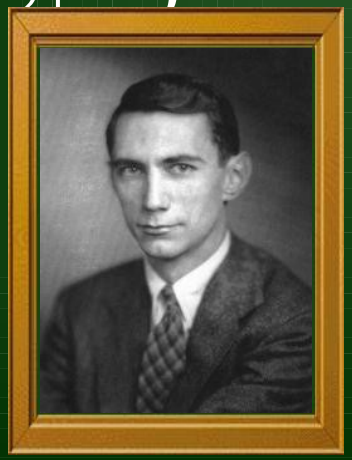
信 息 论

在信息学竞赛中的简单应用

侯启明

信息论简介

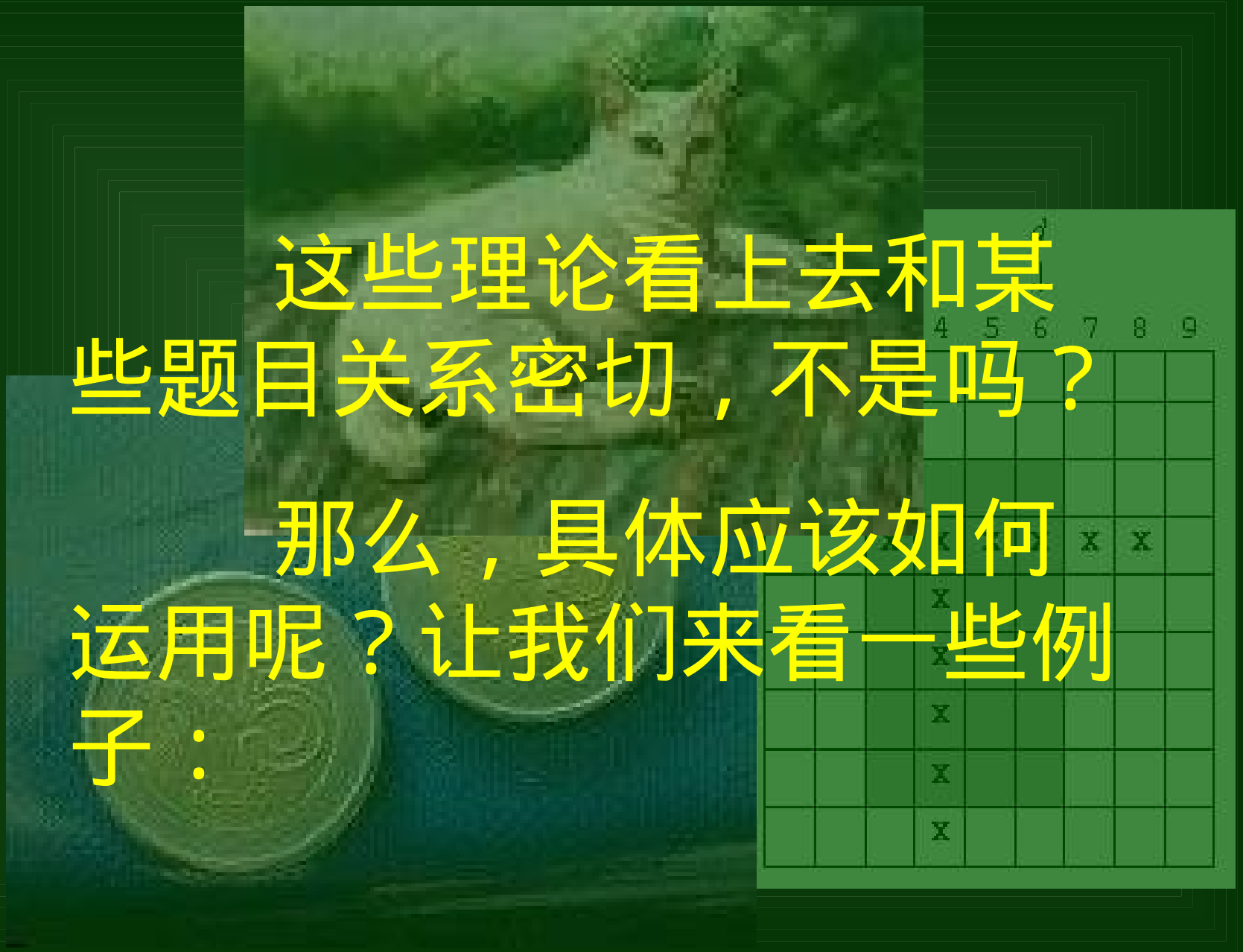
- 信息论是关于信息的本质和传输规律的科学的理论。
- 通过它可以很方便地得到某些交互式问题的一个较好的步数下界 (“ 信息论下界”)



让我们先来看一些信息论的基本理论

理论基础

- 定义：如果一个随机变量 **x** 共有 **n** 种取值，概率分别为 $p_0,p_2,\ldots\ldots,p_n$，则其熵为 $H(x)=f(p_0,p_2,\ldots\ldots,p_n)=-\sum Cp_i\log p_i$
- 定理 **1**：在得到关于随机变量 **x** 的一个熵为 **h** 的信息后，**x** 的熵将会减少 **h**。
- 定理 **2**：当一个随机变量的各种取值概率相等时，它的熵最大。



这些理论看上去和某些题目关系密切，不是吗？

那么，具体应该如何运用呢？让我们来看一些例子：

例 1 ：验证一下定理 1

我们宿舍二楼到三楼之间楼梯的窗户外面是相邻的一个平房的房顶。在那一带栖息着三只浑身雪白，有着一只蓝眼睛和一只绿眼睛的一猫！

例 1 ：验证一下定理 1

在天冷的时候，它们喜欢趴在楼内的暖气上。于是，每只猫就有了两种状态：在屋内和在屋外。因此，三只猫的状态共有 8 种可能情况，假设它们是等概率的。

现在，我在一楼的小卖部。由于种种原因，我希望知道猫当时的状况，因此，我往上看了一眼，结果发现在这个位置只能知道屋内猫的只数.....

例 1 ：验证一下定理 1

问题 1： 把所有猫的情况作为一个随机变量 x，则当我在小卖部的时候， x 的熵是多少?

解答 1： 由于 8 种情况的概率相等，所以:

$$H(x)=f(1/8,1/8,1/8,1/8,1/8,1/8,1/8,1/8)=\log 8$$

问题 2： 我看一眼所得到的信息 y 的熵是多少?

解答 2： 由于猫的只数共有 0,1,2,3 四种情况，概率分别为 (1/8,3/8,3/8,1/8)，所以:

$$H(y)=f(1/8,3/8,3/8,1/8)=\log 8-6\log 3/8$$

例 1 ：验证一下定理 1

问题 3： 我看完之后， x 的熵 H'(x) 是多少?

解答 3： 此时猫的只数为 0,1,2,3 的四种情况的概率依次是 (1/8,3/8,3/8,1/8)， 而每种情况的熵分别为 (0,log3,log3,0)， 所以此时 H'(x) 的数学期望为:

$$H'(x)=1/8*0+3/8*\log3+3/8*\log3+1/8*0=6\log3/8$$

可以发现 H(x)=H(y)+H'(x) 。

定理 1 得到了验证。

例 2 ： Rods(IOI2002)

一个 Rod 是一个由至少 2 个单位正方形连成的水平或竖直的长条。在一个 N*N 的方阵中，放了水平和竖直两个 Rod 。如图 1，其中 Rod 用 X 表示。

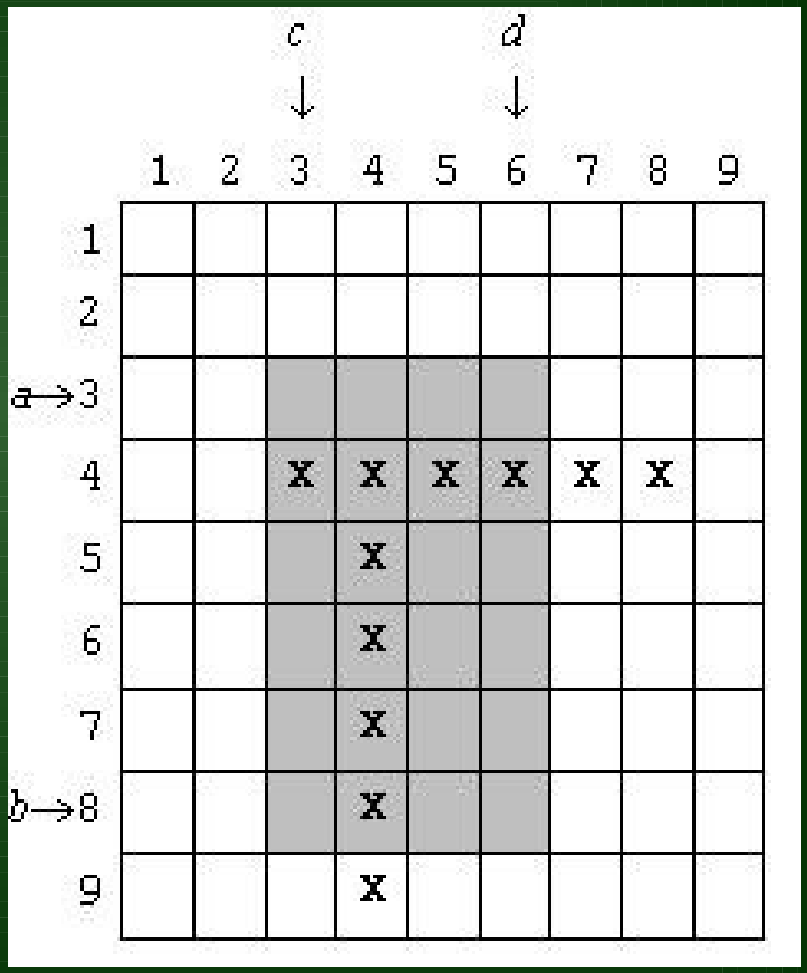


图 1

例 2 ： Rods(IOI2002)

两个 Rod 可能有公共方格，比如在图 1 中，方格（4，4）无法确定是仅属于 1 个 Rod 还是同时属于两个 Rod 。因此，在这种情况下我们假定它同时属于两个 Rod 。这样，图中竖直 Rod 的上端点是（4，4）而不是（5，4）。

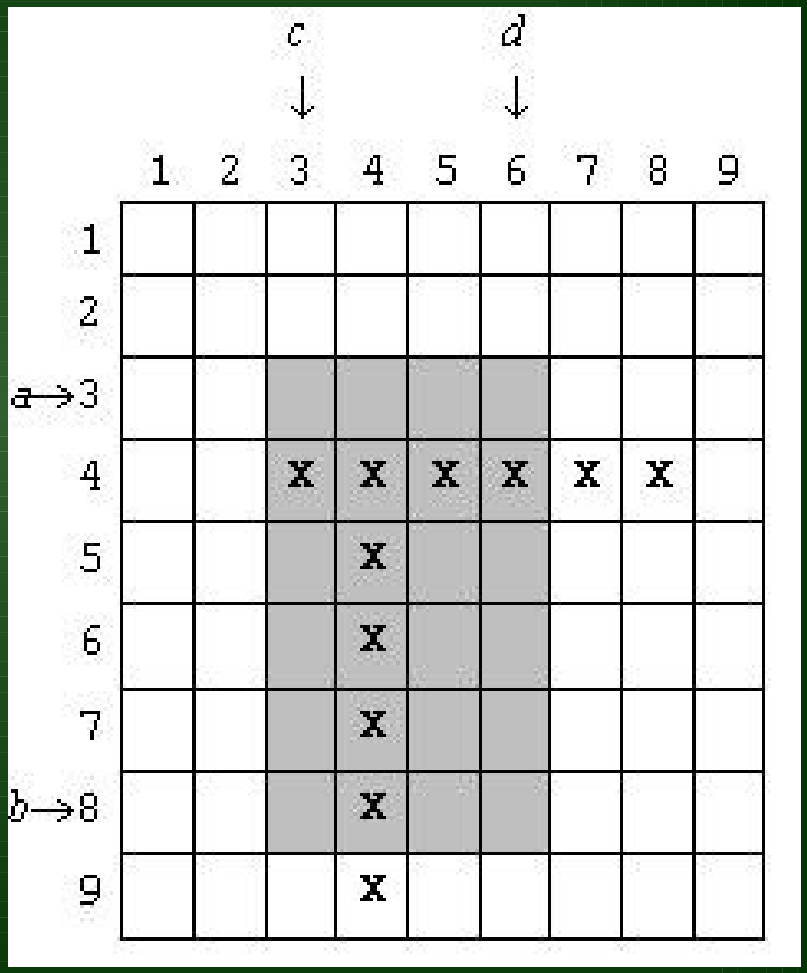


图 1

例 2 ： Rods(IOI2002)

最初我们并不知道两个 Rod 的位置，你的任务是编程序找出它们的位置。你只能通过库函数 rect(a,b,c,d) 来定位两个 Rod 。如果至少一个属于某个 Rod 的方格落在矩形 [a,b]x[c,d] (如图 1 中阴影区域) 内的话， rect 返回 1 ，否则返回 0 。

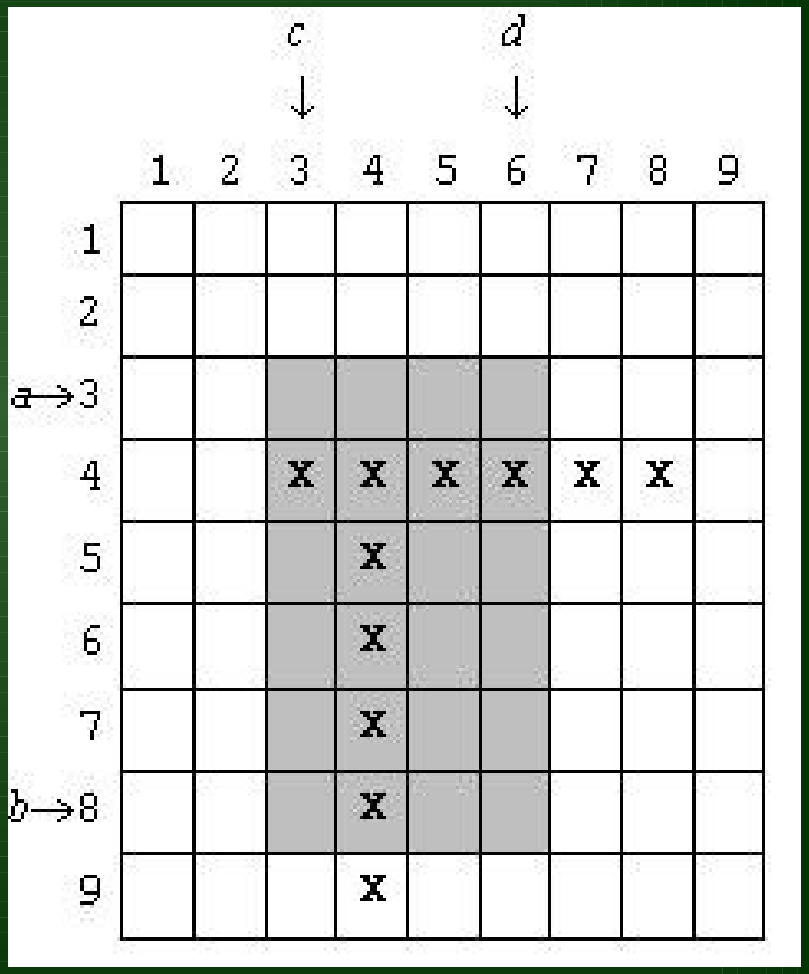


图 1

例 2 ： Rods(IOI2002)

对每个测试点，如果你的程序没有正确确定两个 Rod 的位置或调用 rect 超过 400 次，你将得到 0 分。否则，如果调用 rect 的次数至多为 100，你将得到 5 分；在 101 到 200 间，你将得到 3 分；在 201 到 400 间，你将得到 1 分。

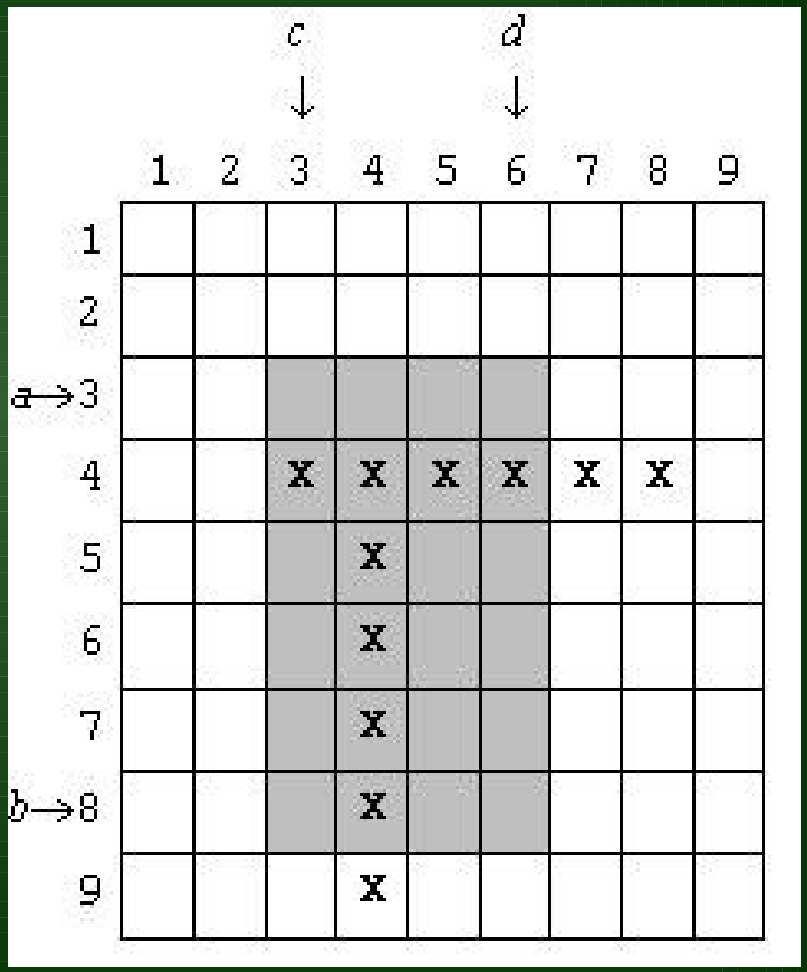


图 1

例 2 ： Rods(IOI2002)

比赛时我很快想到了一个最多调用 rect 函数 $6\log_2 n+C$ （某个常数）次的方法，但是因为这个数差不多刚好达到 100，所以我在这时就开始试图优化上式中 $\log_2 n$ 的系数，结果徒劳无功，反而耽误了时间。因此，看过答案以后，我试着从信息论的角度分析了一下这个问题：

$6\log_2 n+C?$

例 2 ： Rods(IOI2002)

由于题目中没有涉及到概率，因此假设所有情况都是等概率的。所以，设 Rod 的摆放方法为随机变量 x ， x 所有可能的取值数为 f(n) ，那么 x 的熵 H(x) 就等于 log(f(n)) 。而由于库函数只有两种返回值，其熵最大为 $H_{max}(y)=\log 2$ 。因此， rect 调用次数的信息论下界就是

$$L=H(x)/H_{max}(y)=\log(f(n))/\log 2=\log_2 f(n)$$

例 2 ： Rods(IOI2002)

下面讨论 f(n) 的值:

在 n*n 的方阵中放 1 个 Rod （无论横竖）共有 n*C(n+1,2) 种方案，放两个相交的 Rod 共有 $C^2(n+2,3)$ 种方案，所以:

$$f(n)=(n^2(n+1)/2)2-((n+2)(n+1)n/6)^2$$

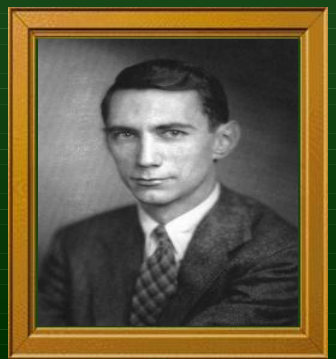
$$=(2n^6+3n^5-n^4-3n^3-n^2)/9$$

当 n 充分大时:

$$L=\log(f(n))/\log 2>\log_2(2n^6/9)\approx 6\log_2 n-2.2$$

例 2 ： Rods(IOI2002)

由于各种原因，不一定总是使两种返回值概率相等，所以最坏情况下的调用次数往往达不到信息论下界，两者大约相差一个常数，因此，可以认为 $6\log_2 n+C$ 是 rect 函数最大调用次数的下界。这样，在得到一个这样的算法之后，就没有什么必要再去徒劳地优化步数了。



$6\log_2 n+C!$

例 3 ： Coins(选手推荐题 0024 ，推荐者饶向荣)

有一堆 n 个硬币，其中有 n-1 个好的，一个坏的。所有好的硬币的质量是相同的，但坏的硬币的质量却不一样，现在告诉你某一枚是好的，能否用一架天平在 k 次以内称出哪个是坏的硬币。

输入 n 和 k ，如果能在 k 此比较中找到 n 枚硬币中的哪枚为坏的，就输出 'POSSIBLE' ，否则输出 'IMPOSSIBLE' 。

例 3 ： Coins(选手推荐题 0024 ，推荐者饶向荣)

两年前我的一位远房亲戚曾给我出过一个类似的题目（ n=14 ， k=3 ），当时我苦苦思索了一晚上，终于想出来一个可行解法。于是，那位亲戚加大了数据规模 (n=1101 ， k=7 ， IMPOSSIBLE) ，我想了大概一周，觉得应该无解，但苦于无法证明我的解法的最优性，始终不能理直气壮地回答 "IMPOSSIBLE" 。

例 3 ： Coins(选手推荐题 0024 ，推荐者饶向荣)

后来她给了我一个 " 说明 " ，但我始终觉得不太严密；拿来问我们班的 IMO 金牌，回答是 " 显然 " ，我觉得也不严密 :(。于是，这件事就成了我这两年来的一个遗憾。

现在，有了信[illegible]这个遗憾终于得到了解决！！

初步分析

首先，对硬币用 1 到 n 进行编号，设坏硬币的编号为 x 。可以认为 x 的所有取值情况概率相等:

∴H(x)=logn 。

∵ 用天平称一次的结果 y 只有 3 种可能情况（左边较重，右边较重，平衡）

∴$H_{max}(y)=log3$ 。

∴ 从 n 个硬币中通过天平找出一个坏硬币至少需要 $H(x)/H_{max}(y)=log_3 n$ 步

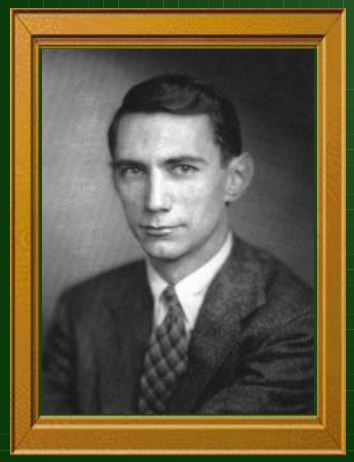
初步分析

下面通过构造证明当知道坏硬币比好硬币轻还是重的时候，这个下界是可以达到的：

每次把所有硬币分成三等份，比较其中两份，如果平衡，说明坏硬币在第三份中，否则坏硬币就在重的一份中，这样每次比较得到三种结果的概率相等，$H(y)\equiv H_{max}(y)$。所以此时，从 n 个硬币中找出一个坏硬币只需 $\log_3 n$ 步。

初步分析

虽然这样，但是在原题的条件下，这个信息论下界是达不到的，不过如果没有这个结论，真正最优的解法的最优性就无从证明。得出这个结论后，后面的困难就迎刃而解了。



请看进一步的分析：

进一步的分析

通过转化，发现只要计算出给出一枚好硬币，k 次比较最多在多少枚硬币（不包括给出的好硬币）中找出一枚坏硬币，就可以解决原问题。

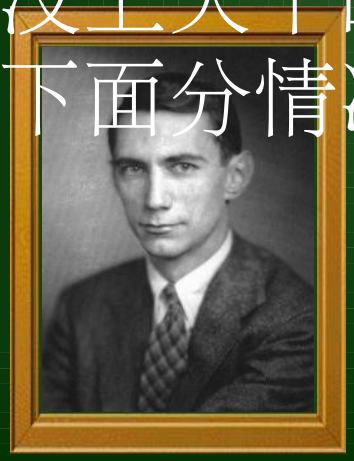
不过，转化之后的问题仍然难以解决，一枚好硬币实在太少，因此，不妨先将原问题“放大”一下，考虑一下有无穷枚好硬币的情形。

进一步的分析

设在有无穷枚好硬币时，k 次比较最多从 g(k) 个硬币中找出一个坏硬币。

当 k=1 时，通过枚举可以发现，g(1)=2 。

当 k>1 时：考虑第一次比较，设 t 为这次没上天平的尚未确定好坏的硬币的个数，下面分情况讨论：



信息论不行的时候，枚举也是必要的。

进一步的分析

- 如果比较结果是“平衡”。

由于可以通过剩下的 k-1 次比较把坏硬币从这 t 个硬币中找出来，所以 t<=g(k-1)。

- 如果比较结果不是“平衡”。

此时可以确定坏硬币在上了天平的 g(k)-t 个硬币中，同样，根据结论 1，得到 $g(k)-t<=3^{k-1}$，故 $g(k)<=g(k-1)+3^{k-1}$:

进一步的分析

现在通过构造来证明 $g(k)=g(k-1)+3^{k-1}$：

第一次比较第 1 到 3^{k-1} 号硬币和 3^{k-1} 个好硬币，分以下情况讨论：

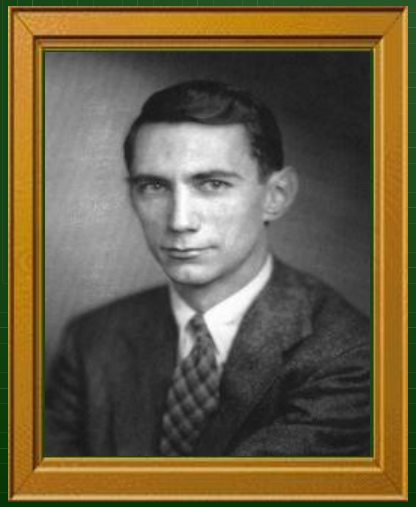
- 平衡：说明坏硬币在剩下的 g(k-1) 个硬币中，由 g 的定义，可以在 k-1 步内找出。
- 好球较轻：说明坏硬币就在这些硬币中，且较重，由上文结论，可以在 k-1 步内找出。
- 好球较重：与上一种情况类似，不再赘述。

进一步的分析

这样，根据 $g(k)=g(k-1)+3^{k-1}$，计算得出：$g(k)=(3^k+1)/2$ 。

“放大”后的问题解决了，那么原问题呢？我们可以猜想，一个好硬币和无穷多个好硬币是等效的，也就是说，如果设在有一个已知的好硬币的情况下，k 次比较最多从 f(k) 个硬币中找出一个坏硬币，那么 $f(k)\equiv g(k)$ 。

下面进行构造证明



最后的构造

根据计算 g(k) 时的推理，第一次比较应该有 3^{k-1} 枚“有嫌疑”的硬币上天平。由于这个数是奇数，所以只好把唯一一枚没有嫌疑的硬币也放上天平，这样就确定了第一次比较的方案。

如果比较结果是“平衡”，那么我们就把嫌疑缩小到了 g(k-1) 个硬币中，而且有了足够的好硬币，很容易通过 k-1 次比较把坏硬币找出来。

最后的构造

但假如比较结果是“不平衡”呢？此时，坏硬币编号的熵为 $\log 3^{k-1}$，如果要在 k-1 次比较内找出坏硬币，那么此后每一次比较结果的熵都得是 log3 。所以，这次比较得到三种结果的概率必须相等。

既然这样，我们不妨来看看这次比较可能得到的三种结果都意味着什么。

最后的构造

- 和第一次比较相同

  说明两次比较中坏硬币在天平同侧。

- 和第一次比较不同

  说明两次比较中坏硬币在天平异侧。

- 平衡

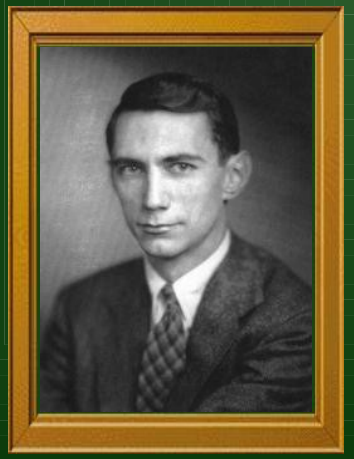
  说明第三次比较中坏硬币没上天平。

最后的构造

因为这三种情况出现的概率相同且必居其一。所以第二次比较时，相对于第一次比较，有 3^{k-2} 枚嫌疑硬币保持原位，3^{k-2} 枚嫌疑硬币换到了另一侧，3^{k-2} 枚嫌疑硬币换成了好硬币。而原来那枚好硬币，为了保持天平两边硬币数相等，也只好换到另一侧。这样，第二次比较的方案也被唯一确定了。

最后的构造

稍加分析不难得出后面的步骤。但由于整个过程的形式化描述过于繁琐，故这里不再赘述。



构造完毕！

问题的解决

就这样，在信息论的帮助下，这个困扰了我两年的问题终于解决了。虽然这道题的重点在构造而不是信息论，但信息论在证明解法的最优性时是十分必要的，而且，信息论的分析也在本题如理乱麻的构造过程中起着关键性的指导作用。

总结

细心的读者应该会注意到，本文中的例题不用信息论的知识都可以解决。那么，信息论在这里的意义是什么呢？实际上，作为一种纯粹的理论，信息论是一种可以用来对一类问题进行分析的工具。它可以为我们的解法提供强有力的理论依据，更可以通过估计上下界来指导解法的构造。综上所述，信息论在信息学竞赛中是大有用武之地的。

谢谢大家！