

神奇的模式概率与“鞅”

李硕彦



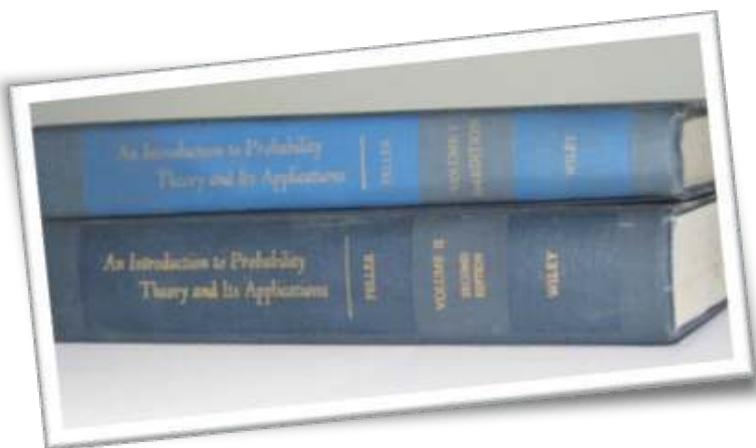
违反直觉的模式概率

William Feller 的经典教科书《概率论及其应用》提到一个试验：重复抛掷一个标准硬币，一直到连续 6 次出现人头。用 H (= Head) 代表硬币的人头那面，T (= Tail) 代表尾巴，连续出现 6 个 H 所构成的模式就表示为 HHHHHH。抛掷硬币的次数就叫作这个模式的“等待时间”，它是一个随机变量，最小值是 6，最大值无限。Feller 问：这个等待时间的

均值有多大？

标准硬币的 H 和 T 出现的概率均等，所有 $2^6 = 64$ 个由 H 和 T 构成的长度为 6 的模式都有相同的机会出现。如果抛掷硬币 64,000,000 次，每一个模式都会出现 1,000,000 次左右。换句话说，平均来讲每掷 64 次就每一个模式都会出现一次。所以我们试着推论：

- 所有 64 个模式的平均等待时间都是 64。



William Feller 生前并没有完成这部上下两册的概率论经典。他死后，学生帮忙作最后的编辑、付印

真的可以这样推论吗？Feller 不放心，于是将这个问题直接当成马尔科夫链，细心地作了冗长的计算，结果是：

- HHHHHH 模式的平均等待时间不是 64，而是 126。

这否定了前面的推论。Feller 又尝试另一个长度为 6 的模式：

- HHTTHH 模式的平均等待时间不是 64，也不是 126，而是 70。

Feller 称自己这些计算的结果“违反直觉”。

真理和童话愉快地吻合，被冷落一旁的是直觉

当科学真理与直觉相违背的时候，问题自然是在直觉那一方。平均来说，每抛掷硬币 64 次，任何一个长度为 6 的模式都会见到一次，这是正确的。用较为严谨的数学语言来说：

- 每一个长度为 6 的模式平均“刷新时间”都是 64。

刷新时间指的是从这个模式出现到下一次再次出现的时间，它也是一个随机变量。譬如说，抛掷硬币出现了这个序列：

...HTHTTHHTHTHTTTTHTHTTTTTHHHHTTHHTTHHH...

其中三个红色的 H 代表 HHTTHH 模式完成的时候。从第一次完成这个模式之后，又掷了硬币 21 次才出现第二次，这时候，刷新时间这个随机变量就取值为 21。之后又掷了 4 次就出现第三次。这回，刷新时间取值为 4。

HHTTHH 模式的刷新时间的最小值是 4，而等待时间的最小值是 6。两个随机变量明显地不同，而等待时间只可能偏大。只要你的直觉将两者区分开来，结论就应该是：

- HHTTHH 模式的平均等待时间 > 64 ，其他等长模式的平均等待时间 ≥ 64 。

Feller 算出 HHTTHH 的平均等待时间为 70，完全没有矛盾。

计算概率的时候，直觉是最重要的，但也是最易引入陷阱的。底下我们再用一个例子来看一般的直觉和科学真理可以相去多远。还是连续抛掷一个标准硬币，看 THTH 和 HTHH 这两个模式之间“哪一个先出现，也就是说，让这两个模式“赛跑”。从平均等待时间来看，THTH 是 20，而 HTHH 是 18。也就是说，THTH 跑得比较慢，HTHH 比较快。那么，

- 这场赛跑的胜算比 (odds) 是不是应该很接近于 9 比 10 呢？

答案是否定的，其实这是一面倒的竞赛，正确的胜算比是 9 比 5。也就是说，双方赢的概率各自是 $9/14$ 和 $5/14$ 。看完本文之后，读者就能很容易地算出这个胜算比，当然也能立刻得出 THTH、HTHH、HHHHHH、HHTTHH 的平均等待时间，而且完全不需要马尔科夫链的冗长计算。

在解释这个一面倒的现象之前，还有一点小小的补充：

- 这个一面倒的胜算比是倒向比较慢的那个模式。平均每 14 场赛跑，THTH 赢 9 次，比较快的模式 HTHH 反倒只赢 5 次。

模式的概率就是这么神奇。有些时候，科学的真理乍听之下好像是童话故事。事实上，用童话故事来解释这个胜算比就再简单不过了：参与赛跑的 THTH 模式以 T 开头，代表乌龟 (Tortoise)。另一个参赛者 HTHH 以 H 开头，代表野兔 (Hare)。童话故事里，龟兔赛跑通常是乌龟胜，当然

也有说是野兔赢的，两种说法的比例大约是 9 比 5。所以真理和童话愉快地吻合了，被冷落一旁的是直觉。当然，这只是文字游戏，不是科学的解释，用来取悦小孩还可以。



参与赛跑的 THTH 模式以 T 开头, 代表乌龟 (Tortoise)。对手 HTHH 模式以 H 开头, 代表野兔 (Hare)。乌龟的平均等待时间是 20, 而野兔是 18, 所以乌龟跑得比较慢, 野兔比较快。但是, 绝大多数的时候是乌龟跑赢

古今中外的田忌赛马

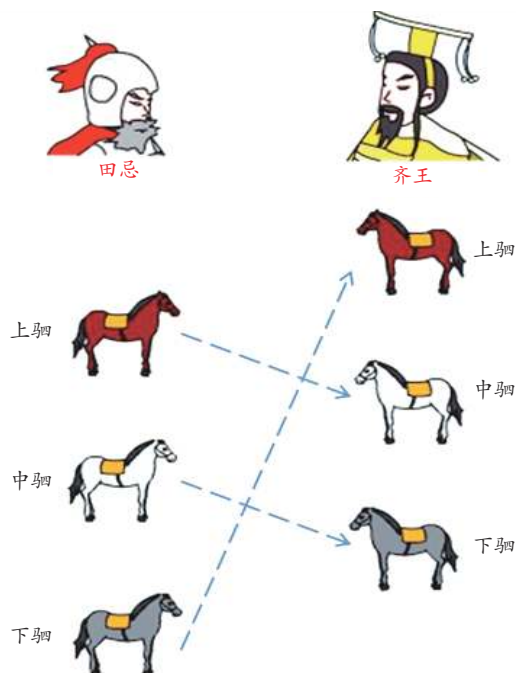
上述受冷落的直觉只是未受过训练的直觉。正确的直觉是什么？底下我们来看一个随机的长序列：

[illegible]

大致上，有十六分之一的字母是绿色的 H，用于表示 THTH 模式完成的时候。有十六分之一是红色的 H，用于表示 HTHH 模式完成的时候。两者的个数大致相等，但是分布的方式不同，有一半的绿 H 其后紧跟着红 H，但是在每一个红 H 之后的 3 个字母绝对不会是绿 H。所以，如果从长序列中随机取一点开始寻找有颜色的 H，第一个碰到的比较可能是绿的。如果用跑马场的语言来说，每次野兔跑赢的时候，都至少要赢一个马身，而乌龟赢的时候有一半是仅赢一个马颈。野兔赢的少，但都是大胜，乌龟赢的多，但常是小胜。合计起来，双方的实力其实相等。

战国时期，齐国的国王总是用赛马来赢公子田忌的钱。有一天，齐王又邀请田忌赛三场马，当时齐王拥有全国第1、第3、第5快马，而田忌有的是第2、第4、第6。不敢拒绝国王的邀请，但是明摆的又要输三笔钱，田忌实感为难。这时，军事谋略家孙臧是田忌的座上客，他教田忌向齐王建议：加大赌注，然后三赛二胜的一方独赢。齐王欣然同意，等着收取一大笔横财，而且，即使意外地输一场也无妨，等于是买了双保险。比赛当天，田忌祭出孙臧的战略：“以上驷对中驷、中驷对下驷、下驷对上驷”。结果，齐王的上驷先胜

一场，然后田忌小胜两场，从此齐王不再当田忌为自动提款机。这就是成语“田忌赛马”的由来。



THTH 常跑赢 HTHH 凭的就是孙臧的小胜多胜战略。1969 年的数学消遣杂志 (J. Recreational Math) 中, Walter Penney 也提到了掷硬币的赌博: 两个人各选一个长度为 3 的模式, 谁的模式先出现就赢。技巧是要让对手先选模式, 然后针对它选一个的最优的克星。那个年代, 美国有一个中学生就将这个游戏理论付诸实践, 赢了同学们很多钱。只可惜, 他的实践成果太过辉煌, 差点被学校开除。

抛掷硬币，两个人各选一个长度为3的模式，谁的模式先出现就赢。技巧是要让对手先选模式，然后针对它选一个的最优的尅星，也就是箭头所指向的模式

