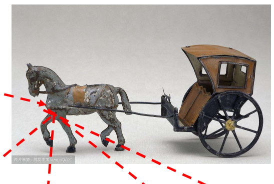
第5章　鞅

骆吉洲
计算机科学与技术学院

鞅：martingle



四川老乡读出来：
“鞅起弟弟/妹妹一路走...”

含义：带弟弟/妹妹一起走，以他/她的速度来限制你的速度

随机变量序列：$X_0,X_1,\ldots,X_n,\ldots$

限制$X_0,X_1,\ldots,X_n,\ldots$增长速度的内在机制

估计硬币投掷过程中
- 正面向上次数与背面向上次数之差的大小

估计随机图（社交网络、路由网络）中
- 路径长短
- 围长
- 色数
- 团大小
- …

估计长序列（DNA序列、字符串序列）中
- 模式串出现次数

- 有依赖关系的随机变量序列$X_0,X_1,X_2,\ldots$

$$\Pr[\,|X_n-X_0\,|>t]=?$$

- 随机变量$X_1,X_2,\ldots,X_n$的函数$f(X_1,X_2,\ldots,X_n)$ 不仅仅是和

$$\Pr[\,|\,f(X_1,X_2,\ldots,X_n)-E[f(X_1,X_2,\ldots,X_n)]\,|>t]=?$$

哪种结构有助于…　　鞅！

提纲

5.1 鞅的定义和基本性质

HIT CS&E

鞅(Martingle)

定义

随机变量序列 $X_0, X_1, X_2, \ldots$

如果

$E[X_i | X_0, X_1, \ldots, X_{i-1}] = X_{i-1}$ $\forall i \geq 1$

则

称$X_0, X_1, X_2, \ldots$是一个鞅

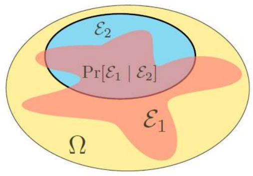
我不知道你在说啥

HIT CS&E

条件概率

定义. “在随机事件$\mathcal{E}_2$发生的条件下$\mathcal{E}_1$发生”的条件概率定义为

$$\Pr[\mathcal{E}_1 | \mathcal{E}_2] = \frac{\Pr[\mathcal{E}_1 \wedge \mathcal{E}_2]}{\Pr[\mathcal{E}_2]}$$



如果$\mathcal{E}_1$和 $\mathcal{E}_2$是独立的

$$\Pr[\mathcal{E}_1 | \mathcal{E}_2] = \frac{\Pr[\mathcal{E}_1 \wedge \mathcal{E}_2]}{\Pr[\mathcal{E}_2]} = \frac{\Pr[\mathcal{E}_1]\Pr[\mathcal{E}_2]}{\Pr[\mathcal{E}_2]} = \Pr[\mathcal{E}_1]$$

HIT CS&E

条件期望

定义. Y关于X的条件期望定义为

$$E[Y | X] = \sum y \cdot \Pr[Y=y | X]$$

例. 均匀随机地抽取一个人

X: 国籍 Y: 身高 Z: 性别

$E[Y | X=\text{“中国”}] = \sum y \cdot \Pr[Y=y | X=\text{“中国”}]$

给定一个“国籍”x,我们就可以计算$E[Y|X=x]$

$E[Y|X]$: X的变化范围 $\to R$

$E[Y|X]$可以视为一个函数$f(X)$

$f(x) = E[Y|X=x]$ $E[Y|X]$也是一个随机变量

HIT CS&E

条件期望

例. 均匀随机地抽取一个人

X: 国籍 Y: 身高 Z: 性别

$E[Y|X,Z]$

$E[Y|X=\text{“中国”}, Z=\text{“男”}]$ = 中国男性平均身高

$E[Y|X=\text{“中国”}, Z=\text{“女”}]$ = 中国女性平均身高

$E[Y|X=\text{“美国”}, Z=\text{“男”}]$ = 美国男性平均身高

$E[Y|X=\text{“美国”}, Z=\text{“女”}]$ = 美国女性平均身高

……

$E[Y|X,Z]$也是一个随机变量

HIT CS&E

条件期望的基本性质

性质1. $E[Y] = E[\, E[Y | X]\,]$

例：所有人身高均值 = 各国身高均值的均值

性质2. $E[Y | Z] = E[\, E[Y | X,Z] \,|\, Z\,]$

例：所有男人身高均值 = 各国男人身高均值的均值

性质3. $E[\, E[f(X)g(X,Y) | X]\,] = E[\, f(X)\, E[g(X,Y) | X]\,]$

一旦给定$X=x$，则$f(X)$的取值是一个实数

HIT CS&E

鞅

定义

随机变量序列 $X_0, X_1, X_2, \ldots$

如果

$E[X_i | X_0, X_1, \ldots, X_{i-1}] = X_{i-1}$ $\forall i \geq 1$

则

称$X_0, X_1, X_2, \ldots$是一个鞅

$\forall x_0, x_1, \ldots, x_{i-1}$,

$$E[X_i \mid X_0 = x_0, X_1 = x_1, \ldots, X_{i-1} = x_{i-1}] = x_{i-1}$$

$$E[X_i - X_{i-1} \mid X_0, \ldots, X_{i-1}] = 0$$

HIT CS&E

例1：公平赌博

$E[X_i|X_0,X_1,\dots,X_{i-1}] = X_{i-1}$

公平赌局
- 初始赌资X_0，X_i表示第i轮之后的赌资
- 每轮以1/2的概率赢，赢得的赌资是a
- 以1/2的概率输,输掉的赌资是a
- 每局的输赢是相互独立的

$E[X_i|X_0,X_1,\dots,X_{i-1}]$
$=E[X_i|X_{i-1}]$
$=(X_{i-1}+a)\Pr[$第i轮赢$] + (X_{i-1}-a)\Pr[$第i轮输$]$
$=(X_{i-1}+a)/2+ (X_{i-1}-a)/2$
$=X_{i-1}$

赌资序列是鞅

HIT CS&E

例2：例1的抽象化

$E[X_i|X_0,X_1,\dots,X_{i-1}] = X_{i-1}$

均匀硬币投掷n次
- $X_0=0$
- X_i = HEADs$_i$-TAILs$_i$
- HEADs$_i$—i次投掷后头面向上的总次数
- TAILs$_i$ —i次投掷后背面向上的总次数

$E[X_i|X_0,X_1,\dots,X_{i-1}]$
$=E[X_i|X_{i-1}]$
$=(X_{i-1}+1)\Pr[$第i次头面向上$] + (X_{i-1}-1)\Pr[$第i次背面向上$]$
$=(X_{i-1}+1)/2+ (X_{i-1}-1)/2$
$=X_{i-1}$

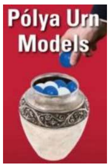
实验成败总次数之差的序列是鞅

HIT CS&E

例3: Polya壶

$E[X_i|X_0,X_1,\dots,X_{i-1}] = X_{i-1}$

Polya壶
- 壶中有b个篮球，w个白球, X_0
- 第i次操作
  - 均匀、随机、独立地从壶中抓出一个球
  - 放回与所抓球同色的c个球
  - 第i次操作后篮球的比例记为X_i
- $X_0,X_1,\dots X_n,\dots$



HIT CS&E

例3: Polya壶

$E[X_i|X_0,X_1,\dots,X_{i-1}] = X_{i-1}$

操作前：球总数a 蓝球比例X_{i-1}

抓到蓝球，$\Pr=X_{i-1}$ 抓到白球，$\Pr=1-X_{i-1}$

操作后：球总数a-1+c
蓝球个数aX_{i-1}-1+c

操作后：球总数a-1+c
蓝球个数aX_{i-1}

$E[X_i|X_0,X_1,\dots,X_{i-1}]$
$= E[X_i|X_{i-1}]$
$$= \frac{aX_{i-1}-1+c}{a-1+c} X_{i-1} + \frac{aX_{i-1}}{a-1+c}(1-X_{i-1})$$
$= X_{i-1}$ Polya壶中蓝球比例是鞅

HIT CS&E

鞅尾不等式

给定鞅
$X_0,X_1,\dots,X_i,\dots$

$\Pr[\ |X_n-X_0| >t\ ] =?$

- 鞅中随机变量偏离X_0超过阈值的概率
- 鞅中随机变量偏离数学期望超过阈值的概率

限制速度
火车跑得快，全靠车头带

HIT CS&E

Azuma不等式

Azuma不等式
如果鞅$X_0,X_1,X_2\dots$对$k\geq1$满足
$|X_k-X_{k-1}|\leq c_k$
则
$$\Pr\left[|X_n - X_0| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\sum_{k=1}^{n} c_k^2}\right)$$

对于随机变量序列，如果每步
- 从平均看，不会偏离当前的值(鞅)
- 取值不会有大的跳跃

则其最终取值不会偏离初始值太远

HIT CS&E

Azuma不等式

Azuma不等式

如果鞅$X_0,X_1,X_2\ldots$对$k\geq 1$满足

$$|X_k-X_{k-1}|\leq c_k$$

则

$$\Pr[|X_n-X_0|\geq t]\leq 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n c_k^2}\right)$$

推论：如果鞅$X_0,X_1,X_2\ldots$对$k\geq 1$满足

$$|X_k-X_{k-1}|\leq c$$

则

$$\Pr\left[|X_n-X_0|\geq ct\sqrt{n}\right]\leq 2e^{-t^2/2}$$

HIT CS&E

Azuma不等式证明思路

Azuma不等式

如果鞅$X_0,X_1,X_2\ldots$对$k\geq 1$满足

$$|X_k-X_{k-1}|\leq c_k$$

则

$$\Pr[|X_n-X_0|\geq t]\leq 2\exp\left(-\frac{t^2}{2\sum_{k=1}^n c_k^2}\right)$$

第1步：将总差表示为分步差之和

$Y_i=X_i-X_{i-1}$ $\qquad X_n-X_0=\sum_{i=1}^n Y_i$

第2步：将markov不等式应用到矩生成函数

$$\Pr[\textstyle\sum_{i=1}^n Y_i\geq t]=\Pr[e^{\lambda\sum_{i=1}^n Y_i}\geq e^{\lambda t}]\leq\frac{\mathrm{E}[e^{\lambda\sum_{i=1}^n Y_i}]}{e^{\lambda t}}$$

第3步：利用鞅的性质和矩生成函数的凸性质

HIT CS&E

例

推论：如果鞅$X_0,X_1,X_2\ldots$对$k\geq 1$满足

$$|X_k-X_{k-1}|\leq c$$

则

$$\Pr\left[|X_n-X_0|\geq ct\sqrt{n}\right]\leq 2e^{-t^2/2}$$

公平赌局

- 初始赌资X_0，X_i表示第i轮之后的赌资
- 每轮以1/2的概率赢，赢得的赌资是a
- 以1/2的概率输,输掉的赌资是a
- 每局的输赢是相互独立的

$|X_i-X_{i-1}|\leq a$ $\qquad \Pr[\,|X_n-X_0|>atn^{1/2}]\leq 2\exp(-t^2/2)$

公平赌局中，有很大的输赢是不太可能的！

HIT CS&E

推论：如果鞅$X_0,X_1,X_2\ldots$对$k\geq 1$满足

$$|X_k-X_{k-1}|\leq c$$

则

$$\Pr\left[|X_n-X_0|\geq ct\sqrt{n}\right]\leq 2e^{-t^2/2}$$

均匀硬币投掷n次

- $X_0=0$
- X_i = HEADs$_i$-TAILs$_i$
- HEADs$_i$—i次投掷后头面向上的总次数
- TAILs$_i$ —i次投掷后背面向上的总次数

$|X_i-X_{i-1}|\leq 1$ $\qquad \Pr[\,|X_n|>tn^{1/2}]\leq 2\exp(-t^2/2)$

输赢次数之差的很大也是不太可能的！

HIT CS&E

5.2 鞅的一般形式

停时定理
瓦尔德方程
鞅尾不等式

HIT CS&E

鞅

定义

$Y_0,Y_1,Y_2,\ldots$称为随机变量序列 $X_0,X_1,X_2,\ldots$的鞅

如果

- Y_i是$X_0,X_1,X_2,\ldots X_i$的函数 $\qquad\forall i\geq 1$
- $\mathrm{E}[Y_i|X_0,X_1,\ldots,X_{i-1}]=Y_{i-1}$ $\qquad\forall i\geq 1$

HIT CS&E

例1

随机变量的和

设独立随机变量序列 $X_1, X_2, \ldots$满足$E[X_i]=0\ (\forall i\geq 1)$

定义

$$Y_i = X_1+X_2+\ldots+X_i \qquad \forall i\geq 1$$

$$
\begin{aligned}
&E[Y_i\,|X_1,\ldots,X_{i-1}] \\
&=E[X_i+Y_{i-1}\,|X_1,\ldots,X_{i-1}] \\
&=E[X_i|X_1,\ldots,X_{i-1}] + E[Y_{i-1}\,|X_1,\ldots,X_{i-1}] \\
&=E[X_i] + E[Y_{i-1}\,|X_1,\ldots,X_{i-1}] \\
&=0+Y_{i-1} \\
&=Y_{i-1}
\end{aligned}
$$

均值为0的随机变量之和是一个鞅

HIT CS&E

例2

随机变量和的平方

设$X_1, X_2, \ldots$是均值为0，方差为σ^2的独立随机变量

定义

$$Y_i = [X_1+X_2+\ldots+X_i\,]^2 - i\sigma^2 \qquad \forall i\geq 1$$

$$
\begin{aligned}
&E[Y_i\,|X_1,\ldots,X_{i-1}] \\
&=E[X_i^2+2X_i(\sum_{k=1}^{i-1}X_k)+(\sum_{k=1}^{i-1}X_k)^2 - i\sigma^2\,|X_1,\ldots,X_{i-1}] \\
&=E[X_i^2-\sigma^2|X_1,\ldots,X_{i-1}] + 2E[X_i\,(\sum_{k=1}^{i-1}X_k)\,|X_1,\ldots,X_{i-1}]+E[Y_{i-1}\,|X_1,\ldots,X_{i-1}] \\
&=(E[X_i])^2 + 2E[X_i]E[(\sum_{k=1}^{i-1}X_k)|X_1,\ldots,X_{i-1}] + E[Y_{i-1}\,|X_0,X_1,\ldots,X_{i-1}] \\
&=Y_{i-1}
\end{aligned}
$$

均值为0的随机变量和的平方是一个鞅

HIT CS&E

杜比鞅

随机变量函数的Doob序列

设$f(X_1, X_2,\ldots X_n)$是随机变量$X_1, X_2,\ldots X_n$的函数，定义

$$Y_i = E[f(X_1,X_2,\ldots,X_n)|\ X_1, X_2,\ldots X_i\,] \quad \forall i=0,1,\ldots,n$$

为函数$f(X_1, X_2,\ldots X_n)$的Doob序列

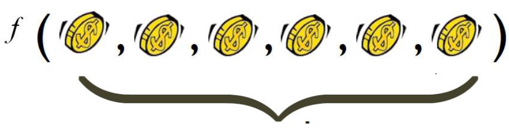
$$
\begin{aligned}
&E[Y_i\,|X_1,\ldots,X_{i-1}] \\
&= E[\ E[f(X_1,X_2,\ldots,X_n)|\ X_1, X_2,\ldots X_i\ |X_1,\ldots,X_{i-1}] \\
&= E[\ E[f(X_1,X_2,\ldots,X_n)\ \ |X_1,\ldots,X_{i-1}] \\
&=Y_{i-1}
\end{aligned}
$$

性质 $E[Y|\,Z] = E[\,E[Y\,|X,Z]\,|\,Z\,]$

Doob序列是一个鞅

HIT CS&E

$$Y_0 = E[\,f(X_1,\ldots,X_n)\,]$$



取均值

HIT CS&E

$$Y_1 = E[\,f(X_1,\ldots,X_n)\,|\,X_1\,]$$

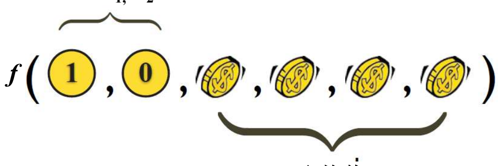
取定X_1的值

取均值

HIT CS&E

$$Y_2 = E[\,f(X_1,\ldots,X_n)\,|\,X_1, X_2\,]$$

取定X_1, X_2的值

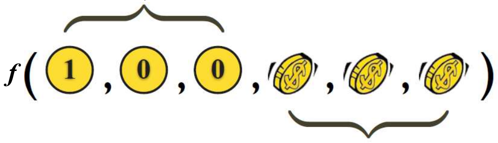


取均值

HIT CS&E

$Y_3 = \mathrm{E}[\, f(X_1,\ldots,X_n) \mid X_1,X_2,X_3\,]$

取定X_1, X_2,X_3的值

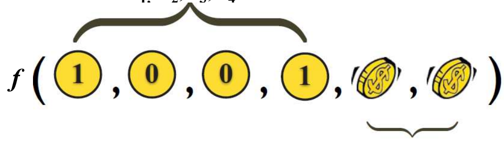


取均值

HIT CS&E

$Y_4 = \mathrm{E}[\, f(X_1,\ldots,X_n) \mid X_1,X_2,X_3,X_4\,]$

取定X_1, X_2,X_3,X_4的

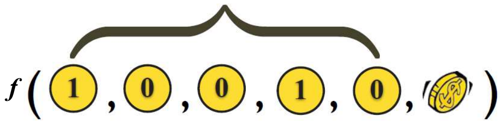


取均值

HIT CS&E

$Y_5 = \mathrm{E}[\, f(X_1,\ldots,X_n) \mid X_1,X_2,X_3,X_4,X_5\,]$

取定X_1, X_2,X_3, X_4,X_5的值

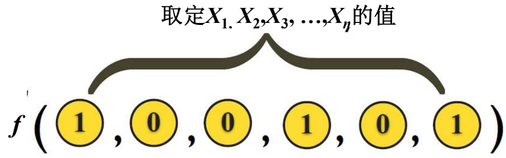


取均值

HIT CS&E

$Y_n = \mathrm{E}[\, f(X_1,\ldots,X_n) \mid X_1,X_2,X_3,\ldots,X_n\,]$

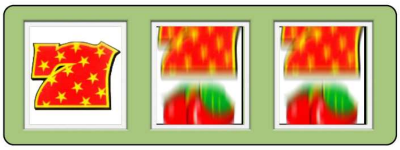
取定$X_1, X_2,X_3, \ldots,X_n$的值



$f(X_1,\ldots,X_n)$的一个具体取值

$Y_n - Y_0 = f(X_1,\ldots,X_n) - \mathrm{E}[f(X_1,\ldots,X_n)]$

HIT CS&E



Y_1 = 看到第1个示数之后的平均收益

HIT CS&E

鞅的性质

定理： 如果$Y_0,Y_1,Y_2,\ldots$是随机变量序列 $X_0,X_1,X_2,\ldots$的**鞅**，则 $\mathrm{E}[Y_i]=\mathrm{E}[Y_0]$

证明： 因为$Y_0,Y_1,Y_2,\ldots$是随机变量序列 $X_0,X_1,X_2,\ldots$的鞅，

$$Y_{i-1} = \mathrm{E}[Y_i \mid Y_0,Y_1,\ldots,Y_{i-1}]$$

两端同时求数学期望得，

$$\mathrm{E}[Y_{i-1}] = \mathrm{E}[\, \mathrm{E}[Y_i \mid Y_0,Y_1,\ldots,Y_{i-1}]\,] = \mathrm{E}[Y_i]$$

$\mathrm{E}[Y_i] = \mathrm{E}[Y_{i-1}] = \ldots = \mathrm{E}[Y_0]$

HIT CS&E

鞅的停时定理

定义：设$Y_0,Y_1,Y_2,\ldots$是随机变量序列 $X_0,X_1,X_2,\ldots$的鞅，如果随机变量$T=n$仅依赖于$Y_0,Y_1,Y_2,\ldots,Y_n$的取值 则称T是鞅$\{Y_i|i\geq 0\}$的一个停时

定理（鞅的停时定理）：设$Y_0,Y_1,Y_2,\ldots$是随机变量序列$X_0,X_1,X_2,\ldots$的鞅，T是鞅$\{Y_i|i\geq 0\}$的一个停时，如果T是有限的，则$E[Y_T]=E[Y_0]$

HIT CS&E

应用1

公平赌局

- 初始赌资$X_0=b$，X_i表示第i轮之后的赌资
- 每轮以1/2的概率赢，赢得的赌资是a
- 以1/2的概率输，输掉的赌资是a
- 每局的输赢是相互独立的
- 玩家赢得L_1或者输掉L_2之后停止游戏
- 问：玩家赢得L_1并停止游戏的概率q是多大？

游戏停止时间记为T

$T=n \Leftrightarrow X_i\text{-}X_0=L_1$ $(X_i\geq X_0)$或$X_0\text{-}X_i=L_2$ $(X_i<X_0)$在第n轮首次发生

T是鞅$\{X_i\}$的停时 $E[X_T] = E[X_0] = b$

$(b+L_1)q+ (b-L_2)(1-q)= b$

$q= L_2/(L_1+L_2)$

HIT CS&E

应用2

选举定理

- 两人竞选，A得a票，B得b票 $(a>b)$
- 计票过程是$a+b$张选票的所有排列中均匀独立选取的
- 问：计票过程中A始终领先B的概率有多大？

$n=a+b$

S_i=统计i张选票之后A领先B的票数 $S_n=a-b$

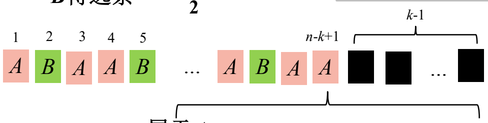
HIT CS&E

A得选票 $\dfrac{n\text{-}k+1+S_{n\text{-}k+1}}{2}$

B得选票 $\dfrac{n\text{-}k+1\text{-}S_{n\text{-}k+1}}{2}$

A得x票 $x+y= n\text{-}k+1$

B得y票 $x\text{-}y = S_{n\text{-}k+1}$



属于A

$S_{n\text{-}k}=S_{n\text{-}k+1}\text{-}1$

概率 $\dfrac{n\text{-}k+1+S_{n\text{-}k+1}}{2(n\text{-}k+1)}$

属于B

$S_{n\text{-}k}=S_{n\text{-}k+1}+1$

概率 $\dfrac{n\text{-}k+1\text{-}S_{n\text{-}k+1}}{2(n\text{-}k+1)}$

$E[S_{n\text{-}k}| S_{n\text{-}k+1}] = S_{n\text{-}k+1}\dfrac{n\text{-}k}{n\text{-}k+1}$

$E[\dfrac{S_{n\text{-}k}}{n\text{-}k} | \dfrac{S_{n\text{-}k+1}}{n\text{-}k+1}]= \dfrac{S_{n\text{-}k+1}}{n\text{-}k+1}$

HIT CS&E

选举定理

- 两人竞选，A得a票，B得b票 $(a>b)$
- 计票过程是a+b张选票的所有排列中均匀独立选取的
- 问：计票过程中A始终领先B的概率有多大？

$n=a+b$

S_i=统计i张选票之后A领先B的票数 $S_n=a-b$

$E[\dfrac{S_{n\text{-}k}}{n\text{-}k} | \dfrac{S_{n\text{-}k+1}}{n\text{-}k+1}]= \dfrac{S_{n\text{-}k+1}}{n\text{-}k+1}$

$0\leq k \leq n\text{-}1$

$X_k = \dfrac{S_{n\text{-}k}}{n\text{-}k}$

$E[X_k| X_0,X_1,\ldots,X_{k\text{-}1}] = X_{k\text{-}1}$

$X_0,X_1,\ldots,X_{n\text{-}1}$是鞅

HIT CS&E

选举定理

- 两人竞选，A得a票，B得b票 $(a>b)$
- 计票过程是a+b张选票的所有排列中均匀独立选取的
- 问：计票过程中A始终领先B的概率有多大？

$n=a+b$

S_i=统计i张选票之后A领先B的票数 $S_n=a-b$

$X_k = \dfrac{S_{n\text{-}k}}{n\text{-}k}$ $X_0,X_1,\ldots,X_{n\text{-}1}$是鞅

T=k 是满足X_k=0的最小k,如果存在这样的k值

T=n-1 如果不存在上述k值

T=i 仅依赖于$X_0,\ldots,X_i$的取值，故$E[X_T]=E[X_0]= (a\text{-}b)/(a+b)$

第一种停时的特征：T<n-1

存在k使得X_k=0, T就是这种k值的最小值。 故X_T=0

第二种停时的特征：T=n-1

不存在k使得X_k=0 且X_0=(a-b)/n>0, 故X_i>0恒成立

$X_k = \frac{S_{n-k}}{n-k}$

$S_1,S_2,\ldots$>0恒成立

A得第1票并一直保持领先

$X_{n-1}=X_T=S_1=1$

选举定理

- 两人竞选，A得a票，B得b票 (a>b)
- 计票过程是a+b张选票的所有排列中均匀独立选取的
- 问：计票过程中A始终领先B的概率有多大？$\frac{a-b}{a+b}$

n=a+b

S_i=统计i张选票之后A领先B的票数 S_n=a-b

$X_k = \frac{S_{n-k}}{n-k}$ $X_0,X_1,\ldots,X_{n-1}$是鞅

T=k 是满足X_k=0的最小k,如果存在这样的k值

T=n-1 如果不存在上述k值

T=i 仅依赖于$X_0,\ldots,X_i$的取值，故E[X_T]=E[X_0]= (a-b)/(a+b)

Pr(A一直领先) · 1 + [1- Pr(一直领先)]·0 = (a-b)/(a+b)

瓦尔德方程

定理(瓦尔德方程):设$X,X_1,X_2,\ldots$是独立同分布的随机变量，T是$\{X_i|i\geq 1\}$的一个停时, 如果E[T]和E[X]均是有限的，则

$$E[\sum_{i=1}^{T} X_i] = E[T]\cdot E[X]$$

$\{X_i-E[X_i] \mid i\geq 1\}$ 是均值为0的随机变量序列

$Y_i=\sum_{j=1}^{i}(X_j-E[X_j])$ $Y_1, Y_2,\ldots$是鞅 T是停时

$E[Y_T] = E[Y_1]= 0$ 停时定理

$E[Y_T] = E[\sum_{i=1}^{T} X_i - T\cdot E[X]] = E[\sum_{i=1}^{T} X_i] - E[T]\cdot E[X] = 0$

应用1

两轮骰子赌局

- 第一轮：投掷均匀骰子的点数X
- 第二轮：投掷均匀骰子X次得点数$Y_1,\ldots,Y_X$
- 玩家收益$Z= Y_1+\ldots+Y_X$
- 问：玩家平均收益E[Z]=?

$Y_1, Y_2,\ldots$独立同分布 $E[Y_i] = 7/2$

X是随机变量序列$Y_1, Y_2,\ldots$的停时 $E[X] = 7/2$

$E[Z] = E[\sum_{i=1}^{X} Y_i] = E[\sum_{i=1}^{X} Y_i] - E[X]\cdot E[Y_i] = 0$

$E[Z] = E[X]\cdot E[Y_i] = 49/4$

应用2：Las Vegas算法的期望运行时间

LAZYSELECT(S, k)

1. R=独立、均匀、可放回地从S随机选取的$n^{3/4}$元素;
2. 在$O(n)$时间内排序R;
3. $x=(k/n)n^{3/4}$;
4. l=max{ , 0}; h= min{ , $n^{3/4}$};
5. $L=min(R, l)$; $H=min(R, h)$;
6. $L_p=Rank(S,L)$, $H_p=Rank(S,H)$; (参见第2章)
7. $P=\{y\in S \mid L\leq y\leq H\}$;
8. If $min(S, k)\in P$ and $|P|\leq 4n^{3/4}+1$
9. Then 排序P, $min(S, k)=min(P, (k-L_p))$, 算法结束;
10. ELSE goto 1.

算法的期望运行时间是多少？

Las Vegas算法的期望运行时间

Las Vegas算法$A(x)$
期望运行时间为$T(n)$
找到解的概率为$p(n)$

算法$B(x)$
1. while(true)
2. $y=A(x)$
3. If y是问题的解 Then 返回y

Y_i是第i遍调用算法A的实际运行时间

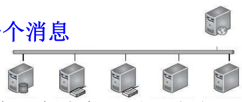
$Y_1,Y_2,\ldots$是均值为$T(n)$的独立同分布的随机变量

算法终止时刻是其停时，仅依赖于前面运行是否找到解

算法的期望运行时间 = 期望运行遍数 × $T(n)$ = $T(n)/p(n)$

应用3:共享总线服务器通信

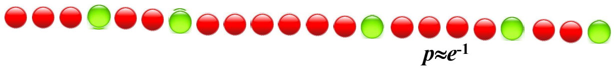
共享总线服务器通信
- n台服务器，共享总线，各自用缓存区缓存消息
- 时间分片，每个时间片段各服务器至多发送一个消息
- 每个时间片
  - 各服务器以$1/n$的概率发送一个消息
  - 有冲突则所有消息发送失败
  - 无冲突则消息发送成功
- 问：每台服务器都至少发送成功一个消息，需要多少时间片段？



在一个时间片段内，若只有一台服务器发消息，则成功

$$p = \binom{n}{1}\frac{1}{n}\left(1-\frac{1}{n}\right)^{n-1} \approx e^{-1}$$

参数为$p≈e^{-1}$的几何分布
每次成功前期望轮数为$1/p = e$



$p≈e^{-1}$

2 1 1 2 3

参数为n, $1/n$的赠券收集过程
n台机器均发消息成功
期望成功次数 = $n\ln n$

所有机器均至少成功发送1次消息前的时间片段数
=期望成功次数 × 每次成功前期望轮数
=$n\ln n \times e$
=$en\ln n$

请您构造随机变量序列，利用瓦尔德方程将分析过程精确化

Azuma-Hoeffding不等式

Azuma不等式
如果$Y_0,Y_1,Y_2\ldots$是随机变量序列$X_0,X_1,X_2\ldots$的鞅，且对$k\geq1$满足$|Y_k-Y_{k-1}|\leq c_k$，则

$$\Pr[|Y_n - Y_0| \geq t] \leq 2\exp\left(-\frac{t^2}{2\sum_{k=1}^{n} c_k^2}\right)$$

推论：如果$Y_0,Y_1,Y_2\ldots$是随机变量序列$X_0,X_1,X_2\ldots$的鞅且对$k\geq1$满足$|Y_k-Y_{k-1}|\leq c$
则

$$\Pr[|Y_n-Y_0| \geq tcn^{1/2}] \leq 2\exp\{-t^2/2\}$$

5.3 鞅的应用

5.3.1 模式匹配
5.3.2 球和箱子模型中空箱子个数
5.3.3 随机图的色数

5.3.1 模式匹配

模式匹配

模式匹配问题
- 输入：大小为s的字符集Σ上长度为n的串$X=(X_1,\ldots,X_n)$
  Σ上长度为k的模式串$B=(B_1,\ldots,B_k)$
- 输出：B在X中的所有出现位置
- 问：这种匹配“有意义”吗？

大量的DNA片段看上去像随机片段
不包含任何特殊的信息
如果B是这种随机片段，模式匹配将毫无意义

B有意义⇔它在随机选定的DNA序列的频率显著地偏离均值

那么一个固定的片段在随机DNA中的平均频率是多少呢？

HIT CS&E

模式的期望频率

模式的期望频率

- 固定大小为s的字符集Σ
- 固定Σ上长度为k的模式串$B=(B_1,\ldots,B_k)$
- 随机取长度为n的串$X=(X_1,\ldots,X_n)$ $(n>>k)$
- 问：B在X出现频率F的期望值是多少？

$E[F] = (n-k+1)(1/s)^k$

本节将证明：F高概率地集中分布于$E[F]$附近

这意味着：F频率偏离$E[F]$后，模式B才是有意义的

HIT CS&E

$E[F] = (n-k+1)(1/s)^k$ $X_1,X_2,\ldots,X_i,X_{i+1}$

在B中的对齐位置

$Z_0=E[F]$

$Z_i=E[F \mid X_1,X_2,\ldots,X_i]$ $1\le i \le n$

$Z_n=F$ 即B在X中的出现次数

$|Z_{i+1}-Z_i|\le k$ 即 $|E[F|X_1,\ldots,X_{i+1}]-E[F|X_1,\ldots,X_i]|\le k$

推论：如果$Y_0,Y_1,Y_2\ldots$是随机变量序列$X_0,X_1,X_2\ldots$的鞅且对$k\ge1$满足$|Y_k-Y_{k-1}|\le c$
则 $\Pr[|Y_n-Y_0| \ge tcn^{1/2}] \le 2\exp\{-t^2/2\}$

$\Pr[|Z_n-Z_0|\ge \varepsilon] \le 2\exp\{-\varepsilon^2/(2nk^2)\}$

HIT CS&E

5.3.2 球和箱子模型中空箱子个数

HIT CS&E

重审球和箱子模型

m个球

均匀独立地将球投入箱子：空箱子有几个?

n个箱子

$$X_i = \begin{cases} 1 & \text{第}i\text{个箱子是空的} \\ 0 & \text{否则} \end{cases} \quad \Pr[X_i=1]=(1-1/n)^m$$

$$X=\sum_{i=1}^{n} X_i$$

$$E[X] = n(1-\frac{1}{n})^m$$

下面证明：X高概率地集中分布于$E[X]$附近

HIT CS&E

Y_i表示第i个球落入的箱子 $i=1,2,\ldots,n$

X表示所有球投掷完成后空箱子的个数

$Z_0=E[X]$

$Z_i=E[X \mid Y_1,Y_2,\ldots,Y_i]$ $1\le i \le n$ Doob鞅

$Z_n=X$ 即所有球投掷完成后空箱子的个数

前i个球落入的箱子确定后

如果第i+1个球落入空箱子，则$Z_{i+1}-Z_i=-1$

如果第i+1个球未落入空箱子，则$Z_{i+1}-Z_i=0$

$|Z_{i+1}-Z_i|\le 1$ $\Pr[|Z_n-Z_0| \ge\varepsilon]\le 2\exp\{-\varepsilon^2/(2n)\}$

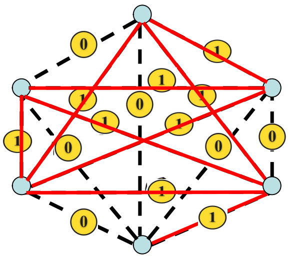
推论：如果$Y_0,Y_1,Y_2\ldots$是随机变量序列$X_0,X_1,X_2\ldots$的鞅且对$k\ge1$满足$|Y_k-Y_{k-1}|\le c$
则 $\Pr[|Y_n-Y_0| \ge tcn^{1/2}] \le 2\exp\{-t^2/2\}$

HIT CS&E

5.3.3 随机图的色数

HIT CS&E

随机图模型

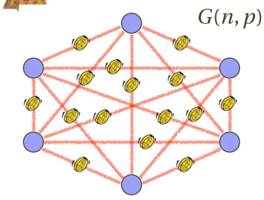


$G(n,p)$

- n个顶点
- 任意顶点对成边的概率为p

HIT CS&E

随机图的参数



随机图的参数$f(G)$

- 直径：最短路径的最大值
- 围长：最短环的边数
- 色数：真着色所需最少颜色数

Doob鞅
边暴露鞅
Edge Explosoure Martingle

估计$f(G)$集中性的一般方法

将所有可能的边依次编号为$1,2,\dots,\binom{n}{2}$，X_i表示第i边的取舍

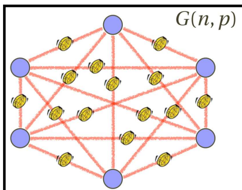
$Y_0 = E[f(G)]$

$Y_i = E[f(G) \mid X_1,\dots,X_i]$

$Y_{n(n-1)/2} = E[f(G) \mid X_1,\dots,X_i,\dots,X_{n(n-1)/2}] = f(G)$

$\Pr[\,|f(G)-E[f(G)]| > t] = \Pr[\,|Y_{n(n-1)/2}-Y_0| > t] = ?$

随机图的色数



随机图的色数$f(G)$

- 色数：真着色所需最少颜色数
- $\chi(G)$

将所有可能的边依次编号为$1,2,\dots,\binom{n}{2}$，X_i表示第i边的取舍

$$X_i = \begin{cases} 1 & 第i条边在G中 \\ 0 & 否则 \end{cases}$$

$Y_0 = E[\chi(G)]$

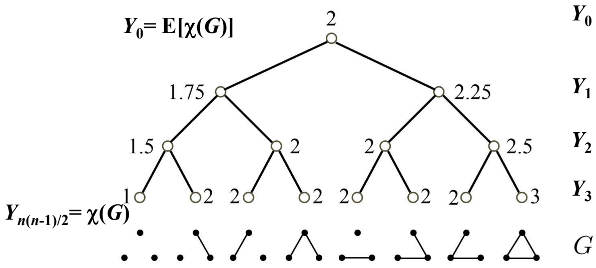
$Y_i = E[\chi(G) \mid X_1,\dots,X_i]$

$Y_{n(n-1)/2} = \chi(G)$

HIT CS&E

边暴露鞅示例

$Y_i = E[\chi(G) \mid X_1,\dots,X_i]$



容易直接验证： $E[Y_{i+1} \mid X_1,\dots,X_i] = Y_i$

HIT CS&E

$X_1,\dots,X_i$取定之后得G_i

在G_i上根据X_{i+1}添加第$i+1$条边

情形1: $X_{i+1}=0$, $\chi(G_{i+1}) = \chi(G_i)$

情形2: $X_{i+1}=1$, 第$i+1$条边连接两个无边顶点

$\chi(G_{i+1}) \le \chi(G_i) + 1$

情形3: $X_{i+1}=1$, 第$i+1$条边连接无边顶点和有边顶点

$\chi(G_{i+1}) \le \chi(G_i) + 1$

情形4: $X_{i+1}=1$, 第$i+1$条边连接两个有边顶点

$\chi(G_{i+1}) \le \chi(G_i) + 1$

$|Y_{i+1}-Y_i| \le 1$

HIT CS&E

$Y_0 = E[\chi(G)]$

$Y_i = E[\chi(G) \mid X_1,\dots,X_i]$

$|Y_{i+1}-Y_i| \le 1$

$Y_{n(n-1)/2} = \chi(G)$

推论：如果$Y_0, Y_1, Y_2\dots$是随机变量序列$X_0, X_1, X_2\dots$的鞅且对$k \ge 1$满足$|Y_k - Y_{k-1}| \le c$
则 $\Pr[|Y_n - Y_0| \ge tcn^{1/2}] \le 2\exp\{-t^2/2\}$

$\Pr[|\chi(G) - E[\chi(G)]| \ge t(n(n-1)/2)^{1/2}] \le 2\exp\{-t^2/2\}$