--- 第 1 页 ---

信息量：

定义

2

本讲和下一讲将介绍最重要的信息量的定义和关键性质，包括熵、相对熵和互信息。 这些信息量为后续讲座中工程问题的研究提供了基本工具，同时它们本身也构成了一个连贯的数学统一体。

2.1 离散概率论

我们对离散随机变量的基本概率论进行简要回顾。 我们对信息论的处理将局限于离散情况，直到第10讲和第11讲，届时将考虑连续值随机变量。

一个概率空间由三部分组成：

► **样本空间** Ω，是所讨论的随机实验所有可能结果的集合。 对于离散情况，Ω 是一个有限或可数无限集，即 Ω={ω1,ω2,...,ωn}，其中 n 可能为 ∞。 ► **事件集** F，其每个元素都是 Ω 的一个子集，满足：

• 至少 Ω（必然事件）和 ∅（不可能事件）包含在 F 中； • 如果 A∈F，那么它的补集 Ac∈F； • 对于 F 的任何有限元素集合 A1,A2,...,An，∪i=1nAi∈F，并且对于 F 的任何元素序列 A1,A2,...，∪i=1∞Ai∈F。

► **概率测度** P，它是一个函数，为 F 的每个元素赋予一个在 [0, 1] 区间内的实数，满足：

• P(Ω)=1; • P(∪i=1nAi)=Σi=1nP(Ai)，对于所有 A1,A2,...,An∈F，使得 ∀i\=j,Ai∩Aj=∅; • P(∪i=1∞Ai)=Σi=1∞P(Ai)，对于所有 A1,A2,...∈F，使得 ∀i\=j，Ai∩Aj=∅。

定义 2.1 给定一个概率空间 (Ω, F, P)，**随机变量**是从样本空间 Ω 到一个指定的有限或可数无限集的映射。

例 2.1 设 Ω={ω1,ω2}。我们可以定义以下随机变量。

--- 第 2 页 ---

8

2 信息量：定义

► **伯努利**：X(ω1)=0,X(ω2)=1

► **抛硬币**：Y(ω1)= 正面, Y(ω2)= 反面。

► **水果**：Z(ω1)= 苹果, Z(ω2)= 香蕉。

所以我们可以将概率空间理解为一个抽象结构，并通过定义合适的随机变量来赋予它任何具体的解释。 一个概率空间可以导出不同的随机变量，而它们底层的概率测度是相同的。

我们可以根据随机变量及其底层的概率测度来定义一个**概率分布**。

定义 2.2 给定概率空间 (Ω,F,P) 上的一个随机变量 X，其**概率分布**为

PX(x)=P({ω:X(ω)=x}), (2.1)

其中 x 取值于 X 的值域 X(Ω)。由于我们考虑的是离散概率空间，PX 也被称为 X 的**概率质量函数** (pmf)。在后面的讲座中，有时我们也会用 X 表示 X(Ω)，并称之为 X 的**字母表**。

很明显，PX 满足

1. $P_X(x) \geq 0, \ \forall x \in X(\Omega)$
2. $\Sigma_{x \in X(\Omega)} P_X(x) = 1$.

当考虑在同一概率空间上的一组随机变量时，研究它们的联合行为是很有意义的。

定义 2.3 给定概率空间 $(\Omega, \mathcal{F}, P)$ 上的一组随机变量 $X_1, X_2, \ldots, X_n$，它们的**联合概率分布**定义为

$$P_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = P(\{\omega : X_1(\omega) = x_1\} \cap \{\omega : X_2(\omega) = x_2\} \cap \ldots \ldots \cap \{\omega : X_n(\omega) = x_n\}), \quad (2.2)$$

其中 $(x_1, x_2, \ldots x_n)$ 取值于 $X_1(\Omega) \times X_2(\Omega) \times \ldots \times X_n(\Omega)$。

联合概率分布满足

1. $P_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) \geq 0, \ \forall (x_1,x_2,\ldots x_n) \in X_1(\Omega) \times X_2(\Omega) \times \ldots \times X_n(\Omega)$;
2. $\Sigma_{(x_1,x_2,\ldots x_n) \in X_1(\Omega) \times X_2(\Omega) \times \ldots \times X_n(\Omega)} P_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = 1$;

1. **(边缘化)** 对于任何 $1 \leq i \leq n$,

$$\sum_{x_i \in X_i(\Omega)} P_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = P_{X_1,X_2,\ldots,X_{i-1},X_{i+1},\ldots X_n}(x_1,x_2,\ldots,x_{i-1},x_{i+1},\ldots,x_n).$$

对于在同一概率空间上的一组随机变量，研究其中一些变量在其余变量固定时的行为也很有意义。 这由**条件概率分布**来描述。

定义 2.4 给定概率空间 $(\Omega, \mathcal{F}, P)$ 上的随机变量 X 和 Y，对于 $x \in X(\Omega)$ 且 $P_X(x) > 0$，Y 以 $\{\omega : X(\omega) = x\}$ 为条件的**条件概率分布**定义为

$$P_{Y|X}(y \mid x) = P_{X}(x) P_{X,Y}(x,y), \quad (2.3)$$

对于 $x \in X(\Omega)$ 且 $P_X(x) = 0$，$P_{Y|X}(y \mid x)$ 未定义。

从条件概率分布的定义，著名的**贝叶斯定理**是其直接推论。

定理 2.1 考虑概率空间 $(\Omega, \mathcal{F}, P)$ 上的随机变量 X 和 Y，贝叶斯定理指出，对于任何 $y \in Y(\Omega)$ 且 $P_Y(y) > 0$,

$$P_{X|Y}(x \mid y) = P_Y(y) P_{Y|X}(y \mid x) P_X(x). \quad (2.4)$$

独立性是随机变量间一种特殊的联合行为，定义如下。

定义 2.5 概率空间 $(\Omega, \mathcal{F}, P)$ 上的随机变量 $X_1, X_2, \ldots, X_n$ 被称为**相互独立**，如果

$$P_{X_1,X_2,\ldots,X_n}(x_1,x_2,\ldots,x_n) = P_{X_1}(x_1) P_{X_2}(x_2) \ldots P_{X_n}(x_n), \quad (2.5)$$

对于任何 $(x_1, x_2, \ldots x_n) \in X_1(\Omega) \times X_2(\Omega) \times \ldots \times X_n(\Omega)$; 并且被称为**两两独立**，如果对于任何一对 $i \neq j$

$$P_{X_i,X_j}(x_i,x_j) = P_{X_i}(x_i) P_{X_j}(x_j), \quad (2.6)$$

对于任何 $(x_i, x_j) \in X_i(\Omega) \times X_j(\Omega)$。

注意，只有当 $n = 2$ 时，两两独立和相互独立才是等价的。 举一个简单的例子，让 X 和 Z 是相互独立的伯努利随机变量，其中 $P_X(0) = P_X(1) = 21$, $P_Z(0) = P_Z(1) = 21$, 并让 Y 是模二和

X 和 Z 的关系，Y=X⊕Z。可以很容易地验证 X, Y, Z 是两两独立的，但不是相互独立的。

一个重要的情况是 X1,X2,...,Xn 是相互独立的，并且 PX1,PX2,...,PXn 是相同的。 我们称这 n 个随机变量为**独立同分布 (i.i.d.)**。

条件独立是一个至关重要的概念，在信息论中有重要应用。

定义 2.6 对于概率空间 (Ω,F,P) 上的随机变量 X, Y 和 Z，如果对于任何 (x,y,z)∈X(Ω)×Y(Ω)×Z(Ω)

PX,Z ∣ Y(x,z ∣ y)=PX ∣ Y(x ∣ y)PZ ∣ Y(z ∣ y) (2.7)

则称 X 和 Z 在给定 Y 的条件下是**条件独立**的。这种关系可以表示为 X↔Y↔Z，并被称为**马尔可夫链**。

接下来我们介绍随机变量的期望。

定义 2.7 对于概率空间 (Ω,F,P) 上的一个随机变量 X，以及一个函数 F:X(Ω)↦R，F(X) 的**期望**定义为

E[F(X)]=∑x∈X(Ω)F(x)PX(x)。 (2.8)

类似地，对于概率空间 (Ω,F,P) 上的一组随机变量 X1,X2,...,Xn 和一个函数

F:X1(Ω)×X2(Ω)×...×Xn(Ω)↦R,

F(X1,X2,...,Xn) 的期望定义为

E[F(X1,X2,...,Xn)]= (2.9) ∑(x1,x2,...,xn)∈X1(Ω)×X2(Ω)×...×Xn(Ω)F(x1,x2,...,xn)PX1,X2,...,Xn(x1,x2,...,xn). (2.10)

以下期望的基本性质将非常有用。

定理 2.2 对于概率空间 (Ω,F,P) 上的随机变量 X 和 Y，以及函数 F:X(Ω)↦R 和 G:Y(Ω)↦R，我们有

► **线性性**:

E[F(X)+G(X)]=E[F(X)]+E[G(X)]. (2.11)

► **缩放**: 对于任何 c∈R,

E[cF(X)]=cE[F(X)]. (2.12)

► 如果 X 和 Y 是独立的，

E[F(X)G(Y)]=E[F(X)]E[G(Y)] (2.13)

关于第三个性质 (2.13)，如果 E[XY]=E[X]E[Y]，那么我们说 X 和 Y 是**不相关**的。 注意，独立性意味着不相关性，但反之通常不成立。

指示函数通常便于推导。 考虑一个概率空间 (Ω, F, P)。对于任何事件 A∈F，在 Ω 上定义一个函数 EA，如果 ω∈A 则 EA(ω)=1，否则为 0。 这也可以写成 EA(ω)=1{ω∈A}。那么 EA 是一个指示事件 A 是否发生的随机变量，并且我们有 E[EA]=P(A)。

我们可以进一步定义**条件期望**如下。

定义 2.8 对于概率空间 (Ω,F,P) 上的随机变量 X 和 Y，以及一个函数 F:X(Ω)↦R，F(X) 在事件 {ω:Y(ω)=y}，y∈Y(Ω), 条件下的**条件期望**定义为

E[F(X) ∣ y]=∑x∈X(Ω)F(x)PX ∣ Y(x ∣ y)。 (2.14)

应该记住，E[F(X) ∣ Y] 本身是一个由随机变量 Y 导出的随机变量。

以下关于条件期望的性质，称为**全期望定律**，非常有用。

定理 2.3 对于概率空间 (Ω,F,P) 上的随机变量 X 和 Y，我们有

E[X]=E[E[X ∣ Y]] (2.15)

此时，我们介绍随机变量的收敛。 随机变量有几种不同的收敛概念，但对于我们讲义中的大部分目的，我们只需要一种弱形式的收敛，如下所示。

定义 2.9 对于一个随机变量序列 X1,X2,..., 如果存在一个随机变量 X，使得对于任何 ϵ>0，

limn→∞P(∣Xn−X∣<ϵ)=1, (2.16)

则 X1,X2,... **依概率收敛**于 X，记为 "XnP imgX" 或 "Xn→X in probability"。

当 X1,X2,... 是 i.i.d. 随机变量时，以下**弱大数定律 (WLLN)** 成立。

定理 2.4 对于一个 i.i.d. 随机变量序列 X1,X2,...，定义 Yn=(X1+X2+...+Xn)/n, n=1,2,...。那么我们有 Yn→E[X] 依概率收敛。

WLLN 仅表明对于任何足够大的 n，Yn 接近 E[X] 的概率任意接近于 1，但它不排除 Yn 的某些特定轨迹，n=1,2,..., 最终偏离 E[X] 的可能性。 这种可能性被**强大数定律 (SLLN)** 所排除，该定律指出轨迹 Yn，n=1,2,..., 以概率 1 收敛于极限 E[X]。 SLLN 超出了我们讲义的范围。

2.2 熵、联合熵和条件熵

考虑概率空间 (Ω, F, P) 上的一个随机变量 X。 我们以概率 PX(x)=P({ω:X(ω)=x}) 观测到 X=x，并将这次观测的**信息**或“**惊奇度**”称为

i(x)=logPX(x)1, ∀x∈X(Ω). (2.17)

我们将 X 的**熵**定义为 i(X) 的期望。

定义 2.10 对于概率空间 (Ω,F,P) 上的一个随机变量 X，其**熵**定义为

H(X)=E[i(X)]=−∑x∈X(Ω)PX(x)logPX(x). (2.18)

信息论中的熵概念与其在统计物理学中的同名概念有密切联系，但如下一讲所示，它有其自身的工程解释，超越了其统计物理学的对应部分。 熵定义中对数的底可以是任意的。常用的底包括 2 和 e。 当底为 2 时，熵的单位是**比特 (bit)**； 当底为 e 时，熵的单位是**奈特 (nat)**。 可以很容易地验证 1 奈特等于 log2e≈1.443 比特。 另一个很少使用的底是 10，对应的熵单位是**哈特利 (hartley)**。 出于实际目的，工程师可能更喜欢使用比特，因为在数字电路中二进制状态是

通用的，但对于信息论研究，使用奈特通常更方便，这得益于 ln 函数良好的分析性质。

在熵的定义中，求和可以跳过 PX(x)=0 的 x，这由极限 limp→0+plogp=0 来证明是合理的。换句话说，我们可以采用 0log0=0 的约定。

例 2.2 对于伯努利随机变量 X，X(Ω)={0,1}，PX(0)=1−ϵ 和 PX(1)=ϵ, 我们有

$$H(X)=-\sum_{x\in\{0,1\}}P_X(x)\log P_X(x) =-(1-\epsilon)\log(1-\epsilon)-\epsilon\log\epsilon. \quad (2.19)$$

我们通常用 $h_2(\epsilon)$ 表示这个熵，其中下标 2 表示随机变量是二元的。 我们在图 2.1 中绘制了 $h_2(\epsilon)$（单位为比特）。注意 $h_2(\epsilon)$ 关于 $\epsilon=0.5$ 是对称的，即 $h_2(\epsilon)=h_2(1-\epsilon)$。图上的以下特殊点可能有用：$h_2(0)=h_2(1)=0$，$h_2(0.5)=1$，以及 $h_2(0.11)=h_2(0.89)\approx 0.5$（均以比特为单位）。 在 $\epsilon=0$ 附近展开 $h_2(\epsilon)$，我们有 $h_2(\epsilon)=\epsilon\ln\epsilon 1+\epsilon+o(\epsilon)$（以奈特为单位），其第一项是主导项。 在图形上，这意味着 $h_2(\epsilon)$ 在 0 处的斜率是无穷大的。

图 2.1: 伯努利随机变量的熵（单位：比特）的图像

图 2.1: 伯努利随机变量的熵（单位：比特）。

例 2.3 考虑一个几何随机变量 X，$X(\Omega)=\{1,2,...\}$，其 $P_X(x)=\epsilon(1-\epsilon)x-1$。这样的随机变量可以解释为在一系列 i.i.d. 伯努利(e)随机变量中第一次出现 "1" 时的试验次数。 可以验证 X 的期望是 $\epsilon 1$。 X 的熵可以计算为

图 2.2: 几何随机变量的熵（单位：比特）的图像

图 2.2: 几何随机变量的熵（单位：比特）。

计算为

$$H(X)=-E[\log P_X(X)] =-E[\log\epsilon(1-\epsilon)X-1] =-\log\epsilon-\log(1-\epsilon)\cdot E[X-1] =-\log\epsilon-\log(1-\epsilon)\cdot(\epsilon 1-1) =\epsilon-\epsilon\log\epsilon-(1-\epsilon)\log(1-\epsilon)=\epsilon h_2(\epsilon). \quad (2.20)$$

我们可以验证 $H(X)$ 在 $\epsilon\in(0,1]$ 中是单调递减的，并且 $\lim_{\epsilon\to 0+}H(X)=\infty$, 如图 2.2 所示。

当有多个随机变量时，我们可以将它们一起处理并定义它们的**联合熵**如下。

定义 2.11 对于概率空间 (Ω, F, P) 上的随机变量 X 和 Y，它们的**联合熵**定义为

$$H(X,Y)=E[i(X,Y)] =-\sum_{(x,y)\in X(\Omega)\times Y(\Omega)}P_{X,Y}(x,y)\log P_{X,Y}(x,y). \quad (2.21)$$

对于两个以上的随机变量，它们的联合熵可以用类似的方式定义。

现在我们考虑一个随机变量在另一个随机变量条件下的熵。

定义 2.12 对于概率空间 (Ω,F,P) 上的随机变量 X 和 Y，在事件 $\{\omega:X(\omega)=x\}$ 发生条件下 Y 的熵是

$$H(Y\mid X=x)=-\sum_{y\in Y(\Omega)}P_{Y\mid X}(y\mid x)\log P_{Y\mid X}(y\mid x)。 \quad (2.22)$$

那么，给定 X 的 Y 的**条件熵**定义为

$$H(Y\mid X)=\sum_{x\in X(\Omega)}P_X(x)H(Y\mid X=x) =-\sum_{(x,y)\in X(\Omega)\times Y(\Omega)}P_{X,Y}(x,y)\log P_{Y\mid X}(y\mid x) =-E[\log P_{Y\mid X}(Y\mid X)] \quad (2.23)$$

其中期望是关于联合概率分布 $P_{X,Y}(x,y)$ 的。

例 2.4 考虑独立的伯努利随机变量 X 和 Z，每个都以等概率 0.5 取值 1 和 0，并设 $Y=X\cdot Z$，即 X 和 Z 之间的乘积。所以我们可以计算

$$H(X,Y)=\sum_{(x,y)\in X(\Omega)\times Y(\Omega)}P_{X,Y}(x,y)\log P_{X,Y}(x,y)1 =21\log 2+0+41\log 4+41\log 4 =23\log 2, \quad (2.24)$$

$$H(Y\mid X=0)=h_2(0)=0, \quad (2.25)$$

H(Y ∣ X=1)=h2(0.5)=log2, (2.26)

H(Y ∣ X)=21H(Y ∣ X=0)+21H(Y ∣ X=1) =21log2. (2.27)

2.3 相对熵和互信息

给定两个概率分布，**相对熵**（也称为**库尔贝克-莱布勒距离**或**I-散度**）是描述它们之间差异的广泛使用的量。

定义 2.13 两个概率分布 P(x) 和 Q(x) 之间的**相对熵**定义为

D(P ∣ ∣ Q)=∑x∈XP(x)logQ(x)P(x), (2.28)

其中 X 表示 x 的字母表。

通常情况下，对于某些 x∈x，P(x) 或 Q(x) 可能为零，我们在计算 D(P ∣ ∣ Q) 时采用以下约定：

► 0log00=0; ► 0logq0=0 对于 q>0;

► plog0p=∞ 对于 p>0。

例 2.5 考虑两个伯努利分布 P 和 Q，其中 PX(0)=1−ϵ，PX(1)=ϵ，QX(0)=1−δ，QX(1)=δ。我们有

D(P ∣ ∣ Q)=(1−ϵ)log1−δ1−ϵ+ϵlogδϵ. (2.29) D(Q ∣ ∣ P)=(1−δ)log1−ϵ1−δ+δlogϵδ. (2.30)

有趣的是，每当 ϵ>0 且 δ=0 时，D(P ∣ ∣ Q)=∞。 这个例子也清楚地表明，通常情况下，D(P ∣ ∣ Q)\=D(Q ∣ ∣ P)，也就是说，相对熵对于其两个概率分布是不对称的。

例 2.6 考虑三个伯努利分布 P, Q, 和 Q'，其中 PX(0)=1, PX(1)=0, QX(0)=21, QX(1)=21, QX'(0)=41, QX'(1)=43。 那么我们有 D(P ∣ ∣ Q)=log2, D(P ∣ ∣ Q')=2log2, 和 D(Q ∣ ∣ Q')=log2−21log3。因此，D(P ∣ ∣ Q)+D(Q ∣ ∣ Q') −D(P ∣ ∣ Q')=−21log3<0。这个例子因此表明，相对熵通常不满足三角不等式。

现在我们准备好将具有联合概率分布的两个随机变量之间的**互信息**定义为一个特殊的相对熵。

定义 2.14 对于概率空间 (Ω,F,P) 上的随机变量 X 和 Y，它们的**互信息**定义为

I(X;Y)=D(PX,Y ∣ ∣ PXPY) (2.31) =∑(x,y)∈X(Ω)×Y(Ω)PX,Y(x,y)logPX(x)PY(y)PX,Y(x,y) (2.32)

对于任何 (x,y)∈X(Ω)×Y(Ω)，我们定义其相关的**信息密度**为*

i(x;y)=logPX(x)PY(y)PX,Y(x,y). (2.33)

我们看到互信息 I(X;Y) 是 i(X;Y) 的期望。

当存在一个额外的随机变量 Z 时，给定 Z 的 X 和 Y 之间的**条件互信息**定义如下。

* 注意，我们在这里重用了符号 i，它既用于 (2.17) 中的信息或“惊奇度”，也用于这里的信息密度。

定义 2.15 给定 Z 的 X 和 Y 之间的**条件互信息**是

I(X;Y ∣ Z)=∑(x,y,z)∈X(Ω)×Y(Ω)×Z(Ω) PX,Y,Z(x,y,z)logPX ∣ Z(x ∣ z)PY ∣ Z(y ∣ z)PX,Y ∣ Z(x,y ∣ z) (2.34)

2.4 熵率

在我们的讲义中，我们满足于将离散时间随机过程看作是某个概率空间上的一系列随机变量。 从时间索引 1 开始，一个随机过程 X 表示为 X1,X2,...。X 中的前 n 个随机变量的联合熵为 H(X1,X2,...,Xn)，我们研究当 n 无限增长时 H(X1,X2,...,Xn) 的趋势。 这引出了随机过程熵率的定义。

定义 2.16 对于一个随机过程 X，其**熵率**定义为

H(X)=limn→∞n1H(X1,X2,...,Xn) , (2.35)

当该极限存在时。

类似地，对于两个随机过程 X 和 Y，我们可以定义它们的**互信息率**为

I(X;Y)=limn→∞n1I(X1,...,Xn;Y1,...,Yn) , (2.36)

当该极限存在时。

熵率的性质和例子将在下一讲中提供。

注释

熵和互信息都首次出现在香农的里程碑式论文[1]中，但其中并未使用“互信息”这一术语。 相反，香农称条件熵为“含糊度 (equivocation)”，并使用了熵 H(X) 和含糊度 H(X∣Y) 之间的差，这正如下一讲将要展示的，就是互信息 I(X;Y)。 “互信息”这个名称在[1]发表几年后出现（例如，见[4]）。 有传言说约翰·冯·诺伊曼建议使用“熵”这个名称，但香农在一次采访中澄清

--- 第 12 页 ---

事实并非如此。 “比特 (bit)”是“二进制数字 (binary digit)”的简称，这个名字是由快速傅里叶变换的发明者之一约翰·W·图基建议的。

在香农之前，哈里·奈奎斯特在 1924 年左右提出，通信系统的传输速率应与单位时间内信号电平数量的对数成正比，而拉尔夫·哈特利在 1928 年左右提议将变量的信息定量地度量为其字母表大小的对数。 例如，对于一个装有两种颜色（黑和白）球的罐子，从中抽一个球的信息量是 log 2。这些思想可以被看作是熵的先驱，但它们肯定是不够的，因为它们没有考虑到概率性质。

熵有许多推广。例如，以匈牙利数学家阿尔弗雷德·雷尼命名的 **α 阶雷尼熵**，其中 α≥0 且 α\=1，对于一个随机变量 X，定义为

Hα(X)=1−α1log(∑x∈X(Ω)PX(x)α) (2.37)

当 α→1 时，雷尼熵收敛到我们在定义 2.10 中介绍的熵。

信息密度 i(x;y) 最早由俄罗斯数学家马克·S·平斯克在他对信息量的研究中使用[5]，多年来，这导致了信息论的信息谱方法（例如，见[6]）。 **相对熵**最早由美国数学家所罗门·库尔贝克和理查德·莱布勒引入[7]，并在信息论、统计学和机器学习中得到了广泛应用。

练习

1. 对于一个概率空间 (Ω, F, P)，证明以下性质：

   a) P(∅)=0。 b) 对于任何 A, B∈F，如果 A⊆B 则 P(A)≤P(B)。 c) 对于任何 A, B∈F，P(A∪B)=P(A)+P(B)−P(A∩B)。

2. 对于离散随机变量 X 和 Y 在一个概率空间 (Ω, F, P) 上，

a) 证明全期望定律，

E[X] = E[E[X|Y]]。

--- 第 13 页 ---

b) 证明全方差定律，

varX=E[var[X ∣ Y]]+varE[X ∣ Y]

1. 设 X1, X2,X3,X4 是随机变量，使得 X1↔(X2,X3)↔X4 和 X1↔(X2,X4)↔X3 同时成立。 a) 如果对于任何 (x1,x2,x3,x4)∈X1(Ω)×X2(Ω)×X3(Ω)×X4(Ω)，PX1,X2,X3,X4(x1,x2,x3,x4)>0，证明 X1↔X2↔(X3,X4) 成立。 b) 你能给出一个例子，其中对于某个 (x1,x2,x3,x4)，PX1,X2,X3,X4(x1,x2,x3,x4)=0, 且 X1↔X2↔(X3,X4) 不成立吗？这说明了具有严格零概率的概率分布的微妙之处 [8, Prop. 2.12]。
2. 证明以下基本不等式：

   a) **马尔可夫不等式**：对于一个具有有限期望的非负随机变量 X 和任何 a>0,

   P(X≥a)≤aE[X]. (2.38)

   b) **切比雪夫不等式**：对于一个具有有限期望和方差的随机变量 X 和任何 a>0

   P(∣X−E[X]∣≥a)≤a2varX. (2.39)

   c) **切诺夫不等式**：对于一个随机变量 X 和任何 a，

   P(X≥a)≤minλ≥0e−λaE[eλX]. (2.40)
3. 如果我们用 PX,Y(x,y)=PX(x)PY(y)(1+ϵ(x,y)) 来模拟一对弱依赖的随机变量 X 和 Y，使得存在 δ<1 满足 ∣ϵ(x,y)∣≤δ, ∀(x,y)∈X(Ω)×Y(Ω)，你能提供一个关于 H(X,Y) 和 H(X)+H(Y) 之间差异的上界吗？
4. **交叉熵**是机器学习中的一个重要概念，通常在训练神经网络进行分类任务时用作目标函数。 对于定义域为 X 的两个概率分布 P(x) 和 Q(x)，Q(x) 相对于 P(x) 的交叉熵定义为

   Hc(P,Q)=−∑x∈XP(x)logQ(x).

   当 P(x) 和 Q(x) 分别是参数为 ϵP 和 ϵQ 的几何分布时，计算交叉熵 Hc(P,Q)。

--- 第 14 页 ---

1. 对于一个值域大小为 m 的随机变量，我们可以将其 pmf 表示为一个包含 {pi}i=1m 元素的向量，并将其熵表示为 H(p1,p2,...,pm)=−Σi=1mpilogpi。 验证熵的以下性质：

   a) **可扩展性**：H(p1,p2,...,pm,0)=H(p1,p2,...,pm)。 b) **可加性**：

   H(p1,p2,...,pm)+H(q1,q2,...,qn) =H(p1q1,...,p1qn,p2q1,...,pmq1,...,pmqn)

   c) **分组性**：

   H(p1,p2,...,pm,q1,q2,...,qn)=H(∑i=1mpi,∑j=1nqj) +(∑i=1mpi)H(∑i=1mpip1,∑i=1mpip2,...,∑i=1mpipm) +(∑j=1nqj)H(∑j=1nqjq1,∑j=1nqjq2,...,∑j=1nqjqn)

   这些是各种“公理化”性质中的一部分。 可以证明，从一组合适的此类公理化性质出发，熵函数的定义除了一个缩放因子外是唯一的； 参见 [9] 中对此类结果的总结。
2. 考虑独立的随机变量 X 和 Y，每个都在 {1,2,...,n} 上均匀分布。 a) 使用计算机数值研究 H(X+Y) 并绘制其随 n 的增长图。 b) 使用计算机数值研究 H(X·Y) 并绘制其随 n 的增长图。

--- 第 15 页 ---

信息量： 性质 3

本讲继续我们对信息量的探索，发展它们的关键性质。 当给出随机变量时，它们的熵和互信息可以根据第2讲中介绍的定义进行计算。然而，正如本讲将要展示的，熵和互信息有许多有用的性质，这些性质不仅极大地简化了计算，而且在我们后续讲座中研究信息论问题时也起着关键作用。

3.1 链式法则

链式法则为将涉及多个随机变量的熵或互信息分解为多个各自涉及较少随机变量的项提供了有用的工具。 从数学上讲，这种分解并不奇怪，它是对数函数性质的直接结果； 也就是说，几个变量乘积的对数等于各个变量对数的和。

定理 3.1 熵的链式法则：

► **（基本形式）** 对于两个随机变量 X 和 Y，我们有

H(X,Y)=H(X)+H(Y∣X)=H(Y)+H(X∣Y). (3.1)

► **（条件形式）** 对于三个随机变量 X, Y 和 Z，我们有

H(X,Y∣Z)=H(X∣Z)+H(Y∣X,Z). (3.2)

► **（一般形式）** 对于 n 个随机变量 X1,X2,...,Xn，我们有

H(X1,X2,...,Xn)=∑i=1nH(Xi∣Xi−1,...,X1), (3.3)

其中 X0 被理解为一个退化的随机变量，比如一个常数，因此 H(X1∣X0)=H(X1)。

证明：我们只证明基本形式。 另外两种形式可以类似地证明，并留作练习。根据联合熵的定义

--- 第 16 页 ---

联合熵（定义 2.11），我们有

H(X,Y)=E[logPX,Y(X,Y)1] =E[logPX(X)PY∣X(Y∣X)1] =E[logPX(X)1+logPY∣X(Y∣X)1] =
(b)E[logPX(X)1]+E[logPY∣X(Y∣X)1] =H(X)+H(Y∣X), (3.4)

其中 (a) 是因为注意到 i(x,y)=logPX,Y(x,y)1，(b) 来自期望的线性性质（定理 2.2），在 (c) 中我们将第一个期望中的概率分布从 PX,Y 改为 PX。这表明 H(X,Y)=H(X)+H(Y∣X)；另一个恒等式 H(X,Y)=H(Y)+H(X∣Y) 可以通过交换 X 和 Y 的角色以同样的方式证明。

现在来考察互信息。 互信息的以下性质是基本的。

定理 3.2 互信息满足以下基本性质：

► **（对称性）** I(X;Y)=I(Y;X), ► **（自信息）** I(X;X)=H(X)。 ► **（分解）**

I(X;Y)=H(X)−H(X∣Y) (3.5) =H(Y)−H(Y∣X) (3.6) =H(X)+H(Y)−H(X,Y)。 (3.7)

► **（带条件的分解）**

I(X;Y∣Z)=H(X∣Z)−H(X∣Y,Z) (3.8) =H(Y∣Z)−H(Y∣X,Z) (3.9) =H(X∣Z)+H(Y∣Z)−H(X,Y∣Z)。 (3.10)

证明：所有这些基本性质都是互信息定义（定义 2.14）的直接推论，因此留作练习。

互信息的链式法则由以下定理给出。

--- 第 17 页 ---

定理 3.3 对于随机变量 X1,X2,...,Xn,Y 我们有

I(X1,X2,...,Xn;Y)=∑i=1nI(Xi;Y ∣ Xi−1,...,X1). (3.11)

证明： 我们从定理 3.2 中的分解性质开始得到

I(X1,X2,...,Xn;Y) =H(X1,X2,...,Xn)−H(X1,X2,...,Xn ∣ Y). (3.12)

然后，根据熵的链式法则（定理 3.1），我们有

H(X1,X2,...,Xn)=∑i=1nH(Xi ∣ Xi−1,...,X1). (3.13) H(X1,X2,...,Xn ∣ Y)=∑i=1nH(Xi ∣ Xi−1,...,X1,Y). (3.14)

因此，将这两个分解结合起来，我们得到

I(X1,X2,...,Xn;Y) =∑i=1n[H(Xi ∣ Xi−1,...,X1)−H(Xi ∣ Xi−1,...,X1,Y)] =∑i=1nI(Xi;Y ∣ Xi−1,...,X1). (3.15)

这证明了互信息的链式法则。

3.2 非负性

信息论中最基本的非负性是相对熵总是非负的。

定理 3.4 相对熵 D(P ∣ ∣ Q) 总是非负的，且当且仅当 P(x)=Q(x) ∀x∈x 时等于零。

证明： 使用自然对数不失一般性，根据其定义（定义 2.13），P 和 Q 之间的相对熵可以写成

D(P ∣ ∣ Q)=∑x∈XP(x)lnQ(x)P(x) =∑x∈X'P(x)lnQ(x)P(x), (3.16)

其中 X'=X{x:P(x)=0}。

--- 第 18 页 ---

应用不等式 lnt≤t−1, ∀t≥0，当且仅当 t=1 时等号成立，我们有

D(P ∣ ∣ Q)=−∑x∈X'P(x)lnP(x)Q(x) ≥−∑x∈X'P(x)[P(x)Q(x)−1] =−∑x∈X'[Q(x)−P(x)] =−∑x∈X'Q(x)+∑x∈X'P(x) ≥−∑x∈XQ(x)+∑x∈XP(x) =−1+1=0。 (3.17)

检查上述步骤中的两个不等式，第一个不等式当且仅当对于任何 x∈X' 都有 P(x)=Q(x) 时等号成立，第二个不等式当且仅当每当 P(x)=0 时都有 Q(x)=0 时等号成立。所以总而言之，我们有 D(P ∣ ∣ Q)≥0，当且仅当 P(x)=Q(x), ∀x∈x 时等号成立。

相对熵的非负性对熵和互信息有许多重要且有用的推论。 首先，我们给出熵的一般下界和上界如下。

推论 3.1 熵 H(X) 的界为 0≤H(X)≤log ∣ X(Ω) ∣ 。此外，H(X)=0 成立当且仅当 X 是一个确定性常数，而 H(X)=log ∣ X(Ω) ∣ 成立当且仅当 X 服从 X(Ω) 上的均匀分布。

证明： 下界为零是显而易见的，注意到 H(X)=E[i(X)] 和 i(x)=logPX(x)1≥0 ∀x∈X(Ω)。 达到下界零的条件是，对于 ∀x∈X(Ω)，要么 i(x)=0 要么 PX(x)=0 成立，这要求 X 是一个确定性常数。 为了证明上界 log ∣ X(Ω) ∣ ，我们考察 PX 和 X(Ω) 上均匀分布 PX,u 之间的相对熵。

D(PX ∣ ∣ PX,u)=∑x∈X(Ω)PX(x)logPX,u(x)PX(x) =∑x∈X(Ω)PX(x)log(∣ X(Ω) ∣ PX(x)) =−H(X)+log ∣ X(Ω) ∣ 。 (3.18)

根据定理 3.4，D(PX ∣ ∣ PX,u)≥0 成立，且等号成立当且仅当 PX 也是均匀分布

--- 第 19 页 ---

X(Ω)。所以我们有 H(X)≤log ∣ X(Ω) ∣ , 等号成立当且仅当 X 服从在 X(Ω) 上的均匀分布。

相对熵非负性的另一个推论是以下的最大熵结果。

推论 3.2 对于一个在 X(Ω)={1,2,...} 上满足 EX=A>1 的随机变量 X，具有 pmf PX,g(x)=A1(1−A1)x−1 的几何分布最大化了熵 H(X)。

证明： 对于一个期望为 A 的几何随机变量，正如在第 2 讲的例 2.3 中计算的，其熵为

H(X)=1/Ah2(1/A) =AlogA−(A−1)log(A−1)。 (3.19)

对于一个满足 EX=A>1 的在 X(Ω)={1,2,...} 上的任意随机变量 X，让我们考察其概率分布 PX 和几何分布 PX,g 之间的相对熵。

D(PX ∣ ∣ PX,g)=∑x=1∞PX(x)logPX,g(x)PX(x) =−H(X)−∑x=1∞PX(x)log[A1(1−A1)x−1] =−H(X)+logA−logAA−1E[X−1] =−H(X)+AlogA−(A−1)log(A−1)。(3.20)

因此，由于 D(PX ∣ ∣ PX,g) 的非负性，H(X) 总是被几何分布的熵所上界，这便完成了证明。

由于互信息是一种特殊的相对散度，非负性也适用于互信息。

推论 3.3 对于随机变量 X 和 Y，互信息满足 I(X;Y)≥0, 等号成立当且仅当 X 和 Y 是独立的。

证明： 回忆 I(X;Y)=D(PX,Y ∣ ∣ PXPY)，I(X;Y) 的非负性立即得出。 D(PX,Y ∣ ∣ PXPY)=0 的充要条件是 PX,Y(x,y)=PX(x)PY(y), ∀(x,y)∈X(Ω)×Y(Ω)，这正是 X 和 Y 独立的条件。

条件互信息也满足非负性。

--- 第 20 页 ---

推论 3.4 对于随机变量 X, Y 和 Z，条件互信息 I(X;Y ∣ Z)≥0，等号成立当且仅当 X 和 Y 在给定 Z 的条件下是条件独立的； 也就是说，马尔可夫链 X↔Z↔Y 成立。

证明： 回忆在第 2 讲的定义 2.15 中，I(X;Y ∣ Z) 定义为

I(X;Y ∣ Z)= ∑(x,y,z)∈X(Ω)×Y(Ω)×Z(Ω)PX,Y,Z(x,y,z)logPX ∣ Z(x ∣ z)PY ∣ Z(y ∣ z)PX,Y ∣ Z(x,y ∣ z) (3.21)

可以重写为

I(X;Y ∣ Z)=∑z∈Z(Ω)PZ(z)(∑(x,y)∈X(Ω)×Y(Ω)PX,Y ∣ Z(x,y ∣ z)logPX ∣ Z(x ∣ z)PY ∣ Z(y ∣ z)PX,Y ∣ Z(x,y ∣ z)) (3.22)

对于每个 z∈Z(Ω)，内层求和正是 PX,Y ∣ Z 和 PX ∣ ZPY ∣ Z 之间的相对熵。 因此，I(X;Y ∣ Z) 的非负性直接源于相对熵的非负性。 此外，为了使 I(X;Y ∣ Z)=0，除非 PZ(z)=0，否则对于每个 z∈Z(Ω)，PX,Y ∣ Z 和 PX ∣ ZPY ∣ Z 之间的相对熵必须为零。这等价于 X 和 Y 在给定 Z 的条件下是条件独立的（定义 2.6）。

推论 3.4 在可以识别出多个随机变量之间的马尔可夫链时，与互信息的链式法则结合使用尤其有用。

条件互信息非负性的一个重要推论是**数据处理不等式 (DPI)**，它断言互信息沿马尔可夫链递减。

定理 3.5 对于满足马尔可夫链 X↔Y↔Z 的三个随机变量 X, Y 和 Z，我们有 I(X;Y)≥I(X;Z)，等号成立当且仅当 X↔Z↔Y 也构成一个马尔可夫链。

证明： 让我们用链式法则以两种方式展开互信息 I(X;Y,Z)：

I(X;Y,Z)=I(X;Y)+I(X;Z ∣ Y) (3.23) =I(X;Z)+I(X;Y ∣ Z) (3.24)

因为 X↔Y↔Z，根据推论 3.4，我们有 I(X;Z ∣ Y)=0。 这直接导致 I(X;Y)≥I(X;Z)。为了使等号

--- 第 21 页 ---

成立，我们需要 I(X;Y ∣ Z)=0, 这根据推论 3.4 等价于 X↔Z↔Y 也构成一个马尔可夫链的条件。

作为一个特例，如果我们通过一个映射 f 处理 Y 以获得 f(Y)，那么总是有 I(X;f(Y))≤I(X;Y)。 直观地说，处理原始数据（即 Y）会降低我们提取其信息内容（即 X）的能力。 如果 DPI 中等号成立，Z 被称为 Y 的一个**充分统计量**，这个概念在统计学中起着关键作用。

3.3 条件化的影响

从推论 3.3 和 3.4，我们可以得到熵和条件熵之间的以下关系，通常称为"**条件作用降低熵**"。

定理 3.6 对于随机变量 X, Y 和 Z，

H(X)≥H(X ∣ Y), (3.25)

等号成立当且仅当 X 和 Y 是独立的，并且

H(X ∣ Z)≥H(X ∣ Y,Z), (3.26)

等号成立当且仅当 X 和 Y 在给定 Z 的条件下是条件独立的。

3.3 条件化的影响 27

证明：因为

I(X;Y)=H(X)−H(X ∣ Y), (3.27) I(X;Y ∣ Z)=H(X ∣ Z)−H(X ∣ Y,Z), (3.28)

不等式 (3.25) 和 (3.26) 分别直接源于互信息（推论 3.3）和条件互信息（推论 3.4）的非负性。

那么，从定理 3.6 和熵的链式法则（定理 3.1）可以明显得出以下推论。

推论 3.5 对于随机变量 X1,X2,...,Xn，我们有

H(X1,X2,...,Xn)≤∑i=1nH(Xi), (3.29)

等号成立当且仅当 X1,X2,...,Xn 是相互独立的。

--- 第 22 页 ---

28

3 信息量：性质

作为熵的非负性和"条件作用降低熵"性质的应用，我们有以下关于随机变量及其映射之间关系的有用结果。

推论 3.6 考虑一个随机变量 X，

► 存在一个随机变量 Y 使得 H(Y ∣ X)=0 成立，当且仅当 Y 在给定 X 的条件下是确定性的，即，存在一个映射 f 使得 Y=f(X)。

► 对于任何映射 f，H(f(X))≤H(X) 成立，等号成立当且仅当 f 是一个双射，即，它拥有一个逆映射 f−1 使得 f−1(f(X))=X。

证明：根据条件熵的定义（定义 2.12），我们有

I(Y ∣ X)=∑x∈X(Ω)PX(x)H(Y ∣ X=x)。 (3.30)

因此，根据熵的非负性（推论 3.1），H(Y ∣ X)=0 等价于对于每个 x∈X(Ω)（除了那些 PX(x)=0 的情况），H(Y ∣ X=x)=0。此外，根据推论 3.1，H(Y ∣ X=x)=0 当且仅当在给定 X=x 的条件下 Y 是一个确定性常数。这意味着 Y 是由 X 决定的；也就是说，存在一个从 X(Ω) 到 Y(Ω) 的映射 f。

为了证明第二个论断，我们从 H(X,f(X)) 开始，并用熵的链式法则（定理 3.1）以两种方式展开它：

H(X,f(X))=H(X)+H(f(X) ∣ X) (3.31) =H(f(X))+H(X ∣ f(X))。(3.32)

因为我们刚刚证明了 H(f(X) ∣ X)=0，通过注意到 H(X ∣ f(X))≥0，我们立即得到 H(f(X))≤H(X)。 当等号成立时，我们需要 H(X ∣ f(X))=0，这导致了 f 是一个双射的要求。

例 3.1 在一个保密系统中，有三方：一个发送方，一个预期的接收方和一个窃听者，如图 3.1 所示。发送方和预期的接收方事先商定一个密钥，这个密钥对窃听者是未知的。 发送方使用密钥对他想要与预期接收方分享的明文进行加密。 预期的接收方在收到加密的码字后，使用密钥解密明文。 窃听者观察加密的码字并试图在没有密钥的情况下解密。 设明文为随机变量 X，密钥为与 X 无关的随机变量 Z。有一个映射 f

--- 第 23 页 ---

图 3.1: 一个保密系统。图中显示了发送方（Sender）、接收方（Receiver）和窃听者（Eavesdropper）。发送方使用明文 X 和密钥 Z 通过函数 f 生成密文 Y=f(X,Z)。接收方使用密文 Y 和密钥 Z 通过函数 g 恢复明文 X=g(Y,Z)。窃听者只能观察到密文 Y。

将 X 加密为加密码字 Y，即 Y=f(X,Z)，还有另一个映射 g 将 Y 解密为明文 X，即 X=g(Y,Z)。一个具体的例子是所谓的“一次性密码本”：X 和 Z 都是二进制字符串，Y 是 X 和 Z 的模2和，Y=X⊕Z。预期的接收方只需取 Y 和 Z 的模2和即可重现 X，X=Y⊕Z。

一个**完美保密**系统要求窃听者无法做得比纯粹猜测更好，这个要求归结为 I(X;Y)=0 的条件，即 鉴于推论 3.3，X 和 Y 是独立的。 对于这个设置，因为 I(X;Y)=0, 我们有

H(X)=(a)H(X ∣ Y) ≤H(X,Z ∣ Y) =(c)H(Z ∣ Y)+H(X ∣ Y,Z) ≤(d)H(Z ∣ Y) ≤(e)H(Z), (3.33)

其中，(a) 是由于 I(X;Y)=H(X)−H(X ∣ Y) (定理 3.2)，(b) 是通过熵的链式法则 (定理 3.1) 展开 H(X,Z ∣ Y)=H(X ∣ Y)+H(Z ∣ X,Y) 并使用熵的非负性 (推论 3.1)，(c) 也是通过熵的链式法则，(d) 是通过将推论 3.6 应用于 X=g(Y,Z)，(e) 是由于“条件作用降低熵”的性质 (定理 3.6)。

在这一点上，我们只需启发式地将关系 H(X)≤H(Z) 解释为这样一个事实：对于一个完美的保密系统，密钥中的信息量必须不小于明文中的信息量。

在本节的最后，我们讨论互信息的凸性和凹性，并基于我们到目前为止发展的基本性质提供它们的证明。 回顾我们对互信息 I(X;Y) 的定义是关于一对随机变量 X 和 Y 的，由它们的联合概率分布 PX,Y=PXPY ∣ X 来表征，这是有帮助的。

图 3.1: 一个保密系统。

--- 第 24 页 ---

定理 3.7 互信息 I(X;Y) 对于任何固定的 PY ∣ X 来说是关于 PX 的凹函数，对于任何固定的 PX 来说是关于 PY ∣ X 的凸函数。

证明：

(a) I(X;Y) 关于 PX 的凹性的含义如下：取 X 上的两个任意概率分布 PX(i)，i∈{0,1}，并将联合概率分布 PX(i)PY∣X 的互信息表示为 I(i)(X;Y)。 对于任意 λ∈[0,1], 定义 PX(λ)=(1−λ)PX(0)+λPX(1)，并将联合概率分布 PX(λ)PY∣X 的互信息表示为 I(λ)(X;Y)。那么 I(X;Y) 的凹性意味着

(1−λ)I(0)(X;Y)+λI(1)(X;Y)≤I(λ)(X;Y). (3.34)

为了证明 (3.34)，我们构建一个马尔可夫链如下。 设 Q 是一个伯努利随机变量，以概率 1−λ 取 0，以概率 λ 取 1，X 在给定 Q 的条件下具有概率分布 PX(Q)，Y 在给定 X 的条件下具有条件概率分布 PY∣X。因此 Q↔X↔Y 构成一个马尔可夫链，且 (X, Y) 的联合概率分布恰好是 PX(λ)PY∣X。 现在，用互信息的链式法则以两种方式展开互信息 I(X,Q;Y)：

I(X,Q;Y)=I(X;Y)+I(Q;Y∣X) (3.35) =I(Q;Y)+I(X;Y∣Q) (3.36)

根据我们之前的约定，I(X;Y)=I(λ)(X;Y); 将推论 3.4 应用于马尔可夫链 Q↔X↔Y, I(Q;Y∣X)=0; 并且根据条件互信息的定义 (定义 2.15)，我们可以按如下方式评估 I(X;Y∣Q)

I(X;Y∣Q)=PQ(0)I(X;Y∣Q=0)+PQ(1)I(X;Y∣Q=1) =(1−λ)I(0)(X;Y)+λI(1)(X;Y). (3.37)

因此，将 (3.35) 和 (3.36) 放在一起，可以得到

I(λ)(X;Y)≥(1−λ)I(0)(X;Y)+λI(1)(X;Y), (3.38)

这正是 (3.34)。

(b) 与 (a) 类似，I(X;Y) 关于 PY∣X 的凸性的含义如下：取 Y 在给定 X 条件下的两个任意条件概率分布 PY∣X(i),i∈{0,1}, 并将联合概率分布 PXPY∣X(i) 的互信息表示为 I(i)(X;Y)。 对于任意 λ∈[0,1], 定义

PY∣X(λ)=(1−λ)PY∣X(0)+λPY∣X(1)，并将联合概率分布 PXPY∣X(λ) 的互信息表示为 I(λ)(X;Y)。那么 I(X;Y) 的凸性意味着

(1−λ)I(0)(X;Y)+λI(1)(X;Y)≥I(λ)(X;Y). (3.39)

为了证明 (3.39)，我们引入一个伯努利随机变量 Q，以概率 1−λ 取 0，以概率 λ 取 1，且与 X 无关。设 Y 在给定 Q 的条件下具有条件概率 PY∣X(Q)。显然，(X, Y) 的联合概率分布恰好是 PXPY∣X(λ)。 现在，用链式法则以两种方式展开互信息 I(X;Q,Y)：

I(X;Q,Y)=I(X;Y)+I(X;Q∣Y) (3.40) =I(X;Q)+I(X;Y∣Q). (3.41)

根据定义，I(X;Y)=I(λ)(X;Y); 由于 Q 和 X 之间的独立性，I(X;Q)=0（见推论 3.3）； 我们可以根据以下方式评估 I(X;Y∣Q)：

I(X;Y∣Q)=PQ(0)I(X;Y∣Q=0)+PQ(1)I(X;Y∣Q=1) =(1−λ)I(0)(X;Y)+λI(1)(X;Y). (3.42)

因此，通过将 (3.40) 和 (3.41) 放在一起，可以得到

I(λ)(X;Y)≤(1−λ)I(0)(X;Y)+λI(1)(X;Y), (3.43)

这正是 (3.39)。

3.4 费诺不等式

对于具有联合概率分布 PX,Y 的一对随机变量 X 和 Y，如果我们只能观察到 Y 并希望确定 X 的值，我们能做得多好？ 从数学上讲，我们的决策，记为 X^，是根据某个条件概率分布 PX^∣Y 由 Y 导出的一个随机变量。 因此存在一个马尔可夫链 X↔Y↔X^。

为了继续，我们需要具体说明如何评估一个决策的质量。 让我们使用 X^ 不等于 X 的概率作为性能度量，这通常被称为**错误概率**，Pe=P(X^\=X)。

最小化 Pe 的决策由以下定理给出。

3.4 费诺不等式 31

--- 第 26 页 ---

32

3 信息量：性质

定理 3.8 以下**最大后验概率 (MAP)** 决策对于观察 Y=y∈Y(Ω) 最小化了 Pe=P(X^\=X)：

X^=arg maxx∈X(Ω)PX∣Y(x∣y). (3.44)

证明： 我们将 Pe=P(X^\=X) 展开如下： P(X^\=X) =∑x∈X(Ω)PX(x)P(X^\=x∣X=x)
=∑x∈X(Ω)PX(x)∑y∈Y(Ω)P(X^\=x,Y=y∣X=x) =∑x∈X(Ω)PX(x)∑y∈Y(Ω)P(X^\=x∣Y=y,X=x)PY∣X(y∣x)
=∑x∈X(Ω)PX(x)∑y∈Y(Ω)[1−P(X^=x∣Y=y,X=x)]PY∣X(y∣x) =(a)∑x∈X(Ω)PX(x)∑y∈Y(Ω)
[1−PX^∣Y(x∣y)]PY∣X(y∣x) =∑(x,y)∈X(Ω)×Y(Ω)PX(x)PY∣X(y∣x)[1−PX^∣Y(x∣y)]
=1−∑(x,y)∈X(Ω)×Y(Ω)PX,Y(x,y)PX^∣Y(x∣y), (3.45)

其中 (a) 是由于马尔可夫链 X↔Y↔X~.

所以我们转而最大化 Σ(x,y)∈X(Ω)×Y(Ω)PX,Y(x,y)PX^∣Y(x∣y), 对此我们有

∑(x,y)∈X(Ω)×Y(Ω)PX,Y(x,y)PX^∣Y(x∣y) =∑y∈Y(Ω)PY(y)∑x∈X(Ω)PX∣Y(x∣y)PX^∣Y(x∣y). (3.46)

因此，对于每个 y∈Y(Ω)，我们需要最大化内层求和，即 Σx∈X(Ω)PX∣Y(x∣y)PX^∣Y(x∣y)。 因为这是 ∣X(Ω)∣ 个非负项的和，在约束

∑x∈X(Ω)PX^∣Y(x∣y)=1, (3.47)

下，最优解显然是让对于能达到最大 PX∣Y(x∣y) 的 x∈X(Ω) 的 PX^∣Y(x∣y)=1，并让其余的 PX^∣Y(x∣y)=0。当有多个 x∈X(Ω) 最大化 PX∣Y(x∣y) 时，我们可以任意选择其中一个。 这正是 MAP 决策 (3.44)。

使用贝叶斯法则，MAP 决策 (3.44) 可以重写为

--- 第 27 页 ---

为

X^=arg maxx∈X(Ω)PX(x)PY∣X(y∣x). (3.48)

注意，分母 PY(y) 在 (3.48) 中已被移除，因为最大化是针对 X(Ω) 的，并且不依赖于 PY(y) 的值。 此外，如果 X 在 X(Ω) 上是均匀分布的，决策就变成了 X^=arg maxx∈X(Ω)PY∣X(y∣x)，这被称为**最大似然 (ML) 决策**。

尽管我们从一个相当普遍的要求开始，即 X^ 是由 Y 导出的满足 X↔Y↔X^ 的随机变量，但最终的 MAP 决策 (3.44) 是一个基于 Y 的确定性规则。当考虑除 Pe 以外的其他标准时，这可能不成立，并且可能需要随机化决策。 MAP 决策对于许多统计推断问题是基础的。 然而，如果问题的维度变大，其性能通常难以获得封闭形式，甚至难以数值计算。 此外，肯定存在其他类型的决策可能因某些原因而被使用，因此我们也想评估它们的错误概率。 **费诺不等式**，建立在熵的基本性质之上，为任何（不一定是 MAP）决策 X^ 的 Pe 提供了一个下界。

定理 3.9 对于任何决策 X^ 使得 X↔Y↔X^，Pe=P(X^\=X) 满足

H(X∣Y)≤H(X∣X^)≤h2(Pe)+Pelog(∣X(Ω)∣−1). (3.49)

证明： 第一个不等式 H(X∣Y)≤H(X∣X^) 直接由 DPI 得出，随后我们证明第二个不等式。 定义一个指示随机变量 E 如下：如果 X^\=X，则 E=1，否则 E=0。 很明显，E 是一个以概率 Pe 取值为 1 的伯努利随机变量。 应用熵的链式法则（定理 3.1），我们得到

H(X,E ∣ X^)=H(X ∣ X^)+H(E ∣ X,X^) (3.50) =H(E ∣ X^)+H(X ∣ X^,E) (3.51)

因为 E 是由 (X,X^) 决定的，根据推论 3.6，H(E ∣ X,X^)=0； 因为条件作用会降低熵（定理 3.6），
H(E ∣ X^)≤H(E)=h2(Pe)。对于 H(X ∣ X^,E)，我们可以将其展开为

H(X ∣ X^,E)=PE(0)H(X ∣ X^,E=0)+PE(1)H(X ∣ X^,E=1) ≤(1−Pe)·0+Pe·log(∣ X(Ω) ∣ −1) (3.52)

其中

3.4 费诺不等式 33

--- 第 28 页 ---

34

3 信息量：性质

► 当 E=0 时，H(X ∣ X^,E=0)=0，因为此时 X^=X 成立； ► 当 E=1 时，H(X ∣ X^,E=1)≤log(∣ X(Ω) ∣ −1)，因为此时 X 不能等于 X^，因此只能在集合 X(Ω){X^} 中。

所以我们从 (3.50) 和 (3.51) 得到，

H(X ∣ X^)+H(E ∣ X,X^)=H(E ∣ X^)+H(X ∣ X^,E) H(X ∣ X^)+0≤h2(Pe)+(1−Pe)·0 +Pe·log(∣ X(Ω) ∣ −1)
H(X ∣ X^)≤h2(Pe)+Pelog(∣ X(Ω) ∣ −1)

这便完成了证明。 (3.53)

由于费诺不等式对任何决策 X^ 都成立，它将被发现对于证明不可能性结果非常有用，即错误概率无论使用何种决策都不能低于某个水平。我们将在第 6 讲中用它来证明香农信道编码基本定理的逆定理。检查 (3.49)，我们看到当 ∣ X(Ω) ∣ <∞, 如果 Pe=0，即无差错决策，那么根据推论 3.6，必须有 H(X ∣ Y)=0, 即 X 在给定 Y 的情况下是确定性的。

3.5 平稳过程的熵率

在第 2 讲中，我们已经定义了随机过程的熵率。 在本讲中，我们研究其性质。

让我们先看一些例子。

例 3.2 考虑一个随机过程 X:X1,X2,... 其中所有元素都是 i.i.d. 随机变量。 这样的随机过程被称为“无记忆的”，因为对于任何 i，Xi 不依赖于其“历史”{Xi−1,...,X1}。 那么，因为

H(X1,X2,...,Xn)=(a)∑i=1nH(Xi) =nH(X1) (3.54)

其中 (a) 是由于推论 3.5，我们立即得到 X 的熵率为

H(X)=limn→∞n1H(X1,X2,...,Xn) =limn→∞n1nH(X1) =H(X1). (3.55)

--- 第 29 页 ---

3.5 平稳过程的熵率 35

例 3.3 考虑一个随机过程 X: X1,X2,... 其中所有元素都是相互独立的随机变量，且对于所有 i，Xi 服从参数为 2−i 的几何分布。

那么，我们有

n1H(X1,X2,...,Xn)=(a)n1∑i=1nH(Xi) =(b)n1∑i=1n2−ih2(2−i) >n1∑i=1n2i2−ilog2i =nlog2∑i=1ni =2log2(n+1) →∞

当 n→∞ 时，其中 (a) 是由于推论 3.5，(b) 来自第 2 讲的例 2.3，(c) 是通过展开 h2(ϵ) 并只保留项 ϵlogϵ1 得到的。所以这个随机过程的熵率，即当 n→∞ 时 n1H(X1,X2,...,Xn) 的极限，不存在。

例 3.4 考虑一个随机过程 X:X1,X2,... 生成如下：使用一个参数为 1/2 的伯努利随机变量 Z 作为开关，当 Z=1 时，X1,X2,... 是 i.i.d. 随机变量，每个都服从参数为 1/2 的伯努利分布，当 Z=0 时，X1,X2,... 是常数零。 对于这个随机过程，我们有

PX1,...,Xn(x1,...,xn) =21PX1,...,Xn ∣ Z(x1,...,xn ∣ Z=1)+21PX1,...,Xn ∣ Z(x1,...,xn ∣ Z=0) =2−(n+1)+211{(x1,...,xn)=(0,...,0)}. (3.57)

所以我们可以通过一些计算得到，

H(X1,...,Xn)=1+21−2−nn−21+2−nlog2(1+2−n) (3.58)

单位为比特，因此

H(X)=limn→∞n1H(X1,...,Xn) =21 (比特)。 (3.59)

这个结果可以直观地理解如下：随机过程 X 是两个分量随机过程的混合，一个

是无记忆伯努利过程，熵率为1比特，另一个是确定性过程，熵率为零，所以总的熵率就是这两个分量随机过程熵率的平均值。

随后，我们关注平稳随机过程。

定义 3.1 一个离散时间随机过程 X 是**平稳**的，如果 X 的任何子集的联合概率分布对于时间平移是不变的； 也就是说，

P(Xi1=x1,Xi2=x2,...,Xin=xn) =P(Xi1+l=x1,Xi2+l=x2,...,Xin+l=xn) , (3.60)

对于任何 n，任何下标索引 i1,i2,...,in, 任何时间平移 l，以及任何 x1,x2,...,xn∈X(Ω) 的集合。

对于一个平稳随机过程 X，其熵率由以下定理给出。

定理 3.10 对于一个满足 H(X1)<∞ 的平稳随机过程 X，其熵率存在，并且满足

H(X)=limn→∞H(Xn ∣ Xn−1,...,X1) 。 (3.61)

证明： 让我们从对 H(X1,...,Xn) 应用熵的链式法则（定理 3.1）开始：

H(X1,...,Xn)=∑i=1nH(Xi ∣ Xi−1,...,X1) 。 (3.62)

求和中的各项构成一个单调非增序列，因为

H(Xn+1 ∣ Xn,...,X1) ≤(a)H(Xn+1 ∣ Xn,...,X2) =(b)H(Xn ∣ Xn−1,...,X1), (3.63)

其中 (a) 是由于条件作用降低熵的性质（定理 3.6），(b) 是由于 X 的平稳性。所以序列 {H(Xi ∣ Xi−1,...,X1)}i=1∞ 是非负单调非增的，因此存在一个极限，可以表示为

H'(X)=limn→∞H(Xn ∣ Xn−1,...,X1). (3.64)

因此，对于任何 ϵ>0, 存在一个整数 Nϵ 使得对于任何 i>Nϵ', ∣ H(Xi ∣ Xi−1,...,X1)−H'(X) ∣ <ϵ。

现在回到 (3.62)，让我们考察差距

∣ n1H(X1,...,Xn)−H'(X) ∣

如下：

∣n1H(X1,...,Xn)−H'(X)∣ =(a)∣n1∑i=1nH(Xi∣Xi−1,...,X1)−H'(X)∣ ≤(b)n1∑i=1n∣H(Xi∣Xi−1,...,X1)−H'(X)∣ =n1∑i=1Nϵ∣H(Xi∣Xi−1,...,X1)−H'(X)∣ +n1∑i=Nϵ+1n∣H(Xi∣Xi−1,...,X1)−H'(X)∣ <n1∑i=1Nϵ∣H(Xi∣Xi−1,...,X1)−H'(X)∣+n1(n−Nϵ)ϵ <n1∑i=1Nϵ∣H(Xi∣Xi−1,...,X1)−H'(X)∣+ϵ, (3.65)

其中 (a) 是通过熵的链式法则（定理 3.1），(b) 是绝对值的三角不等式。 (3.65) 的右侧对于所有足够大的 n 都可以被 2ε 上界。 因此总结来说，极限 limn→∞n1H(X1,...,Xn) 存在且等于 H'(X)。

根据定理 3.10，平稳随机过程的熵率是“当前”状态在“所有过去”状态条件下的条件熵。

另一种特殊类型的随机过程是马尔可夫链。一个随机过程 X:X1,X2,... 是一个马尔可夫链，如果

(X1,...,Xn−1)↔Xn↔Xn+1 (3.66)

对任何 n 成立。 如果 PXn+1∣Xn 不依赖于 n，则该马尔可夫链是时不变的。 因此，一个时不变马尔可夫链由条件概率分布 PX2∣X1(b∣a),∀a,b∈X1(Ω) 来描述。

对于一个时不变马尔可夫链，如果存在一个在 X1 上的概率分布 PX1 使得对于所有 b∈X1(Ω)，

PX1(b)=∑a∈X1(Ω)PX1(a)PX2∣X1(b∣a), (3.67)

成立，那么得到的马尔可夫链也是平稳的，且 PX1

被称为马尔可夫链的**平稳分布**。

如果一个马尔可夫链既是平稳的又是时不变的，其熵率由定理 3.10 的以下推论给出。

推论 3.7 一个平稳时不变马尔可夫链 X 的熵率满足

H(X)=H(X2∣X1). (3.68)

证明：根据定理 3.10，

H(X)=limn→∞H(Xn∣Xn−1,...,X1), (3.69)

因此我们应用马尔可夫链关系 (X1,...,Xn−2)↔Xn−1↔Xn 来得到

H(X)=limn→∞H(Xn∣Xn−1,...,X1) =limn→∞H(Xn∣Xn−1)。 (3.70)

因为 X 也是平稳的，我们有

H(X)=H(X2∣X1), (3.71)

从而完成了证明。

例 3.5 考虑一个时不变的两状态马尔可夫链，其 X1(Ω)={0,1}，转移概率分布由 PX2∣X1(0∣0)=1−α, PX2∣X1(1∣0)=α. PX2∣X1(0∣1)=β 和 PX2∣X1(1∣1)=1−β 给出，如图 3.2 所示。 平稳分布 PX1 可以通过 (3.67) 解出，为

PX1(0)=PX1(0)PX2∣X1(0∣0)+PX1(1)PX2∣X1(0∣1) =(1−α)PX1(0)+βPX1(1) ,
PX1(1)=PX1(0)PX2∣X1(1∣0)+PX1(1)PX2∣X1(1∣1) =αPX1(0)+(1−β)PX1(1), PX1(0)+PX1(1)=1, (3.72)

并且由 PX1(0)=α+ββ 和 PX1(1)=α+βα 给出。所以根据推论 3.7，这个随机过程的熵率为

H(X)=H(X2 ∣ X1)=α+ββh2(α)+α+βαh2(β). (3.73)

--- 第 33 页 ---

图 3.2：一个时不变的两状态马尔可夫链。状态0转移到状态1的概率是α，停留在状态0的概率是1-α。状态1转移到状态0的概率是β，停留在状态1的概率是1-β。

图 3.2：一个时不变的两状态马尔可夫链。

注释

本讲中的大多数性质都可以在信息论的标准教科书中找到。 相对熵的非负性是其他非负性质的基础； 在许多教科书中，性质的阐述遵循一种不同的方法，从凸函数的詹森不等式开始。

费诺不等式归功于罗伯特·费诺，他是信息论形成的先驱，他在 1950 年代初开发了第一门课程，并在麻省理工学院撰写了关于该主题的最早的综合性教科书之一 [10]。

关于信息量的更深入和系统的处理，请参考 [8, 第 6 章]，其中发展了信息量和集合操作之间的一一对应关系。 本讲中遇到的不等式属于所谓的香农型不等式，但也存在非香农型不等式； 见 [8, 第 13 和 14 章]。

例 3.1 中的完美保密系统由香农在其基础性文章《保密系统的通信理论》[11] 中处理。 一个完美保密系统要求密钥的熵不小于明文的熵这一结论，在一定程度上阻碍了信息论密码学的发展，直到 1970 年代末； 见 [12]。

历史上，ML 决策是在将 X 视为确定性参数，而不考虑 X 的任何概率结构的情况下提出的。当 X 是一个在 X(Ω) 上具有均匀概率分布的随机变量时，得到的 MAP 决策与 ML 决策的形式相同。

练习

1. 对于随机变量 X 和 Y，证明 H(X+Y)≤H(X)+H(Y) 成立。
2. 对于随机变量 X1,X2,...,Xn,Y1,Y2,...,Yn，何时

H(X1,X2,...,Xn ∣ Y1,Y2,...,Yn) =H(X1 ∣ Y1)+H(X2 ∣ Y2)+...+H(Xn ∣ Yn)

--- 第 34 页 ---

成立?

1. 我们从定理 3.6 知道条件作用会降低熵。 对于互信息 I(X;Y) 和条件互信息 I(X;Y ∣ Z)，是否存在类似的性质?
2. 使用相对散度的非负性，证明对数和不等式：对于非负数 {ai}i=1,...,n 和 {bi}i=1,...,n，

∑i=1nailogbiai≥a logba

其中 a=∑i=1nai 和 b=Σi=1nbi, 等号成立当且仅当存在 c 使得对于所有 i 都有 ai=cbi。

1. 在这个练习中，我们将推论 3.2 应用于詹姆斯·L·梅西提出的一个猜测问题 [13]。 假设我们想猜测一个随机变量 X 的值，其值域为 X(Ω)={1,2,...}。平均而言，我们需要猜测多少次? 不失一般性，我们总可以重新标记随机变量，使得 PX(1)≥PX(2)≥... 成立。 证明平均而言，我们需要猜测的次数不少于 eH(X)−1 次，其中熵的单位是奈特。
2. 考虑一个随机变量 X，其值域为 X(Ω)={1,2,...}。

   a) 证明如果 EX 是有限的，那么 H(X) 也是有限的。 b) 证明如果 E log X 是有限的，那么 H(X) 也是有限的。 c) 证明如果 H(X) 是有限的且 PX(x) 随 x 单调非增，那么 Elog X 是有限的。 d) 给出一个例子来说明前述陈述中 PX(x) 的单调非增条件是必要的。

3. 考虑一个在 {0,1,...,m−1} 上均匀分布的随机变量 X，其观测值 Y 从 {(X−1)(modm),X,(X+1)(modm)} 中均匀抽取。定义 Pe=P(Y\=X)。

   a) 使用费诺不等式给出 Pe 的一个下界。 b) 找到该下界与 MAP 决策的精确 Pe 值之间的差距。 c) 你能通过检查费诺不等式的证明并加以改进来解决这个差距吗?

4. 构造一个在费诺不等式中等号成立的例子。
5. 如果估计 X^ 是 X(Ω) 的一个大小为 L 的子集，并将错误事件定义为 {X∈/X^}，建立费诺不等式的一个扩展。

--- 第 35 页 ---

3.5 平稳过程的熵率 41

1. 证明 Csiszár 恒等式：

∑i=1nI(Xi+1,...,Xn;Yi ∣ Y1,...,Yi−1) =∑i=1nI(Y1,...,Yi−1;Xi ∣ Xi+1,...,Xn),

其中 Xn+1 和 Y0 被理解为退化的。

1. 在这个练习中，我们提供一个信息论证明，证明著名的数论结果——存在无穷多个素数。 为此，考虑一个任意整数 n，并用 π(n) 表示不大于 n 的素数个数。 取一个在 {1,2,...,n} 上均匀分布的随机变量 N，并将其写成其唯一的素数分解形式，N=p1X1p2X2...pπ(n)Xπ(n)，其中 {p1,p2,...,pπ(n)} 是不大于 n 的素数，每个 Xi 是使得 pik 整除 N 的最大幂 k≥0。通过考察 H(N)，证明当 n→∞ 时 π(n)→∞。进一步阅读请参考 [14]。
2. 对于整数集 [n]:={1,2,...,n}，以概率 p 独立地抽取其每个元素，会得到一个 [n] 的随机子集。 对于两个这样独立生成的子集 A 和 B，计算 H(A) 和 H(A∪B)，并证明当 p≤23−5  img 时 H(A∪B)>H(A)。 这与所谓的并集封闭集猜想有关，该猜想的第一个常数下界是使用信息论论证建立的； 进一步阅读请参考 [15]。
3. 考虑一个如下生成的随机变量 X：以一个取值于 {1,2,...} 的随机变量 Z 为条件，让 X 是一个参数为 2−Z 的几何随机变量（见例 2.3）。 a) 证明如果 E[Z]= ∞ 则 H(X)=∞。 b) 定义一个随机变量 Y 如下：Y 以概率 1−ϵ 为 0，以概率 ϵ 为 X。 设 Y^=0 的概率为 1。证明如果 H(X)=∞，则无论决策错误概率 Pe=P(Y\=Y^)=ϵ>0 有多小，H(Y ∣ Y^) 都不趋于零。 这个例子说明了当字母表是无限时应用费诺不等式的微妙之处（[8, 例 2.49]）。
4. 证明熵的次模性：对于任何两个随机变量集 S1 和 S2，H(S1∪S2)+H(S1∩S2)≤H(S1)+H(S2)。
5. 对于随机变量 X 和 Y 以及一个映射 f，在什么条件下 H(X ∣ f(Y))=H(X ∣ Y) 成立?
6. 假设 Θ∈(0,1) 是单位区间上的一个随机变量，

--- 第 36 页 ---

42

3 信息量：性质

区间，并以此为条件，X=(X1,...,Xn) 由 n 个 i.i.d. 随机变量 Xi∼Bernoulli(Θ) 组成。 定义 T=Σi=1nXi。T 是 Θ 的充分统计量吗?

1. 对于例 3.5 中的两状态马尔可夫链，如果我们对其进行欠采样以获得一个新的随机过程 X1,X3,X5,...，它还是一个马尔可夫链吗? 在平稳条件下，评估其熵率并与原始马尔可夫链 X1,X2,X3,... 的熵率进行比较。
2. 为三个随机变量 (X,Y,Z) 定义一个“近似马尔可夫”关系，如果它们满足

p(z ∣ x,y)=p(z ∣ y)(1+ϵ(x,y,z)),

其中对于任何 (x,y,z) 元组，∣ϵ(x,y,z)∣≤δ。 证明对于这样的“近似马尔可夫”关系，我们有以下“δ2-近似 DPI”成立：

I(X;Z)≤I(X;Y)+δ2;

1. 对于随机变量 V, W1,W2,...,Wn, 证明

$$H(V) \geq \sum_{i=1}^{n} I(V;W_i),$$

当 $W1,W2,\ldots,Wn$ 相互独立时。