量现在可以轻松地被限制住。函数 ψ 的作用仅仅是通过这个论证获得更好的界限。

下一个定理适用于任何满足 Blackwell 条件的预测器（因此不仅仅是加权平均预测器）。然而，它会给不同版本的加权平均预测器提供一些有趣的界限。

定理 2.1 假设一个预测器满足某个势函数 $\Phi(\mathbf{u}) = \psi \left(\sum_{i=1}^N \phi(u_i) \right)$ 的 Blackwell 条件。那么，对于所有 $n = 1, 2, \dots$,

$$\Phi(\mathbf{R}_n) \leq \Phi(\mathbf{0}) + \frac{1}{2} \sum_{t=1}^n C(\mathbf{r}_t)$$

其中，

$$C(\mathbf{r}_t) = \sup_{\mathbf{u} \in \mathbb{R}^N} \psi' \left(\sum_{i=1}^N \phi(u_i) \right) \sum_{i=1}^N \phi''(u_i) r_{i,t}^2$$

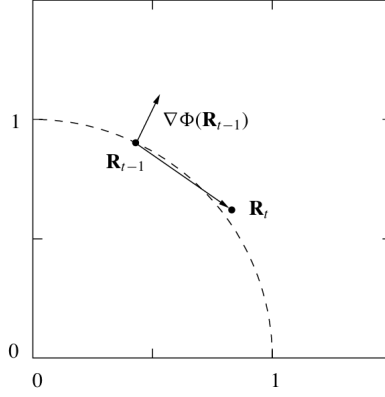


图 2.1: 这是一个展示 Blackwell 条件的例子，其中 $N = 2$ 。虚线表示在遗憾空间中，潜力等于 1 的点。预测在时间 t 改变了潜力，从 $\Phi(\mathbf{R}_{t-1}) = 1$ 变为 $\Phi(\mathbf{R}_t) = \Phi(\mathbf{R}_{t-1} + \mathbf{r}_t)$ 。尽管 $\Phi(\mathbf{R}_t) > \Phi(\mathbf{R}_{t-1})$ ，但 \mathbf{r}_t 和梯度 $\nabla\Phi(\mathbf{R}_{t-1})$ 之间的内积是负的，因此 Blackwell 条件得以满足。

证明. 我们利用泰勒定理来估计 $\Phi(\mathbf{R}_t)$ 与 $\Phi(\mathbf{R}_{t-1})$ 之间的关系。因此，我们得到

$$\begin{aligned}\Phi(\mathbf{R}_t) &= \Phi(\mathbf{R}_{t-1} + \mathbf{r}_t) \\ &= \Phi(\mathbf{R}_{t-1}) + \nabla\Phi(\mathbf{R}_{t-1}) \cdot \mathbf{r}_t + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \frac{\partial^2 \Phi}{\partial u_i \partial u_j} \bigg|_{\xi} r_{i,t} r_{j,t}\end{aligned}$$

(其中 ξ 是 \mathbb{R}^N 中的某个向量)

$$\leq \Phi(\mathbf{R}_{t-1}) + \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \frac{\partial^2 \Phi}{\partial u_i \partial u_j} \bigg|_{\xi} r_{i,t} r_{j,t}$$

不等式是根据 Blackwell 条件得出的。现在通过简单的计算可以得到

$$\begin{aligned}& \sum_{i=1}^N \sum_{j=1}^N \frac{\partial^2 \Phi}{\partial u_i \partial u_j} \bigg|_{\xi} r_{i,t} r_{j,t} \\ &= \psi'' \left(\sum_{i=1}^N \phi(\xi_i) \right) \sum_{i=1}^N \sum_{j=1}^N \phi'(\xi_i) \phi'(\xi_j) r_{i,t} r_{j,t} \\ & \quad + \psi' \left(\sum_{i=1}^N \phi(\xi_i) \right) \sum_{i=1}^N \phi''(\xi_i) r_{i,t}^2\end{aligned}$$

进一步化简为

$$\begin{aligned}&= \psi'' \left(\sum_{i=1}^N \phi(\xi_i) \right) \left(\sum_{i=1}^N \phi'(\xi_i) r_{i,t} \right)^2 + \psi' \left(\sum_{i=1}^N \phi(\xi_i) \right) \sum_{i=1}^N \phi''(\xi_i) r_{i,t}^2 \\ &\leq \psi' \left(\sum_{i=1}^N \phi(\xi_i) \right) \sum_{i=1}^N \phi''(\xi_i) r_{i,t}^2 \quad (\text{由于 } \psi \text{ 是凹函数}) \\ &\leq C(\mathbf{r}_t)\end{aligned}$$

在最后一步，我们使用了 $C(\mathbf{r}_t)$ 的定义。因此，我们得到了 $\Phi(\mathbf{R}_t) - \Phi(\mathbf{R}_{t-1}) \leq C(\mathbf{r}_t)/2$ 。最后通过对 $t = 1, \dots, n$ 求和即可完成证明。

定理 2.1 可以如下应用。由于 ψ 和 ϕ 的单调性,

$$\psi\left(\phi\left(\max_{i=1,\dots,N} R_{i,n}\right)\right) = \psi\left(\max_{i=1,\dots,N} \phi(R_{i,n})\right) \leq \psi\left(\sum_{i=1}^N \phi(R_{i,n})\right) = \Phi(\mathbf{R}_n)$$

注意, 根据势函数的定义, ψ 是可逆的。如果 ϕ 也是可逆的, 那么我们可以得到

$$\max_{i=1,\dots,N} R_{i,n} \leq \phi^{-1}(\psi^{-1}(\Phi(\mathbf{R}_n)))$$

其中 $\Phi(\mathbf{R}_n)$ 被定理 2.1 给出的界限所替代。然而, 在接下来的两个例子中, 第一个例子中 ϕ 不是可逆的, 因此 $\max_{i=1,\dots,N} R_{i,n}$ 是直接用定理 2.1 给出的界限的一个函数来加以限制的。

2.1.1 多项式加权平均预测器

考虑基于势函数

$$\Phi_p(\mathbf{u}) = \left(\sum_{i=1}^N (u_i)_+^p\right)^{2/p} = \|\mathbf{u}_+\|_p^2 \quad (\text{多项式的势})$$

其中 $p \geq 2$ 。这里 \mathbf{u}_+ 表示 \mathbf{u} 的正部分的向量。分配给专家的权重由以下公式给出:

$$w_{i,t-1} = \nabla \Phi_p(\mathbf{R}_{t-1})_i = \frac{2(R_{i,t-1})_+^{p-1}}{\|(\mathbf{R}_{t-1})_+\|_p^{p-2}}$$

预测器的预测就是专家预测的加权平均值:

$$\hat{p}_t = \frac{\sum_{i=1}^N \left(\sum_{s=1}^{t-1} (\ell(\hat{p}_s, y_s) - \ell(f_{i,s}, y_s))\right)_+^{p-1} f_{i,t}}{\sum_{j=1}^N \left(\sum_{s=1}^{t-1} (\ell(\hat{p}_s, y_s) - \ell(f_{j,s}, y_s))\right)_+^{p-1}}$$

推论 2.1 假设损失函数 ℓ 在其第一个参数上是凸的, 并且其值在 $[0, 1]$ 范围内。则对于任何结果序列 $y_1, y_2, \dots \in \mathcal{Y}$ 和任何 $n \geq 1$, 多项式加权平均预测器的遗憾满足

$$\hat{L}_n - \min_{i=1,\dots,N} L_{i,n} \leq \sqrt{n(p-1)N^{2/p}}$$

这表明, 对于所有 $p \geq 2$, 每轮的遗憾以 $O(1/\sqrt{n})$ 的速率均匀地收敛到零, 无论是对结果序列还是对专家建议。选择 $p = 2$ 会产生一个特别简单的算法。另一方面, 选择 $p = 2 \ln N$ (当 $N > 2$ 时), 该选择大致上最小化了上界, 得到

$$\hat{L}_n - \min_{i=1,\dots,N} L_{i,n} \leq \sqrt{ne(2 \ln N - 1)}$$

这在专家数量 N 上有显著更好的依赖性。

证明. 使用定理 2.1 结合多项式势函数。设 $\phi(x) = x_+^p$ 和 $\psi(x) = x^{2/p}$, 其中 $x \geq 0$ 。此外,

$$\psi'(x) = \frac{2}{px^{(p-2)/p}} \quad \text{以及} \quad \phi''(x) = p(p-1)x_+^{p-2}$$

根据 Hölder 不等式,

$$\begin{aligned} \sum_{i=1}^N \phi''(u_i) r_{i,t}^2 &= p(p-1) \sum_{i=1}^N (u_i)_+^{p-2} r_{i,t}^2 \\ &\leq p(p-1) \left(\sum_{i=1}^N \left((u_i)_+^{p-2} \right)^{p/(p-2)} \right)^{(p-2)/p} \left(\sum_{i=1}^N |r_{i,t}|^p \right)^{2/p} \end{aligned}$$

因此,

$$\psi' \left(\sum_{i=1}^N \phi(u_i) \right) \sum_{i=1}^N \phi''(u_i) r_{i,t}^2 \leq 2(p-1) \left(\sum_{i=1}^N |r_{i,t}|^p \right)^{2/p}$$

满足定理 2.1 的条件, 其中 $C(\mathbf{r}_t) \leq 2(p-1) \|\mathbf{r}_t\|_p^2$ 。由于 $\Phi_p(\mathbf{0}) = 0$, 定理 2.1 以及损失函数的有界性表明,

$$\left(\sum_{i=1}^N (R_{i,n})_+^p \right)^{2/p} = \Phi_p(\mathbf{R}_n) \leq (p-1) \sum_{t=1}^n \|\mathbf{r}_t\|_p^2 \leq n(p-1)N^{2/p}$$

最后, 由于

$$\hat{L}_n - \min_{i=1,\dots,N} L_{i,n} = \max_{i=1,\dots,N} R_{i,n} \leq \left(\sum_{i=1}^N (R_{i,n})_+^p \right)^{1/p}$$

结果得证。

备注 2.1 我们定义了多项式势函数为 $\Phi_p(\mathbf{u}) = \|\mathbf{u}_+\|_p^2$, 这对应于选择 $\psi(x) = x^{2/p}$ 。请注意, ψ 对预测没有任何影响, 它只在分析中发挥作用。这里分析的特定形式是为了方便选择的, 但也有其他可能性可以得到类似的结果。例如, 可以认为选择 $\psi(x) = x^{1/p}$ 更自然, 这会导致势函数 $\Phi(\mathbf{u}) = \|\mathbf{u}_+\|_p$ 。

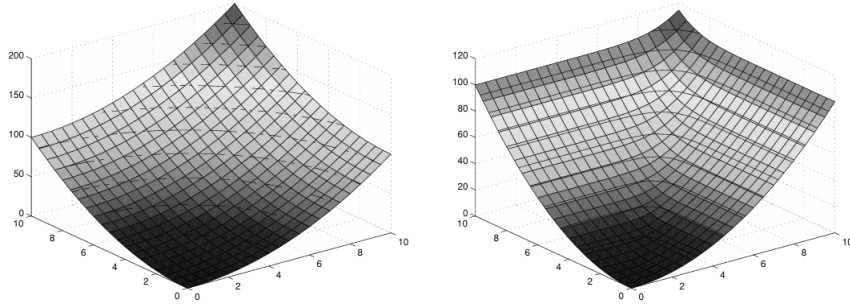


图 2.2: $N = 2$ 名专家的多项式势函数 $\Phi_p(\mathbf{u})$ 的图示, 其中指数为 $p = 2$ 和 $p = 10$ 。

我们留作练习, 以根据这个选择推导出类似于 Corollary 2.1 的界限。

2.1.2 指数加权平均预报器

我们的第二个主要示例是基于势函数的指数加权平均预报器, 其势函数为

$$\Phi_\eta(\mathbf{u}) = \frac{1}{\eta} \ln \left(\sum_{i=1}^N e^{\eta u_i} \right) \quad (\text{指数势函数})$$

其中 η 是一个正参数。在这种情况下, 分配给专家的权重形式为

$$w_{i,t-1} = \nabla \Phi_\eta(\mathbf{R}_{t-1})_i = \frac{e^{\eta R_{i,t-1}}}{\sum_{j=1}^N e^{\eta R_{j,t-1}}}$$

而加权平均预报器简化为

$$\hat{p}_t = \frac{\sum_{i=1}^N \exp \left(\eta \left(\hat{L}_{t-1} - L_{i,t-1} \right) \right) f_{i,t}}{\sum_{j=1}^N \exp \left(\eta \left(\hat{L}_{t-1} - L_{j,t-1} \right) \right)} = \frac{\sum_{i=1}^N e^{-\eta L_{i,t-1}} f_{i,t}}{\sum_{j=1}^N e^{-\eta L_{j,t-1}}}$$

指数加权平均预报器的优点在于，它只依赖于专家的去表现，而使用其他一般势函数的预报器则还依赖于过去的预测 $\hat{p}_s, s < t$ 。此外，预报器分配给专家的权重可以通过简单的递增方式计算：设 $w_{1,t-1}, \dots, w_{N,t-1}$ 是在回合 t 用于计算预测 $\hat{p}_t = \sum_{i=1}^N w_{i,t-1} f_{i,t}$ 的权重。那么，可以很容易验证，

$$w_{i,t} = \frac{w_{i,t-1} e^{-\eta \ell(f_{i,t}, y_t)}}{\sum_{j=1}^N w_{j,t-1} e^{-\eta \ell(f_{j,t-1}, y_t)}}$$

定理 2.1 的简单应用揭示了指数加权平均预报器的以下性能界限。

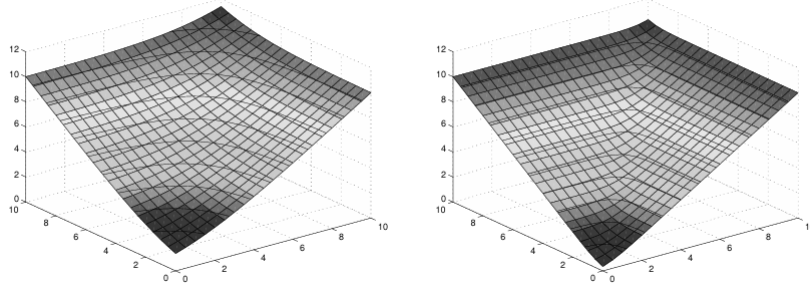


图 2.3: $N = 2$ 名专家的指数势函数 $\Phi_\eta(\mathbf{u})$ 的图示，其中 $\eta = 0.5$ 和 $\eta = 2$ 。

推论 2.2 假设损失函数 ℓ 在其第一个参数上是凸的，并且取值范围在 $[0, 1]$ 之间。对于任意 n 和 $\eta > 0$ ，以及所有 $y_1, \dots, y_n \in \mathcal{Y}$ ，指数加权平均预报器的遗憾满足

$$\hat{L}_n - \min_{i=1, \dots, N} L_{i,n} \leq \frac{\ln N}{\eta} + \frac{n\eta}{2}.$$

优化上界建议选择 $\eta = \sqrt{2 \ln N / n}$ 。在这种情况下，上界变为 $\sqrt{2n \ln N}$ ，这比使用 $\phi(x) = x_+^p$ 且 $p = 2 \ln N$ 时获得的最佳上界略好。在下一节中，我们通过直接分析改进了推论 2.2 的上界。指数加权的一个缺点是，最优的参数 η 调节需要提前知道时间跨度 n 。在接下来的两节中，我们将描述几种不受此缺陷影响的指数加权平均预报器的变体。

证明. 使用指数势函数应用定理 2.1。此时 $\phi(x) = e^{\eta x}$ ， $\psi(x) = (1/\eta) \ln x$ ，并且

$$\psi' \left(\sum_{i=1}^N \phi(u_i) \right) \sum_{i=1}^N \phi''(u_i) r_{i,t}^2 \leq \eta \max_{i=1, \dots, N} r_{i,t}^2 \leq \eta.$$

利用 $\Phi_\eta(\mathbf{0}) = (\ln N)/\eta$ ，定理 2.1 推导出

$$\max_{i=1, \dots, N} R_{i,n} \leq \Phi_\eta(\mathbf{R}_n) \leq \frac{\ln N}{\eta} + \frac{n\eta}{2}.$$

2.2 最优界限

本节目的是展示即使对于一般的凸损失函数，推论 2.2 中的界限也可以对指数加权平均预报器进行改进。以下结果通过一个常数因子改进了推论 2.2 的界限。在第 3.7 节中，我们将看到此处获得的界限无法进一步改进。

定理 2.2 假设损失函数 ℓ 在其第一个参数上是凸的，并且取值范围在 $[0, 1]$ 之间。对于任意 n 和 $\eta > 0$ ，以及所有 $y_1, \dots, y_n \in \mathcal{Y}$ ，指数加权平均预报器的遗憾满足

$$\hat{L}_n - \min_{i=1, \dots, N} L_{i,n} \leq \frac{\ln N}{\eta} + \frac{n\eta}{8}.$$

特别地, 当 $\eta = \sqrt{8 \ln N / n}$ 时, 上界变为 $\sqrt{(n/2) \ln N}$ 。

证明的思路与推论 2.2 相似, 但现在我们不是对 $(1/\eta) \ln (\sum_i e^{\eta R_{i,t}})$ 进行界限估计, 而是对相关量 $(1/\eta) \ln (W_t/W_{t-1})$ 进行界限估计, 其中

$$W_t = \sum_{i=1}^N w_{i,t} = \sum_{i=1}^N e^{-\eta L_{i,t}}$$

对于 $t \geq 1$, 且 $W_0 = N$ 。在证明中, 我们使用了 Hoeffding 的以下经典不等式 [161]。

引理 2.2 设 X 是一个随机变量, 且 $a \leq X \leq b$ 。那么对于任何 $s \in \mathbb{R}$,

$$\ln \mathbb{E} [e^{sX}] \leq s\mathbb{E}X + \frac{s^2(b-a)^2}{8}.$$

证明见附录 A.1 部分。

证明. (定理 2.2) 首先观察到

$$\begin{aligned} \ln \frac{W_n}{W_0} &= \ln \left(\sum_{i=1}^N e^{-\eta L_{i,n}} \right) - \ln N \\ &\geq \ln \left(\max_{i=1, \dots, N} e^{-\eta L_{i,n}} \right) - \ln N \\ &= -\eta \min_{i=1, \dots, N} L_{i,n} - \ln N. \end{aligned}$$

另一方面, 对于每个 $t = 1, \dots, n$,

$$\begin{aligned} \ln \frac{W_t}{W_{t-1}} &= \ln \frac{\sum_{i=1}^N e^{-\eta \ell(f_{i,t}, y_t)} e^{-\eta L_{i,t-1}}}{\sum_{j=1}^N e^{-\eta L_{j,t-1}}} \\ &= \ln \frac{\sum_{i=1}^N w_{i,t-1} e^{-\eta \ell(f_{i,t}, y_t)}}{\sum_{j=1}^N w_{j,t-1}}. \end{aligned}$$

现在使用引理 2.2, 我们观察到上述量可以通过以下方式找到上界:

$$\begin{aligned} -\eta \frac{\sum_{i=1}^N w_{i,t-1} \ell(f_{i,t}, y_t)}{\sum_{j=1}^N w_{j,t-1}} + \frac{\eta^2}{8} &\leq -\eta \ell \left(\frac{\sum_{i=1}^N w_{i,t-1} f_{i,t}}{\sum_{j=1}^N w_{j,t-1}}, y_t \right) + \frac{\eta^2}{8} \\ &= -\eta \ell(\hat{p}_t, y_t) + \frac{\eta^2}{8}, \end{aligned}$$

我们使用了损失函数在其第一个参数上的凸性以及指数加权平均预报器的定义。对 $t = 1, \dots, n$ 进行求和, 我们得到

$$\ln \frac{W_n}{W_0} \leq -\eta \hat{L}_n + \frac{\eta^2}{8} n.$$

将此结果与下界结合并解出 \hat{L}_n , 我们得到

$$\hat{L}_n \leq \min_{i=1, \dots, N} L_{i,n} + \frac{\ln N}{\eta} + \frac{\eta}{8} n,$$

2.3 对时间的均匀界限

正如我们在上一节中提到的，指数加权平均预报器的缺点是推论 2.2 中的后悔界限并不在任意长度的序列上均匀有效，而仅对一个特定长度 n 的序列有效，其中 n 是用来选择参数 η 的值。为了解决这个问题，可以使用所谓的“倍增技巧”。这个方法的思路是将时间划分为长度指数增长的时期。在每个时期内，使用一个针对该时期长度最优选择的参数 η 的加权平均预报器。当一个时期结束时，加权平均预报器被重置，并在下一个时期以新的 η 值重新开始。如果使用倍增技巧与指数加权平均预报器配合，那么它对于任何序列 $y_1, y_2, \dots \in \mathcal{Y}$ 和任意 $n \geq 1$ 将实现

$$\hat{L}_n - \min_{i=1, \dots, N} L_{i,n} \leq \frac{\sqrt{2}}{\sqrt{2}-1} \sqrt{\frac{n}{2} \ln N},$$

(见练习 2.8)。这个界限比定理 2.2 的界限差了一个 $\sqrt{2}/(\sqrt{2}-1)$ 的因子，大约为 3.41。

考虑到倍增技巧在每个时期后重置了基础预报器的权重，人们可能会想知道是否可以通过更直接的论证获得更好的界限。事实上，我们可以完全避免倍增技巧，通过使用具有时间变化潜力的加权平均预报器来实现这一点。也就是说，我们让指数潜力的参数 η 依赖于轮次编号 t 。由于对指数潜力的最佳非均匀界限是通过选择 $\eta = \sqrt{8(\ln N)/n}$ 获得的，因此对时间变化的指数潜力的自然选择是 $\eta_t = \sqrt{8(\ln N)/t}$ 。通过适应用于证明定理 2.2 的方法，我们可以为这种 η_t 的选择获得一个主要项为 $2\sqrt{(n/2) \ln N}$ 的遗憾界限，因此比倍增技巧的界限更好。更准确地说，我们证明了以下结果。

定理 2.3 假设损失函数 ℓ 在其第一个参数上是凸的，并且取值范围在 $[0, 1]$ 之间。对于所有 $n \geq 1$ 和所有 $y_1, \dots, y_n \in \mathcal{Y}$ ，具有时间变化参数 $\eta_t = \sqrt{8(\ln N)/t}$ 的指数加权平均预报器的遗憾满足

$$\hat{L}_n - \min_{i=1, \dots, N} L_{i,n} \leq 2\sqrt{\frac{n}{2} \ln N} + \sqrt{\frac{\ln N}{8}}.$$

时间变化的潜力函数下的指数加权平均预测器的预测公式为 $\hat{p}_t = \sum_{i=1}^N f_{i,t} w_{i,t-1} / W_{t-1}$ ，其中 $W_{t-1} = \sum_{j=1}^N w_{j,t-1}$ ， $w_{i,t-1} = e^{-\eta_t L_{i,t-1}}$ 。潜力参数选择为 $\eta_t = \sqrt{a(\ln N)/t}$ ，其中 $a > 0$ 由分析确定。我们用 $w'_{i,t-1} = e^{-\eta_{t-1} L_{i,t-1}}$ 来表示权重 $w_{i,t-1}$ ，其中参数 η_t 被替换为 η_{t-1} 。最后，我们用 k_t 来表示在前 t 轮后遗憾最小的专家（若有并列，则选择索引最小的专家）。也就是说， $L_{k_t,t} = \min_{i \leq N} L_{i,t}$ 。在定理的证明中，我们还用到以下技术引理。

引理 2.3 对于所有 $N \geq 2$ ，对所有 $\beta \geq \alpha \geq 0$ ，以及所有 $d_1, \dots, d_N \geq 0$ 满足 $\sum_{i=1}^N e^{-\alpha d_i} \geq 1$ ，

$$\ln \frac{\sum_{i=1}^N e^{-\alpha d_i}}{\sum_{j=1}^N e^{-\beta d_j}} \leq \frac{\beta - \alpha}{\alpha} \ln N$$

证明。 我们首先写出

$$\ln \frac{\sum_{i=1}^N e^{-\alpha d_i}}{\sum_{j=1}^N e^{-\beta d_j}} = \ln \frac{\sum_{i=1}^N e^{-\alpha d_i}}{\sum_{j=1}^N e^{(\alpha-\beta)d_j} e^{-\alpha d_j}} = -\ln \mathbb{E} \left[e^{(\alpha-\beta)D} \right] \leq (\beta - \alpha) \mathbb{E} D$$

根据 Jensen 不等式，其中 D 是一个随机变量，其取值为 d_i 的概率为 $e^{-\alpha d_i} / \sum_{j=1}^N e^{-\alpha d_j}$ ，对每个

$i = 1, \dots, N$ 。由于 D 取值最多为 N 个不同的值，它的熵 $H(D)$ 至多为 $\ln N$ （见附录 A.2）。因此，

$$\begin{aligned}\ln N &\geq H(D) \\ &= \sum_{i=1}^N e^{-\alpha d_i} \left(\alpha d_i + \ln \sum_{k=1}^N e^{-\alpha d_k} \right) \frac{1}{\sum_{j=1}^N e^{-\alpha d_j}} \\ &= \alpha \mathbb{E}D + \ln \sum_{k=1}^N e^{-\alpha d_k} \\ &\geq \alpha \mathbb{E}D\end{aligned}$$

最后的不等式成立因为 $\sum_{i=1}^N e^{-\alpha d_i} \geq 1$ 。因此， $\mathbb{E}D \leq (\ln N)/\alpha$ 。由于假设 $\beta > \alpha$ ，我们可以在上述推导中代入 $\mathbb{E}D$ 的上界，从而完成证明。

现在我们准备证明主要定理。

证明. (定理 2.3) 与定理 2.2 的证明类似，我们研究 $\ln(W_t/W_{t-1})$ 的演变。然而，在这里，我们需要将其与 $\ln(w_{k_{t-1},t-1}/w_{k_t,t})$ 结合起来，同时在两个项中都包含时间变化参数 η_t 。通过跟踪当前最佳的专家，我们用 k_t 来下界权重 $\ln(w_{k_t,t}/W_t)$ 。事实上，总体最佳专家的权重（经过 n 轮之后）在预测过程中可能变得任意小。因此，我们写出如下：

$$\begin{aligned}&\frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t-1}}{W_{t-1}} - \frac{1}{\eta_{t+1}} \ln \frac{w_{k_t,t}}{W_t} \\ &= \underbrace{\left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \ln \frac{W_t}{w_{k_t,t}}}_{(A)} + \underbrace{\frac{1}{\eta_t} \ln \frac{w'_{k_t,t}/W'_t}{w_{k_t,t}/W_t}}_{(B)} + \underbrace{\frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t-1}/W_{t-1}}{w'_{k_t,t}/W'_t}}_{(C)}\end{aligned}$$

我们现在分别对右边的三个项进行界定。项 (A) 很容易界定，因为 $\eta_{t+1} < \eta_t$ ，并且利用 k_t 是在前 t 轮后损失最小的专家这一事实。因此， $w_{k_t,t}/W_t$ 必须至少为 $1/N$ 。因此我们有

$$(A) = \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \ln \frac{W_t}{w_{k_t,t}} \leq \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \ln N$$

我们接着如下界定项 (B)：

$$\begin{aligned}(B) &= \frac{1}{\eta_t} \ln \frac{w'_{k_t,t}/W'_t}{w_{k_t,t}/W_t} = \frac{1}{\eta_t} \ln \frac{\sum_{i=1}^N e^{-\eta_{t+1}(L_{i,t} - L_{k_t,t})}}{\sum_{j=1}^N e^{-\eta_t(L_{j,t} - L_{k_t,t})}} \\ &\leq \frac{\eta_t - \eta_{t+1}}{\eta_t \eta_{t+1}} \ln N = \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \ln N\end{aligned}$$

这个不等式通过应用引理 2.3 证明，其中 $d_i = L_{i,t} - L_{k_t,t}$ 。注意到 $d_i \geq 0$ ，因为 k_t 是前 t 轮后损失最小的专家，并且 $\sum_{i=1}^N e^{-\eta_{t+1}d_i} \geq 1$ ，因为对于 $i = k_{t+1}$ ，我们有 $d_i = 0$ 。项 (C) 首先分解如下：

$$(C) = \frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t-1}/W_{t-1}}{w'_{k_t,t}/W'_t} = \frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t-1}}{w'_{k_t,t}} + \frac{1}{\eta_t} \ln \frac{W'_t}{W_{t-1}}$$

我们分别处理右边的两个子项。对于第一个子项，我们有

$$\frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t-1}}{w'_{k_t,t}} = \frac{1}{\eta_t} \ln \frac{e^{-\eta_t L_{k_{t-1},t-1}}}{e^{-\eta_t L_{k_t,t}}} = L_{k_t,t} - L_{k_{t-1},t-1}$$

对于第二个子项，我们按照定理 2.2 的证明方法，对随机变量 Z 应用 Hoeffding 界限（引理 2.2），该随机变量的值为 $\ell(f_{i,t}, y_t)$ ，其概率为 $w_{i,t-1}/W_{t-1}$ ，对于每个 $i = 1, \dots, N$ ：

$$\begin{aligned} \frac{1}{\eta_t} \ln \frac{W'_t}{W_{t-1}} &= \frac{1}{\eta_t} \ln \sum_{i=1}^N \frac{w_{i,t-1}}{W_{t-1}} e^{-\eta_t \ell(f_{i,t}, y_t)} \\ &\leq - \sum_{i=1}^N \frac{w_{i,t-1}}{W_{t-1}} \ell(f_{i,t}, y_t) + \frac{\eta_t}{8} \\ &\leq -\ell(\hat{p}_t, y_t) + \frac{\eta_t}{8} \end{aligned}$$

其中在最后一步中，我们利用了损失函数 ℓ 的凸性。最后，我们将项 (A) 和 (B) 的界限以及项 (C) 的两个子项的界限代入主方程。通过解出 $\ell(\hat{p}_t, y_t)$ ，我们得到：

$$\begin{aligned} \ell(\hat{p}_t, y_t) &\leq (L_{k_t,t} - L_{k_{t-1},t-1}) + \frac{\sqrt{a \ln N}}{8} \frac{1}{\sqrt{t}} \\ &\quad + \frac{1}{\eta_{t+1}} \ln \frac{w_{k_t,t}}{W_t} - \frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t-1}}{W_{t-1}} \\ &\quad + 2 \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) \ln N \end{aligned}$$

我们将上述不等式应用于每个 $t = 1, \dots, n$ 并求和，利用 $\sum_{t=1}^n \ell(\hat{p}_t, y_t) = \hat{L}_n$ ， $\sum_{t=1}^n (L_{k_t,t} - L_{k_{t-1},t-1}) = \min_{i=1, \dots, N} L_{i,n}$ ， $\sum_{t=1}^n 1/\sqrt{t} \leq 2\sqrt{n}$ ，以及

$$\begin{aligned} \sum_{t=1}^n \left(\frac{1}{\eta_{t+1}} \ln \frac{w_{k_t,t}}{W_t} - \frac{1}{\eta_t} \ln \frac{w_{k_{t-1},t-1}}{W_{t-1}} \right) &\leq -\frac{1}{\eta_1} \ln \frac{w_{k_0,0}}{W_0} = \sqrt{\frac{\ln N}{a}} \\ \sum_{t=1}^n \left(\frac{1}{\eta_{t+1}} - \frac{1}{\eta_t} \right) &= \sqrt{\frac{n+1}{a(\ln N)}} - \sqrt{\frac{1}{a(\ln N)}} \end{aligned}$$

因此，我们可以写出界限：

$$\hat{L}_n \leq \min_{i=1, \dots, N} L_{i,n} + \frac{\sqrt{an \ln N}}{4} + 2\sqrt{\frac{(n+1) \ln N}{a}} - \sqrt{\frac{\ln N}{a}}$$

最后，通过过度估计并选择 $a = 8$ 来平衡两个主要项，我们得到：

$$\hat{L}_n \leq \min_{i=1, \dots, N} L_{i,n} + 2\sqrt{\frac{n}{2} \ln N} + \sqrt{\frac{\ln N}{8}}.$$

2.4 小损失的提升

对于指数加权平均预测器的遗憾界限（在定理 2.2 中给出）可以显著改进，特别是当事先知道最佳专家的累计损失会很小时。在某些情况下，正如我们将看到的，即使没有任何先验知识，这种改进也可能实现。

要理解为什么在最佳专家的累计损失较小时可以期望更好的遗憾界限，请回忆一下引言中描述的简单示例，其中 $\mathcal{Y} = \mathcal{D} = \{0, 1\}$ 和 $\ell(\hat{p}, y) = |\hat{p} - y| \in \{0, 1\}$ （这个示例违反了我们假设的 \mathcal{D} 是凸集，但有助于理解基本现象）。如果预测器事先知道 N 个专家中有一个将会遭受零损失，即 $\min_{i=1, \dots, N} L_{i,n} = 0$ ，那么他可以使用以下简单的“多数投票”方法进行预测。在时间 $t = 1$ 时，如果至少一半的 N 个专家预

测为 0，则预测 $\hat{p}_1 = 0$ 。在第一个比特 y_1 揭示后，丢弃所有预测值 $f_{i,1} \neq y_1$ 的专家。在时间 $t = 2$ 时，如果至少一半的剩余专家预测为 0，则预测 $\hat{p}_2 = 0$ ，在 y_2 揭示后丢弃所有错误预测的专家，以此类推。因此，每次预测器犯错时，至少会丢弃一半的存活专家（因为预测器总是根据剩余专家的多数进行投票）。如果只剩下一个专家，预测器就不会再犯错。因此，预测器的总错误次数（在这种情况下与其遗憾一致）最多为 $\lfloor \log_2 N \rfloor$ 。

在本节中，我们展示了当预测器获得信息表明最佳专家将不会遭受任何损失时，所有有界和凸损失在凸决策空间上可以获得 $O(\ln N)$ 形式的遗憾界限。请注意，这种界限明显优于 $\sqrt{(n/2) \ln N}$ ，后者与最佳专家的损失无关。

为了简化表示，我们写 $L_n^* = \min_{i=1,\dots,N} L_{i,n}$ 。我们现在展示，当事先知道 L_n^* 时，可以选择指数加权平均预测器的参数 η ，使得他的遗憾由 $\sqrt{2L_n^* \ln N} + \ln N$ 进行界定，当 $L_n^* = 0$ 时，这等于 $\ln N$ ，而当 L_n^* 为 n 的数量级时，这个界限为 $\sqrt{n \ln N}$ 。我们主要使用以下定理 2.4 的改进版作为工具。

定理 2.4 假设损失函数 ℓ 在其第一个参数上是凸的，并且取值在 $[0, 1]$ 之间。那么对于任意 $\eta > 0$ ，指数加权平均预测器的遗憾满足：

$$\hat{L}_n \leq \frac{\eta L_n^* + \ln N}{1 - e^{-\eta}}$$

可以很容易地看到，在某些情况下，即使没有事先选择 η 的信息，选择一个非正式的 η 仍然可以得到良好的遗憾界限。

推论 2.3 假设使用 $\eta = 1$ 的指数加权平均预测器。那么，在定理 2.4 的条件下，

$$\hat{L}_n \leq \frac{e}{e-1} (L_n^* + \ln N)$$

如果 $L_n^* \ll \sqrt{n}$ ，这个界限比定理 2.2 的一般界限要好得多，但在其他情况下可能会差得很多。

我们现在通过调整定理 2.4 中的 η ，根据最佳专家的总损失 L_n^* ，推导出一个新的界限。

推论 2.4 假设使用 $\eta = \ln \left(1 + \sqrt{(2 \ln N)/L_n^*} \right)$ 的指数加权平均预测器，其中 $L_n^* > 0$ 假定事先已知。那么，在定理 2.4 的条件下，

$$\hat{L}_n - L_n^* \leq \sqrt{2L_n^* \ln N} + \ln N$$

证明。 利用定理 2.4，我们只需证明，对于我们选择的 η ，

$$\frac{\ln N + \eta L_n^*}{1 - e^{-\eta}} \leq L_n^* + \ln N + \sqrt{2L_n^* \ln N}$$

我们从基本不等式 $(e^\eta - e^{-\eta})/2 = \sinh(\eta) \geq \eta$ 开始，该不等式对所有 $\eta \geq 0$ 都成立。将左边 (2.2) 的分子中的 η 替换为这个上界，我们发现 (2.2) 可以由以下不等式推导：

$$\frac{\ln N}{1 - e^{-\eta}} + \frac{1 + e^{-\eta}}{2e^{-\eta}} L_n^* \leq L_n^* + \ln N + \sqrt{2L_n^* \ln N}$$

证明完毕，通过注意到上述不等式在我们选择的 η 下成立时等号也成立。

当然， L_n^* 这个量只有在第 n 轮预测之后才能得到。没有这个信息可以通过让 η 根据当前最佳专家的损失来变化来进行补偿，类似于第 2.3 节中所展示的方法（参见习题 2.10）。通过这种方法得到的遗憾界限是形如 $2\sqrt{2L_n^* \ln N} + c \ln N$ 的，其中 $c > 1$ 是一个常数。需要注意的是，与定理 2.3 类似，使用时间变化的参数 η_t 会导致界限的主项是固定选择并且根据预测范围 n （如定理 2.2 中）或最佳专家的损失 L_n^* （如推论 2.4 中）最优选择的 η 得到的主项的两倍。

证明. (定理 2.4) 证明过程是对定理 2.2 的简单修改。唯一的不同是，霍夫丁不等式现在被引理 A.3（见附录）替代。回忆定理 2.2 的证明，我们有

$$-\eta L_n^* - \ln N \leq \ln \frac{W_n}{W_0} = \sum_{t=1}^n \ln \frac{W_t}{W_{t-1}} = \sum_{t=1}^n \ln \frac{\sum_{i=1}^N w_{i,t-1} e^{-\eta \ell(f_{i,t}, y_t)}}{\sum_{j=1}^N w_{j,t-1}}$$

我们对随机变量 X_t 应用引理 A.3, 该随机变量在每个 $i = 1, \dots, N$ 下的值是 $\ell(f_{i,t}, y_t)$, 其概率是 $w_{i,t-1}/W_{t-1}$ 。注意，由于损失函数的凸性和詹森不等式， $\mathbb{E}X_t \geq \ell(\hat{p}_t, y_t)$ ，因此根据引理 A.3，

$$\ln \frac{\sum_{i=1}^N w_{i,t-1} e^{-\eta \ell(f_{i,t}, y_t)}}{\sum_{j=1}^N w_{j,t-1}} = \ln \mathbb{E} [e^{-\eta X_t}] \leq (e^{-\eta} - 1) \mathbb{E}X_t \leq (e^{-\eta} - 1) \ell(\hat{p}_t, y_t)$$

因此，

$$-\eta L_n^* - \ln N \leq \sum_{t=1}^n \ln \frac{\sum_{i=1}^N w_{i,t-1} e^{-\eta \ell(f_{i,t}, y_t)}}{\sum_{j=1}^N w_{j,t-1}} \leq (e^{-\eta} - 1) \hat{L}_n$$

解出 \hat{L}_n 得到结果。

2.5 使用损失梯度的预测器

再考虑一次指数加权平均预测器，其预测定义为

$$\hat{p}_t = \frac{\sum_{i=1}^N w_{i,t-1} f_{i,t}}{\sum_{j=1}^N w_{j,t-1}}$$

其中，专家 i 在第 t 轮的权重 $w_{i,t-1}$ 定义为 $w_{i,t-1} = e^{-\eta L_{i,t-1}}$ 。在这一节中，我们引入并分析了一种不同的指数加权平均预测器，在这种预测器中， $w_{i,t-1}$ 指数中的累计损失 $L_{i,t-1}$ 被替换为损失的梯度，这些梯度在时间 $t-1$ 累加。这个新的预测器类将在第 11 章中进行推广，在那里我们还提供了广泛的分析和动机。

回顾一下，决策空间 \mathcal{D} 是线性空间的一个凸子集。在本节中，我们还假设 \mathcal{D} 是有限维的，尽管这个假设可以很容易地放宽。如果 ℓ 是可微的，我们用 $\nabla \ell(\hat{p}, y)$ 来表示相对于第一个参数 $\hat{p} \in \mathcal{D}$ 的梯度。

定义基于梯度的指数加权平均预测器为：

$$\hat{p}_t = \frac{\sum_{i=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \nabla \ell(\hat{p}_s, y_s) \cdot f_{i,s}\right) f_{i,t}}{\sum_{j=1}^N \exp\left(-\eta \sum_{s=1}^{t-1} \nabla \ell(\hat{p}_s, y_s) \cdot f_{j,s}\right)}$$

要理解这个预测器背后的直觉，注意到如果内积 $\nabla \ell(\hat{p}_s, y_s) \cdot f_{i,s}$ 的总和在过去较大，专家 i 被分配的权重就会较小。当专家 i 的建议 $f_{i,s}$ 指向损失函数增大的方向时，这个内积会较大。这样的大值意味着，如果给该专家分配稍大的权重， s 时刻的损失就会增加。根据基于梯度的指数加权平均预测器的原理，这样的专家的权重应该减少。

当然，这个预测器的预测通常与标准的指数加权平均预测器有所不同。不过，注意到在二元预测且使用绝对损失的特殊情况下（即如果 $\mathcal{D} = [0, 1]$, $\mathcal{Y} = \{0, 1\}$ 且 $\ell(x, y) = |x - y|$ ），这两个预测器的预测是相同的（参见习题）。

我们现在展示，在适当的梯度范数条件下，新预测器的遗憾（regret）可以用与推论 2.2 中用于限制标准指数加权平均预测器遗憾的相同数量来界定。

推论 2.5 假设决策空间 \mathcal{D} 是欧几里得单位球 $\{q \in \mathbb{R}^d : \|q\| \leq 1\}$ 的凸子集, 损失函数 ℓ 在其第一个参数上是凸的, 且其梯度 $\nabla \ell$ 存在并满足 $\|\nabla \ell\| \leq 1$ 。对于任意的 n 和 $\eta > 0$, 以及任意的 $y_1, \dots, y_n \in \mathcal{Y}$, 基于梯度的指数加权平均预测器的遗憾满足

$$\widehat{L}_n - \min_{i=1, \dots, N} L_{i,n} \leq \frac{\ln N}{\eta} + \frac{n\eta}{2}$$

证明. 该预测器使用的权重向量 $\mathbf{w}_{t-1} = (w_{1,t-1}, \dots, w_{N,t-1})$ 的分量为:

$$w_{i,t-1} = \exp \left(-\eta \sum_{s=1}^{t-1} \nabla \ell(\widehat{p}_s, y_s) \cdot f_{i,s} \right)$$

注意, 这些权重对应于基于损失函数 ℓ' 的指数加权平均预测器, ℓ' 在时刻 t 定义为:

$$\ell'(q, y_t) = q \cdot \nabla \ell(\widehat{p}_t, y_t), \quad q \in \mathcal{D}$$

根据假设, ℓ' 的取值范围为 $[-1, 1]$ 。在将 ℓ' 缩放至 $[0, 1]$ 后 (参见第 2.6 节), 应用定理 2.2, 我们得到:

$$\begin{aligned} \max_{i=1, \dots, N} \sum_{t=1}^n (\widehat{p}_t - f_{i,t}) \cdot \nabla \ell(\widehat{p}_t, y_t) &= \sum_{t=1}^n \ell'(\widehat{p}_t, y_t) - \min_{i=1, \dots, N} \sum_{t=1}^n \ell'(f_{i,t}, y_t) \\ &\leq \frac{\ln N}{\eta} + \frac{n\eta}{2}. \end{aligned}$$

通过如下展开 $\ell(f_{i,t}, y_t)$, 完成证明:

$$\ell(\widehat{p}_t, y_t) - \ell(f_{i,t}, y_t) \leq (\widehat{p}_t - f_{i,t}) \cdot \nabla \ell(\widehat{p}_t, y_t),$$

由此推出:

$$\widehat{L}_n - \min_{i=1, \dots, N} L_{i,n} \leq \sum_{t=1}^n \ell'(\widehat{p}_t, y_t) - \min_{i=1, \dots, N} \sum_{t=1}^n \ell'(f_{i,t}, y_t).$$

2.6 缩放损失和带符号的博弈

到目前为止, 我们一直假设损失函数 ℓ 的取值范围是单位区间 $[0, 1]$ 。现在, 我们来探讨一下对该取值范围的缩放和平移如何影响预测策略及其表现。

首先考虑一个损失函数 $\ell \in [0, M]$ 的情况。如果 M 是已知的, 我们可以对缩放后的损失 ℓ/M 使用加权平均预测器, 并直接应用针对 $[0, 1]$ 情况的分析。例如, 对于指数加权平均预测器, 推论 2.4 应用于这些缩放后的损失时, 得出的遗憾上界为:

$$\widehat{L}_n - L_n^* \leq \sqrt{2L_n^* M \ln N} + M \ln N$$

其中, $M \ln N$ 这一加法项是必要的。事实上, 如果 ℓ 对于所有 $p \in \mathcal{D}$, 存在 $p' \in \mathcal{D}$ 和 $y \in \mathcal{Y}$ 使得 $\ell(p, y) = M$ 且 $\ell(p', y) = 0$, 那么可以选择专家建议, 使得任意预测器在某个结果序列上遭受的累计损失至少为 $M \log N$, 而此时 $L_n^* = 0$ 。

接下来考虑平移后的区间 $[-M, 0]$ 。如果我们将负的损失视为收益, 我们可以引入遗憾 $G_n^* - \widehat{G}_n$, 其衡量的是最佳专家的累计收益 $G_n^* = -L_n^* = \max_{i=1, \dots, N} (-L_{i,n})$ 与预测器的累计收益 $\widehat{G}_n = -\widehat{L}_n$ 之间的差异。同样, 如果 M 是已知的, 我们可以对缩放后的收益 $(-\ell)/M$ 使用加权平均预测器, 并应用针对 $[0, 1]$ 区间损失函数的分析。通过适应推论 2.4, 我们可以得到如下形式的界:

$$G_n^* - \widehat{G}_n \leq \sqrt{2G_n^* M \ln N} + M \ln N$$

注意，现在遗憾与最大的累计收益 G_n^* 成比例。

接下来，我们讨论更一般的情况，即使用通用收益函数 $h: \mathcal{D} \times \mathcal{Y} \rightarrow \mathbb{R}$ 来对预测器进行评分，该收益函数在其第一个参数上是凹的。预测器的目标是最大化其累计收益。相应的遗憾定义为：

$$\max_{i=1, \dots, N} \sum_{t=1}^n h(f_{i,t}, y_t) - \sum_{t=1}^n h(\hat{p}_t, y_t) = H_n^* - \hat{H}_n$$

如果收益函数 h 的取值范围是 $[0, M]$ ，那么这是一个收益函数，预测器就参与了收益博弈。类似地，如果 h 的取值范围在 $[-M, 0]$ ，那么它就是损失函数的相反数，预测器参与的是损失博弈。最后，如果 h 的取值范围包含 0 的邻域，则预测器参与的是带符号的博弈。

转化为上述术语，本节开始提出的论点表明，在任何无符号博弈中（即任何损失博弈 $[-M, 0]$ 或收益博弈 $[0, M]$ ），当 M 是已知的情况下，对收益进行重新缩放会产生数量级为 $\sqrt{|H_n^*| M \ln N}$ 的遗憾界。然而，对于带符号的博弈，单纯的缩放并不够。如果 $h \in [-M, M]$ ，那么可以通过线性变换 $h \mapsto (h + M)/(2M)$ 来将问题转化为 $[0, 1]$ 情况。应用该变换并采用推论 2.4 的分析，我们得到的遗憾边界为：

$$H_n^* - \hat{H}_n \leq \sqrt{4(H_n^* + Mn)(M \ln N)} + 2M \ln N = O(M\sqrt{n \ln N})$$

这表明，对于带符号的博弈，将其简单地转化为 $[0, 1]$ 的情况可能并不是最好的策略。理想情况下，我们希望将领先项中的 n 替换为类似 $|h(f_{i,1}, y_1)| + \dots + |h(f_{i,n}, y_n)|$ 这样的量。接下来的章节中，我们将展示，在某些情况下，我们甚至可以得到比这更好的结果。

2.7 多线性预测器

势函数为设计加权平均预测器提供了一个便利的工具。然而，如本节所示，针对带符号的博弈，设计优秀的预测器也可以不使用势函数。

首先固定一个带符号的博弈，其收益函数为 $h: \mathcal{D} \times \mathcal{Y} \rightarrow \mathbb{R}$ ，接着我们考虑一个加权平均预测器，它在时间 t 的预测为：

$$\hat{p}_t = \frac{\sum_{i=1}^N w_{i,t-1} f_{i,t}}{W_{t-1}}$$

其中 $W_{t-1} = \sum_{j=1}^N w_{j,t-1}$ 。该预测器的权重 $w_{i,t-1}$ 是递归定义的，如下所示：

$$w_{i,t} = \begin{cases} 1 & \text{如果 } t = 0 \\ w_{i,t-1} (1 + \eta h(f_{i,t}, y_t)) & \text{否则} \end{cases}$$

其中 $\eta > 0$ 是预测器的一个参数。由于 $w_{i,t}$ 是收益的多线性形式，我们称这个预测器为多线性预测器。

需要注意的是，权重 $w_{1,t}, \dots, w_{N,t}$ 不能用各个成分的遗憾 \mathbf{R}_t ，即 $H_{i,n} - \hat{H}_n$ ，来表达。另一方面，由于 $(1 + \eta h) \approx e^{\eta h}$ ，多线性预测器的遗憾可以通过类似于定理 2.2 中用于指数加权平均预测器的技术来界定。我们只需要用到下面这个简单引理（证明留作练习）。

引理 2.4 对所有 $z \geq -1/2$ ，有 $\ln(1 + z) \geq z - z^2$ 。

接下来的结果表明，多线性预测器的遗憾自然可以用任意专家的收益平方和来表示。

定理 2.5 假设收益函数 h 在其第一个参数上是凹的，并且满足 $h \in [-M, \infty)$ 。对于任意 n 和 $0 < \eta < 1/(2M)$ ，且对于所有 $y_1, \dots, y_n \in \mathcal{Y}$ ，多线性预测器的遗憾满足：

$$H_{i,n} - \hat{H}_n \leq \frac{\ln N}{\eta} + \eta \sum_{t=1}^n h(f_{i,t}, y_t)^2 \quad \text{对每个 } i = 1, \dots, N$$

证明. 对于任意 $i = 1, \dots, N$, 注意到 $h(f_{i,t}, y_t) \geq -M$ 且 $\eta \leq 1/(2M)$ 意味着 $\eta h(f_{i,t}, y_t) \geq -1/2$ 。因此, 我们可以将引理 2.4 应用于 $\eta h(f_{i,t}, y_t)$, 并得到:

$$\begin{aligned} \ln \frac{W_n}{W_0} &= -\ln N + \ln \prod_{t=1}^n (1 + \eta h(f_{i,t}, y_t)) \\ &= -\ln N + \sum_{t=1}^n \ln (1 + \eta h(f_{i,t}, y_t)) \\ &\geq -\ln N + \sum_{t=1}^n \left(\eta h(f_{i,t}, y_t) - \eta^2 h(f_{i,t}, y_t)^2 \right) \\ &= -\ln N + \eta H_{i,n} - \eta^2 \sum_{t=1}^n h(f_{i,t}, y_t)^2 \end{aligned}$$

另一方面,

$$\begin{aligned} \ln \frac{W_n}{W_0} &= \sum_{t=1}^n \ln \frac{W_t}{W_{t-1}} \\ &= \sum_{t=1}^n \ln \left(\sum_{i=1}^N \hat{p}_{i,t} (1 + \eta h(f_{i,t}, y_t)) \right) \\ &\leq \eta \sum_{t=1}^n \sum_{i=1}^N \hat{p}_{i,t} h(f_{i,t}, y_t) \quad (\text{因为 } \ln(1+x) \leq x \text{ 对 } x > -1) \\ &\leq \eta \hat{H}_n \quad (\text{因为 } h(\cdot, y) \text{ 是凹函数}). \end{aligned}$$

结合 $\ln(W_n/W_0)$ 的上下界, 并除以 $\eta > 0$, 我们得到了所述的边界。

设 $Q_n^* = h(f_{k,1}, y_1)^2 + \dots + h(f_{k,n}, y_n)^2$, 其中 k 满足 $H_{k,n} = H_n^* = \max_{i=1, \dots, N} H_{i,n}$ 。如果 η 是根据 Q_n^* 选择的, 那么定理 2.5 直接推导出以下结果。

推论 2.6 假设多线性预测器的 η 取为:

$$\eta = \min \left\{ \frac{1}{2M}, \sqrt{\frac{\ln N}{Q_n^*}} \right\}$$

其中 $Q_n^* > 0$ 是事先已知的。在定理 2.5 的条件下,

$$H_n^* - \hat{H}_n \leq 2\sqrt{Q_n^* \ln N} + 4M \ln N$$

为了更好地理解这个结果, 考虑一个损失博弈, 其中 $h \in [-M, 0]$, 并且设 $L_n^* = -\max_i H_{i,n}$ 。因为 $Q_n^* \leq ML_n^*$, 多线性预测器的表现保证至多比指数加权平均预测器的遗憾大 $\sqrt{2}$ 倍, 而在该情况下, 指数加权平均预测器的遗憾主要项为 $\sqrt{2L_n^* M \ln N}$ (参见 2.4 节)。然而, 在某些情况下, Q_n^* 可能显著小于 ML_n^* , 因此推论 2.6 提供的界是一个实际的改进。在第 2.8 节中, 我们将展示更仔细地分析指数加权平均预测器时, 可以得到类似的 (虽然不可比的) 二阶遗憾边界。

至今, 如何在不依赖于序列 y_1, \dots, y_n 的先验知识的情况下, 获得秩序为 $\sqrt{Q_n^* \ln N}$ 的遗憾界限仍然是一个开放问题 (见习题 2.14)。实际上, 自适应调节技术的分析, 比如倍增技巧 (doubling trick) 或时间变化的 η , 依赖于决定调节策略的量的单调性。然而, 序列 Q_1^*, Q_2^*, \dots 并不一定是单调的, 因为在第 t 轮和第 $t+1$ 轮中获得最大累积收益的专家可能不同, 导致 Q_t^* 和 Q_{t+1}^* 之间无法直接关联。

2.8 带符号游戏的指数预测器

通过对之前分析的轻微修改可以证明, 指数加权平均预测器在带符号的游戏中也能够获得较小的遗憾。与 2.7 节的多线性预测器类似, 这个新的遗憾界限也是通过专家损失相对于预测器权重分布的方差相

关的二次和来表达的。此外，使用时间变化的势函数使得我们不再需要对最佳累积收益 H_n^* 预先有所了解。

我们首先为使用收益函数的设置重新定义 2.3 节中引入的带时间变化势函数的指数加权平均预测器。给定收益函数 h ，此预测器通过 $\hat{p}_t = \sum_{i=1}^N f_{i,t} w_{i,t-1} / W_{t-1}$ 进行预测，其中 $W_{t-1} = \sum_{j=1}^N w_{j,t-1}$ ， $w_{i,t-1} = e^{\eta_1 H_{i,t-1}}$ ， $H_{i,t-1} = h(f_{i,1}, y_1) + \dots + h(f_{i,t-1}, y_{t-1})$ ，并且我们假设参数序列 η_1, η_2, \dots 是正值。注意， η_1 的值无关紧要，因为对所有 i 来说 $H_{i,0} = 0$ 。

设 X_1, X_2, \dots 是随机变量，使得 $X_t = h(f_{i,t}, y_t)$ 的概率为 $w_{i,t-1} / W_{t-1}$ ，对所有 i 和 t 均如此。接下来的结果（证明留作习题）在以 X_1, \dots, X_n 表示的过程中，基于任何递减的势参数序列，限制了指数预测器的遗憾。需要注意的是，该引理并未对收益函数做任何有界性的假设。

引理 2.5 设 h 是一个对其第一个参数是凹函数的收益函数。指数加权平均预测器在任何递减序列 η_1, η_2, \dots 的参数下运行，满足对于任意 $n \geq 1$ 和任意序列 y_1, \dots, y_n 的结果，

$$H_n^* - \hat{H}_n \leq \left(\frac{2}{\eta_{n+1}} - \frac{1}{\eta_1} \right) \ln N + \sum_{t=1}^n \frac{1}{\eta_t} \ln \mathbb{E} \left[e^{\eta_t (X_t - \mathbb{E} X_t)} \right]$$

设

$$V_t = \sum_{s=1}^t \text{var}(X_s) = \sum_{s=1}^t \mathbb{E} \left[(X_s - \mathbb{E} X_s)^2 \right]$$

我们的下一个结果表明，选择适当的 η_t 序列后，指数预测器在时间 n 的遗憾最多是 $\sqrt{V_n \ln N}$ 级别。然而需要注意的是，该界限并不是闭合形式的，因为 V_n 依赖于所有 i 和 t 的预测器权重 $w_{i,t}$ 。

定理 2.6 设 h 是取值在 $[-M, M]$ 之间，并且对其第一个参数是凹的收益函数。假设指数加权平均预测器运行时使用

$$\eta_t = \min \left\{ \frac{1}{2M}, \sqrt{\frac{2(\sqrt{2}-1)}{e-2}} \sqrt{\frac{\ln N}{V_{t-1}}} \right\}, \quad t = 2, 3, \dots$$

那么对于任意 $n \geq 1$ 和任意结果序列 y_1, \dots, y_n ，

$$H_n^* - \hat{H}_n \leq 4\sqrt{V_n \ln N} + 4M \ln N + (e-2)M$$

证明. 为简洁起见，设

$$C = \sqrt{\frac{2(\sqrt{2}-1)}{e-2}}$$

我们首先应用引理 2.5（假设 $\eta_1 = \eta_2$ ）

$$\begin{aligned} H_n^* - \hat{H}_n &\leq \left(\frac{2}{\eta_{n+1}} - \frac{1}{\eta_1} \right) \ln N + \sum_{t=1}^n \frac{1}{\eta_t} \ln \mathbb{E} \left[e^{\eta_t (X_t - \mathbb{E} X_t)} \right] \\ &\leq 2 \max \left\{ 2M \ln N, \frac{1}{C} \sqrt{V_n \ln N} \right\} + \sum_{t=1}^n \frac{1}{\eta_t} \ln \mathbb{E} \left[e^{\eta_t (X_t - \mathbb{E} X_t)} \right] \end{aligned}$$

由于 $\eta_t \leq 1/(2M)$ ， $\eta_t (X_t - \mathbb{E} X_t) \leq 1$ ，我们可以应用不等式 $e^x \leq 1 + x + (e-2)x^2$ 对所有 $x \leq 1$ 。因此我们得到

$$H_n^* - \hat{H}_n \leq 2 \max \left\{ 2M \ln N, \frac{1}{C} \sqrt{V_n \ln N} \right\} + (e-2) \sum_{t=1}^n \eta_t \text{var}(X_t)$$

接着，我们将 T 定义为 $V_t > M^2$ 的第一个时间步 t 。由于对所有 t ， $\eta_t \leq 1/(2M)$ 且 $V_T \leq 2M^2$ ，我们得到

$$\sum_{t=1}^n \eta_t \text{var}(X_t) \leq M + \sum_{t=T+1}^n \eta_t \text{var}(X_t)$$

我们用 $\eta_t \leq C\sqrt{(\ln N)/V_{t-1}}$ 来界定对 $t \geq 2$ 的和（注意，对于 $t > T, V_{t-1} \geq V_T > M^2 > 0$ ）。这得到

$$\sum_{t=T+1}^n \eta_t \text{var}(X_t) \leq C\sqrt{\ln N} \sum_{t=T+1}^n \frac{V_t - V_{t-1}}{\sqrt{V_{t-1}}}$$

设 $v_t = \text{var}(X_t) = V_t - V_{t-1}$ 。因为 $V_t \leq V_{t-1} + M^2$ 且 $V_{t-1} \geq M^2$ ，我们有

$$\frac{v_t}{\sqrt{V_{t-1}}} = \frac{\sqrt{V_t} + \sqrt{V_{t-1}}}{\sqrt{V_{t-1}}} (\sqrt{V_t} - \sqrt{V_{t-1}}) \leq (\sqrt{2} + 1) (\sqrt{V_t} - \sqrt{V_{t-1}})$$

因此，

$$\sum_{t=T+1}^n \eta_t \text{var}(X_t) \leq \frac{C\sqrt{\ln N}}{\sqrt{2}-1} (\sqrt{V_n} - \sqrt{V_T}) \leq \frac{C}{\sqrt{2}-1} \sqrt{V_n \ln N}$$

将我们选定的 C 代入并进行简单的近似操作后，证明完成。

备注 2.2 定理 2.5、推论 2.6 和定理 2.6 的分析表明，无论是否需要转换收益，线性组合预报器和指数加权平均预报器都能在有符号和无符号的博弈中起作用。此外，这些结果中的后悔界限可能远优于通过显式的收益转换 $h \mapsto (h + M)/(2M)$ 从有符号博弈转化为无符号博弈时得到的不变界限 $M\sqrt{n \ln N}$ （见第 2.6 节）。然而，这些界限并不适用于没有任何预先信息的收益序列。

定理 2.6 中界限的主要项包含 V_n 。这一量小于以下形式的所有量：

$$\sum_{t=1}^n \sum_{i=1}^N \frac{w_{i,t-1}}{W_{t-1}} (h(f_{i,t}, y_t) - \mu_t)^2$$

其中 μ_1, μ_2, \dots 是可以根据后见之明选择的一系列实数，且不需要用于预报器的定义。这为我们提供了一个完整的上界族，并且我们可以选择最方便的 μ_t 序列来进行分析。

为了提供一个具体的例子，设第 t 轮的有效收益范围为 $R_t = \max_{i=1, \dots, N} h(f_{i,t}, y_t) - \min_{j=1, \dots, N} h(f_{j,t}, y_t)$ ，并考虑选择 $\mu_t = \min_{j=1, \dots, N} h(f_{j,t}, y_t) + R_t/2$ 。

推论 2.7 在与定理 2.6 相同的假设下，

$$H_n^* - \hat{H}_n \leq 2\sqrt{(\ln N) \sum_{t=1}^n R_t^2 + 4M \ln N + (e-2)M}$$

内容...

在一个损失博弈中，若 h 的取值范围为 $[-M, 0]$ ，推论 2.7 说明后悔被一个由 $2M\sqrt{n \ln N}$ 主导的量所限制。将此与定理 2.3 的界限进行比较，我们可以看到，为了得到更广泛的结果，后悔仅损失了一个 $\sqrt{2}$ 的系数。

2.9 可模拟专家

到目前为止，我们将专家视为未指定的实体，在每一轮 t 提供建议，预测者可以访问这些建议。另一种情况是，当专家本身可以被预测者访问时，预测者可以进行任意实验，以揭示专家未来的行为。在这种情况下，我们可以使用函数序列 $f_{E,t} : \mathcal{Y}^{t-1} \rightarrow \mathcal{D}, t = 1, 2, \dots$ 来定义一个专家 E ，在每一时刻 t ，专家根据 $f_{E,t}(y^{t-1})$ 进行预测。我们还假设预测者可以访问这些函数，因此可以在任何时刻假设未来的结果并计算

所有专家对特定结果序列的未来预测。这样，预测者可以“模拟”专家的未来反应，我们称这些专家为可模拟专家。例如，对于一个预测问题中 $\mathcal{D} = \mathcal{Y}$ 的可模拟专家是 E ，使得 $f_{E,t}(y^{t-1}) = (y_1 + \dots + y_{t-1}) / (t-1)$ 。请注意，我们假设在时间 t ，可模拟专家的预测仅依赖于过去观察到的结果序列 y^{t-1} 。对于更一般的专家来说，这并不成立，他们的建议可能依赖于预测者不可见的任意信息来源。更重要的是，一般专家在时间 t 的预测可能依赖于预测者过去的决策 $\hat{p}_1, \dots, \hat{p}_{t-1}$ （回想一下专家建议的预测博弈协议），而可模拟专家的输出仅依赖于过去的结果序列。由于这里我们不涉及计算问题，因此允许 $f_{E,t}$ 是 y^{t-1} 的任意函数，并假设预测者总能计算出这样的函数。

一种特殊类型的可模拟专家是静态专家。当专家 E 的预测 $f_{E,t}$ 仅依赖于轮次 t 而不依赖于 y^{t-1} 时，我们称其为静态专家。换句话说，函数 $f_{E,t}$ 都是常值函数。因此，静态专家 E 可以通过其在每一轮 t 的预测序列 $f_{E,1}, f_{E,2}, \dots$ 完全描述。这个序列不依赖于实际观察到的结果。因此，我们用 $f = (f_1, f_2, \dots)$ 来表示一个任意的静态专家。稍作滥用符号，我们也用 f 来表示可模拟专家。

可模拟和静态专家为预测者提供了额外的能力。因此，考虑预测者是否可以利用这种额外的能力来减少后悔是很有趣的。对于具体的损失函数，这将在第 8 章和第 9 章中进行深入探讨。

2.10 Minimax 遗憾

在专家建议的预测模型中，目前为止得到的最佳遗憾界限（适用于所有取值为 $[0,1]$ 的凸损失）是 $\sqrt{(n/2) \ln N}$ 。这一结果由指数加权平均预测器在固定的 $n \geq 1$ 时取得。那么，这是否是最优的统一界限？哪种预测器能够为每个特定的损失函数提供最佳的遗憾界限？为了严谨地回答这些问题，我们引入了 *minimax* 遗憾的概念。固定一个损失函数 ℓ 并考虑 N 个通用的专家。定义在时间范围 n 的 minimax 遗憾为

$$V_n^{(N)} = \sup_{(f_{1,1}, \dots, f_{N,1}) \in \mathcal{D}^N} \inf_{\hat{p}_1 \in \mathcal{D}} \sup_{y_1 \in \mathcal{Y}} \sup_{(f_{1,2}, \dots, f_{N,2}) \in \mathcal{D}^N} \inf_{\hat{p}_2 \in \mathcal{D}} \sup_{y_2 \in \mathcal{Y}} \dots \sup_{(f_{1,n}, \dots, f_{N,n}) \in \mathcal{D}^N} \inf_{\hat{p}_n \in \mathcal{D}} \sup_{y_n \in \mathcal{Y}} \left(\sum_{t=1}^n \ell(\hat{p}_t, y_t) - \min_{i=1, \dots, N} \sum_{t=1}^n \ell(f_{i,t}, y_t) \right).$$

Minimax 遗憾的一个等价但更简单的定义可以通过使用静态专家给出。定义一个预测策略为每一轮 t 的预测 \hat{p}_t ，该预测基于前 $t-1$ 轮的结果 y_1, \dots, y_{t-1} 和专家建议 $(f_{1,s}, \dots, f_{N,s})$ ，其中 $s = 1, \dots, t$ 。形式上，一个预测策略 P 是一系列函数 $\hat{p}_1, \hat{p}_2, \dots$

$$\hat{p}_t : \mathcal{Y}^{t-1} \times (\mathcal{D}^N)^t \rightarrow \mathcal{D}$$

现在，固定任意的 N 个静态专家类 \mathcal{F} ，并让 $\hat{L}_n(P, \mathcal{F}, y^n)$ 表示预测策略 P 对序列 y^n 使用专家建议 \mathcal{F} 时的累积损失。那么 minimax 遗憾 $V_n^{(N)}$ 可以等价定义为

$$V_n^{(N)} = \inf_P \sup_{\{\mathcal{F} : |\mathcal{F}|=N\}} \sup_{y^n \in \mathcal{Y}^n} \max_{i=1, \dots, N} \left(\hat{L}_n(P, \mathcal{F}, y^n) - \sum_{t=1}^n \ell(f_{i,t}, y_t) \right)$$

其中下确界取于所有预测策略 P ，上确界取于所有可能的 N 个静态专家类（见习题 2.18）。

Minimax 遗憾衡量了对于一个预测算法而言，可以保证在面对所有可能的 N 个专家类和长度为 n 的结果序列时能获得的最佳表现。*Minimax* 遗憾的上界表明存在一个预测策略，无论专家类和结果序列如何，该策略的遗憾都不会超过这个上界。而 *minimax* 遗憾的下界则表明，对于任意预测策略，都会存在某个 N 个专家类和一个结果序列，使得预测者的遗憾至少与这个下界相等。

在本章和接下来的章节中,我们对多种损失函数推导了 **minimax** 遗憾的上界,包括对对数损失 $\ell(x, y) = -\mathbb{I}_{\{y=1\}} \ln x - \mathbb{I}_{\{y=0\}} \ln(1-x)$ (其中 $x \in [0, 1], y \in \{0, 1\}$) 的 $V_n^{(N)} \leq \ln N$ 和对于所有取值为 $[0, 1]$ 的凸损失的 $V_n^{(N)} \leq \sqrt{(n/2) \ln N}$, 两者均由指数加权平均预测器实现。在第 3、8、9 章中, 我们补充了这些结果, 证明了对数损失在 $n \geq \log_2 N$ 时的 $V_n^{(N)} = \ln N$, 且绝对损失 $\ell(x, y) = |x - y|$ 的 *minimax* 遗憾渐进为 $\sqrt{(n/2) \ln N}$, 与我们为凸损失推导的上界相匹配。这意味着指数加权平均预测器在渐进意义上对对数和绝对损失是 **minimax** 最优的。

上面定义的 **minimax** 遗憾基于任何预测器在面对最糟糕的专家类时的表现。然而, 很多情况下人们关心的是一个预测器相对于某个固定专家类的最佳表现。这引出了对于固定(可模拟)专家类的 **minimax** 遗憾定义。固定一个损失函数 ℓ 并设 \mathcal{F} 为一个(不一定有限的)可模拟专家类。基于 \mathcal{F} 的预测策略 P 现在只是一系列函数 $\hat{p}_1, \hat{p}_2, \dots$, 其中 $\hat{p}_t: \mathcal{Y}^{t-1} \rightarrow \mathcal{D}$ 。注意, \hat{p}_t 隐式地依赖于 \mathcal{F} , 而 \mathcal{F} 是固定的。因此, 由于 \mathcal{F} 中的专家是可模拟的, \hat{p}_t 不需要显式地依赖于专家建议。在时间范围 n 上, 相对于 \mathcal{F} 的 **minimax** 遗憾定义为

$$V_n(\mathcal{F}) = \inf_P \sup_{y^n \in \mathcal{Y}^n} \left(\sum_{t=1}^n \ell(\hat{p}_t(y^{t-1}), y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^n \ell(f_t(y^{t-1}), y_t) \right)$$

这种遗憾的概念将在第 8 章和第 9 章中针对特定的损失进行研究。

给定一个可模拟专家的类 \mathcal{F} , 我们还可以定义一个密切相关的量:

$$U_n(\mathcal{F}) = \sup_Q \inf_P \int_{\mathcal{Y}^n} \left(\sum_{t=1}^n \ell(\hat{p}_t(y^{t-1}), y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^n \ell(f_t(y^{t-1}), y_t) \right) dQ(y^n),$$

其中, 上确界取于所有在 \mathcal{Y}^n 上的概率测度, \mathcal{Y}^n 表示长度为 n 的结果序列。 $U_n(\mathcal{F})$ 被称为相对于 \mathcal{F} 的 **maximin** 遗憾。当然, 要定义在 \mathcal{Y}^n 上的概率测度, 结果集 \mathcal{Y} 应该满足某些正则性条件。为了简化, 我们假设 \mathcal{Y} 是 \mathbb{R}^d 的紧子集。这个假设适用于本书中的大多数例子, 当然必要时可以显著放宽。第 7 章中的一个一般的 **minimax** 定理表明, 如果决策空间 \mathcal{D} 是凸的, 并且损失函数 ℓ 在第一个变量上是凸且连续的, 那么

$$V_n(\mathcal{F}) = U_n(\mathcal{F})$$

这个等式的成立仅仅是由于以下事实: 函数

$$F(P, Q) = \int_{\mathcal{Y}^n} \left(\sum_{t=1}^n \ell(\hat{p}_t(y^{t-1}), y_t) - \inf_{f \in \mathcal{F}} \sum_{t=1}^n \ell(f_t(y^{t-1}), y_t) \right) dQ(y^n)$$

在其第一个变量上是凸的, 而在第二个变量上是凹的(实际上是线性的)。这里我们定义了两种预测策略 $P^{(1)} = (\hat{p}_1^{(1)}, \hat{p}_2^{(1)}, \dots)$ 和 $P^{(2)} = (\hat{p}_1^{(2)}, \hat{p}_2^{(2)}, \dots)$ 的凸组合 $\lambda P^{(1)} + (1 - \lambda)P^{(2)}$, 其方法是预测者在时间 t 时根据

$$\lambda \hat{p}_t^{(1)}(y^{t-1}) + (1 - \lambda) \hat{p}_t^{(2)}(y^{t-1})$$

进行预测。关于验证定理 7.1 的条件, 我们将细节留给读者自行完成(参见习题 2.19)。

2.11 折旧遗憾

在许多应用中, 假设过去的损失比最近的损失不那么重要是合理的。因此, 我们可以考虑如下形式的折旧遗憾:

$$\rho_{i,n} = \sum_{t=1}^n \beta_{n-t} r_{i,t}$$

其中折旧因子 β_t 通常随着 t 递减, $r_{i,t} = \ell(\hat{p}_{i,t}, y_t) - \ell(f_{i,t}, y_t)$ 是与专家 i 在第 t 轮的即时遗憾。具体来说, 我们假设 $\beta_0 \geq \beta_1 \geq \beta_2 \geq \dots$ 是一个不减序列, 并且不失一般性地设 $\beta_0 = 1$ 。因此, 在 $t = n$ 时, 当前的遗憾 $r_{i,t}$ 权重为 1, 而过去的遗憾权重较小; 过去的时间越远, 权重越小。

在这种设置下，预测者的目标是确保无论结果序列如何，平均折旧累积遗憾

$$\max_{i=1,\dots,N} \frac{\sum_{t=1}^n \beta_{n-t} r_{i,t}}{\sum_{t=1}^n \beta_{n-t}}$$

尽可能小。更确切地说，希望将平均折旧遗憾界定为一个随 n 收敛至零的函数。本节的目的是探讨对哪些折旧因子的序列可以实现这一目标。当 $\beta_t = 1$ 对所有 t 都成立时，对应于本章其余部分讨论的情形。其他自然的选择包括指数折旧序列 $\beta_t = a^{-t}$ （其中 $a > 1$ ）或 $\beta_t = (t+1)^{-a}$ 形式的序列（ $a > 0$ ）。

首先我们观察到，如果折旧序列下降得太快，除了一些平凡的情况外，无法证明任何有意义的界限。

定理 2.7 假设存在一个正数常数 c ，使得对于每个 n ，存在结果 $y_1, y_2 \in \mathcal{Y}$ 和两个专家 $i \neq i'$ ，使得 $i = \operatorname{argmin}_j \ell(f_{j,n}, y_1)$, $i' = \operatorname{argmin}_j \ell(f_{j,n}, y_2)$ ，并且 $\min_{y=y_1, y_2} |\ell(f_{i,n}, y) - \ell(f_{i',n}, y)| \geq c$ 。如果 $\sum_{t=0}^{\infty} \beta_t < \infty$ ，则存在一个常数 C ，使得对于任何预测策略，都存在一个结果序列使得

$$\max_{i=1,\dots,N} \frac{\sum_{t=1}^n \beta_{n-t} r_{i,t}}{\sum_{t=1}^n \beta_{n-t}} \geq C$$

对于所有 n 都成立。

证明。 下界仅仅通过观察在最后一个时间步 $t = n$ 的遗憾权重大到与整个过去的总权重相当这一事实得出。形式上，

$$\max_{i=1,\dots,N} \sum_{t=1}^n \beta_{n-t} r_{i,t} \geq \max_{i=1,\dots,N} \beta_0 r_{i,n} = \ell(\hat{p}_n, y_n) - \min_{i=1,\dots,N} \ell(f_{i,n}, y_n)$$

因此，

$$\begin{aligned} & \sup_{y \in \mathcal{Y}} \max_{i=1,\dots,N} \frac{\sum_{t=1}^n \beta_{n-t} r_{i,t}}{\sum_{t=1}^n \beta_{n-t}} \\ & \geq \frac{\sup_{y \in \mathcal{Y}} (\ell(\hat{p}_n, y) - \min_{i=1,\dots,N} \ell(f_{i,n}, y))}{\sum_{t=0}^{\infty} \beta_t} \\ & \geq \frac{C}{2 \sum_{t=0}^{\infty} \beta_t}. \end{aligned}$$

接下来，我们通过对比的方式展示，当折旧因子下降得足够慢，以至于 $\sum_{t=0}^{\infty} \beta_t = \infty$ 时，能够使得平均折旧遗憾在 n 足够大时消失。这一结论可以通过简单应用定理 2.1 得到。我们可以通过在加权平均预测器的定义中，将 $r_{i,t}$ 替换为 $\tilde{r}_{i,t} = \beta_{n-t} r_{i,t}$ ，基于折旧遗憾定义加权平均策略。当然，要使用这种预测器，必须提前知道时间跨度 n 。我们得到如下结论。

定理 2.8 考虑一个多项式加权平均折旧预测器，其定义为，对于 $t = 1, \dots, n$ ，

$$\hat{p}_t = \frac{\sum_{i=1}^N \phi' \left(\sum_{s=1}^{t-1} \tilde{r}_{i,s} \right) f_{i,t}}{\sum_{j=1}^N \phi' \left(\sum_{s=1}^{t-1} \tilde{r}_{j,s} \right)} = \frac{\sum_{i=1}^N \phi' \left(\sum_{s=1}^{t-1} \beta_{n-s} r_{i,s} \right) f_{i,t}}{\sum_{j=1}^N \phi' \left(\sum_{s=1}^{t-1} \beta_{n-s} r_{j,s} \right)}$$

其中 $\phi'(x) = (p-1)x^p$ ，且 $p = 2 \ln N$ 。则平均折旧遗憾满足

$$\max_{i=1,\dots,N} \frac{\sum_{t=1}^n \beta_{n-t} r_{i,t}}{\sum_{t=1}^n \beta_{n-t}} \leq \sqrt{2e \ln N} \frac{\sqrt{\sum_{t=1}^n \beta_{n-t}^2}}{\sum_{t=1}^n \beta_{n-t}}$$

（类似的上界也可以证明适用于折旧指数加权平均预测器）。特别地，如果 $\sum_{t=0}^{\infty} \beta_t = \infty$ ，则

$$\max_{i=1,\dots,N} \frac{\sum_{t=1}^n \beta_{n-t} r_{i,t}}{\sum_{t=1}^n \beta_{n-t}} = o(1).$$

证明. 显然, 该预测器满足 Blackwell 条件, 并且对每个 t , $|\tilde{r}_{i,t}| \leq \beta_{n-t}$ 。然后, 定理 2.1 蕴含了类似于推论 2.1 证明中的结果,

$$\max_{i=1,\dots,N} \rho_{i,n} \leq \sqrt{2e \ln N} \sqrt{\sum_{t=1}^n \beta_{n-t}^2}$$

因此得出第一个结论。要证明第二个结论, 只需注意

$$\begin{aligned} \sqrt{2e \ln N} \frac{\sqrt{\sum_{t=1}^n \beta_{n-t}^2}}{\sum_{t=1}^n \beta_{n-t}} &= \sqrt{2e \ln N} \frac{\sqrt{\sum_{t=1}^n \beta_{t-1}^2}}{\sum_{t=1}^n \beta_{t-1}} \\ &\leq \sqrt{2e \ln N} \frac{\sqrt{\sum_{t=1}^n \beta_{t-1}}}{\sum_{t=1}^n \beta_{t-1}} \\ &= \frac{\sqrt{2e \ln N}}{\sqrt{\sum_{t=1}^n \beta_{t-1}}} = o(1) \end{aligned}$$

考虑 $\beta_t = (t+1)^{-a}$ 的特例, 其中 $0 < a \leq 1$ 。(回想定理 2.7 中提到, 对于 $a > 1$, 无法导出有意义的界限。)如果 $a = 1$, 定理 2.8 蕴含

$$\max_{i=1,\dots,N} \frac{\sum_{t=1}^n \beta_{n-t} r_{i,t}}{\sum_{t=1}^n \beta_{n-t}} \leq \frac{C}{\log n}$$

其中常数 $C > 0$ 。这一慢速收敛速率并不令人意外, 考虑到定理 2.7, 因为级数 $\sum_t 1/(t+1)$ 是“几乎不可求和”的。事实上, 该上界无法被显著改善 (见习题 2.20)。然而, 对于 $a < 1$, 收敛速度更快。事实上, 简单计算表明, 定理的上界蕴含了

$$\max_{i=1,\dots,N} \frac{\sum_{t=1}^n \beta_{n-t} r_{i,t}}{\sum_{t=1}^n \beta_{n-t}} = \begin{cases} O(1/\log n) & \text{如果 } a = 1 \\ O(n^{a-1}) & \text{如果 } 1/2 < a < 1 \\ O(\sqrt{(\log n)/n}) & \text{如果 } a = 1/2 \\ O(1/\sqrt{n}) & \text{如果 } a < 1/2 \end{cases}$$

不出意料, 折旧因子下降得越慢, 平均折旧累积遗憾收敛到零的速度就越快。然而, 有趣的是, $a = 1/2$ 时会出现“相变”: 对于所有 $a < 1/2$, 平均遗憾以 $n^{-1/2}$ 的速率下降, 这种行为与不考虑折旧时的情况在定量上非常相似。

2.12 书目备注

我们的专家建议序列预测模型源自重复博弈理论。具有固定损失矩阵的零和重复博弈是博弈论中的经典课题。在这些博弈中, 经过 n 次博弈后的遗憾被定义为行玩家相较于已知列玩家在这 n 次博弈中动作的经验分布情况下, 所承受的额外损失。Hannan [141] 在他的开创性工作中设计了一种随机游戏策略, 其每轮的期望遗憾以 $N\sqrt{3nm/2}$ 的速率增长, 其中 N 是行的数量, m 是损失矩阵的列数, n 是博弈次数。正如第 7 章所展示的, 我们的多项式和指数加权平均预测器可以用来进行零和重复博弈, 获得的遗憾为 $\sqrt{(n/2) \ln N}$ 。与 Hannan 的遗憾相比, 我们获得了相同的 n 次博弈的依赖关系, 但显著改进了对损失矩阵维度 N 和 m 的依赖。另一种具有消失的每轮遗憾的随机玩家策略可以通过著名的 Blackwell 可接近性定理 [28] 推导出来, 该定理将冯·诺伊曼的极小极大定理推广到向量值收益。这一结果在第 7 章中被重新推导出来, 基于与我们多项式加权平均预测器 (当 $p = 2$ 时) 等效的混合策略。Cover [68] 首次给出了极小极大遗憾的精确渐近常数 (针对特殊情况)。在我们的术语中, Cover 研究了使用两个静态专家预测一系列二元结果的问题, 一个专家总是预测 0, 另一个总是预测 1。他展示了在这种特殊情况下, 绝对损失的极小极大遗憾为 $(1 + o(1))\sqrt{n/(2\pi)}$ 。

序列预测问题在没有任何概率假设的情况下，与个体数据序列的压缩问题有着深刻的联系。在这一领域的开创性研究由 Ziv [317,318] 以及 Lempel 和 Ziv [197,317,319] 展开，他们解决了如何几乎与最优有限状态自动机一样好地压缩个体数据序列的问题。Feder、Merhav 和 Gutman [95] 展示，Lempel-Ziv 压缩器可以作为一种随机预测器（针对绝对损失），并且对于所有有限状态专家类，具有消失的每轮遗憾。考虑到该类专家的丰富结构，这一结果令人惊讶。此外，Feder、Merhav 和 Gutman 为同一专家类设计了一种预测器，其收敛速率优于 Lempel-Ziv 预测器可证明的速率（参见 Merhav 和 Feder [213] 的进一步结果）。在第 9 节中，我们继续探讨预测与压缩之间的关系，展示了在简单条件下，使用对数损失的预测在极小极大意义上等价于自适应数据压缩。Vovk 和 Watkins [303] 研究了专家建议预测与个体序列信息内容的关系，他们引入了数据序列的预测复杂性概念，针对对数损失，该量与序列的 Kolmogorov 复杂性相关。我们推荐阅读 Li 和 Vitányi [198] 的书，该书对算法随机性理论提供了出色的介绍。

在 Hannan 和 Blackwell 奠定预测博弈论方法基础的同时，Solomonoff 提出了将人类归纳推理现象形式化为一个序列预测过程的想法。这一研究最终引导他提出了一个用于贝叶斯推理的普遍先验概率 [273-275]。Solomonoff 普遍先验的一个重要“副产品”是算法随机性的概念，他独立于 Kolmogorov 提出了这一概念。虽然我们承认 Solomonoff 在序列预测理论领域，尤其是在与 Kolmogorov 复杂性相关的领域中发挥了关键作用，但本书从不同的角度看待预测问题。尽管如此，我们认为，通过预测的统一概念，探索算法随机性与博弈论之间的联系是一个非常激动人心的研究方向。

归纳推理领域研究当专家来自一个大的类（可能包括所有递归语言或所有部分递归函数）时的序列预测问题，任务是最终确定一个与无限观察序列一致（或几乎一致）的专家。这个学习范式由 Gold [130] 于 1967 年引入，至今仍在积极研究中。与本书中以博弈论为根基的理论不同，归纳推理使用的主要思想和分析工具来自递归理论（参见 Odifreddi [227]）。

在计算机科学中，与专家预测相关的一个领域是在线算法的竞争性分析（参见 Borodin 和 El-Yaniv [36] 的综述专著）。Blum 和 Burch [32] 的文章是一个很好的例子，展示了如何在竞争性分析中使用预测算法。

专家建议预测的范式最早由 Littlestone 和 Warmuth [203] 以及 Vovk [297] 提出，随后由 Cesa-Bianchi、Freund、Haussler、Helmbold、Schapire 和 Warmuth [48] 以及 Vovk [298] 进一步发展，尽管其一些核心成分早在 De Santis、Markowski 和 Wegman [260]，Littlestone [200]，以及 Foster [103] 的论文中就已经出现。序列预测中使用势函数的做法源于 Hart 和 Mas-Colell [146]，他们在博弈论背景下使用了 Blackwell 条件，而 Grove、Littlestone 和 Schuurmans [133] 则在感知器算法某些变体的分析中使用了完全相同的条件（见第 11 章）。我们的定理 2.1 受到 Hart 和 Mas-Colell 关于重复博弈中 Λ 策略分析的启发 [146]，同时部分基于 Grove、Littlestone 和 Schuurmans [133] 的准加性算法分析。我们在此描述的基于势函数的序列预测统一框架最早由 Cesa-Bianchi 和 Lugosi [54] 提出。基于指数势的预测在博弈论中作为光滑拟态博弈的变体被使用（参见如 Fudenberg 和 Levine [119] 的书）。在学习理论中，指数加权平均预测器由 Littlestone 和 Warmuth [203]（加权多数算法）以及 Vovk [297]（聚合算法）提出并进行分析。将多项式势的参数 p 设为 $2 \ln N$ 的技巧归功于 Gentile [123]。第 2.2 节的分析基于 Cesa-Bianchi 的工作 [46]。第 2.3 节中的倍增技巧出现在 Cesa-Bianchi、Freund、Haussler、Helmbold、Schapire 和 Warmuth [48] 以及 Vovk [298] 的文章中，而定理 2.3 的分析则改编自 Auer、Cesa-Bianchi 和 Gentile [13]。第 2.4 节中的数据依赖界基于两个来源：定理 2.4 来自 Littlestone 和 Warmuth [203] 的工作，推论 2.4 则由 Freund 和 Schapire [112] 提出。对具有时间变化 η_t 的指数加权平均预测器的更复杂分析来自 Yaroshinski、El-Yaniv 和 Seiden [315]，他们展示了一个遗憾界为 $(1 + o(1))\sqrt{2L_n^* \ln N}$ ，其中 $o(1) \rightarrow 0$ 当 $L_n^* \rightarrow \infty$ 。Hutter 和 Poland [165] 使用“跟随受扰领导者”的随机预测器证明了与习题 2.10 相似的结果，该随机预测器将在第 4 章中进行分析。

第 2.8 节中的多线性预测器及其结果归功于 Cesa-Bianchi、Mansour 和 Stoltz [57]。推论 2.7 的较弱版本由 Allenberg-Neeman 和 Neeman [7] 证明。第 2.5 节中的基于梯度的预测器由 Kivinen 和 Warmuth [181]

提出。推论 2.5 的证明来自 Cesa-Bianchi [46]。关于可模拟专家和专家框架下最坏情况遗憾的概念最早由 Cesa-Bianchi 等人 [48] 研究。更一般损失函数的结果可以在 Chung 的论文 [60] 中找到。Fudenberg 和 Levine [121] 研究了一个与这里讨论的模型略有不同的折旧遗憾模型。

专家建议预测模型与贝叶斯决策理论相关。例如，当使用绝对损失时，基于指数势的加权平均预测器的归一化权重与简单随机生成模型的后验分布非常接近（见习题 2.7）。从这个角度看，我们的遗憾分析展示了一个贝叶斯决策在某种强意义下具有鲁棒性的例子，因为它们的表现不仅可以相对于序列的随机抽取的期望进行界定，还可以对每一个个体序列进行界定。

2.13 练习

2.1 假设你需要预测一个序列 $Y_1, Y_2, \dots \in \{0, 1\}$ ，这是具有未知分布的独立同分布随机变量。你的决策空间是 $[0, 1]$ ，损失函数为 $\ell(\hat{p}, y) = |\hat{p} - y|$ 。你会如何进行预测？尝试估计你的预测器的累积损失，并将其与两个专家中的最佳专家的累积损失进行比较，其中一个专家始终预测 1，另一个始终预测 0。哪些分布是最“困难”的？你的（期望的）遗憾与加权平均算法相比如何（假设加权平均算法并不知道结果序列是独立同分布的）？

2.2 考虑一个基于势函数的加权平均预测器

$$\Phi(\mathbf{u}) = \psi \left(\sum_{i=1}^N \phi(u_i) \right)$$

进一步假设，定理 2.1 中提到的量 $C(\mathbf{r}_t)$ 对于所有 \mathbf{r}_t 的取值都由常数限制，并且函数 $\psi(\phi(u))$ 是严格凸的。证明存在一个非负序列 $\varepsilon_n \rightarrow 0$ ，使得预测器的累积遗憾满足，对于任意 n 和任意结果序列 y^n ，

$$\frac{1}{n} \left(\max_{i=1, \dots, N} R_{i,n} \right) \leq \varepsilon_n$$

2.3 使用定理 2.1 分析多项式加权平均预测器，但使用的势函数为 $\Phi(\mathbf{u}) = \|\mathbf{u}_+\|_p$ ，而不是推论 2.1 证明中使用的 $\Phi_p(\mathbf{u}) = \|\mathbf{u}_+\|_p^2$ 。推导出与推论 2.1 中相同形式的界，可能会有不同的常数。

2.4 令 $\mathcal{Y} = \{0, 1\}$ ， $\mathcal{D} = [0, 1]$ ，且 $\ell(\hat{p}, y) = |\hat{p} - y|$ 。证明指数加权平均预测器的累积损失 \hat{L} 总是至少等于最佳专家的累积损失 $\min_{i \leq N} L_i$ 。证明对于其他损失函数，例如平方损失 $(\hat{p} - y)^2$ ，这不一定成立。提示：尝试逆向证明定理 2.2。

2.5（非均匀初始权重）根据定义，加权平均预测器使用均匀的初始权重 $w_{i,0} = 1$ ，对于所有 $i = 1, \dots, N$ 。然而，这个选择并没有特别之处，对于这个预测器的遗憾分析可以使用任意非负初始权重集。

考虑一个带有任意初始权重 $w_{1,0}, \dots, w_{N,0} > 0$ 的指数加权平均预测器，其定义为，对于所有 $t = 1, 2, \dots$ ，

$$\hat{p}_t = \frac{\sum_{i=1}^N w_{i,t-1} f_{i,t}}{\sum_{j=1}^N w_{j,t-1}}, \quad w_{i,t} = w_{i,t-1} e^{-\eta \ell(f_{i,t}, y^t)}$$

在定理 2.2 的相同条件下，证明对于任意 n 和任意结果序列 y^n ，

$$\hat{L}_n \leq \min_{i=1, \dots, N} \left(L_{i,n} + \frac{1}{\eta} \ln \frac{1}{w_{i,0}} \right) + \frac{\ln W_0}{\eta} + \frac{\eta}{8} n$$

其中 $W_0 = w_{1,0} + \dots + w_{N,0}$ 。

2.6（许多优秀专家）如果有许多专家的损失较小，那么直观上这些结果序列会更容易预测。改写定理 2.2 的证明，证明指数加权预测器满足以下性质：对于任意 n ，任意结果序列 y^n ，以及任意 $L > 0$ ，

$$\hat{L}_n \leq L + \frac{1}{\eta} \ln \frac{N}{N_L} + \frac{\eta}{8} n$$

其中 N_L 是集合 $\{1 \leq i \leq N : L_{i,n} \leq L\}$ 的基数。

2.7 (结果序列的随机生成) 考虑指数加权平均预测器, 并为生成序列 $y^n \in \{0, 1\}^n$ 定义以下概率模型, 其中我们现在将每个位 y_t 视为一个伯努利随机变量 Y_t 的实现。专家 I 从 N 个专家中随机抽取。对于每一个 $t = 1, \dots, n$, 首先从 $\{0, 1\}$ 中抽取 X_t , 其中 $X_t = 1$ 的概率为 $f_{I,t}$ 。然后, Y_t 以概率 β 设为 X_t , 以概率 $1 - \beta$ 设为 $1 - X_t$, 其中 $\beta = 1/(1 + e^{-\eta})$ 。证明预测器的权重 $w_{i,t}/(w_{1,t} + \dots + w_{N,t})$ 等于后验概率 $P[I = i | Y_1 = y_1, \dots, Y_{t-1} = y_{t-1}]$, 即在观察到序列 y_1, \dots, y_{t-1} 后, 专家 i 被抽取的概率。

2.8 (倍增技巧) 考虑以下预测策略 (“倍增技巧”): 时间被分为多个区间 $(2^m, \dots, 2^{m+1} - 1)$, 其中 $m = 0, 1, 2, \dots$ 。在区间 $(2^m, \dots, 2^{m+1} - 1)$ 内, 策略使用初始时刻为 2^m 的指数加权平均预测器, 参数 $\eta_m = \sqrt{8(\ln N)/2^m}$ 。因此, 每当时间为 2 的整数次幂时, 加权平均预测器会重置, 并使用一个新的 η 值。使用定理 2.2 证明, 对于任意结果序列 $y_1, y_2, \dots \in \mathcal{Y}$, 以及任意 $n \geq 1$, 该预测器的遗憾至多为

$$\hat{L}_n - \min_{i=1, \dots, N} L_{i,n} \leq \frac{\sqrt{2}}{\sqrt{2}-1} \sqrt{\frac{n}{2} \ln N}$$

2.9 (倍增技巧, 续) 在习题 2.8 中, 我们任意地将时间划分为长度为 $2^m, m = 1, 2, \dots$ 的区间。研究如果区间长度变为 $\lfloor a^m \rfloor$, 其中 $a > 0$, 会发生什么。哪个 a 的选择能够使界中的常数在渐近情况下最小? 与文本中给出的界相比, 你可以获得多少改进?

2.10 结合定理 2.4 和练习 2.8 的翻倍技巧, 构建一个预测器, 使得在没有预先了解 L^* 的情况下, 对于所有的 n , 都能达到以下结果:

$$\hat{L}_n - L_n^* \leq 2\sqrt{2L_n^* \ln N} + c \ln N$$

其中, 当损失函数在其第一个变量上是有界且凸的, c 是一个正常数。

2.11 (另一种时间变化的潜在函数) 考虑自适应的指数加权平均预测器, 该预测器在时间 t 使用

$$\eta_t = c \sqrt{\frac{\ln N}{\min_{i=1, \dots, N} L_{i,t-1}}}$$

其中 c 是一个正常数。证明, 当 ℓ 是一个在其第一个变量上凸的 $[0, 1]$ 取值损失函数时, 存在一个合适的 c 使得:

$$\hat{L}_n - L_n^* \leq 2\sqrt{2L_n^* \ln N} + \kappa \ln N$$

其中 $\kappa > 0$ 是一个合适的常数 (Auer, Cesa-Bianchi 和 Gentile [13])。提示: 按照定理 2.3 的证明思路进行。本题不容易。

2.12 考虑 $\mathcal{Y} = \mathcal{D} = [0, 1]$ 且绝对损失 $\ell(\hat{p}, y) = |\hat{p} - y|$ 的预测问题。证明在这种情况下, 基于梯度的指数加权平均预测器与指数加权平均预测器一致。(注意, 当 $\hat{p} = y$ 时, 损失的导数不存在, 因此需要适当调整基于梯度的指数加权平均预测器的定义。)

2.13 证明引理 2.4。

2.14 使用翻倍技巧证明一个不需要预先了解结果序列的推论 2.6 的变体 (但与该推论一样, 我们仍然假设收益函数 h 的取值范围为 $[-M, M]$, 并且在其第一个变量上是凹的)。将遗憾界表示为序列 Q_1^*, Q_2^*, \dots 上最小单调上界的形式 (参见 Cesa-Bianchi, Mansour 和 Stoltz [57])。

2.15 证明定理 2.6 的一个变体, 该变体不需要预先了解收益函数的范围 $[-M, M]$ (参见 Cesa-Bianchi, Mansour 和 Stoltz [57])。提示: 在定义 η_t 时, 将 $1/(2M)$ 替换为 $2^{-(1+k_t)}$, 其中 k 是使得 $\max_{s=1, \dots, t-1} \max_{i=1, \dots, N} |h(f_{i,s}, y_s)| \leq 2^k$ 的最小非负整数。

2.16 使用更新公式 $w_{i,t} = w_{i,t-1}(1 + \eta r_{i,t})$, 其中 $r_{i,t} = h(f_{i,t}, y_t) - h(\hat{p}_t, y_t)$ 是即时遗憾, 证明多线性预测器的遗憾界。你能对专家总权重 $W_t = w_{1,t} + \dots + w_{N,t}$ 的演化得出什么结论?

2.17 证明引理 2.5。提示: 调整定理 2.3 的证明。

2.18 证明第 2.10 节中的极小极大遗憾 $V_n^{(N)}$ 的两种表达式是等价的。

2.19 考虑一个可模拟专家类 \mathcal{F} 。假设结果集 \mathcal{Y} 是 \mathbb{R}^d 的一个紧子集, 决策空间 \mathcal{D} 是凸的, 且损失函数 ℓ 在其第一个变量上是凸且连续的。证明 $V_n(\mathcal{F}) = U_n(\mathcal{F})$ 。提示: 检查定理 7.1 的条件。

2.20 考虑折旧因子 $\beta_t = 1/(t+1)$, 并假设存在一个正数 c 使得对于每一个 n , 存在结果 $y_1, y_2 \in \mathcal{Y}$ 和两个专家 $i \neq i'$, 满足 $i = \operatorname{argmin}_j \ell(f_{j,n}, y_1), i' = \operatorname{argmin}_j \ell(f_{j,n}, y_2)$, 并且 $\min_{y=y_1, y_2} |\ell(f_{i,n}, y) - \ell(f_{i',n}, y)| \geq c$ 。证明存在一个常数 C , 使得对于任何预测策略, 存在一个结果序列满足

$$\max_{i=1, \dots, N} \frac{\sum_{t=1}^n \beta_{n-t} r_{i,t}}{\sum_{t=1}^n \beta_{n-t}} \geq \frac{C}{\log n}$$

对所有 n 成立。