--- PAGE 1 ---

信源表示：无损压缩 5

在上一讲中，我们已经描述了信源表示中率与失真之间的基本权衡，这由率失真函数给出。 在许多应用中，对信源消息进行完美再现是可取的，甚至是强制性的，即无损压缩。 当期望失真为零时，率失真函数特化为渐进无损压缩。 此外，如果要求精确无损压缩，则可变长度编码变得必要。 有趣的是，渐进无损压缩和精确无损压缩共享相同的最终基本性能极限，该极限由信源的熵给出。 本讲将探讨这些主题。

5.1 零期望失真下的率失真函数

回想一下，在第 4 讲中处理的有损情况，在失真度量 d 下，离散无记忆信源 S 的率失真函数由香农的信源编码基本定理（定理 4.1）表征为

R(D)=minPS^ I SI(S;S^)

(5.1)

服从约束条件 E[d(S,S^)]≤D

(5.2)

现在，假设 S=δ^ 并且失真度量是汉明距离。 让我们关注 R(0)。

当 D=0 时，由于失真度量是汉明距离，约束条件 E[d(S,S^)]≤D=0 变为 P(S\=S^)=0。因此，在求解 R(0) 时，我们有 S^=S 的概率为 1。 这直接导致

(5.3)

I(S;S)=I(S;S)=H(S)。

因此我们得出结论，R(0)=H(S)，即离散无记忆信源 S 的熵。

备注 5.1 注意，在第 4 讲第一节的问题表述下，D=0 意味着

limn→∞E[d(S,S)]=0

(5.4)

--- PAGE 2 ---

也就是说，对于任何 ϵ>0, 对于所有足够大的 n,

n1∑i=1nP(Si\=S^i)≤ϵ . (5.5)

这仅意味着随着信源消息长度无限增长，成功再现的信源符号的比例渐进地接近 100%。 但这不足以让我们断言精确无损压缩，即

P(S=S)=1,

(5.6)

在有损信源表示问题表述下是可能的。 记住这一细微差别非常重要。

通过重新审视定理 4.1 的可达性部分的证明，我们可以获得一些更深入的见解。 其中，码本 C 的元素是服从概率分布 PS^ 的独立同分布 (i.i.d.) 随机变量。 现在，由于 S^=S, 我们有 PS^=PS。所以 C 简单地由 S 的独立同分布样本构成。

让我们检查 D=0 时的编码过程。由于失真度量是汉明距离且在 (4.39) 中 S=S^，我们有

d(s,C(ϖ))=n1∑i=1nd(si,Ci(w)) =n1∑i=1n1{si\=Ci(w)},

(5.7)

即 S 和 C(w) 不一致的位置的比例，并且准则 (4.39) 要求这个比例很小，具体来说，不大于 e。 另一方面，在 (4.40) 中，由于 S^=S，s 和 C(w) 之间的信息密度为

i(g;C(w))=n1∑i=1nlogPS(si)PS(Ci(w))PS,S(si,Ci(w)) =n1∑i=1nlogPS(si)1,

(5.8)

如果 s=C(w)，否则为无穷大。 这是因为即使只有一个位置 si\=Ci(:v) 也会导致 PS,S(si,Ci(w))=0. 所以准则 (4.40) 要求 s=C(w) 成立，并且 n1∑i=1nlogPS(si)1 位于 [H(S)−ϵ,H(S)+ϵ] 区间内。

在这里，很明显 (4.40) 支配了 (4.39)。 因此，给定一个信源消息 S, 编码过程在码本 C 中搜索一个与 s 完全相同的码字，如果 s 在 C 中未找到，则声明编码失败。第 4.4 节的证明确保了，通过选择

--- PAGE 3 ---

5.2 用于精确无损压缩的可变长度编码 67

任何速率 R>R(0)=H(S)，对于任何 ϵ>0，当 n→∞ 时，编码失败发生的概率趋于零。 如何直观地理解这个结果？当然，使用速率 R<log丨δ丨 的码本不可能详尽地包含所有可能的信源消息，但是通过根据 PS 随机生成码本，可以确保那些未包含在码本中的信源消息出现的概率任意小，只要信源消息的长度足够大。 5.2 用于精确无损压缩的可变长度编码

如前一节所述，R(0)=H(S) 意味着对于任何大于 H(S) 的速率，渐进无损压缩是可能的。但如果要求精确无损压缩，即 S=S 以概率 1 成立，我们需要一种不同的方法。 基本思想是利用 S 的概率分布通常不均匀这一事实。 因此，可变长度编码就派上了用场。回想一下第 4.1 节中的问题表述，其中信源消息 S 通过编码映射被编码为索引 W，然后 W 通过解码映射被解码为再现的消息 S^。对于可变长度编码，我们将 W 视为一个二进制或 q-元字符串，并允许其长度对不同的信源消息而变化。 直观上，通过为概率较高（或较低）的信源消息分配较短（或较长）的索引字符串，可以提高编码的整体效率，该效率由期望索引字符串长度来衡量。 考虑一个离散无记忆信源 S，不失一般性，将其字母表 S={a1,a2,...} 排列，使得概率不增加，即 PS(a1)≥PS(a2)≥... 我们将 S 编码为索引 W=f(S)，当表示为二进制字符串时，它取自

∀∀∗={∅,0,1,00,01,10,11,000,001,...},

(5.9)

即所有有限长度二进制字符串的集合，包括空字符串 0。在本节中，我们只考虑二进制情况，扩展到 q-元情况留作练习。 为了最小化期望索引长度，可以说编码规则是总是为更频繁出现的信源消息分配更短的索引字符串。 也就是说，f(a1)=∅, f(a2)=0, f(a3)=1,f(a4)=00, f(a5)=01,... 我们可以很容易地验证，对于信源消息 ai，其索引字符串的长度为 l(ai)=⌊log2i⌋。 让我们分析期望索引字符串长度，l=∑s∈SPS(s)l(s)。

--- PAGE 4 ---

68

5 信源表示：无损压缩

关于 l 的上界，注意到对于 ai, PS(ai)≤1/i。 所以我们有

l(ai)=⌊log2i⌋≤log2i≤−log2PS(ai),

并且

l=∑s∈SPS(s)l(s)

≤−∑s∈SPS(s)log2PS(s)=H(S) (以比特为单位)。 (5.10)

(5.11)

关于 l 的下界，我们按如下方式进行

H(S)H(S)+H(l(S)|S)

H(S, l(S))

H(S|l(S)) +H(l(S)),

(5.12)

其中 (a) 是由于推论 3.6，因为在给定 S 的条件下 l(S) 是确定性的，而 (b) 和 (c) 都来自熵的链式法则，即定理 3.1。 对于 H(S丨l(S))，注意到对于每个 l(S)=l, S 最多有 2l 种不同的可能性，否则至少有两个信源消息将被分配给相同的索引字符串，编码将不是精确无损的。 所以

H(S丨l(S))=∑l=0∞P(l(S)=l)H(S丨l(S)=l) ≤∑l=0∞P(l(S)=l)l =E[l(S)]=l (以比特为单位)。

对于 H(l(S)), 我们有

H(l(S))≤(b)(a)log1→∞(l++λλλλ≤(l+1)log2(l+1)−llog2l

log2[e(+1)]

(5.13)

(c)

Vl

log2[e(H(S) +1)] (以比特为单位),

(5.14)

其中 (a) 是通过使用几何分布在给定均值下使熵最大化的性质，即推论 3.2，(b) 来自于 (x+1)ln(x+1)−x ln x<ln(x+1)+1 对于 x>0, 并且 (c) 来自于我们刚刚在 (5.11) 中获得的 l 的上界。

--- PAGE 5 ---

5.3 唯一可译性与克拉夫特不等式 69

回到 (5.12)，我们有

H(S)=H(S丨l(S))+H(l(S)) <l+log2[e(H(S)+1)];

也就是说，

(5.15)

l>H(S)−log2[e(H(S)+1)]. (5.16)

总结一下，我们对 l 有如下估计：

H(S)−log2[e(H(S)+1)]<l≤H(S) (以比特为单位)。 (5.17)

现在我们可以将这个结果应用到一个长度为 n 的信源消息 S，并相应地将期望索引字符串长度除以 n，以估计每个信源符号的期望索引字符串长度，这与第 4 讲第 4.1 节中问题表述中的速率概念一致。我们对 R=l/n 的上界 (5.11) 和下界 (5.16) 因此变为

R≤nH(S)=H(S), R>nH(S)−nlog2[e(H(S)+1)] =H(S)−nlog2[e(nH(S)+1)]

(5.18)

(5.19)

分别以比特为单位。 当 n→∞ 时，我们有 R→H(S) (假设 H(S)<∞)。所以我们得出结论，用于精确无损压缩的可变长度编码的基本性能极限是 H(S)，这与渐进无损压缩的极限相同，即上一节中获得的 R(0)。 5.3 唯一可译性与克拉夫特不等式

上一节描述的编码在信源消息大小预先规定时是可行且最优的。 然而，如果要对一个大小未预先规定的信源消息进行编码，它可能会失败。 例如，考虑对两个不同的信源消息 s=[a2,a2,a2,a3] 和 s'=[a4,a5] 进行精确无损压缩。 根据上一节的编码规则，我们发现 S 和 s' 都被编码成相同的索引字符串 0001。因此很明显，当将该编码应用于信源消息本身长度可变的“流式”场景时，存在一些模糊性，因此需要对可行的编码施加一些进一步的限制。

--- PAGE 6 ---

定义 5.1 对于一个离散无记忆信源 S，其 q-元索引 W=f(S) 是一个取自 {0,1,...,q−1}∗:=Un=1∞{0,1,...,q−1}n 的有限长度字符串。 W=f(S) 的长度表示为 l(S)。映射 f:δ↦{0,1,...,q−1}∗ 称为 S 的基本码。

注意，与上一节不同，这里空字符串被排除在码之外。 定义 5.2 给定一个离散无记忆信源 S 及其基本码 f，对于任何有限长度的信源消息 s=[s1,s2,...,sn]，我们将 f(s) 定义为基本码的逐符号级联，即，

f(s)=[f(s1),f(s2),...,f(sn)],

(5.20)

对于任何整数 n≥1 和任何 s∈Sn。 这样，映射 f 被扩展到 δ∗，即所有有限长度的信源消息，因此我们可以称之为 S 的码。

定义 5.3 如果一个离散无记忆信源 S 的码 f 在定义 5.1 的上下文中是单射的，即对于任何 s\=s'∈S，有 f(s)\=f(s')，则该码称为非奇异码。 备注 5.2 尽管在无损情况下，非奇异性似乎是自然的，但它并非完全不重要，因为对于第 4 讲中一般情况下的有损信源表示，正如其可达性证明中所说明的，编码通常不能是非奇异的。 不幸的是，当我们将码 f 从 δ 扩展到 S∗ 时，非奇异性是不够的。因此，我们需要根据定义 5.2 将非奇异性的概念扩展到 δ∗。 定义 5.4 如果一个离散无记忆信源 S 的码 f 在定义 5.2 的上下文中是单射的，即对于任何两个不同的有限长度信源消息 S 和 s'，都有 f(s)\=f(s')，则该码称为唯一可译码。 存在一些算法（例如，[23]）可以检查给定的码是否唯一可译，但一个通用的唯一可译码使用起来可能仍然不太方便。 接下来我们关注一类特殊的唯一可译码，即那些满足所谓的无前缀特性的码。定义 5.5 如果一个离散无记忆信源 S 的码对于任何 s\=s'∈δ，f(s) 都不是 f(s') 的前缀，则该码称为无前缀码。 例如，如果一个码由索引字符串 {00, 01, 001} 组成，那么它就不是无前缀的，因为 00 是 001 的前缀； 如果我们将这个码修改为 {00, 01, 101}，那么它就变成无前缀的了。

下面的引理是显而易见的。

--- PAGE 7 ---

引理 5.1 如果一个码是无前缀码，那么它就是唯一可译码。 无前缀码的一个便利特性是，当我们从头开始扫描以解码级联字符串 f(s)=[f(s1),f(s2),...,f(sn)] 时，每当识别出一个属于 {f(s):s∈δ} 的字符串，我们就可以立即解码其对应的信源符号。 因此，无前缀码也称为即时码。 示例 5.1 对于一个由 {0, 01, 11} 组成的唯一可译但非无前缀码，在解码 01101（由 [a1,a3,a2] 编码而来）时，我们只有在扫描到 011 时才能确定第一个 0 对应于 a1； 但是对于一个由 {0, 10, 11} 组成的无前缀码，在解码 01110（同样由 [a1,a3,a2] 编码而来）时，由于码中没有字符串 01，在读到第一个 0 时我们就可以立即解码出 a1。 无前缀码的另一个便利特性是，每个用于离散无记忆信源 S 且 |S| < ∞ 的无前缀码都可以通过一个有根的 q-元树来可视化， 并且其每个索引字符串都对应于树的一个唯一的叶子。 可能会有一些未使用的叶子，但由于无前缀特性，没有非叶子节点可以是索引字符串。 对于 s∈S，f(s) 的长度 l(s) 正是树中对应于 f(s) 的叶子的深度。 图 5.1 给出了一个二进制无前缀码的有根二叉树可视化示例。 根-

0 000

0

1

001

0

0

010

1

1

未使用

0

10

1

1

-未使用

显然，对于一个离散无记忆信源 S，可以有许多不同的无前缀码。它们的索引字符串长度是否存在任何基本限制？ 答案由以下定理给出。

定理 5.1 对于一个离散无记忆信源 S，其中 |S| < ∞， 它的任何无前缀码

图 5.1：一个二进制无前缀码的有根树可视化。

--- PAGE 8 ---

码必须满足克拉夫特不等式

∑s∈Sq−l(s)≤1 ; (5.21)

反之，对于任何满足 (5.21) 的 {l(s):s∈S}，都存在一个具有这些索引字符串长度的无前缀码。 证明：回想一下，无前缀码可以用一个有根的 q-元树来可视化。 树中最深的叶子对应于最长的索引字符串，我们用 lmax 表示这个深度，由于 ∣δ∣<∞ 的假设，这个深度是有限的。对于每个 s∈S，其索引字符串 f(s) 在树中的深度为 l(s)。 如果我们从叶子 f(s) 开始生长树直到达到最大深度 lmax，那么这个过程将在深度 lmax 处产生 qlmax−l(s) 个叶子。 由于无前缀特性，由任何两个不同索引字符串扩展出的所有叶子都不会重叠。 因此，我们有

∑s∈Sqlmax−l(s)≤qlmax ,

(5.22)

其中右侧项 qlmax 是当我们完全生长树时深度为 lmax 的叶子总数。 两边同除以 qlmax 即可得到克拉夫特不等式 (5.21)。 对于反证部分，考虑任何满足 (5.21) 的 {l(s):s∈δ}，并重新排列 的元素，使得 l(a1)≤l(a2)≤...≤l(a∣S∣)=lmax。现在，让我们从一个所有叶子都在深度 lmax 的有根 q-元树开始。对于 a1，我们搜索深度为 l(a1) 的第一个节点，移除其所有子节点使其成为一个叶子，并将其标记为 f(a1)；然后对于 a2，我们搜索深度为 l(a2) 的第一个节点，移除其所有子节点使其成为一个叶子，并将其标记为 f(a2)； 我们对 a3,a4,... 重复这个过程，直到 a∣δ∣。在上述过程中，当标记 f(ai) 时，我们移除了原始完全生长的树中的 qlmax−l(ai 个叶子。由于 {l(s):s∈S} 满足 (5.21)，该过程保证能够完成，并且它为 S 生成了一个无前缀码。

备注 5.3 从证明中我们可以看出，克拉夫特不等式取等号等价于相应的有根 q-元树没有未使用的叶子。

我们可能想知道，如果我们考虑不一定是无前缀码的唯一可译码，是否可以对索引字符串的长度有更宽松的限制。下面的定理排除了这种可能性。

定理 5.2 对于一个离散无记忆信源 S 且 |S| < ∞，其任何唯一可译码也必须满足克拉夫特不等式 (5.21)。

--- PAGE 9 ---

证明：让我们取(5.21)式左边的k次幂，并展开为

(∑s∈Sq−l(s))k=∑s1,s2,...,sk∈Skq−l(s1)−l(s2)−...−l(sk) =∑s∈δkq−l(s) ,

(5.23)

其中 s=[s1,s2,...,sk] 且 l(s) 正是 f(s) 的长度。 我们可以通过枚举 l(s) 来改写求和 (5.23)，其取值范围从 k 到 klmax'，如下：

$\sum_{s\in S^{k}}q^{-l(\underline{s})}=\sum_{m=k}^{kl_{mxx}}A(m)q^{-m},$

(5.24)

其中 A(m) 表示其索引字符串长度为 m 的长度为 k 的信源消息的数量。 由于该码是唯一可译的，我们有 A(m)≤qm。因此，

∑m=kklmaxA(m)q−m≤∑m=kklmaxqmq−m=k(lmax−1).

总而言之，我们有

当 k→∞

∑s∈δq−l(s)≤[k(lmax−1)]1/k→1

这就完成了证明。 (5.25)

(5.26)

备注 5.4 定理 5.2 的一个关键推论是，对于任何唯一可译码，总存在一个无前缀码，其索引字符串长度的集合完全相同。 因此，为了研究信源可能的索引字符串长度，只关注无前缀码不会失去一般性。 5.4 香农-范诺码和霍夫曼码

克拉夫特不等式为任何唯一可译码的索引字符串长度提供了一个基本约束，但它没有反映信源概率分布的影响。 现在，考虑一个离散无记忆信源 S，以及其任何 q-元唯一可译码，其长度为 {l(s):s∈S}。 让我们计算 PS 与另一个概率分布 Q 之间的相对熵，其中 Q 定义为

Q(s)=∑s'∈Sq˙−l(s')'q−l(s) s∈S,

(5.27)

--- PAGE 10 ---

以获得

D(PS ∣ ∣ Q)=∑s∈SPS(s)logqQ(s)PS(s)

=−H(S)−∑s∈SPS(s)logq∑s'∈Sq−l(s')q−l(s) =−H(S)−∑s∈SPS(s)[−l(s)−logq(∑s'∈Sq−l(s'))]
=−H(S)+∑s∈SPS(s)l(s)+logq(∑s'∈Sq−l(s'))

≤−H(S)+l,

(5.28)

其中最后一个不等式是由于克拉夫特不等式。 从相对熵的非负性（见定理 3.4），我们立即得出任何唯一可译码都应满足 l≥H(S)（以 q 为底）。 此外，当且仅当 l(s)=−logqPs(s) 对所有 s∈S 成立时，等号成立。虽然索引字符串长度 l(s) 对每个 s∈δ 都应为整数，但通常情况下 −logqPS(s) 并非对每个 s∈S 都是整数。如果对于一个离散无记忆信源恰好是这种情况，则称该信源为 q-adic，并且存在无前缀码，其期望索引字符串长度恰好达到下界 H(S)（以 q 为底）。 示例 5.2 假设一个离散无记忆信源 S 具有 PS(a1)=1/2,PS(a2)=1/4 和 PS(a3)=PS(a4)=1/8。当 q=2 时，我们有 −log2PS(a1)=1，−log2PS(a2)=2 和 −log2PS(a3)=−log2PS(a4)=3。所以 S 是 2-adic 的。 但对于任何 q>2 它都不是 q-adic 的。

当一个离散无记忆信源不是 q-adic 时，不可能将非整数值 −logqPS(s) 赋给 l(s)。 一个简单而合理的想法是将 −logqPS(s) 向上取整以获得一个整数值的索引字符串长度； 也就是说，对于一个离散无记忆信源 S，令 l(s)=⌈−logqPS(s)⌉，∀s∈δ。这里 [] 是不小于 t 的最小整数。 显然，这种分配满足克拉夫特不等式，相应的无前缀码称为香农-范诺码。 从 −logqPS(s)≤l(s)=⌈−logqPS(s)⌉<−logqPS(s)+1, 我们得到香农-范诺码的期望索引长度满足

H(S)≤l<H(S)+1,

(5.29)

其中 H(S) 是以 q 为底的。 类似于 5.2 节的最后部分，当将香农-范诺码应用于长度为 n 的源消息 S 时，速率 R=l/n 满足

H(S)≤R<H(S)+n1,

(5.30)

--- PAGE 11 ---

因此当 n→∞ 时，R→H(S)。 我们再次看到，精确无损压缩的基本性能极限是 H(S)，即使在唯一可译性的限制下也是如此。 备注 5.5 有趣的是，这里的速率 R 从上方逼近 H(S)，而在 5.2 节中速率 R 从下方逼近 H(S)。 这种微小的差异反映了由于唯一可译性限制而产生的额外开销。 尽管香农-范诺码简单，但通常是次优的。下面是一个例子。 例子 5.3 考虑一个参数为 1/16 的伯努利离散无记忆信源 S。二进制香农-范诺码的索引字符串长度为 1 和 4。显然，使用长度为 4 的索引字符串是浪费的，因为对于 ∣S∣=2，我们可以简单地让码由 {0, 1} 组成，两者长度均为 1。

接下来我们介绍霍夫曼码，它是在最小化期望索引字符串长度意义上的最优无前缀码。 让我们先考虑二进制情况 (q=2)。假设 K=∣δ∣<∞。 不失一般性，我们重新排列信源消息，使得 PS(a1)≥PS(a2)≥...≥PS(aK)

对于一个最优的无前缀码，考虑其二叉树可视化。 以下两个事实成立。

1. 二叉树中的所有叶子都对应于索引字符串。
2. 索引字符串总是可以被安排成这样：两个最不可能的信源消息的索引字符串在二叉树的最大深度处是兄弟节点，也就是说，它们仅在最后一个位置上有所不同。 这两个事实都可以通过反证法来证明，并留作练习。 所以一个最优无前缀码的期望索引字符串长度可以写成如下形式：

= ∑Ps(s)l(s)

SES

=∑s∈S{aK−1,aK}PS(s)l(s)+PS(aK−1)l(aK−1)+PS(aK)l(aK)

= ∑s∈S{aK−1,aK}PS(s)l(s)+[PS(aK−1)+PS(aK)]l(aK−1).(5.31)

这表明我们可以通过将 f(aK−1) 和 f(aK) 的父节点转换为一个叶子来“合并”它们的叶子，这个叶子对应于一个以概率 PS(aK−1)+PS(aK) 出现的新消息。因此，我们获得一个新的离散无记忆信源 S'，其字母表为 S'= {a1,a2,...,aK−2,aK−1'}，PS'(s')=PS(s') 对于 s'∈{a1,a2,...,aK−2}

--- PAGE 12 ---

并且 PS'(aK−1')=PS(aK−1)+PS(aK)。此外，S 的二叉树在合并操作后对应于 S' 的一个无前缀码。 那么我们可以将 (5.31) 改写为

l=∑s'∈S'PS'(s')l(s')+[PS(aK−1)+PS(aK)],

(5.32)

其中右边的求和是 S' 码的期望索引字符串长度。 为了使 l 最小化，用于 S' 的无前缀码应该是最优的，并且其对应的二叉树也应满足事实 1 和 2。 所以我们可以重复前面的论证，进一步合并 S' 中两个最不可能的源消息。 递归地继续这个过程，直到不可能再进行合并，即获得一个只有一个源消息的信源，我们就得到了一个对应于 S 的最优无前缀码的二叉树。

然后考虑一般的 q-元情况。 对于最优无前缀码，让我们检查其 q-元树可视化的事实 1 和 2。 这两个事实现在变成如下：我们可以安排索引字符串，使得相应的 q-元树最多有 q−2 个未使用的叶子，所有这些叶子都位于最大深度，在由单个父节点生长的同一分支中。 二进制情况下的“排序和合并”论证仍然有效，但在第一步中需要进行一些修改，因为在最大深度要合并的叶子数量可能少于 q。 确实，我们需要确定最大深度处 0≤r≤q−2 个未使用叶子的数量。 为此，请注意，有限 q-元树中的叶子总数总是可以写成 q+n(q−1)，其中 n 是一个对应于非叶节点（不包括根）数量的整数。 因此我们有

K+r=q+n(q−1)

即 q−K=−n(q−1)+r

(5.33)

(5.34)

这是一个余数问题。 因此，我们可以将r作为将q−K除以q−1的余数来获得。为方便起见，我们将(K−q)(q−1)加到两边，得到

$$(K-q)(q-2)=(K-q-n)(q-1)+r$$

(5.35)

这决定了 r，即最大深度的未使用叶子数。 在后续步骤中，没有未使用的叶子，我们总是合并 q 个叶子，直到到达根节点。 总结一下，霍夫曼 q-元树的生成过程如下。

--- PAGE 13 ---

1. 创建 K 个节点作为 u1,u2,...,uK，并为 uk 分配概率 PS(ak), k=1,...,K。将这些节点初始化为“活动”状态。 计算 r 作为 (K−q)(q−2) 除以 q−1 的余数。对于 q=2, r 始终为零。
2. 将概率最小的 q−r 个活动节点连接在一起，创建一个新节点。 将 q−r 个连接的节点标记为“非活动”，并将新节点标记为“活动”。 为新节点分配 q−r 个连接节点的概率之和。
3. 如果只剩下一个活动节点，则将其标记为根并停止，否则，设置 r=0 并重复步骤 2。

生成霍夫曼 q-元树后，相应的无前缀码立即得出，该码最小化了期望索引字符串长度。 请注意，霍夫曼码仅适用于 ∣S∣<∞ 的情况。 如果 ∣δ∣=∞，通常不清楚最优码是什么样的。 然而，对于服从几何分布的离散无记忆信源，最优码是已知的；参见[24]。 笔记

在我们的讲义中，关于无损压缩的这一讲主要定位为第 4 讲之后信源表示一般理论的一个特例。然而，从历史上看，对精确无损压缩的研究早于率失真理论。 克拉夫特不等式最早出现在 Leon Kraft 1949 年左右在麻省理工学院的硕士论文[25]中，针对的是无前缀码。 其在定理 5.2 中针对唯一可译码的更一般形式是由 Brockway McMillan [26] 发现的，而本讲中的证明是由 Jack Karush [27] 找到的。 香农-范诺码最早在香农的原始文章[1]中间接描述，并由罗伯特·范诺独立发现。 无损压缩最优码的构造最初被认为是一个棘手的问题，但由大卫·霍夫曼在 1951 年解决[28]，这是罗伯特·范诺在他麻省理工学院信息论课上布置的一篇学期论文。 我们对霍夫曼码的解释基于 James L. Massey 的讲义[29]。 无损压缩也与随机数生成密切相关；参见，例如，[18, 第 5.11 章]。 由于可靠性在计算机文件系统中的巨大重要性，已经发明了许多无损压缩码。 与将固定长度的信源符号序列编码为可变长度字符串的霍夫曼码不同，Tunstall 码[30]将可变长度的信源符号序列编码为固定长度的字符串。 基于香农-范诺-埃利亚斯码构建的算术码已被广泛使用；参见，例如，[18, 第 13.3 章]。

--- PAGE 14 ---

游程编码 [31] 对于压缩具有长串全零和全一子序列的信源符号序列非常有效。 Burrows-Wheeler 变换 [32] 是一种巧妙的重新排列信源符号序列以产生长游程的方法，可用作游程编码的预处理步骤。 实践中广泛使用的一类码被称为通用码，因为它们不需要了解信源的概率分布，并且当信源符号序列的长度无限增长时，能渐进地达到最优压缩效率（对于离散无记忆信源是熵，对于平稳甚至更一般的信源是熵率）。 Lempel-Ziv 码可能是迄今为止最著名的通用编码方案，由 Abraham Lempel 和 Jacob Ziv 在 20 世纪 70 年代发明。 练习

1. 对于一个失真度量为 d(s,s^) 的离散无记忆信源 S，定义一个新的失真度量为 d~(s,s^)=1 如果 d(s,s^)>a，否则为 0，给定某个参数 a>0。描述在 d~ 下，D=0 时 S 的率失真函数。
2. 在第二节中，我们研究了当索引字符串为二进制时的精确无损压缩。 如果索引字符串是 q-元的，q≥2，通过推广第二节中的分析，推导期望索引字符串长度的下界和上界。
3. 对于第二节研究的精确无损压缩码，当 S 是 (a) 在 {1,2,...,M} 上均匀分布，和 (b) 参数为 的几何分布时，数值计算 l。 将这些情况下 l 的精确值与第二节中得到的上界和下界进行比较。
4. 证明一个码是唯一可译的当且仅当对于任何整数 n≥1，以及任何 s\=s'∈Sn，f(s)\=f(s')。
5. 我们可以用以下形式重写克拉夫特不等式，而不是定理 5.1 中的 (5.21)

ΣA ≤1,

=1

(5.36)

其中 Al 表示长度为 l 的索引字符串的数量。 让我们使用这种形式的克拉夫特不等式来证明定理 5.1 的反证部分； 也就是说，对于给定的满足克拉夫特不等式的集合 {Al:l=1,2,...}，我们可以构造一个相应的无前缀码。 从根节点开始，完成以下归纳：

a) 证明从根节点开始，在深度 1 处至少有 A1 个叶子来容纳 A1 个长度为 1 的索引字符串。

--- PAGE 15 ---

5.4 香农-范诺码与霍夫曼码 79

b) 假设我们已经容纳了所有从长度 1 到长度 l−1 的索引字符串。证明在深度 1 处至少有 Al 个未使用的叶子来容纳 Al 个长度为 l 的索引字符串。 6. 如果对于任意 s\=s'∈S，f(s) 都不是 f(s') 的后缀，则该码称为无后缀码；如果既是无前缀码又是无后缀码，则称为无前后缀码。 对于一个离散无记忆信源 S，其中 |丨δ丨<∞, 当 Σs∈Sq−l(s)≤1/2 时，找一种方法来构造一个具有长度 {l(s):s∈δ} 的 q-元无前后缀码。 7. 证明对于一个 丨δ丨=∞ 的离散无记忆信源 S，一个无前缀码仍然满足克拉夫特不等式，反之，对于任何满足克拉夫特不等式的索引字符串长度，都存在一个相应的无前缀码。 (提示：不要使用有根树的可视化方法，而是将无前缀码可视化为单位区间 [0, 1] 的一个划分，使得每个索引字符串对应一个子区间，并且所有子区间都是不相交的；参见 [18, 第 5.5 章])

1. 推导一个在 {1,2,...,10000} 上均匀分布的离散无记忆信源 S 的二进制霍夫曼码，并将得到的期望索引字符串长度与熵界 log2 10000 比特进行比较。
2. 对于一个离散无记忆信源 S，我们设计一个无前缀码，该码最小化加权期望索引字符串长度 l=∑s∈SPS(s)c(s)l(s)，其中 c(s)>0 是信源消息 s 的每个索引位置的成本。 注意，当 c(s)=1, ∀s∈S 时，我们回到了第四节研究的问题，该问题由霍夫曼码解决。 a) 推导 l 的下界，并讨论何时可以达到该下界。 b) 推广霍夫曼码以产生最小化 l. 的无前缀码。
3. 对于一个具有 K 个正概率和一个零概率的离散无记忆信源 S，即 DS(a1)≥PS(a2)≥...≥PS(aK)>PS(aK+1)=0, 我们可以设计一个忽略零概率的霍夫曼码，或者包含它。 找出这两种不同霍夫曼码的期望索引字符串长度之间的关系。
4. 考虑独立的离散无记忆信源 S1 和 S2，它们具有（不一定相同的）有限字母表。 将它们的二进制霍夫曼码分别表示为 fS1 和 fS2。 现在将 (S1,S2) 视为一个单一的离散无记忆信源，并使用级联 [fS1,fS2] 作为 (S1,S2) 的码。例如，如果对于某个 (s1,s2)，fS1(s1)=001 且 fS2(s2)=101，那么 fS1,S2(s1,s2)=001101。 a) 证明 fS1,S2 是一个无前缀码。

b) fS1,S2 的克拉夫特不等式是否总是取等号？ 12. 香农-范诺码采用了一种保守的理念

--- PAGE 16 ---

80

5 信源表示：无损压缩

通过对 −logqPS(s) 的所有非整数值进行向上取整。 通过明智地对 −logqPS(s) 的某些非整数值进行向下取整，有可能获得一个具有更小期望索引字符串长度的无前缀码。 a) 提出一种设计无前缀码的算法，该算法可能通过选择性地对 −logqPS(s) 的某些非整数值进行向下取整来优于香农-范诺码。

b) 找一个例子，其中您设计的码严格差于霍夫曼码。 13. 在第 4 讲有损信源表示的问题表述中，编码后的索引 W∈{1,2,...,Mn} 也可以被看作是一个固定长度为 ⌈log2Mn⌉ 的二进制字符串。现在，如果我们允许 W 是可变长度的，取自所有有限长度二进制字符串的集合 w∗={∅,0,1,00,01,10,11,000,001,...}。将一个码的速率定义为 R=E[l(S)]/n, 其中 l(S) 是编码 S 的 W 的长度，n 是 S 的长度。修改第 4 讲中反证部分的证明，以表明可变长度编码仍然不能优于率失真函数。