# 3 特定损失的紧边界

## 3.1 引言

在第 2 章中，我们已经证明了在一般情况下，存在一种预测器能够针对任意由 $N$ 个专家组成的有限类，在最坏情况下实现 $\sqrt{n \ln N}$ 数量级的遗憾。我们唯一要求的条件是决策空间 $\mathcal{D}$ 是一个凸集，并且损失函数 $\ell$ 在其第一个参数中是有界且凸的。在许多情况下，基于对损失函数和/或专家类的特定假设，可以获得显著更紧的性能界限。本章的目的是回顾在这些假设下可以实现改进的各种情形。我们还通过给出最坏情况遗憾的下界，探讨这些改进的可能极限。

在第 3.2 节中，我们展示了在某些情况下，一个仅根据过去表现选择最优专家的短视策略，在对损失函数和专家类施加某些光滑性假设的条件下，可以实现快速减少的最坏情况遗憾。

我们的主要技术工具，第 2.1 定理，已经被用来将加权平均预测器的势函数 $\Phi(\mathbf{R}_n)$ 与初始势函数 $\Phi(0)$ 以及一系列项的和进行界定，这些项限制了在每个 $t = 1, \dots, n$ 时对 $\Phi(\mathbf{R}_t)$ 取线性近似时所犯的误差。然而，在某些情况下，我们可以直接通过 $\Phi(\mathbf{0})$ 来限制 $\Phi(\mathbf{R}_n)$，而不需要进行任何线性近似。为此，我们可以结合特定的损失函数，利用势函数所表现出的简单几何特性。在本章中，我们将发展几种此类技术，并使用它们推导出针对各种损失函数的更紧遗憾界限。

在第 3.3 节中，我们讨论了损失函数的一个基本性质，即指数凹性，该性质确保了指数加权平均预测器的遗憾界限为 $(\ln N)/\eta$，其中 $\eta$ 必须小于一个依赖于特定指数凹性损失的临界值。在第 3.4 节中，我们采用了一种更加极端的方法，研究一个选择最小化势函数最坏情况增加的预测器。结果表明，这个“贪心”预测器的表现不会比加权平均预测器差。在第 3.5 节中，我们聚焦于指数势函数，并通过引入聚合预测器来进一步优化前几节的分析，聚合预测器包含了贪心预测器。这些适用于指数凹损失的预测器被设计为实现形如 $c \ln N$ 的遗憾，其中 $c$ 是每个损失函数能够实现这种界限的最佳常数。在分析过程中，我们描述了一个重要的可混合损失子类。这些损失在某种意义上是最容易处理的。最后，在第 3.7 节中，我们证明了针对一般损失的 $\Omega(\log N)$ 形式的下界，并且为绝对损失给出了一个与指数加权平均预测器所实现的上界（包含常数）相匹配的下界。

## 3.2 跟随最优专家

在本节中，我们研究可能是最简单的预测策略。这种策略在时间 $t$ 选择一个专家，该专家在前 $t-1$ 个时间点上最小化了累积损失。换句话说，预测器总是跟随迄今为止累积损失最小的专家。或许令人惊讶的是，这种简单的预测器在损失函数和专家的通用条件下表现良好。

形式上，考虑一个专家类 $\mathcal{E}$，并定义预测器 $\widehat{p}_t$ 为在过去时间段内累积损失最小的专家所作出的预测，即

$$\widehat{p}_t = f_{E,t} \quad \text{当且仅当} \quad E = \operatorname*{argmin}_{E' \in \mathcal{E}} \sum_{s=1}^{t-1} \ell\left(f_{E',s}, y_s\right)$$

$\widehat{p}_1$ 被定义为 $\mathcal{D}$ 中的任意元素。在本节中，我们假设该最小值总是可以实现。如果存在多个最小值，$E$ 可以任意选择。

我们的目标与之前相同，即将 $\widehat{p}$ 的表现与该类中表现最好的专家进行比较，具体而言，就是推导遗憾的界限：

$$\widehat{L}_n - \inf_{E \in \mathcal{E}} L_{E,n} = \sum_{t=1}^{n} \ell\left(\widehat{p}_t, y_t\right) - \inf_{E \in \mathcal{E}} \sum_{t=1}^{n} \ell\left(f_{E,t}, y_t\right)$$

考虑如下定义的假设预测器：

$$p_t^* = f_{E,t} \quad \text{当且仅当} \quad E = \operatorname*{argmin}_{E' \in \mathcal{E}} \sum_{s=1}^{t} \ell\left(f_{E',s}, y_s\right)$$

注意，$p_t^*$ 与 $\widehat{p}_t$ 的定义相似，唯一的区别是 $p_t^*$ 还考虑了在时间 $t$ 所遭受的损失。显然，$p_t^*$ 不是一个“合法的”预测器，因为它允许预见未来；它仅作为分析工具定义。下面的简单引理表明，$p_t^*$ 的表现至少与表现最好的专家相当。

**引理 3.1** *对于任意序列 $y_1, \ldots, y_n$ 的结果，有*

$$\sum_{t=1}^{n} \ell\left(p_t^*, y_t\right) \leq \sum_{t=1}^{n} \ell\left(p_n^*, y_t\right) = \min_{E \in \mathcal{E}} L_{E,n}$$

**证明.** 证明采用归纳法。对于 $n = 1$ 的情况，结论显然成立。现在假设

$$\sum_{t=1}^{n-1} \ell\left(p_t^*, y_t\right) \leq \sum_{t=1}^{n-1} \ell\left(p_{n-1}^*, y_t\right)$$

由于根据定义 $\sum_{t=1}^{n-1} \ell\left(p_{n-1}^*, y_t\right) \leq \sum_{t=1}^{n-1} \ell\left(p_n^*, y_t\right)$，因此归纳假设意味着

$$\sum_{t=1}^{n-1} \ell\left(p_t^*, y_t\right) \leq \sum_{t=1}^{n-1} \ell\left(p_n^*, y_t\right)$$

在两边都加上 $\ell\left(p_n^*, y_n\right)$ 即可得出结果。

这一简单性质意味着预测器 $\widehat{p}$ 的遗憾上界可以表示为

$$\widehat{L}_n - \inf_{E \in \mathcal{E}} L_{E,n} \leq \sum_{t=1}^{n} \left(\ell\left(\widehat{p}_t, y_t\right) - \ell\left(p_t^*, y_t\right)\right)$$

回顾 $\widehat{p}_t$ 和 $p_t^*$ 的定义，在某些情形下，可以合理预期 $\widehat{p}_t$ 和 $p_t^*$ 彼此非常接近。例如，如果能够保证对每个 $t$，

$$\sup_{y \in \mathcal{Y}} \left(\ell\left(\widehat{p}_t, y\right) - \ell\left(p_t^*, y\right)\right) \leq \varepsilon_t$$

对于一列实数 $\varepsilon_t > 0$，则不等式表明

$$\widehat{L}_n - \inf_{E \in \mathcal{E}} L_{E,n} \leq \sum_{t=1}^{n} \varepsilon_t$$

接下来，我们将建立一些一般条件，在这些条件下，$\varepsilon_t \sim 1/t$，这意味着遗憾增长得极为缓慢，仅为 $O(\ln n)$。

在所有这些例子中，我们考虑的是“常量”专家；也就是说，我们假设对于每个 $E \in \mathcal{E}$ 和 $y \in \mathcal{Y}$，专家 $E$ 的损失与时间无关。换句话说，对于任何固定的 $y$，都有 $\ell\left(f_{E,1}, y\right) = \cdots = \ell\left(f_{E,n}, y\right)$。为了简化记号，我们将共同值记为 $\ell(E, y)$，此时每个专家由元素 $E \in \mathcal{D}$ 来表征。

### 3.2.1 平方损失

作为第一个例子，考虑一般向量空间中的平方损失。假设 $\mathcal{D} = \mathcal{Y}$ 是 Hilbert 空间 $\mathcal{H}$ 中的单位球 $\{p : \|p\| \leq 1\}$，损失函数定义为：

$$\ell(p, y) = \|p - y\|^2, \quad p, y \in \mathcal{H}$$

令专家类 $\mathcal{E}$ 包含所有由 $\mathcal{D}$ 索引的常量专家。在这种情况下，可以明确地确定预测器 $\widehat{p}$。实际上，由于对任意 $p \in \mathcal{D}$，有

$$\frac{1}{t-1}\sum_{s=1}^{t-1}\|p-y_s\|^2 = \frac{1}{t-1}\sum_{s=1}^{t-1}\left\|\frac{1}{t-1}\sum_{r=1}^{t-1}y_r - y_s\right\|^2 + \left\|\frac{1}{t-1}\sum_{r=1}^{t-1}y_r - p\right\|^2$$

（通过展开平方项可以很容易地验证），我们得到

$$\widehat{p}_t = \frac{1}{t-1}\sum_{s=1}^{t-1}y_s$$

类似地，

$$p_t^* = \frac{1}{t}\sum_{s=1}^{t}y_s$$

则对于任意 $y \in \mathcal{D}$，有：

$$\begin{aligned}\ell(\widehat{p}_t, y) - \ell(p_t^*, y) &= \|\widehat{p}_t - y\|^2 - \|p_t^* - y\|^2 \\ &= (\widehat{p}_t - p_t^*)\cdot(\widehat{p}_t + p_t^* - 2y) \\ &\le 4\|\widehat{p}_t - p_t^*\|\end{aligned}$$

由于集合 $\mathcal{D}$ 是有界的。可以通过以下表达式轻松地对差值的范数进行估计：

$$\widehat{p}_t - p_t^* = \frac{1}{t-1}\sum_{s=1}^{t-1}y_s - \frac{1}{t}\sum_{s=1}^{t}y_s = \left(\frac{1}{t-1}-\frac{1}{t}\right)\sum_{s=1}^{t-1}y_s - \frac{y_t}{t}$$

因此，无论结果序列如何，显然有 $\|\widehat{p}_t - p_t^*\| \le 2/t$。因此，

$$\ell(\widehat{p}_t, y) - \ell(p_t^*, y) \le \frac{8}{t}$$

根据不等式 (3.1)，我们得到

$$\widehat{L}_n - \inf_{E\in\mathcal{E}} L_{E,n} \le \sum_{t=1}^{n}\frac{8}{t} \le 8(1+\ln n)$$

其中我们使用了 $\sum_{t=1}^{n} 1/t \le 1 + \int_1^n \frac{dx}{x} = 1 + \ln n$。请注意，这一性能上界明显小于第二章中得到的 $\sqrt{n}$ 阶的一般界限。此外，令人惊讶的是，专家类的“规模”并未反映在该上界中。例子中的专家类不仅是无限的，还可以是一个无限维向量空间中的单位球！

### 3.2.2 凸损失与常量专家

在本节的剩余部分中，我们通过推导一般的充分条件，将平方损失的例子推广到专家类包含常量专家时的凸损失函数。为了简化叙述，我们将讨论限制在有限维的情况，但该结果可以很容易地推广。

令 $\mathcal{D}$ 为 $\mathbb{R}^d$ 的有界凸子集，假设每个元素 $E \in \mathcal{D}$ 对应一个专家，该专家在时间 $t$ 的损失为 $\ell(E, y_t)$。遵循之前的约定，用粗体字母表示向量，我们记跟随最佳专家的预测器为 $\widehat{\mathbf{p}}_t = (\widehat{p}_{1,t}, \ldots, \widehat{p}_{d,t})$，并将相应的假设预测器记为 $\mathbf{p}_t^* = \left(p_{1,t}^*, \ldots, p_{d,t}^*\right)$。我们对损失函数做如下假设：

1. $\ell$ 在第一个参数上是凸的，并且取值在 $[0,1]$；

2. 对于固定的 $y \in \mathcal{Y}$，$\ell(\cdot, y)$ 在第一个参数上是 Lipschitz 连续的，常数为 $B$；

3. 对于固定的 $y \in \mathcal{Y}$，$\ell(\cdot, y)$ 是二阶可微的。此外，存在常数 $C > 0$，使得对于每个固定的 $y \in \mathcal{Y}$，Hessian 矩阵

$$\left(\frac{\partial^2 \ell(\mathbf{p}, y)}{\partial p_i \partial p_j}\right)_{d\times d}$$

是正定的，且其特征值下界为 $C$；

4. 对于任意的 $y_1, \ldots, y_t$，使得最小化 $\mathbf{p}_t^*$ 满足 $\nabla \Psi_t(\mathbf{p}_t^*) = 0$，其中对每个 $\mathbf{p} \in \mathcal{D}$，定义

$$\Psi_t(\mathbf{p}) = \frac{1}{t} \sum_{s=1}^{t} \ell(\mathbf{p}, y_s)$$

**定理 3.1** 在上述假设下，最佳专家跟随算法的每轮遗憾满足

$$\frac{1}{n} \left( \widehat{L}_n - \inf_{E \in \mathcal{E}} L_{E,n} \right) \leq \frac{4B^2(1 + \ln n)}{Cn}$$

**证明.** 首先，通过在 $\mathbf{p}_t^*$ 处对 $\Psi_t$ 进行泰勒展开，我们有

$$\begin{aligned} \Psi_t\left(\widehat{\mathbf{p}}_t\right) - \Psi_t\left(\mathbf{p}_t^*\right) =& \frac{1}{2} \sum_{i=1}^{d} \sum_{j=1}^{d} \left. \frac{\partial^2 \ell(\mathbf{p}, y)}{\partial p_i \partial p_j} \right|_{\mathbf{p}} \left(\widehat{p}_{i,t} - p_{i,t}^*\right) \left(\widehat{p}_{j,t} - p_{j,t}^*\right) \\ & \text{(对于某个} \mathbf{p} \in \mathcal{D}) \\ \geq & \frac{C}{2} \left\| \widehat{\mathbf{p}}_t - \mathbf{p}_t^* \right\|^2 \end{aligned}$$

其中我们利用了 $\nabla \Psi_t(\mathbf{p}_t^*) = 0$ 以及对损失函数的海森矩阵的假设。另一方面，

$$\begin{aligned} & \Psi_t\left(\widehat{\mathbf{p}}_t\right) - \Psi_t\left(\mathbf{p}_t^*\right) \\ & = \left(\Psi_{t-1}\left(\widehat{\mathbf{p}}_t\right) - \Psi_t\left(\mathbf{p}_t^*\right)\right) + \left(\Psi_t\left(\widehat{\mathbf{p}}_t\right) - \Psi_{t-1}\left(\widehat{\mathbf{p}}_t\right)\right) \\ & \leq \left(\Psi_{t-1}\left(\mathbf{p}_t^*\right) - \Psi_t\left(\mathbf{p}_t^*\right)\right) + \left(\Psi_t\left(\widehat{\mathbf{p}}_t\right) - \Psi_{t-1}\left(\widehat{\mathbf{p}}_t\right)\right) \end{aligned}$$

（根据 $\widehat{\mathbf{p}}_t$ 的定义）

$$\begin{aligned} & = \frac{1}{t(t-1)} \sum_{s=1}^{t-1} \left(\ell\left(\mathbf{p}_t^*, y_s\right) - \ell\left(\widehat{\mathbf{p}}_t, y_s\right)\right) + \frac{1}{t} \left(\ell\left(\widehat{\mathbf{p}}_t, y_t\right) - \ell\left(\mathbf{p}_t^*, y_t\right)\right) \\ & \leq \frac{2B}{t} \left\| \widehat{\mathbf{p}}_t - \mathbf{p}_t^* \right\| \end{aligned}$$

其中最后一步我们使用了 $\ell(\cdot, y)$ 的 Lipschitz 性质。比较我们为 $\Psi_t\left(\widehat{\mathbf{p}}_t\right) - \Psi_t\left(\mathbf{p}_t^*\right)$ 推导出的上界和下界，可以看出对于每个 $t = 1, 2, \ldots$，

$$\left\| \widehat{\mathbf{p}}_t - \mathbf{p}_t^* \right\| \leq \frac{4B}{Ct}$$

因此，

$$\widehat{L}_n - \inf_{E \in \mathcal{E}} L_{E,n} \leq \sum_{t=1}^{n} \left(\ell\left(\widehat{\mathbf{p}}_t, y\right) - \ell\left(\mathbf{p}_t^*, y\right)\right) \leq \sum_{t=1}^{n} B \left\| \widehat{\mathbf{p}}_t - \mathbf{p}_t^* \right\| \leq \frac{4B^2}{C} \sum_{t=1}^{n} \frac{1}{t}$$

如所示。

定理 3.1 为损失函数建立了一般性条件，在这些条件下，遗憾以缓慢的 $\ln n$ 速率增长。就像平方损失的情形一样，专家类别的大小不会显式地出现在上界中。特别地，上界与维度无关。第三个条件基本上要求损失函数在最小值附近具有近似二次的行为。最后一个条件对许多光滑的严格凸损失函数是成立的。例如，如果 $\mathcal{Y} = \mathcal{D}$ 且 $\ell(\mathbf{p}, \mathbf{y}) = \|\mathbf{p} - \mathbf{y}\|^\alpha$，其中 $\alpha \in (1, 2]$，则很容易看出所有假设都得到满足。其他关于最佳专家跟随算法实现对数遗憾的通用条件已经得到了广泛的研究。文献的指引将在本章末尾给出。

### 3.2.3 指数凹损失函数

现在我们回到第 2.1 节的情境。因此，我们考虑一个有限的 $N$ 个专家类别。在本节中，我们介绍一类在与指数势函数 $\Phi_\eta(\mathbf{u}) = \frac{1}{\eta} \ln \left( \sum_{i=1}^{N} e^{\eta u_i} \right)$ 结合使用时，具有某些有用性质的损失函数。对于某个 $\eta > 0$，

如果函数 $F(z)=e^{-\eta\ell(z,y)}$ 对所有 $y\in\mathcal{Y}$ 是凹的，则称损失函数 $\ell$ 为 ** 指数凹 ** 的。指数凹性比 $\ell$ 的第一个参数的凸性更强（见练习）。$\eta$ 的值越大，指数凹性的假设就越严格。

以下结果展示了指数凹函数的一个关键性质。回想一下，指数加权平均预测器定义为

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

其中 $w_{i,t-1}=e^{-\eta L_{i,t-1}}$。

**定理 3.2** *若损失函数 $\ell$ 对某个 $\eta>0$ 是指数凹的，则使用相同 $\eta$ 的指数加权平均预测器的遗憾满足，对所有 $y_1,\dots,y_n\in\mathcal{Y}$，有*

$$\Phi_\eta\left(\mathbf{R}_n\right)\le\Phi_\eta(0)$$

**证明.** 证明的关键是展示势函数的值不会增加，即 $\Phi_\eta\left(\mathbf{R}_t\right)\le\Phi_\eta\left(\mathbf{R}_{t-1}\right)$ 或等价地：

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

这可以改写为：

$$\exp\left(-\eta\ell\left(\widehat{p}_t,y_t\right)\right)\ge\frac{\sum_{i=1}^N w_{i,t-1}\exp\left(-\eta\ell\left(f_{i,t},y_t\right)\right)}{\sum_{j=1}^N w_{j,t-1}}$$

这一结果直接来自 $\widehat{p}$ 的定义、函数 $F(z)$ 的凹性以及詹森不等式。

预测器能够保证 $\Phi_\eta\left(\mathbf{R}_n\right)\le\Phi_\eta(\mathbf{0})$，这意味着预测器的遗憾被限制在一个独立于序列长度 $n$ 的常数范围内。

**命题 3.1** *若对于某个损失函数 $\ell$ 和某个 $\eta>0$，预测器满足对所有 $y_1,\dots,y_n\in\mathcal{Y}$ 有 $\Phi_\eta\left(\mathbf{R}_n\right)\le\Phi_\eta(\mathbf{0})$，则预测器的遗憾满足*

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

**证明.** 利用 $\Phi_\eta\left(\mathbf{R}_n\right)\le\Phi_\eta(\mathbf{0})$，我们立即得到：

$$\widehat{L}_n-\min_{i=1,\dots,N}L_{i,n}=\max_{i=1,\dots,N}R_{i,n}\le\frac{1}{\eta}\ln\sum_{j=1}^N e^{\eta R_{j,n}}=\Phi_\eta\left(\mathbf{R}_n\right)\le\Phi_\eta(\mathbf{0})=\frac{\ln N}{\eta}$$

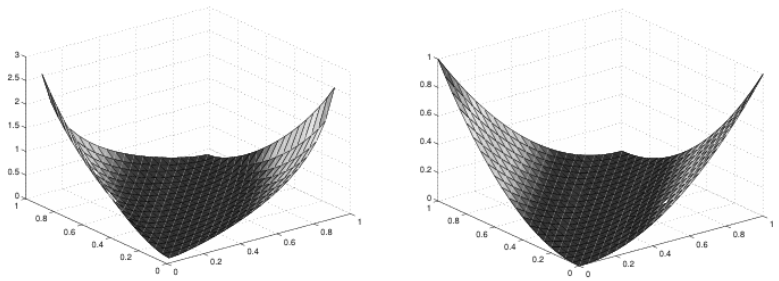

**图 3.1:** 相对熵损失函数和平方损失函数在 $\mathcal{D}=\mathcal{Y}=[0,1]$ 上的图形表示。

显然，$\eta$ 越大，由命题 3.1 保证的上界就越好。对于每个损失函数 $\ell$，都有一个最大值 $\eta$，使得 $\ell$ 是指数凹的（如果存在这样的 $\eta$）。为了优化性能，指数加权平均预测器应使用这个最大的 $\eta$。

对于指数加权平均预测器，将定理 3.2 与命题 3.1 结合，我们得到对于所有指数凹损失，预测器的遗憾被限制在 $(\ln N)/\eta$ 内（其中 $\eta$ 可能取决于损失函数），且这一数量与结果序列的长度 $n$ 无关。注意，除了指数凹性的假设外，我们没有对损失函数做其他假设。特别是，我们没有明确假设损失函数是有界的。接下来的例子展示了一些简单而重要的损失函数是指数凹的（更多例子见练习部分）。

### 3.2.4 相对熵损失

令 $\mathcal{D} = \mathcal{Y} = [0,1]$，考虑相对熵损失 $\ell(\widehat{p}, y) = y \ln(y/\widehat{p}) + (1-y)\ln((1-y)/(1-\widehat{p}))$。通过简单的计算可以验证，对于 $\eta = 1$，函数 $F(z)$ 对所有 $y$ 的值是凹的。注意这是一个无界损失函数。该损失函数的一个特殊情形是，当 $\mathcal{Y} = \{0,1\}$ 且 $\mathcal{D} = [0,1]$ 时，即为对数损失 $\ell(z,y) = -\mathbb{I}_{\{y=1\}} \ln z - \mathbb{I}_{\{y=0\}} \ln(1-z)$。对数损失函数在若干应用中起着核心作用，我们将在第 9 章中专门讨论它。

### 3.2.5 平方损失

该损失函数定义为 $\ell(z,y) = (z-y)^2$，其中 $\mathcal{D} = \mathcal{Y} = [0,1]$。通过直接计算可以得出，当且仅当对于所有 $z$，$(z-y)^2 \le 1/(2\eta)$ 时，$F(z)$ 是凹的。显然，当 $\eta \le 1/2$ 时，这一条件得以保证。

### 3.2.6 绝对损失

设 $\mathcal{D} = \mathcal{Y} = [0,1]$，考虑绝对损失函数 $\ell(z,y) = |z-y|$。很容易看出，对于任何值的 $\eta$，$F(z)$ 都不是凹函数。事实上，正如第 3.7 节所示，对于该损失函数，不存在一个预测器，其累积超额损失可以独立于 $n$ 进行界定。

指数凹损失函数也使得证明适用于可数多个专家的遗憾界限变得容易。我们只需要将指数势函数 $\Phi_\eta(\mathbf{R}) = \frac{1}{\eta}\ln\left(\sum_{i=1}^N e^{\eta R_i}\right)$ 修改为 $\Phi_\eta(\mathbf{R}) = \frac{1}{\eta}\ln\left(\sum_{i=1}^\infty q_i e^{\eta R_i}\right)$，其中 $\{q_i : i = 1,2\ldots\}$ 是正整数集合上的任意概率分布。这确保了级数的收敛。等价地，$q_i$ 表示专家 $i$ 的初始权重（如习题 2.5 所示）。

**推论 3.1** *假设 $(\ell,\eta)$ 满足定理 3.2 的假设。对于任何可数的专家类以及任何在正整数集合上的概率分布 $\{q_i : i = 1,2\ldots\}$，上述定义的指数加权平均预测器满足，对于所有 $n \ge 1$ 和所有 $y_1,\ldots,y_n \in \mathcal{Y}$，*

$$\widehat{L}_n \le \inf_{i\ge 1}\left(L_{i,n} + \frac{1}{\eta}\ln\frac{1}{q_i}\right)$$

**证明.** 证明几乎与定理 3.2 的证明相同；我们只需要将该证明中的 $w_{i,t-1}$ 重新定义为 $q_i e^{-\eta L_{i,t-1}}$。这使我们可以得出 $\Phi_\eta(\mathbf{R}_n) \le \Phi_\eta(\mathbf{0})$。因此，对于所有 $i \ge 1$，

$$q_i e^{\eta R_{i,n}} \le \sum_{j=1}^\infty q_j e^{\eta R_{j,n}} = e^{\eta\Phi_\eta(\mathbf{R}_n)} \le e^{\eta\Phi_\eta(\mathbf{0})} = 1$$

解出 $R_{i,n}$ 得到所需的边界。

因此，预测器的累积损失超过每个专家的损失最多一个常数，但该常数依赖于专家。如果我们将可数多个专家的指数加权预测器写成以下形式：

$$\widehat{p}_t = \frac{\sum_{i=1}^\infty f_{i,t}\exp\left(-\eta\left(L_{i,t-1} + \frac{1}{\eta}\ln\frac{1}{q_i}\right)\right)}{\sum_{j=1}^\infty \exp\left(-\eta\left(L_{j,t-1} + \frac{1}{\eta}\ln\frac{1}{q_j}\right)\right)}$$

我们可以看到，量 $\frac{1}{\eta}\ln(1/q_i)$ 可以视为我们在每个时间 $t$ 添加到专家 $i$ 的累积损失中的“惩罚”。推论 3.1 是一个所谓的“oracle 不等式”，它指出混合预测器实现了与最佳惩罚累积损失匹配的累积损失（参见第 3.5 节）。

### 3.2.7 用于指数凹损失的混合预测器

我们通过展示如何自然地将指数加权平均预测器扩展到处理某些不可数的专家类来结束本节。我们考虑的类由有限数量的“基础”专家的凸包给出。因此，预测器的目标是尽可能预测得与最佳的基础专家凸组合一样好。正式模型描述如下：考虑 $N$ 个（基础）专家，其在时间 $t$ 的预测为 $f_{i,t}\in\mathcal{D}, i=1,\dots,N, t=1,\dots,n$。我们用 $\mathbf{f}_t$ 表示在时间 $t$ 的专家建议向量 $(f_{1,t},\dots,f_{N,t})$。假设决策空间 $\mathcal{D}$ 是一个凸集合，我们假设损失函数 $\ell$ 对于某个值 $\eta>0$ 是指数凹的。定义预测器的遗憾，相对于 $N$ 个基础专家的凸包为：

$$\widehat{L}_n-\inf_{\mathbf{q}\in\Delta}L_{\mathbf{q},n}=\sum_{t=1}^{n}\ell\left(\widehat{p}_t,y_t\right)-\inf_{\mathbf{q}\in\Delta}\sum_{t=1}^{n}\ell\left(\mathbf{q}\cdot\mathbf{f}_t,y_t\right)$$

其中，$\widehat{L}_n$ 是预测器的累积损失，$\Delta$ 表示 $N$ 维向量的单纯形 $\mathbf{q}=(q_1,\dots,q_N)$，满足 $q_i\geq 0,\sum_{i=1}^{N}q_i=1$，而 $\mathbf{q}\cdot\mathbf{f}_t$ 表示由凸组合 $\sum_{i=1}^{N}q_i f_{i,t}$ 给出的 $\mathcal{D}$ 中的元素。最后，$L_{\mathbf{q},n}$ 表示与 $\mathbf{q}$ 相关的专家的累积损失。

接下来我们分析由“混合”定义的指数加权平均预测器的遗憾（按先前定义的意义）

$$\widehat{p}=\frac{\int_\Delta w_{\mathbf{q},t-1}\mathbf{q}\cdot\mathbf{f}\mathrm{d}\mathbf{q}}{\int_\Delta w_{\mathbf{q},t-1}\,\mathrm{d}\mathbf{q}}$$

其中，对于每个 $\mathbf{q}\in\Delta$，$w_{\mathbf{q},t-1}=\exp\left(-\eta\sum_{s=1}^{t-1}\ell\left(\mathbf{q}\cdot\mathbf{f}_s,y_s\right)\right)$。与之前一样，用于权重定义的 $\eta$ 的值使得损失函数是指数凹的。因此，预测器计算了整个单纯形 $\Delta$ 上的加权平均，这个加权是由与每个凸系数向量 $\mathbf{q}$ 相关的过去表现以指数形式加权的。此时我们不考虑计算问题，但在第 9 节中我们将看到，预测器在某些特殊情况下可以很容易地计算出来。下一个定理显示了遗憾的界限是 $N\ln(n/N)$ 级别的。这并不总是最优的；例如，考虑第 3.2 节中研究的平方损失和常量专家的情况，我们为其导出了一个与 $N$ 无关的界限。然而，这里显示的界限要一般得多，并且在某些情况下可以被认为是紧的；例如，对于第 9 章中研究的对数损失。为了简化论证，我们假设损失函数是有界的，尽管在某些情况下这一条件可以放宽。

**定理 3.3** *假设损失函数 $\ell$ 对于 $\eta$ 是指数凹的，并且取值在 $[0,1]$ 范围内。则上述定义的指数加权混合预测器满足：*

$$\widehat{L}_n-\inf_{\mathbf{q}\in\Delta}L_{\mathbf{q},n}\leq\frac{N}{\eta}\ln\frac{e\eta n}{N}$$

**证明.** 定义对于所有 $\mathbf{q}$ 的遗憾为：

$$R_{\mathbf{q},n}=\widehat{L}_n-L_{\mathbf{q},n}$$

我们可以将 $\mathbf{R}_n$ 写作函数 $\mathbf{q}\mapsto R_{\mathbf{q},n}$。引入势函数：

$$\Phi_\eta\left(\mathbf{R}_n\right)=\int_\Delta e^{\eta R_{\mathbf{q},n}}\mathrm{d}\mathbf{q}$$

我们通过模拟定理 3.2 的证明，可以得出 $\Phi_\eta\left(\mathbf{R}_n\right)\leq\Phi_\eta(\mathbf{0})=1/(N!)$。剩下的任务是将超额累积损失与势函数的值 $\Phi_\eta\left(\mathbf{R}_n\right)$ 相关联。

用 $\mathbf{q}^*$ 表示在 $\Delta$ 中使得：

$$L_{\mathbf{q}^*,n}=\inf_{\mathbf{q}\in\Delta}L_{\mathbf{q},n}$$

由于损失函数在其第一个参数中是凸的（见习题 3.4），对于任意 $\mathbf{q}'\in\Delta$ 和 $\lambda\in(0,1)$，

$$L_{(1-\lambda)\mathbf{q}^*+\lambda\mathbf{q}',n}\leq(1-\lambda)L_{\mathbf{q}^*,n}+\lambda L_{\mathbf{q}',n}\leq(1-\lambda)L_{\mathbf{q}^*,n}+\lambda n$$

我们使用了损失函数的有界性。因此，对于任何固定的 $\lambda \in (0,1)$，

$$\begin{aligned}\Phi_\eta(\mathbf{R}_n) &= \int_\Delta e^{\eta R_{\mathbf{q},n}} \mathrm{d}\mathbf{q} \\ &= e^{\eta \hat{L}_n} \int_\Delta e^{-\eta L_{\mathbf{q},n}} \mathrm{d}\mathbf{q} \\ &\geq e^{\eta \hat{L}_n} \int_{\{\mathbf{q}:\mathbf{q}=(1-\lambda)\mathbf{q}^*+\lambda\mathbf{q}',\mathbf{q}'\in\Delta\}} e^{-\eta L_{\mathbf{q},n}} \mathrm{d}\mathbf{q} \\ &\geq e^{\eta \hat{L}_n} \int_{\{\mathbf{q}:\mathbf{q}=(1-\lambda)\mathbf{q}^*+\lambda\mathbf{q}',\mathbf{q}'\in\Delta\}} e^{-\eta\left((1-\lambda)L_{\mathbf{q}^*,n}+\lambda n\right)} \mathrm{d}\mathbf{q} \\ &\quad \text{(由上面的不等式得出)} \\ &= e^{\eta \hat{L}_n} e^{-\eta\left((1-\lambda)L_{\mathbf{q}^*,n}+\lambda n\right)} \int_{\{\mathbf{q}:\mathbf{q}=(1-\lambda)\mathbf{q}^*+\lambda\mathbf{q}',\mathbf{q}'\in\Delta\}} \mathrm{d}\mathbf{q}.\end{aligned}$$

右侧的积分是单纯形的体积，按 $\lambda$ 缩放，并以 $\mathbf{q}^*$ 为中心。显然，这等于 $\lambda^N$ 乘以 $\Delta$ 的体积，即 $\lambda^N/(N!)$。利用 $\Phi_\eta(\mathbf{R}_n) \leq 1/(N!)$，并重新排列得到的不等式，我们得到：

$$\widehat{L}_n - \inf_{\mathbf{q}\in\Delta} L_{\mathbf{q},n} \leq \widehat{L}_n - (1-\lambda)L_{\mathbf{q}^*,n} \leq \frac{1}{\eta}\ln\lambda^{-N} + \lambda n$$

右侧的最小值由 $\lambda = N/\eta n$ 达到，这样得到了定理的边界。

## 3.3 贪婪预测器

在之前分析加权平均预测器性能的若干论证中，证明的关键是对时间 $t$ 时势函数 $\Phi(\mathbf{R}_t)$ 在遗憾上的增加进行界限估计，并与之前的值 $\Phi(\mathbf{R}_{t-1})$ 进行比较。在定理 2.1 中，我们使用了 Blackwell 条件。在某些情况下，可以利用损失函数的特殊性质来推导出更精确的界限。例如，在第 3.3 节中，展示了对于某些损失函数，指数势函数实际上在每一步都减少（见定理 3.2）。因此，人们可能会倾向于构造预测策略，在每个时间点上最小化势函数的最坏情况增加。本节的目的是探索这种可能性。

一个可能的想法是构造一个预测器 $\widehat{p}$，在每个时间点 $t$ 上，预测以最小化最坏情况下的遗憾，即：

$$\begin{aligned}\widehat{p}_t &= \operatorname*{argmin}_{p\in\mathcal{D}} \sup_{y_t\in\mathcal{Y}} \max_{i=1,\dots,N} R_{i,t} \\ &= \operatorname*{argmin}_{p\in\mathcal{D}} \sup_{y_t\in\mathcal{Y}} \max_{i=1,\dots,N} \left(R_{i,t-1} + \ell(p_t,y_t) - \ell(f_{i,t},y_t)\right)\end{aligned}$$

很容易看出，只要损失函数 $\ell$ 在其第一个参数上是有界且凸的，$\widehat{p}_t$ 中的最小值就存在。然而，最小值可能不是唯一的。在这种情况下，我们可以按照任何预先指定的规则选择一个最小值。不幸的是，这种策略在重复博弈的背景下称为虚拟游戏（见第 7 章），未能保证每轮的遗憾逐渐消失（见习题）。

下一步的尝试可能是最小化 $\max_{i\leq N} R_{i,t}$ 的一个“平滑”版本，例如指数势函数：

$$\Phi_\eta(\mathbf{R}_t) = \frac{1}{\eta}\ln\left(\sum_{i=1}^N e^{\eta R_{i,t}}\right)$$

注意，如果 $\eta\max_{i\leq N} R_{i,t}$ 很大，那么 $\Phi_\eta(\mathbf{R}_t) \approx \max_{i\leq N} R_{i,t}$；但 $\Phi_\eta$ 现在是其分量的一个平滑函数。量 $\eta$ 是一种平滑参数。对于较大的 $\eta$ 值，近似更为紧密，尽管 $\Phi_\eta$ 的等值线变得不那么平滑（见图 2.3）。

基于势函数 $\Phi$，我们现在可以引入预测器 $\widehat{p}$，在每个时间点上贪婪地最小化所有可能结果 $y_t$ 的势函数的最大可能增加。即：

$$\widehat{p}_t = \operatorname*{argmin}_{p\in\mathcal{D}} \sup_{y_t\in\mathcal{Y}} \Phi(\mathbf{R}_{t-1} + \mathbf{r}_t) \quad \text{（贪婪预测器）}$$

回顾一下，遗憾向量 $\mathbf{r}_t$ 的第 $i$ 个分量是 $\ell(p, y_t) - \ell(f_{i,t}, y_t)$。注意，对于指数势函数，前述条件等价于：

$$\widehat{p}_t = \operatorname*{argmin}_{p \in \mathcal{D}} \sup_{y_t \in \mathcal{Y}} \left( \ell(p, y_t) + \frac{1}{\eta} \ln \sum_{i=1}^{N} e^{-\eta L_{i,t}} \right)$$

接下来，我们将展示贪婪预测器是定义良好的，且实际上具有与基于相同势函数的加权平均预测器相同的性能保证。

假设势函数 $\Phi$ 是凸的。因为凸函数的上确界是凸的，所以如果 $\ell$ 在其第一个参数中是凸的，则 $\sup_{y_t \in \mathcal{Y}} \Phi(\mathbf{R}_{t-1} + \mathbf{r}_t)$ 是 $p$ 的凸函数。因此，关于 $p$ 的最小值存在，尽管可能不唯一。同样，最小值可以通过任何预先指定的规则来选择。为了计算预测值 $\widehat{p}_t$，在每一步需要最小化一个凸函数。这在许多情况下是计算上可行的，尽管通常不像计算加权平均预测器的预测那样简单。然而，在某些情况下，预测值可以以封闭形式给出。我们提供了一些例子。

以下明显的结果有助于分析贪婪预测器。在第 3.5 节中推导了某些损失的更好界限（见命题 3.3）。

**定理 3.4** *设 $\Phi : \mathbb{R}^N \to \mathbb{R}$ 是一个非负的、二阶可微的凸函数。假设存在一个预测器，其遗憾向量满足：*

$$\Phi(\mathbf{R}'_t) \leq \Phi(\mathbf{R}'_{t-1}) + c_t$$

*对于任何遗憾向量序列 $\mathbf{r}'_1, \ldots, \mathbf{r}'_n$ 和任何 $t = 1, \ldots, n$，其中 $c_t$ 是仅依赖于 $t$ 的常数。那么，贪婪预测器的遗憾 $\mathbf{R}_t$ 满足：*

$$\Phi(\mathbf{R}_n) \leq \Phi(\mathbf{0}) + \sum_{t=1}^{n} c_t$$

**证明.** 只需证明，对于每个 $t = 1, \ldots, n$，

$$\Phi(\mathbf{R}_t) \leq \Phi(\mathbf{R}_{t-1}) + c_t$$

根据贪婪预测器的定义，这等价于说存在一个 $\widehat{p}_t \in \mathcal{D}$，使得：

$$\sup_{y_t \in \mathcal{Y}} \Phi(\mathbf{R}_{t-1} + \mathbf{r}_t) \leq \Phi(\mathbf{R}_{t-1}) + c_t$$

其中 $\mathbf{r}_t$ 是由分量 $\ell(\widehat{p}_t, y_t) - \ell(f_{i,t}, y_t)$ 组成的向量，$i = 1, \ldots, N$。假设保证了这样的 $\widehat{p}_t$ 的存在。

在第 2.1 节和第 3.3 节中分析的加权平均预测器都满足定理 3.4 的条件，因此对应的贪婪预测器继承了在定理 2.1 和 3.2 中证明的性质。以下是使用指数势函数 $\Phi_\eta$ 的一些简单贪婪预测器的例子。

### 3.3.1 绝对损失

考虑绝对损失 $\ell(\widehat{p}, y) = |\widehat{p} - y|$，在简单情况下，$\mathcal{Y} = \{0, 1\}$ 且 $\mathcal{D} = [0, 1]$。考虑基于指数势函数 $\Phi_\eta$ 的贪婪预测器。由于 $y_t$ 是二值的，确定 $\widehat{p}_t$ 就相当于最小化两个凸函数的最大值。经过简单化计算，我们得到：

$$\widehat{p}_t = \operatorname*{argmin}_{p \in [0,1]} \max \left\{ \sum_{i=1}^{N} e^{\eta(\ell(p,0) - \ell(f_{i,t},0) - L_{i,t-1})}, \sum_{i=1}^{N} e^{\eta(\ell(p,1) - \ell(f_{i,t},1) - L_{i,t-1})} \right\}$$

（请记住，$L_{i,t} = \ell(f_{i,1}, y_1) + \cdots + \ell(f_{i,t}, y_t)$ 表示专家 $i$ 在时间 $t$ 的累积损失。）要确定最小值，只需观察两个凸函数的最大值在其最小值处要么是两个函数相等的点，要么是其中一个函数的最小值。因此，$\widehat{p}$ 要么等于 0 或 1，或者

$$\frac{1}{2} + \frac{1}{2\eta} \ln \frac{\sum_{i=1}^{N} e^{-\eta L_{i,t-1} - \eta \ell(f_{i,t},1)}}{\sum_{j=1}^{N} e^{-\eta L_{j,t-1} - \eta \ell(f_{j,t},0)}}$$

具体取决于哪一个值给出较小的势函数最坏情况值。现在，由定理 3.4 和 2.2 可以推导出，这个贪婪预测器的累积损失被界限如下：

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

### 3.3.2 平方损失

接下来考虑前述例子的设置，唯一不同的是损失函数现在是 $\ell(\widehat{p}, y) = (\widehat{p} - y)^2$。计算过程可以同样重复，值得注意的是，贪婪预测器 $\widehat{p}_t$ 的形式与之前完全相同；即，它等于 0 或 1，或者

$$\frac{1}{2} + \frac{1}{2\eta} \ln \frac{\sum_{i=1}^N e^{-\eta L_{i,t-1} - \eta \ell(f_{i,t},1)}}{\sum_{j=1}^N e^{-\eta L_{j,t-1} - \eta \ell(f_{j,t},0)}}$$

具体取决于哪一个值给出较小的势函数最坏情况值。在定理 3.2 中显示，平方损失的特殊性质意味着，如果使用指数加权平均预测器并且 $\eta = 1/2$，则势函数在任何一步中都不会增加。定理 3.2 结合之前的结果表明：如果 $\eta = 1/2$，则贪婪预测器满足：

$$\widehat{L}_n - \min_{i=1,\dots,N} L_{i,n} \le 2 \ln N$$

### 3.3.3 对数损失

同样地，设 $\mathcal{Y} = \{0, 1\}$ 和 $\mathcal{D} = [0, 1]$，考虑对数损失 $\ell(\widehat{p}, y) = -\mathbb{I}_{\{y=1\}} \ln \widehat{p} - \mathbb{I}_{\{y=0\}} \ln(1 - \widehat{p})$ 以及指数势函数。有趣的是，在这种情况下，当 $\eta = 1$ 时，贪婪预测器与指数加权平均预测器完全一致（见习题）。

## 3.4 聚合预测器

第 3.3 节中基于指数凹性的分析依赖于为给定的损失找到某个 $\eta > 0$，使得指数势函数 $\Phi_\eta(\mathbf{R}_n)$ 在预测通过指数加权平均预测器计算时保持被初始势函数 $\Phi_\eta(\mathbf{0})$ 约束。如果找到这样的 $\eta$，则命题 3.1 意味着遗憾保持在 $(\ln N)/\eta$ 之内。然而，正如我们在本节中所展示的，对于所有存在此类 $\eta$ 的损失，可以通过找到一个预测器（不一定基于加权平均）来获得更好的遗憾界限，该预测器保证 $\Phi_\eta(\mathbf{R}_n) \le \Phi_\eta(\mathbf{0})$，其 $\eta$ 值大于加权平均预测器所能承受的最大 $\eta$。因此，我们寻找一个预测器，其预测 $\widehat{p}_t$ 满足 $\Phi_\eta(\mathbf{R}_{t-1} + \mathbf{r}_t) \le \Phi_\eta(\mathbf{R}_{t-1})$，无论下一结果 $y_t \in \mathcal{Y}$ 选择如何（回顾，通常 $\mathbf{r}_t = (r_{1,t}, \dots, r_{N,t})$，其中 $r_{i,t} = \ell(\widehat{p}_t, y_t) - \ell(f_{i,t}, y_t)$）。很容易看出，这等价于条件：

$$\ell(\widehat{p}_t, y_t) \le -\frac{1}{\eta} \ln \left( \sum_{i=1}^N e^{-\eta \ell(f_{i,t}, y_t)} \mathbf{q}_{i,t-1} \right) \quad \text{对于所有} y_t \in \mathcal{Y}$$

分布 $q_{1,t-1}, \dots, q_{N,t-1}$ 是通过与指数势函数相关的权重来定义的。即，$q_{i,t-1} = e^{-\eta L_{i,t-1}} / \left( \sum_{j=1}^N e^{-\eta L_{j,t-1}} \right)$。为了分析那些没有预测器能够防止指数势函数不断增加的损失，我们通过将因子 $1/\eta$ 替换为 $\mu(\eta)/\eta$ 来略微放宽之前的条件。实值函数 $\mu$ 被称为损失 $\ell$ 的混合曲线，其正式定义如下。对于所有 $\eta > 0$，$\mu(\eta)$ 是所有数字 $c$ 的下确界，使得对于所有 $N$、所有概率分布 $(q_1, \dots, q_N)$ 以及所有专家建议 $f_1, \dots, f_N \in \mathcal{D}$，存在一个 $\widehat{p} \in \mathcal{D}$，使得：

$$\ell(\widehat{p}, y) \le -\frac{c}{\eta} \ln \left( \sum_{i=1}^N e^{-\eta \ell(f_i, y)} \mathbf{q}_i \right) \quad \text{对于所有} y \in \mathcal{Y}$$

使用 Vovk 提出的术语，我们称任何预测器为聚合预测器，只要在输入参数 $\eta$ 的情况下，该预测器使用 $\widehat{p}$ 进行预测，并满足 (3.3)，其中 $c = \mu(\eta)$。

混合曲线可以用来为所有值 $\eta > 0$ 的聚合预测器的损失提供界限。

**命题 3.2** 设 $\mu$ 为任意损失函数 $\ell$ 的混合曲线。那么，对于所有 $\eta > 0$，聚合预测器实现：

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

对于所有 $n \geq 1$ 和所有 $y_1, \dots, y_n \in \mathcal{Y}$。

**证明.** 设 $W_t = \sum_{i=1}^N e^{-\eta L_{i,t}}$。然后，根据混合性的定义，对于每个 $t$，存在一个 $\widehat{p}_t \in \mathcal{D}$ 使得：

$$\ell(\widehat{p}_t, y_t) \leq -\frac{\mu(\eta)}{\eta} \ln \left( \frac{\sum_{i=1}^N e^{-\eta L_{i,t-1}} e^{-\eta \ell(f_{i,t}, y_t)}}{\sum_{j=1}^N e^{-\eta L_{j,t-1}}} \right) = -\frac{\mu(\eta)}{\eta} \ln \left( \frac{W_t}{W_{t-1}} \right)$$

将此不等式的左右两边对于 $t = 1, \dots, n$ 求和，我们得到：

$$\begin{aligned}
\widehat{L}_n &\leq -\frac{\mu(\eta)}{\eta} \ln \left( \prod_{t=1}^n \frac{W_t}{W_{t-1}} \right) \\
&= -\frac{\mu(\eta)}{\eta} \ln \frac{W_n}{W_0} \\
&= -\frac{\mu(\eta)}{\eta} \ln \left( \frac{\sum_{j=1}^N e^{-\eta L_{j,n}}}{N} \right) \\
&\leq \frac{\mu(\eta)}{\eta} (\eta L_{i,n} + \ln N)
\end{aligned}$$

对于任何专家 $i = 1, \dots, N$。

在第 2 章的引言中，我们将具有专家建议的预测问题描述为预测器与环境之间的迭代博弈。要进行这个博弈，环境必须使用某种策略来选择每轮的专家建议和结果，基于预测器过去的预测。固定环境的这种策略，并固定预测器的策略（例如，加权平均预测器）。对于每对 $(a, b) \in \mathbb{R}_+^2$，如果预测器的策略实现：

$$\widehat{L}_n \leq a \min_{i=1,\dots,N} L_{i,n} + b \ln N$$

对于所有 $n, N \geq 1$，则称预测器获胜；否则环境获胜（假设专家的数量 $N$ 是由环境在游戏开始时选择的）。

这个博弈是由 Vovk [298] 在开创性的工作中通过混合曲线的概念引入和分析的。在对 $\mathcal{D}, \mathcal{Y}$ 和 $\ell$ 的条件较宽松的情况下，Vovk 显示了以下结果：

- 对于每对 $(a, b) \in \mathbb{R}_+^2$，博弈是决定性的。即，要么存在一种预测策略能够无论环境使用何种策略都获胜，要么环境有一种策略可以击败任何预测策略。

- 预测器仅在那些存在某个 $\eta \geq 0$ 使得 $\mu(\eta) \leq a$ 且 $\mu(\eta)/\eta \leq b$ 的对 $(a, b) \in \mathbb{R}_+^2$ 中获胜。

Vovk 的结果表明，混合曲线正好是所有对 $(a, b)$ 的边界，这些对 $(a, b)$ 满足预测器总是能够保证 $\widehat{L}_n \leq a \min_{i \leq N} L_{i,n} + b \ln N$。可以在对 $\ell$ 的条件较宽松的假设下证明，$\mu \geq 1$。对于使得 $\mu(\eta) = 1$ 的最大值 $\eta$ 尤其与我们的遗憾最小化目标相关。实际上，对于这个 $\eta$，我们可以得到形式为 $(\ln N)/\eta$ 的强遗憾界限。

我们称任何损失函数为 $\eta$-混合的，只要在特殊情况下 $\mathcal{D} = [0, 1]$ 和 $\mathcal{Y} = \{0, 1\}$ 存在一个 $\eta$ 使得 $\mu(\eta) = 1$（0 和 1 的选择是任意的；定理可以证明对任何两个实数 $a, b$，其中 $a < b$）。下一个结果描述了混合损失。

**定理 3.5** （混合性定理）设 $\mathcal{D} = [0, 1]$，$\mathcal{Y} = \{0, 1\}$，并选择一个损失函数 $\ell$。考虑集合 $S \subseteq [0, 1]^2$，其中所有对 $(x, y)$ 满足存在某个 $p \in [0, 1]$ 使得 $\ell(p, 0) \leq x$ 和 $\ell(p, 1) \leq y$。对于每个 $\eta > 0$，引入同胚映射 $H_\eta : [0, 1]^2 \to [e^{-\eta}, 1]^2$，定义为 $H_\eta(x, y) = (e^{-\eta x}, e^{-\eta y})$。则 $\ell$ 是 $\eta$-混合的，当且仅当集合 $H_\eta(S)$ 是凸的。

**证明.** 我们需要找到一个 $\eta > 0$，对于所有 $y \in \{0,1\}$、所有概率分布 $q_1, \ldots, q_N$ 和任何专家建议 $f_1, \ldots, f_N \in [0,1]$，存在一个 $p \in [0,1]$ 满足：

$$\ell(p,y) \le -\frac{1}{\eta} \ln \left( \sum_{i=1}^{N} e^{-\eta \ell(f_i,y)} q_i \right)$$

这个条件可以重写为：

$$e^{-\eta \ell(p,y)} \ge \sum_{i=1}^{N} e^{-\eta \ell(f_i,y)} q_i$$

并且，回顾到 $y \in \{0,1\}$，可以写成：

$$e^{-\eta \ell(p,0)} \ge \sum_{i=1}^{N} e^{-\eta \ell(f_i,0)} q_i \quad \text{和} \quad e^{-\eta \ell(p,1)} \ge \sum_{i=1}^{N} e^{-\eta \ell(f_i,1)} q_i$$

这个条件表明，必须存在一个预测值 $p$，使得每个坐标的 $H_\eta(\ell(p,0), \ell(p,1))$ 都不小于相应坐标的凸组合 $\sum_{i=1}^{N} H_\eta\left(\ell\left(f_i,0\right), \ell\left(f_i,1\right)\right) q_i$。如果 $H_\eta(S)$ 是凸的，那么凸组合属于 $H_\eta(S)$。因此，根据 $S$ 的定义，这样的 $p$ 总是存在（见图 3.2）。显然，这个条件也是必要的。

**备注 3.1** 假设 $\ell_0(p) = \ell(p,0)$ 和 $\ell_1(p) = \ell(p,1)$ 是二次可微的，使得 $\ell_0(0) = \ell_1(1) = 0$，并且对于所有 $0 < p < 1$，$\ell_0'(p) > 0$ 和 $\ell_1'(p) < 0$。那么，对于每个 $\eta > 0$，存在一个二次可微函数 $h_\eta$，使得 $y_\eta(p) = h_\eta\left(x_\eta(p)\right)$，并且根据定理 3.5，$\ell$ 是 $\eta$-混合的当且仅当 $h_\eta$ 是凹的。再次利用对 $\ell$ 的假设，$h_\eta$ 的凹性等价于（见习题 3.11）：

$$\eta \le \inf_{0<p<1} \frac{\ell_0'(p)\ell_1''(p) - \ell_0''(p)\ell_1'(p)}{\ell_0'(p)\ell_1'(p)\left(\ell_1'(p) - \ell_0'(p)\right)}$$

**备注 3.2** 我们可以扩展混合性定理，以提供在 $\mathcal{D} = \mathcal{Y} = [0,1]$ 的情况下混合性的充分条件。为此，我们必须验证 $e^{-\eta \ell(p,0)} \ge \sum_{i=1}^{N} e^{-\eta \ell(f_i,0)} q_i$ 和 $e^{-\eta \ell(p,1)} \ge \sum_{i=1}^{N} e^{-\eta \ell(f_i,1)} q_i$ 共同蕴含 $e^{-\eta \ell(p,y)} \ge \sum_{i=1}^{N} e^{-\eta \ell(f_i,y)} q_i$ 对于所有 $0 \le y \le 1$。当且仅当：

$$e^{-\eta \ell(p,y)} - \sum_{i=1}^{N} e^{-\eta \ell(f_i,y)} q_i$$

是对于每个固定的 $p \in [0,1]$、$f_i$ 和 $q_i, i = 1, \ldots, N$ 而 $y \in [0,1]$ 的凹函数时，这种蕴含关系是满足的。

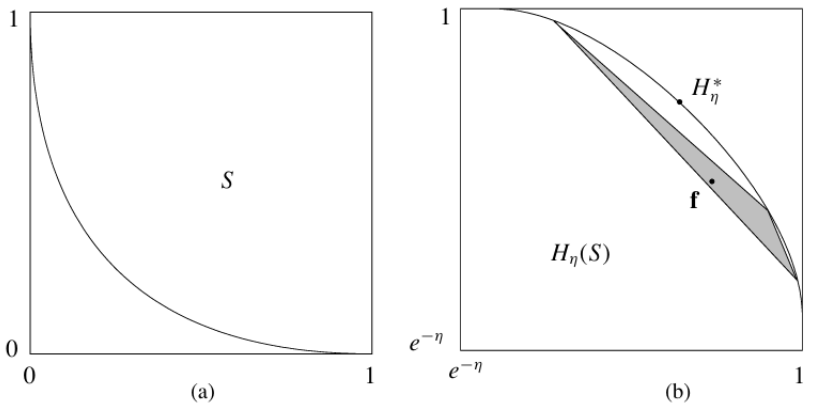


**图 3.2:** 方差损失的混合性。(a) 区域 $S$ 中的点 $(x,y)$ 满足对于某些 $p \in [0,1]$，有 $(p-0)^2 \le x$ 和 $(p-1)^2 \le y$。(b) 预测 $p$ 必须对应于点 $H_\eta^* = H_\eta(\ell(p,0), \ell(p,1))$，该点位于 $\mathbf{f} = \sum_{i=1}^{N} H_\eta\left(\ell\left(f_i,0\right), \ell\left(f_i,1\right)\right) q_i$ 的东北方向。注意到 $\mathbf{f}$ 在凸包（阴影区域）中，其顶点在 $H_\eta(S)$ 中。因此，如果 $H_\eta(S)$ 是凸的，那么 $\mathbf{f} \in H_\eta(S)$，根据 $S$ 的定义，找到这样的 $p$ 总是可能的。

**命题 3.3** 对于任何 $\eta$-混合的损失函数 $\ell$，使用指数势函数 $\Phi_\eta$ 的贪婪预测器是一个聚合预测器。

**证明.** 如果损失函数 $\ell$ 是混合的，那么对于每个 $t$ 和所有 $y_t \in \mathcal{Y}$，存在一个 $\widehat{p}_t \in \mathcal{D}$ 满足：

$$\ell\left(\widehat{p}_t, y_t\right) \leq -\frac{1}{\eta} \ln \frac{\sum_{i=1}^{N} e^{-\eta L_{i,t-1} - \eta \ell\left(f_{i,t}, y_t\right)}}{\sum_{j=1}^{N} e^{-\eta L_{j,t-1}}}$$

这样的 $\widehat{p}_t$ 可以通过以下方式定义：

$$\begin{aligned}
\widehat{p}_t &= \operatorname*{argmin}_{p \in \mathcal{D}} \sup_{y_t \in \mathcal{Y}} \left( \ell\left(p, y_t\right) + \frac{1}{\eta} \ln \frac{\sum_{i=1}^{N} e^{-\eta L_{i,t-1} - \eta \ell\left(f_{i,t}, y_t\right)}}{\sum_{j=1}^{N} e^{-\eta L_{j,t-1}}} \right) \\
&= \operatorname*{argmin}_{p \in \mathcal{D}} \sup_{y_t \in \mathcal{Y}} \left( \ell\left(p, y_t\right) + \frac{1}{\eta} \ln \sum_{i=1}^{N} e^{-\eta L_{i,t}} \right)
\end{aligned}$$

这正是使用指数势函数的贪婪预测器的定义。

### 3.4.1 混合损失的 oracle 不等式

聚合预测器可以通过简单的方式扩展以处理可数无限的专家类。考虑一个序列 $f_1, f_2, \ldots$ 的专家，使得在时间 $t$ 时，专家 $f_i$ 的预测为 $f_{i,t} \in \mathcal{D}$。预测器的目标是预测得与任何专家 $f_i$ 一样好。为实现这一目标，我们为每个专家分配一个正数 $\pi_i > 0$，使得 $\sum_{i=1}^{\infty} \pi_i = 1$。这些数 $\pi_i$ 可以称为先验概率。

现在假设决策空间 $\mathcal{D}$ 是一个紧致的度量空间，且损失函数 $\ell$ 是连续的。令 $\mu$ 为 $\ell$ 的混合性曲线。根据 $\mu$ 的定义，对于每个 $y_t \in \mathcal{Y}$、每个正数序列 $q_1, q_2, \ldots$ 使得 $\sum_i q_i = 1$ 以及 $N > 0$，存在一个 $\widehat{p}^{(N)} \in \mathcal{D}$ 使得

$$\ell\left(\widehat{p}^{(N)}, y_t\right) \leq -\frac{\mu(\eta)}{\eta} \ln \left( \sum_{i=1}^{N} e^{-\eta \ell\left(f_i, y_t\right)} \frac{q_i}{\sum_{j=1}^{N} q_j} \right) \leq -\frac{\mu(\eta)}{\eta} \ln \left( \sum_{i=1}^{N} e^{-\eta \ell\left(f_i, y_t\right)} q_i \right)$$

由于 $\mathcal{D}$ 是紧致的，序列 $\left\{\widehat{p}^{(N)}\right\}$ 存在一个累积点 $\widehat{p} \in \mathcal{D}$。此外，由于 $\ell$ 的连续性，这个累积点满足

$$\ell\left(\widehat{p}, y_t\right) \leq -\frac{\mu(\eta)}{\eta} \ln \left( \sum_{i=1}^{\infty} e^{-\eta \ell\left(f_i, y_t\right)} q_i \right)$$

换句话说，由 $\widehat{p}_t$ 定义的聚合预测器满足

$$\ell\left(\widehat{p}_t, y_t\right) \leq -\frac{\mu(\eta)}{\eta} \ln \frac{\sum_{i=1}^{\infty} \pi_i e^{-\eta L_{i,t-1} - \eta \ell\left(f_{i,t}, y_t\right)}}{\sum_{j=1}^{\infty} \pi_j e^{-\eta L_{j,t-1}}}$$

是定义良好的。然后，根据与引理 3.1 相同的论证，对于所有 $\eta > 0$，我们得到

$$\widehat{L}_n \leq \mu(\eta) \min_{i=1,2,\ldots} \left( L_{i,n} + \frac{1}{\eta} \ln \frac{1}{\pi_i} \right)$$

对于所有 $n \geq 1$ 以及所有 $y_1, \ldots, y_n \in \mathcal{Y}$。

通过将预测器写成

$$\ell\left(\widehat{p}_t, y_t\right) \leq -\frac{\mu(\eta)}{\eta} \ln \frac{\sum_{i=1}^{\infty} \exp\left(-\eta \left(L_{i,t-1} + \frac{1}{\eta} \ln \frac{1}{\pi_i}\right) - \eta \ell\left(f_{i,t}, y_t\right)\right)}{\sum_{j=1}^{\infty} \exp\left(-\eta \left(L_{j,t-1} + \frac{1}{\eta} \ln \frac{1}{\pi_j}\right)\right)}$$

我们可以看出，$\frac{1}{\eta} \ln\left(1/\pi_i\right)$ 可以被视为在每个时间 $t$ 对专家 $i$ 的累计损失添加的“惩罚”。聚合预测器的性能界限是一种所谓的“oracle 不等式”，它表明，聚合预测器的累计损失与专家的最佳加权累计损失相匹配。

## 3.5 特定损失的混合性

在本节中，我们将考察几种具有特殊重要性的损失函数的混合性特性。

### 3.5.1 相对熵损失是 1-混合的

回顾一下，这种损失定义为

$$\ell(p,y)=y\ln\frac{y}{p}+(1-y)\ln\frac{1-y}{1-p},\qquad \text{其中}p,y\in[0,1]$$

首先考虑特殊情况，即 $y\in\{0,1\}$（对数损失）。可以很容易地看出，当 $\eta=1$ 时，混合性定理的条件是满足的。请注意，对于这个 $\eta$ 的选择，满足聚合预测器定义 (3.3) 的唯一预测 $\widehat{p}$，其中 $c=1$，正是指数加权平均预测器的预测。因此，对于对数损失，这种方法在与基于指数凹性的分析（见第 3.3 节）相比，并没有提供任何优势。同样回顾第 3.4 节，对于 $\eta=1$，指数加权平均预测器是势函数的唯一贪婪最小化器。因此，对于对数损失，这种预测器具有各种有趣的性质。实际上，在第 9 章中，我们展示了当专家集合是有限时，指数加权平均预测器是可能的最佳预测器。

回到相对熵损失，验证 (3.3) 对于 $c=1,\eta=1$ 和 $\widehat{p}=\sum_{i=1}^{N}f_iq_i$（即，当 $\widehat{p}$ 是加权平均预测）直接来自于我们在第 3.3 节证明的函数 $F(z)=e^{-\eta\ell(z,y)}$ 的指数凹性。因此，$\eta=1$ 的指数加权平均预测器也满足相对熵损失的混合性定理的条件。另一方面，我们同样没有在与基于指数凹性的分析相比中获得任何改进。

### 3.5.2 平方损失是 2-混合的

对于平方损失 $\ell(p,y)=(p-y)^2,p,y\in[0,1]$，我们首先假设 $y\in\{0,1\}$。在这种情况下，混合性定理的条件很容易用 $\eta=2$ 验证。经过一些额外的工作，我们也可以验证函数 (3.4) 对 $y\in[0,1]$ 是凹的（见习题 3.12）。因此，当 $\mathcal{D}=\mathcal{Y}=[0,1]$ 时，混合性定理的条件也得到了满足。请注意，与相对熵损失的情况不同，在这里我们在遗憾界限上获得了一个因子 4 的提升，这与基于指数凹性的分析相比是有实质性改善的。这种提升是实际的，因为可以证明，指数加权平均预测器一般情况下无法满足混合性定理的条件。此外，可以证明，任何形式为 $\widehat{p}=g\left(\sum_{i=1}^{N}f_iq_i\right)$ 的预测器，无论选择哪种函数 $g$，都无法满足 (3.3) 其中 $c=1$ 和 $\eta=2$（见习题 3.13）。

我们现在推导聚合预测器的预测 $\widehat{p}_t$ 的封闭形式表达式。由于平方损失是混合的，根据命题 3.3，贪婪预测器是一个聚合预测器。回顾第 3.4 节中平方损失的贪婪预测器的预测，我们得到

$$\widehat{p}_t=\begin{cases}0 & \text{如果}r_t<0\\ r_t & \text{如果}0\le r_t\le 1\\ 1 & \text{如果}r_t>1\end{cases}$$

其中

$$r_t=\frac{1}{2}+\frac{1}{2\eta}\ln\frac{\sum_{i=1}^{N}e^{-\eta L_{i,t-1}-\eta\ell(f_{i,t},1)}}{\sum_{j=1}^{N}e^{-\eta L_{j,t-1}-\eta\ell(f_{j,t},0)}}$$

### 3.5.3 绝对损失不是混合的

绝对损失定义为 $\ell(p,y)=|p-y|$，其中 $p,y\in[0,1]$，不满足混合性定理的条件。因此，我们不能期望找到 $\eta>0$ 使得 $\mu(\eta)=1$。然而，我们可以通过应用命题 3.2 来得到所有 $\eta>0$ 的损失界限。为此，我们

需要找到绝对损失的混合性函数，即，找到最小的函数 $\mu:\mathbb{R}_+\to\mathbb{R}_+$，使得对于所有分布 $q_1,\ldots,q_N$，所有 $f_1,\ldots,f_N\in[0,1]$ 和所有结果 $y$，存在一个 $\widehat{p}\in[0,1]$ 满足

$$|\widehat{p}-y|\leq-\frac{\mu(\eta)}{\eta}\ln\left(\sum_{i=1}^{N}e^{-\eta|f_i-y|}q_i\right)$$

这里我们只考虑简单情况，即 $y\in\{0,1\}$（二元结果的预测）。在这种情况下，实现混合性的预测 $\widehat{p}$ 可以表示为指数加权平均预测器的（非线性）函数。在更一般的情况下，当 $y\in[0,1]$ 时，无论选择怎样的函数 $g$，形式为 $\widehat{p}=g\left(\sum_{i=1}^{N}f_iq_i\right)$ 的预测器都无法实现混合性（见习题 3.14）。

利用线性插值 $e^{-\eta x}\leq1-(1-e^{-\eta})\,x$，该不等式对于所有 $\eta>0$ 和 $0\leq x\leq1$ 成立，我们得到

$$\begin{aligned}
&-\frac{\mu(\eta)}{\eta}\ln\left(\sum_{i=1}^{N}e^{-\eta|f_i-y|}q_i\right)\\
&\quad\geq-\frac{\mu(\eta)}{\eta}\ln\left(1-\sum_{i=1}^{N}\left(1-e^{-\eta}\right)|f_i-y|q_i\right)\\
&\quad=-\frac{\mu(\eta)}{\eta}\ln\left(1-\left(1-e^{-\eta}\right)\left|\sum_{i=1}^{N}f_iq_i-y\right|\right)
\end{aligned}$$

其中在最后一步中我们使用了 $y\in\{0,1\}$ 的假设。因此，使用符号 $r=\sum_{i=1}^{N}f_iq_i$，我们只需证明

$$|\widehat{p}-y|\leq-\frac{\mu(\eta)}{\eta}\ln\left(1-\left(1-e^{-\eta}\right)|r-y|\right)$$

或者等价地，再次使用假设 $y\in\{0,1\}$，我们得到

$$1+\frac{\mu(\eta)}{\eta}\ln\left(1-\left(1-e^{-\eta}\right)(1-r)\right)\leq\widehat{p}\leq-\frac{\mu(\eta)}{\eta}\ln\left(1-\left(1-e^{-\eta}\right)r\right)$$

通过在上述不等式中设置 $r=1/2$，我们得到

$$1+\frac{\mu(\eta)}{\eta}\ln\left(\frac{1+e^{-\eta}}{2}\right)\leq-\frac{\mu(\eta)}{\eta}\ln\left(\frac{1+e^{-\eta}}{2}\right)$$

只有以下分配满足这一条件

$$\mu(\eta)=\frac{\eta/2}{\ln\left(2/\left(1+e^{-\eta}\right)\right)}$$

请注意，对于这个取值，不等式可以取等号。简单的计算表明，上述选择的 $\mu$ 满足对所有 $0\leq r\leq1$。进一步注意，对于 $f_1,\ldots,f_N\in\{0,1\}$，线性近似是紧的。如果此外 $r=\sum_{i=1}^{N}f_iq_i=1/2$，则仅存在一个函数 $\mu$ 使得这部分最开始的不等式成立。因此，$\mu$ 实际上是绝对损失的二元结果的混合性曲线。可以证明（见习题 3.15），这个函数也是当 $y\in[0,1]$ 时的混合性曲线。在图 3.3 中，我们展示了作为 $r$ 的函数，预测 $\widehat{p}$ 的混合性实现的上界和下界，以及通过取上下界的平均值获得的曲线 $\widehat{p}=\widehat{p}(r)$。通过比较命题 3.2 给出的混合性实现预测器的遗憾界限与定理 2.2 对加权平均预测器的界限，可以看出前者在 $\eta$ 上的依赖性优于后者。然而，正如我们在第 3.7 节所示，定理 2.2 的界限在渐近上是紧的。因此，随着 $n$ 的增长，使用更复杂的混合性实现预测器的好处会消失。

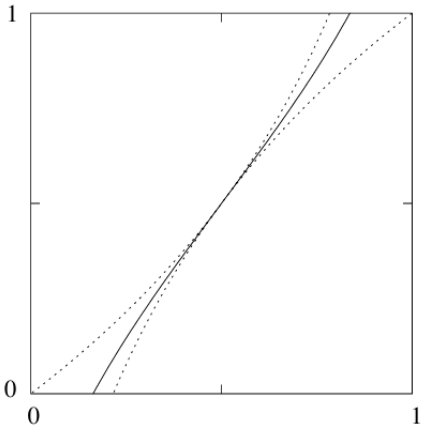


**图 3.3:** 两条虚线显示了作为加权平均值 $r$ 的函数，在绝对损失情况下，当 $\eta = 2$ 时，混合性实现预测 $\widehat{p}$ 的上界和下界；其中的实线是通过取上下界的平均值得到的混合性实现预测 $\widehat{p}$。请注意，在这种情况下，$\widehat{p}$ 可以表示为加权平均值的（非线性）函数。

## 3.6 一般下界

我们现在讨论迄今为止在本章及前一章中获得的上界的紧的性质。我们的目的是推导最坏情况遗憾的下界。更准确地说，我们考察了最小最大遗憾 $V_n^{(N)}$ 的行为。回顾第 2.10 节，给定一个损失函数 $\ell$，$V_n^{(N)}$ 被定义为在最坏情况下 $n$ 个结果和 $N$ 个专家的建议 $f_{i,t}$ 中的最佳可能预测策略的遗憾，其中 $i = 1, \ldots, N$ 和 $t = 1, \ldots, n$。

定理 2.2 的上界表明，如果损失函数在 0 和 1 之间有界，则 $V_n^{(N)} \leq \sqrt{(n/2) \ln N}$。另一方面，本章证明的混合性定理表明，对于任何混合损失 $\ell$，存在更紧的上界 $V_n^{(N)} \leq c_\ell \ln N$，其中 $c_\ell$ 是一个依赖于具体混合损失的参数。（见备注 3.1 以获得该参数的分析特征。）在本节中，我们展示了从某种意义上讲，这两个上界都是紧的。下一个结果表明，除了普通且不感兴趣的情况之外，最小最大损失至少与专家数量 $N$ 的对数成比例。定理提供了任何损失函数的下界。应用于混合损失时，这个下界捕捉了对 $N$ 的对数依赖，但未提供匹配常数。

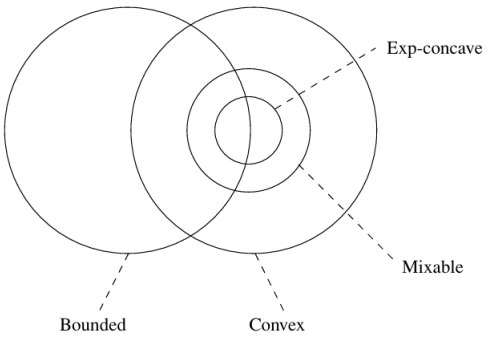


**图 3.4:** 一个维恩图，展示了迄今为止考察的损失函数的层次结构。对于有界且凸的损失函数，其最小最大遗憾为 $\mathrm{O}(\sqrt{n \ln N})$，这一值由加权平均预测器实现。混合损失函数（这些损失函数不总是有界）的最小最大遗憾具有 $c \ln N$ 的形式。这种遗憾不一定由加权平均预测器实现。指数凹损失函数是那些混合损失函数中，加权平均预测器确实保证了 $c \ln N$ 遗憾的损失函数。这些损失函数在以下意义上正确地包含在混合损失集内：对于某些值的 $\eta$，一些损失是 $\eta$-混合的但不是指数凹的（例如，平方损失）。对于不一定是凸的有界损失，最小最大遗憾在第 4 章中使用随机预测器进行研究。

**定理 3.6** 固定任意损失函数 $\ell$。则对于所有 $N \geq 2$ 和所有 $n \geq 1$，有 $V_{n\lfloor \log_2 N \rfloor}^{(N)} \geq \lfloor \log_2 N \rfloor V_n^{(2)}$。

**证明.** 为了不失一般性，假设存在 $N = 2^M$ 个专家，其中 $M \geq 1$。对于任何 $m \leq N/2$，我们称时间 $t$ 的专家建议 $f_{1,t}, \ldots, f_{N,t}$ 是 $m$-耦合的，如果对所有 $i = 1, \ldots, m$ 有 $f_{i,t} = f_{m+i,t}$。类似地，我们称时间 $t$ 的专家建议是 $m$-简单的，如果 $f_{1,t} = \cdots = f_{m,t}$ 且 $f_{m+1,t} = \cdots = f_{2m,t}$。请注意，这些定义对索引 $i > 2m$ 的专家建议没有约束。我们将时间分为 $M$ 个阶段，每个阶段包含 $n$ 个时间步。我们称专家建议在阶段 $s$ 是 $m$-简单的（$m$-耦合的），如果它在该阶段的每个时间步上都是 $m$-简单的（$m$-耦合的）。我们选择专家建议，使得：

1. 在每个阶段 $s = 1, \ldots, M$ 中，建议是 $2^{s-1}$-简单的；

2. 对于每个 $s = 1, \ldots, M-1$，建议在所有时间步中都是 $2^s$-耦合的，直到阶段 $s$ 为止。

注意，我们可以通过在每个阶段 $s = 1, \ldots, M$ 中选择前 $2^s$ 个专家的任意 $2^{s-1}$-简单建议，然后将这些建议复制到剩余专家来获得这样的建议（见图 3.5）。

| | $s=1$ | $s=2$ | $s=3$ |
|---|---|---|---|
| 1 | 0 | 0 | 0 |
| 2 | 1 | 0 | 0 |
| 3 | 0 | 1 | 0 |
| 4 | 1 | 1 | 0 |
| 5 | 0 | 0 | 1 |
| 6 | 1 | 0 | 1 |
| 7 | 0 | 1 | 1 |
| 8 | 1 | 1 | 1 |

**图 3.5:** 专家建议的分配，实现了定理 3.6 对于 $N = 8$ 和 $n = 1$ 的下界。请注意，这些建议在 $s = 1$ 时是 1-简单的，在 $s = 2$ 时是 2-简单的，而在 $s = 3$ 时是 4-简单的。此外，这些建议在 $s = 1$ 时也是 2-耦合的，而在 $s = 1, 2$ 时是 4-耦合的。

考虑一个任意的预测策略 $P$。对于任意固定的序列 $y_1, \ldots, y_M$，以及任意一对阶段 $1 \leq r \leq s \leq M$，定义

$$R_i\left(y_r^s\right) = \sum_{t=n(r-1)+1}^{ns} \left(\ell\left(\widehat{p}_t, y_t\right) - \ell\left(f_{i,t}, y_t\right)\right)$$

其中 $\widehat{p}_t$ 是策略 $P$ 在时间 $t$ 上计算的预测值。为了简化起见，假设 $n = 1$。固定某个任意专家 $i$，并选择任何专家 $j$（可能等于 $i$）。如果 $i, j \leq 2^{M-1}$ 或 $i, j > 2^{M-1}$，则 $R_i\left(y_1^M\right) = R_i\left(y_1^{M-1}\right) + R_j\left(y_M\right)$，因为阶段 $s = M$ 上的建议是 $2^{M-1}$-简单的。否则，假设在不失一般性的情况下，$i \leq 2^{M-1}$ 并且 $j > 2^{M-1}$。由于阶段 $s = 1, \ldots, M-1$ 的建议是 $2^{M-1}$-耦合的，因此存在 $k > 2^{M-1}$ 使得 $R_i\left(y_1^{M-1}\right) = R_k\left(y_1^{M-1}\right)$。此外，由于阶段 $t = M$ 上的建议是 $2^{M-1}$-简单的，$R_j\left(y_M\right) = R_k\left(y_M\right)$。由此我们已经证明，对于任意的 $i$ 和 $j$，总存在 $k$ 使得 $R_i\left(y_1^M\right) = R_k\left(y_1^{M-1}\right) + R_j\left(y_M\right)$。重复这一论证，并利用我们对专家建议的递归假设，我们得到

$$R_i\left(y_1^M\right) = \sum_{s=1}^{M} R_{j_s}\left(y_s\right)$$

其中 $j_1, \ldots, j_M = j$ 是任意专家。这一推理可以很容易地扩展到 $n \geq 1$ 的情况，得到

$$R_i\left(y_1^{nM}\right) = \sum_{s=1}^{M} R_{j_s}\left(y_{n(s-1)+1}^{ns}\right)$$

现在注意到每个阶段 $s = 1, \ldots, M$ 的专家建议是 $2^{s-1}$-简单的，这意味着在每个时间步上我们至少有两

个“未承诺的专家”可用。因此，利用序列 $y_1, \dots, y_{nM}$ 是任意的事实，以及

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

其中 $\mathcal{F}$ 是静态专家的类（参见第 2.10 节），我们有，对于每个阶段 $s$，

$$R_{j_s}\left(y_{n(s-1)+1}^{ns}\right) \geq V_n^{(2)}$$

对于某些结果 $y_t$、专家索引 $j_s$ 和静态专家建议。为了完成证明，注意到，显然有 $V_M^{(N)} \geq R_i\left(y_1^M\right)$。

通过更复杂的论证，可以证明一个参数化的下界，该下界在渐近上（对于 $N, n \to \infty$）匹配上界 $c_\ell \ln N$，其中 $c_\ell$ 是由混合性定理实现的最佳常数。

关于非混合损失的 $V_n^{(N)}$，例如考虑绝对损失 $\ell(p, y) = |p - y|$。根据定理 2.2，指数加权平均预测器的遗憾由 $\sqrt{(n/2)\ln N}$ 有界，这意味着，对于所有 $n$ 和 $N$，

$$\frac{V_n^{(N)}}{\sqrt{(n/2)\ln N}} \leq 1$$

下一个结果显示，从某种意义上说，这个界限无法进一步改进。它还表明，指数加权平均预测器是渐近最优的。

**定理 3.7** *如果 $\mathcal{Y} = \{0, 1\}$，$\mathcal{D} = [0, 1]$ 且 $\ell$ 是绝对损失 $\ell(p, y) = |p - y|$，那么*

$$\sup_{n,N} \frac{V_n^{(N)}}{\sqrt{(n/2)\ln N}} \geq 1$$

**证明.** 显然，$V_n^{(N)} \geq \sup_{\mathcal{F}:|\mathcal{F}|=N} V_n(\mathcal{F})$，其中我们对 $N$ 个静态专家类取上确界（参见第 2.9 节定义静态专家）。我们首先下界 $V_n(\mathcal{F})$ 对于固定的 $\mathcal{F}$。回忆一下，固定专家类的最小最大遗憾 $V_n(\mathcal{F})$ 定义为

$$V_n(\mathcal{F}) = \inf_P \sup_{y^n \in \{0,1\}^n} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \left(|\widehat{p}_t - y_t| - |f_t - y_t|\right)$$

其中 infimum 是对所有预测策略 $P$ 取的。引入 i.i.d. 对称伯努利随机变量 $Y_1, \dots, Y_n$（即，$\mathbb{P}[Y_t = 0] = \mathbb{P}[Y_t = 1] = 1/2$），可以明显得到

$$\begin{aligned} V_n(\mathcal{F}) &\geq \inf_P \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \left(|\widehat{p}_t - Y_t| - |f_t - Y_t|\right) \\ &= \inf_P \mathbb{E} \sum_{t=1}^{n} |\widehat{p}_t - Y_t| - \mathbb{E} \inf_{f \in \mathcal{F}} \sum_{t=1}^{n} |f_t - Y_t| \end{aligned}$$

（在第 8 章中，我们将证明这一点实际上是等式成立的。）由于序列 $Y_1, \dots, Y_n$ 是完全随机的，对于所有预测策略，显然有 $\mathbb{E}\sum_{t=1}^{n} |\widehat{p}_t - Y_t| = n/2$。因此，

$$\begin{aligned} V_n(\mathcal{F}) &\geq \frac{n}{2} - \mathbb{E}\left[\inf_{f \in \mathcal{F}} \sum_{t=1}^{n} |f_t - Y_t|\right] \\ &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \left(\frac{1}{2} - |f_t - Y_t|\right)\right] \\ &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \left(\frac{1}{2} - f_t\right)\sigma_t\right] \end{aligned}$$

其中 $\sigma_t = 1 - 2Y_t$ 是 i.i.d. Rademacher 随机变量（即，$\mathbb{P}[\sigma_t = 1] = \mathbb{P}[\sigma_t = -1] = 1/2$）。我们通过对包含 $N$ 个专家的适当选择的专家类进行平均来下界 $\sup_{\mathcal{F}:|\mathcal{F}|=N} V_n(\mathcal{F})$。这可以通过将每个专家 $f = (f_1, \ldots, f_n)$ 替换为一系列对称 i.i.d. 伯努利随机变量来实现。更精确地说，设 $\{Z_{i,t}\}$ 是一个 $N \times n$ 的 i.i.d. Rademacher 随机变量阵列，其分布为 $\mathbb{P}[Z_{i,t} = -1] = \mathbb{P}[Z_{i,t} = 1] = 1/2$。那么

$$\begin{aligned}\sup_{\mathcal{F}:|\mathcal{F}|=N} V_n(\mathcal{F}) &\geq \sup_{\mathcal{F}:|\mathcal{F}|=N} \mathbb{E}\left[\sup_{f\in\mathcal{F}} \sum_{t=1}^{n} \left(\frac{1}{2} - f_t\right)\sigma_t\right] \\ &\geq \frac{1}{2}\mathbb{E}\left[\max_{i=1,\ldots,N} \sum_{t=1}^{n} Z_{i,t}\sigma_t\right] \\ &= \frac{1}{2}\mathbb{E}\left[\max_{i=1,\ldots,N} \sum_{t=1}^{n} Z_{i,t}\right]\end{aligned}$$

根据中心极限定理，对于每个 $i = 1, \ldots, N$，$n^{-1/2}\sum_{t=1}^{n} Z_{i,t}$ 收敛到一个标准正态随机变量。实际上，证明（参见附录中的引理 A. 11）并不困难，得出

$$\lim_{n\to\infty} \mathbb{E}\left[\max_{i=1,\ldots,N} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} Z_{i,t}\right] = \mathbb{E}\left[\max_{i=1,\ldots,N} G_i\right]$$

其中 $G_1, \ldots, G_N$ 是独立的标准正态随机变量。但众所周知（参见附录中的引理 A. 12），

$$\lim_{N\to\infty} \frac{\mathbb{E}\left[\max_{i=1,\ldots,N} G_i\right]}{\sqrt{2\ln N}} = 1$$

这就结束了证明。

## 3.7 书目备注

”跟随最佳专家” 预测器及其变体在一种稍微更一般的框架中得到了彻底研究，这种框架被称为 Robbins 提出的顺序复合决策问题 [244]；另见 Blackwell [28, 29]、Gilliland [127]、Gilliland 和 Hannan [128]、Hannan [141]、Hannan 和 Robbins [142]、Merhav 和 Feder [213]、van Ryzin [254] 以及 Samuel [256, 257]。这些文献中可以找到一般条件，保证每轮遗憾收敛于 0。引理 3.1 是 Hannan [141] 的贡献。带有常数专家的平方损失示例由 Takimoto 和 Warmuth [284] 研究，他们展示了最小最大遗憾为 $\ln n - \ln\ln n + o(1)$。

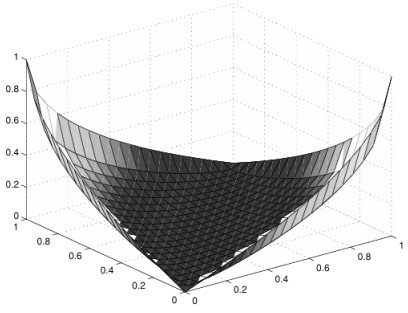

图 3.6: Hellinger 损失。

指数凸性损失函数由 Kivinen 和 Warmuth [182] 研究。定理 3.3 是 Blum 和 Kalai [33] 论证的推广。第 3.5 节的混合性曲线由 Vovk [298, 300] 引入，并且也由 Haussler、Kivinen 和 Warmuth [151] 研究，他们表征了 $\Theta(\ln N)$ 遗憾界限可能的损失函数。备注 3.1 中给出的混合函数的最优 $\eta$ 的表征归功于 Haussler、Kivinen 和 Warmuth [151]。混合性的示例取自 Cesa-Bianchi、Freund、Haussler、Helmbold、Schapire 和 Warmuth

[48]、Haussler、Kivinen 和 Warmuth [151] 以及 Vovk [298, 300]。第 3.5 节的混合性曲线等价于 Yamanishi [313] 引入的“扩展随机复杂性”概念。这个概念推广了 Rissanen 的随机复杂性 [240, 242]。

在二元结果 $\mathcal{Y}=\{0,1\}$ 的情况下，损失函数 $\ell$ 的混合性等价于存在 $\ell$ 的预测复杂性，如 Kalnishkan、Vovk 和 Vyugin [178] 所示，他们还提供了二元情况下混合性的分析表征。对数损失的预测复杂性等价于 Levin 版本的 Kolmogorov 复杂性，参见 Zvonkin 和 Levin [320]。从这个意义上说，预测复杂性可以被视为 Kolmogorov 复杂性的推广。

定理 3.6 归功于 Haussler、Kivinen 和 Warmuth [151]。在同一篇论文中，他们还证明了一个下界，该下界在渐近情况下（对于 $N,n\to\infty$）与任何混合损失 $\ell$ 的上界 $c_\ell \ln N$ 相匹配。这个结果最初由 Vovk [298] 通过更复杂的分析证明。定理 3.7 归功于 Cesa-Bianchi、Freund、Haussler、Helmbold、Schapire 和 Warmuth [48]。其他损失函数的类似结果见 Haussler、Kivinen 和 Warmuth [151]。对绝对损失的更一般下界由 Cesa-Bianchi 和 Lugosi [51] 证明。

## 3.8 练习

3.1 考虑在第 3.2 节中研究的跟随最佳专家预测器。假设 $\mathcal{D}=\mathcal{Y}$ 是一个拓扑向量空间中的凸子集。假设结果序列的性质是 $\lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^n y_t = y$，其中 $y\in\mathcal{Y}$。建立一些弱的一般条件，以保证预测器的每轮遗憾趋近于 0。

3.2 令 $\mathcal{Y}=\mathcal{D}=[0,1]$，并考虑 Hellinger 损失

$$\ell(z,y)=\frac{1}{2}\left((\sqrt{x}-\sqrt{y})^2+(\sqrt{1-x}-\sqrt{1-y})^2\right)$$

（见图 3.6）。确定 $\eta$ 的值，使得第 3.3 节中定义的函数 $F(z)$ 是凹函数。

3.3 令 $\mathcal{Y}=\mathcal{D}$ 为半径 $r>0$ 的闭球，中心在 $\mathbb{R}^d$ 的原点，考虑损失函数 $\ell(z,y)=\|z-y\|^2$（其中 $\|\cdot\|$ 表示 $\mathbb{R}^d$ 中的欧几里得范数）。证明定理 3.2 中的 $F(z)$ 在 $\eta\le 1/\left(8r^2\right)$ 时是凹函数。

3.4 证明如果对于某个 $y\in\mathcal{Y}$ 函数 $F(z)=e^{-\eta\ell(z,y)}$ 是凹的，那么 $\ell(z,y)$ 是 $z$ 的凸函数。

3.5 证明如果一个损失函数对于某个值 $\eta>0$ 是指数凹的，那么它对于任何 $\eta'\in(0,\eta)$ 也是指数凹的。

3.6 在分类问题中，已经使用了各种版本的所谓铰链损失。通常，$\mathcal{Y}=\{-,1\}$，$\mathcal{D}=[-1,1]$，损失函数的形式为 $\ell(\widehat{p},y)=c(-y\widehat{p})$，其中 $c$ 是一个非负、递增且凸的成本函数。推导 $c$ 和参数 $\eta$ 的条件，使得铰链损失是指数凹的。

3.7 （指数凹损失的折旧遗憾）考虑折旧遗憾

$$\rho_{i,n}=\sum_{t=1}^n \beta_{n-t}\left(\ell\left(\widehat{p}_t,y_t\right)-\ell\left(f_{i,t},y_t\right)\right)$$

定义在第 2.11 节中，其中 $1=\beta_0\ge\beta_1\ge\cdots$ 是一个递减的折旧因子序列。假设损失函数 $\ell$ 对于某个 $\eta>0$ 是指数凹的，并考虑折旧指数加权平均预测器

$$\widehat{p}_t=\frac{\sum_{i=1}^N f_{i,t}\exp\left(-\eta\sum_{s=1}^{t-1}\beta_{n-s}\ell\left(f_{i,s},y_s\right)\right)}{\sum_{j=1}^N \exp\left(-\eta\sum_{s=1}^{t-1}\beta_{n-s}\ell\left(f_{j,s},y_s\right)\right)}$$

证明平均折旧遗憾被界定为

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

特别地，证明平均折旧遗憾当且仅当 $\sum_{t=0}^{\infty}\beta_t=\infty$ 时为 $o(1)$。

3.8 证明基于“虚拟博弈”的贪婪预测器（在第 3.4 节开始时定义的）并不保证 $n^{-1}\max_{i\le N}R_{i,n}$ 随着 $n\to\infty$ 对所有结果序列收敛到 0。提示：考虑简单示例，当 $\mathcal{Y}=\mathcal{D}=[0,1]$，$\ell(\widehat{p},y)=|\widehat{p}-y|$，且 $N=2$。

3.9 证明对于 $\mathcal{Y}=\{0,1\}$，$\mathcal{D}=[0,1]$ 和对数损失函数，基于指数势函数的贪婪预测器（$\eta=1$）实际上就是指数加权平均预测器（也是 $\eta=1$）。

3.10 证明对于所有满足以下条件的损失函数 $\ell$，有 $\mu(\eta)\ge 1$：(1) 存在 $p\in\mathcal{D}$ 使得 $\ell(p,y)<\infty$ 对于所有 $y\in\mathcal{Y}$；(2) 不存在 $p\in\mathcal{D}$ 使得 $\ell(p,y)=0$ 对于所有 $y\in\mathcal{Y}$（Vovk [298]）。

3.11 证明以下结论：设 $C\subset\mathbb{R}^2$ 是参数方程 $x=x(t)$ 和 $y=y(t)$ 的曲线，这些方程是二次可微的函数。如果存在一个二次可微的函数 $h$，使得 $y(t)=h(x(t))$ 对于某个开区间内的 $t$ 成立，则

$$\frac{dy}{dx}=\frac{\frac{dy}{dt}}{\frac{dx}{dt}}\quad\text{和}\quad\frac{d^2y}{dx^2}=\frac{\frac{d}{dt}\frac{dy}{dx}}{\frac{dx}{dt}}$$

3.12 验证对于平方损失 $\ell(p,y)=(p-y)^2$，$p,y\in[0,1]$，如果 $\eta\le 1/2$（Vovk [300]），则函数 (3.4) 在 $y$ 上是凹的。

3.13 证明对于平方损失，没有函数 $g$ 使得预测 $\widehat{p}=g\left(\sum_{i=1}^N f_iq_i\right)$ 满足 (3.3) 且 $c=1$。提示：考虑 $N=2$，找到 $f_1,f_2,q_1,q_2$ 以及 $f_1',f_2',q_1',q_2'$ 使得 $f_1q_1+f_2q_2=f_1'q_1'+f_2'q_2'$，但将这些值代入 (3.3) 会导致矛盾（Haussler, Kivinen, 和 Warmuth [151]）。

3.14 证明对于绝对损失，没有函数 $g$ 使得预测 $\widehat{p}=g\left(\sum_{i=1}^N f_iq_i\right)$ 满足 (3.5)，其中 $\mu(\eta)$ 是绝对损失的混合性曲线。提示：设 $\eta=1$ 并按照练习 3.13 中的提示操作（Haussler, Kivinen, 和 Warmuth [151]）。

3.15 证明对于二元结果的绝对损失，混合性曲线也是当结果空间为 [0, 1] 时的混合性函数（Haussler, Kivinen, 和 Warmuth [151]）。警告：这个练习不容易。

3.16 找到以下损失的混合性曲线：$\mathcal{D}$ 是 $\mathbb{R}^N$ 中的概率单纯形，$\mathcal{Y}=[0,1]^N$，且 $\ell(\widehat{p},y)=\widehat{p}\cdot y$（Vovk [298]）。警告：这个练习不容易。