考试题型

简答、判断、算法设计、算法分析

第1章 绪论

随机算法的概念

计算：给定计算模型上的可以机械执行的一系列操作步骤

算法：满足确定性、准确性、终止性且具有输入和输出的计算

随机算法：利用概率和统计方法确定算法某些执行步骤的算法

随机算法的特点：优越性（算法简单、时间复杂性低），随机性（同一实例上多次执行，效果可能完全不同）

minHash算法

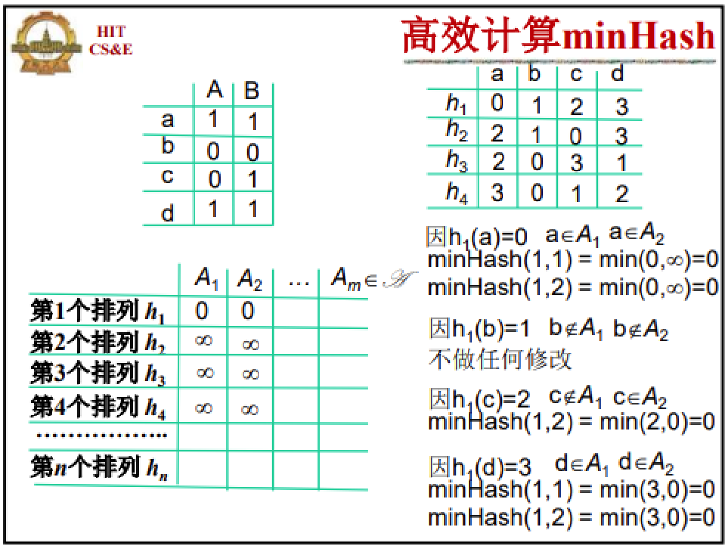
$sim(A,B) = 同时取1的行数/两列之一取1的行数$

$minHash_P(A) = 全集的随机排列P中首个属于A的行$

$Pr[minHash_P(A) = minHash_P(B)] = sim(A,B)$

$sim(A,B) \approx AB两列相等行数/n$

$minHash$的重要特征$sim(A,B)$越大，$minHash$取相同值的概率越高



第2章 随机算法及其分类

概念

样本空间、事件集合、概率测度、事件、概率

容斥原理：
$|A \cup B \cup C| = |A| + |B| + |C| - (|A \cap B| + |A \cap C| + |B \cap C|) + |A \cap B \cap C|$

union bound： $\Pr\left[\bigvee_{1 \le i \le n} \mathcal{E}_i\right] \le \sum_{i=1}^{n} \Pr[\mathcal{E}_i]$

适用于任何依赖关系！

条件概率： 对任意$\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$有： $\Pr\left[\bigwedge_{i=1}^{n} \mathcal{E}_i\right] = \prod_{k=1}^{n} \Pr\left[\mathcal{E}_k \mid \bigwedge_{i<k} \mathcal{E}_i\right]$

全概率公式： 若Ω被划分为$\mathcal{E}_1, \mathcal{E}_2, \ldots, \mathcal{E}_n$（即$\mathcal{E}_i \wedge \mathcal{E}_j = \varnothing$, $\vee_i \mathcal{E}_i = \Omega$），则

$$\Pr[\mathcal{E}] = \sum_{i=1}^{n} \Pr[\mathcal{E} \wedge \mathcal{E}_i] = \sum_{i=1}^{n} \Pr[\mathcal{E} \mid \mathcal{E}_i] \cdot \Pr[\mathcal{E}_i]$$

对任意$\mathcal{E}$成立

概率空间、随机变量

随机变量独立：$Pr[X = x \wedge Y = y] = Pr[X = x]Pr[Y = y]$

数学期望：具有线性性质 $\mathbf{E}[X+Y] = \mathbf{E}[X] + \mathbf{E}[Y]$

$\mathbf{E}[cX] = c\mathbf{E}[X]$

markov不等式： 对于任意非负随机变量X,

$$\Pr[X \ge t] \le \frac{\mathbf{E}[X]}{t}$$

对任意$t>0$成立

方差：$Var[x] = E[(X - E[x])^2] = E[X^2] - (E[X])^2$

二项分布：期望p，方差p(1-p)

Chebyshev不等式： 对任意随机变量X,

$$\Pr[\,|X - E[X]| \ge t\,] \le \frac{\mathrm{Var}[X]}{t^2}$$

对任意$t>0$成立

尾概率界： Tail Bound: $\Pr[\,X > t\,] < \varepsilon$

将n个球放进n个箱子：

$\Pr[$ 第一个箱子内球个数 $> t\,]$ $\le \binom{n}{t}\left(\frac{1}{n}\right)^t$ $\le (\frac{e}{t})^t$

数值随机算法

计算pi值

计算定积分：

$$E(g^*(\xi_i)) = \int_a^b g^*(x) f(x) dx = \int_a^b g(x) dx = I$$

– 令 $f(x)=1/(b-a)$ $\quad a \le x \le b$

– 索求积分可以由如下I'来近似计算I

– 由强大数定律 $\Pr\left(\lim_{n \to \infty} \frac{1}{n} \sum_{i=1}^{n} g^*(\xi_i) = I\right) = 1$

$$I' = \frac{1}{n} \sum_{i=1}^{n} g^*(\xi_i) = \frac{1}{n} \sum_{i=1}^{n} g(\xi_i) / f(\xi_i) = \frac{1}{n} \sum_{i=1}^{n} (b-a) g(\xi_i)$$

随机选择与拉斯维加斯算法

LAZYSELECT算法

拉斯维加斯Las Vegas算法：

- 算法不会产生不正确的解
- 算法一旦得到问题的解，就是正确的
- 得到解的概率p>0
- 算法运行过程可能不能产生问题的解
- 反复运行算法，运行时间不确定，最终可以产生问题的解
- 一般用来刻画yes or no 型问题

素数测试与蒙特卡洛算法

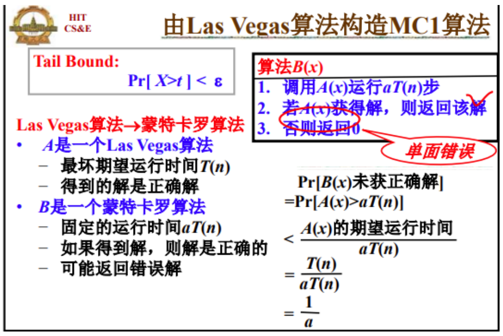
素数测试算法：

给定待测数字N，测试数据$\{1 < b_i < N\}$，对于$\forall b_i$，若满足$b_i^{N-1} \neq 1 mod N$，则b_i是一个合数，若均不满足，也不能说明N就一定是素数

蒙特卡洛算法：

- 用于刻画yes or no型计算问题
- 运行时间是固定的
- 算法得到正确解的概率p>0
- 算法得到错误解的概率1-p>0
- 单面错误蒙特卡洛算法 MC1算法
  - 算法输出yes结论可靠
  - 算法输出no结论可能是错的
- 双面犯错蒙特卡洛算法
  - 算法输出yes和no都可能是错的

由拉斯维加斯算法构造MC1算法



运行aT(n)次拉斯维加斯算法A，若拉斯维加斯算法有解，则返回

随机排序与舍伍德算法

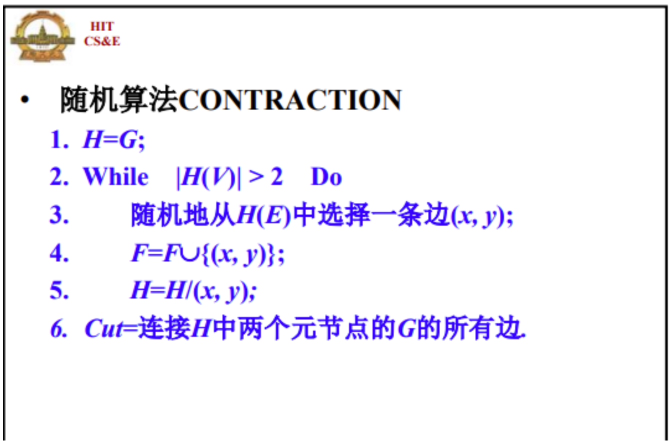
随机排序：期望时间复杂度为O(nlogn)，类似快速排序pivot的思想

舍伍德算法：

- 确定性算法的随机化
- 消除算法在最好实例和最坏实例之间的差别
- 总能找到问题的正确解

最小割与概率放大技术

割：图G的cut是一组边，从G中删除这组边将导致两个或多个连通分量



随机将两个顶点收缩到一起，直到图中只剩两个节点集为止，输出这两个节点集之间的边

概率放大技术：关键操作重复策略

master定理： $T(n) = aT(\frac{n}{b}) + f(n)$

1. 若函数 $n^{log_b a}$ 更大，如情况1，则 $T(n) = \Theta(n^{log_b a})$；
2. 若函数 $f(n)$ 更大，且满足 $af(n/b) \le cf(n)$,如情况3，则 $T(n) = \Theta(f(n))$；
3. 若两函数相等，则 $T(n) = \Theta(n^{log_b a} log^{k+1} n)$

第3章 球和箱子模型

两点分布，几何分布，二项分布

两点分布：$E[X] = p, Var[X] = p(1-p)$

几何分布：$P[X=k] = (1-p)^{k-1}p, Pr[X=n+k|X>k] = Pr[X=n], E[X] = \frac{1}{p}, Var[X] = \frac{1-p}{p^2}$

二项分布：$X = \sum_{i=1}^{n} X_i, E[X] = pn, Var[X] = np(1-p)$

桶排序及其时间复杂度分析

算法BucketSort(A)
Input:数组$A[0:n\text{-}1]$，$0 \le A[i] < 1$
Output: 排序后的数组A

```
for j ← 0 to n-1 do  // 初始化 n个桶
    B[j] ← NULL;
for i ← 0 to n-1 do
    将元素A[i]插入桶 B[⌊nA[i]⌋]中    //链表维护
for i ← 0 to n-1 do
    用InsertionSort排序桶B[i]内的数据
依编号递增顺序将各个桶内的数据回填到A中
```

$$E[\sum_{i=0}^{n-1} X_i^2/2\ ] = \sum_{i=0}^{n-1} E[X_i^2/2] = (2n\text{-}1)/2 = n\text{-}1/2$$ 期望的线性性质

InsertSort排序的最坏时间复杂度为$n^2/2=O(n^2)$

散列完成之后，桶内排序总时间的期望不超过n-1/2

收集排序结果的时间为$O(n)$

跳表及其复杂度分析

分析结论

应用到每种操作上

- 平均处理$E[r]=O(\log n)$层
- 每层平均进行E[Interval] =2次比较

操作的时间复杂度

- Find(x)的期望时间复杂度为$O(\log n)$
- Delete(x)的期望时间复杂度为$O(\log n)$
- Insert(x)的期望时间复杂度为$O(\log n)$

球与箱子模型

[M] -> [N]

单射，生日悖论，N个箱子，不存在含有2个球的箱子

满射，赠券收集，N个箱子均不为空

原像，最大负载，N个箱子球最多的有多少个球

将m个球均匀随机投入n个箱子

随机函数$f:[m]\rightarrow[n]$

单射	生日悖论
满射	赠券收集
原像	负载问题

为确保f是单射，至多投入m个球

$$m = \Theta(\sqrt{n})$$

为确保f是满射，至少投入m个球

$$m = n \ln n + O(n)$$

最大负载max $f^{-1}(x)$是

$$\begin{cases} O(\frac{\ln n}{\ln\ln n}) & \text{for } m = \Theta(n), \\ O(\frac{m}{n}) & \text{for } m = \Omega(n \ln n) \end{cases}$$

通用散列函数

相互独立

定义：事件的**相互独立性**

- 随机事件$E_1,E_2,\ldots,E_n$
- 对于任意$I\subseteq\{1,2,\ldots,n\}$均有

$$\Pr\left[\bigcap_{i\in I} E_i\right]=\prod_{i\in I}\Pr[E_i]$$

则称$E_1,E_2,\ldots,E_n$相互独立

定义：随机变量的**相互独立性**

- 随机变量$X_1,X_2,\ldots,X_n$
- 对于任意$I\subseteq\{1,2,\ldots,n\}$和任意$x_i$均有

$$\Pr\left[\bigcap_{i\in I}(X_i=x_i)\right]=\prod_{i\in I}\Pr[X_i=x_i]$$

则称$X_1,X_2,\ldots,X_n$相互独立

k独立

定义：事件的**k-独立性**

- 随机事件$E_1,E_2,\ldots,E_n$
- 对于任意$I\subseteq\{1,2,\ldots,n\}$, $|I|\le k$均有

$$\Pr\left[\bigcap_{i\in I} E_i\right]=\prod_{i\in I}\Pr[E_i]$$

则称$E_1,E_2,\ldots,E_n$是k-独立的

定义：随机变量的**k-独立性**

- 随机变量$X_1,X_2,\ldots,X_n$
- 对于任意$I\subseteq\{1,2,\ldots,n\}$ ($|I|\le k$) 和任意x_i均有

$$\Pr\left[\bigcap_{i\in I}(X_i=x_i)\right]=\prod_{i\in I}\Pr[X_i=x_i]$$

则称$X_1,X_2,\ldots,X_n$是k-独立的

两两独立

定义：事件的**两两独立性**

- 随机事件$E_1,E_2,\ldots,E_n$
- 对于任意E_i,E_j均有

$$\Pr[E_i\cap E_j]=\Pr[E_i]\cdot\Pr[E_j]$$

则称$E_1,E_2,\ldots,E_n$是两两独立的

定义：随机变量的**两两独立性**

- 随机变量$X_1,X_2,\ldots,X_n$
- 对于任意X_i,X_j和x_i,x_j均有

$$\Pr[(X_i=x_i)\cap(X_j=x_j)]=\Pr[X_i=x_i]\cdot\Pr[X_j=x_j]$$

则称$X_1,X_2,\ldots,X_n$是两两独立的

相互独立 > k独立 > 两两独立（推导关系反向则不成立）

素数模构造两两独立（均匀性、独立性）

- **定理**：设X_1,X_2是$[p]$(p是素数)上的均匀独立随机变量

  $$Y_i=X_1+iX_2 \bmod p \qquad i=0,1,2,\ldots,p-1$$

  则$Y_0,Y_1,\ldots,Y_{p-1}$是$[p]$上均匀的两两独立随机变量

k-通用散列函数族

- **定义**：集合$U\rightarrow\{0,1,2,\ \ldots,n-1\}$的一族函数$\mathcal{H}$满足：

  任意$x_1,x_2,\ldots,x_k\in\{1,2,\ldots,n\}$

  均匀随机选取的$h\in\mathcal{H}$

  $$\Pr[h(x_1)=h(x_2)=\ldots=h(x_k)]\le 1/n^{k-1}$$

  则称是一个**k-通用散列函数族**

  4

k-强通用散列函数族

- **定义**：集合$U\rightarrow\{0,1,2,\ \ldots,n-1\}$的一族函数$\mathcal{H}$满足：

  任意$x_1,x_2,\ldots,x_k\in U$

  任意$y_1,y_2,\ldots,y_k\in\{0,1,2,\ldots,n-1\}$

  均匀随机选取的$h\in\mathcal{H}$

  $$\Pr[(h(x_1)=y_1)\cap\ldots\cap(h(x_k)=y_k)]\le 1/n^k$$

  则称是一个**k-强通用散列函数族**

k-强通用蕴含k通用

综合应用

散列表

- 拉链技术

  拉链技术
  - 将哈希值相同的元素组织成链表
  - $Find(x)$—在$h(x)$对应的链表中查找x
    - 最好时间复杂度$O(1)$
    - 最坏时间复杂度$O(\ln n/\ln\ln n)$ $m=o(n\log n)$
    - 最坏时间复杂度$O(m/n)$ $m=\Omega(n\log n)$

  最大负载的结论

第4章 Chernoff界

切尔诺夫界以及常用形式

矩生成函数 $M(\lambda)=\mathrm{E}\left[e^{\lambda X}\right] = \sum_{k=0}^{\infty}\frac{\lambda^k}{k!}\mathrm{E}[X^k]$

矩生成函数的三点性质

- 两个随机变量的矩生成函数相同，则这两个随机变量相同
- 两个随机变量的各阶矩相同，则这两个随机变量相同
- 两个独立随机变量之和的矩生成函数等于这两个随机变量的矩生成函数之积

Chernoff界：

Chernoff界
定理：$X_1,\ldots,X_n$是独立泊松实验，$\Pr[X_i=1]=p_i$，$X=\sum_{i=1}^{n}X_i$，$\mu=\mathrm{E}[X]$，则对任意$\delta>0$有

$$\Pr[X\geq(1+\delta)\mu]<\left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu}$$

两个尾不等式

$$\Pr[X\geq(1+\delta)\mu]<\left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu}\leq e^{-\mu\delta^2/3}$$

$$\Pr[X\leq(1-\delta)\mu]<\left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^{\mu}\leq e^{-\mu\delta^2/2}$$

共四个应用：

Chernoff界
定理：$X_1,\ldots,X_n$是独立泊松实验，$\Pr[X_i=1]=p_i$，$X=\sum_{i=1}^{n}X_i$，$\mu=\mathrm{E}[X]$，则对任意$1>\delta>0$有

$$\Pr[X\leq(1-\delta)\mu]<e^{-\mu\delta^2/2}$$

$$\Pr[X\geq(1+\delta)\mu]<e^{-\mu\delta^2/3}$$

$$\Pr[|X-\mu|\geq\delta\mu]<2e^{-\mu\delta^2/3}$$

对任意$t>2e\mu$有

$$\Pr[X>t]<2^{-t}$$

$\Pr[\ p \notin [q-\delta,\ q+\delta]\ ] < \exp(-n\delta^2/3\ ) + \exp\{-n\delta^2/2\}$

参数估计

- 已知n,δ，可以计算置信水平
- 已知n,γ，可以计算δ，即置信区间
- 已知δ,γ，可以计算实验次数n

特殊情况

Chernoff界

定理：$X_1,\ldots,X_n$是独立随机变量，$\Pr[X_i=1]=\Pr[X_i=-1]=1/2$，$X=\sum_{i=1}^{n} X_i$，$\mu=E[X]$，则对任意$t>0$有

$$\Pr[\ X \ge t\ ] < e^{-t^2/2n}$$

Chernoff界

定理：$X_1,\ldots,X_n$是独立随机变量，$\Pr[X_i=1]=\Pr[X_i=-1]=1/2$，$X=\sum_{i=1}^{n} X_i$，$\mu=E[X]$，则对任意$t>0$有

$$\Pr[\ |X| \ge t\ ] < 2e^{-t^2/2n}$$

Chernoff界

定理：$X_1,\ldots,X_n$是独立随机变量，$\Pr[X_i=1]=\Pr[X_i=0\ ]=1/2$，$X=\sum_{i=1}^{n} X_i$，$\mu=E[X]=n/2$，则

对任意$t>0$有

$$\Pr[\ X \ge \mu+t\ ] < e^{-2t^2/n}$$

对任意$\mu>t>0$有

$$\Pr[\ X \le \mu-t\ ] < e^{-2t^2/n}$$

对任意$\delta>0$有

$$\Pr[\ X \ge (1+\delta)\mu\ ] < e^{-\delta^2\mu}$$

对任意$1>\delta>0$有

$$\Pr[\ X \le (1-\delta)\mu\ ] < e^{-\delta^2\mu}$$

集合平衡配置问题

定理：对于任意0-1矩阵$A_{n\times m}$和任意均匀随机独立选取的向量$x\in\{-1,+1\}^m$，有

$$\Pr[\ \|Ax\|_\infty > \sqrt{12m\ln n}\ ] < \frac{2}{n}$$

随机路由算法

Maurer不等式

定理：$X_1,\ldots,X_n$是独立的非负随机变量且$E[X_i^2]<\infty$
令$X=\sum_i X_i$，则对任意$t>0$有

$$\Pr[\, E[X]-X\geq t\,] < exp\{-\frac{t^2}{2\sum_i E[X_i^2]}\}$$

[1]Maurer. A. A Bound on the Deviation Probabilities for Sums of non-negative Random Variables. Journal of Inequalities in Pure and Applied Mathematics, 4, 2003.

Bernstein不等式

定理：$X_1,\ldots,X_n$是独立随机变量且$X_i-E[X_i]\leq M$对任意i成立
$\sigma_i^2=E^2[X_i]-E[X_i^2]$.令 $X=\sum_i X_i$ 则对任意$t>0$有

$$\Pr[\, X\geq E[X]+t\,] < exp\{-\frac{t^2}{2\sum_i \sigma_i^2+2Mt/3}\}$$

[2]Bernstain. S. Theory of Probability. Moscow, 1927

第5章 鞅

鞅的定义和基本性质

定义

随机变量序列 $X_0,X_1,X_2,\ldots$
如果

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

则

称$X_0,X_1,X_2,\ldots$是一个**鞅**

性质

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

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

鞅尾不等式

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

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

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

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

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

鞅的一般形式

定义

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 $\quad \forall i \geq 1$
- $E[Y_i \mid X_0, X_1, \ldots, X_{i-1}] = Y_{i-1} \quad \forall i \geq 1$

几种形式

- 均值为0的随机变量之和是一个鞅
  - $$Y_i = X_1 + X_2 + \ldots + X_i \qquad \forall i \geq 1$$

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

    均值为0的随机变量之和是一个鞅
- 均值为0的随机变量和的平方是一个鞅
  - $$Y_i = [X_1 + X_2 + \ldots + X_i]^2 - i\sigma^2 \qquad \forall i \geq 1$$

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

    均值为0的随机变量和的平方是一个鞅
- DOOB序列是一个鞅
  - 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
    $$Y_i = E[f(X_1, X_2, \ldots, X_n) \mid X_1, X_2, \ldots X_i] \quad \forall i = 0, 1, \ldots, n$$
    为函数 $f(X_1, X_2, \ldots X_n)$ 的Doob序列

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

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

    Doob序列是一个鞅
- 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 $E[Y_i]=E[Y_0]$

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 $E[Y_T]=E[Y_0]$

通俗解释

> **鞅的停时定理：** 设 T 是鞅过程 X_t 的停止时间，则当下面三个条件之一成立时，有 $E(X_T)=X_0$：
> 1. T 几乎一定有界；
> 2. 赌注 $|X_{t+1}-X_t|$ 一致有界，且 T 的期望有限；
> 3. 赌本 X_t 一致有界，且 T 几乎一定有限。

两种停时的特征

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

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

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

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

$\{X_i - E[X_i] \mid i \geq 1\}$ 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**

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

HIT
CS&E

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　　$E[X]=7/2$

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

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

HIT CS&E

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$k \geq 1$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$k \geq 1$满足$|Y_k - Y_{k-1}| \leq c$

则

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

鞅的应用

模式匹配

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$k \geq 1$满足$|Y_k - Y_{k-1}| \leq c$

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

空箱子个数

随机图的色数

第6章 随机抽样和随机舍入

随机游走

放缩 $q_j \geq \max_k \binom{j+2k}{k}\left(\frac{2}{3}\right)^k\left(\frac{1}{3}\right)^{j+k} \geq \binom{3j}{j}\left(\frac{2}{3}\right)^j\left(\frac{1}{3}\right)^{2j} \geq \frac{\sqrt{3}}{2\sqrt{\pi}}\frac{1}{\sqrt{j}}\frac{1}{2^j}$

收获之一：一种典型的随机算法设计过程

- 先处理简单的2SAT
- 推广过程简单算法去处理难解问题
- 设法克服推广过程中遇到的困难

收获之二：一种可能值得一般化的工具—随机游走

- 2SAT时用过
- 算法推广到3SAT时用过
- 改进推广的3SAT随机赋值算法时也用过

收获之三：工具的综合应用

- 基本概率计算
- 几何分布
- 马尔科夫不等式
- 概率放大
- 参数化设计

马尔可夫链

概率分布

定理：给定非二分无向连通图$G=(V=\{1,2,\ldots,n\},E)$, G上的随机游走的稳定分布是 $\pi_v = d(v)/2|E|$, $d(v)$表示顶点v的度

$\pi_v>0$ 且 $\Sigma_v \pi_v = 1$， 故π是一个概率分布

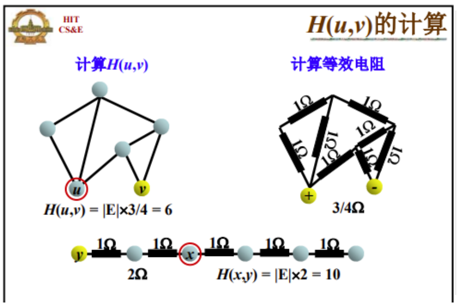
如果将随机游走的状态转移矩阵记为P，则$\pi = \pi P$

$$\pi_v = \sum_{u\in N(v)} \frac{d(u)}{2|E|}\frac{1}{d(u)} = \frac{d(v)}{2|E|}$$

从u出发首次访问v的期望时间$H(u,v)=\frac{2|E|}{d(v)}\leq 2|E|$

覆盖时间上界：找到一棵生成树，共2|V|-1条边，每条边最多2|E|时间内可以访问到，故随机游走的覆盖时间COVER(u)<4|E||V|

等效电阻法：



定理[Chandra et al. 1989 STOC]

将图$G(V,E)$上随机游走的从u到v的Hitting time记为$H(u,v)$

将图$G(V,E)$视为电路，输入点为u,输出点为v, $R(u,v)$表示两点间的电阻

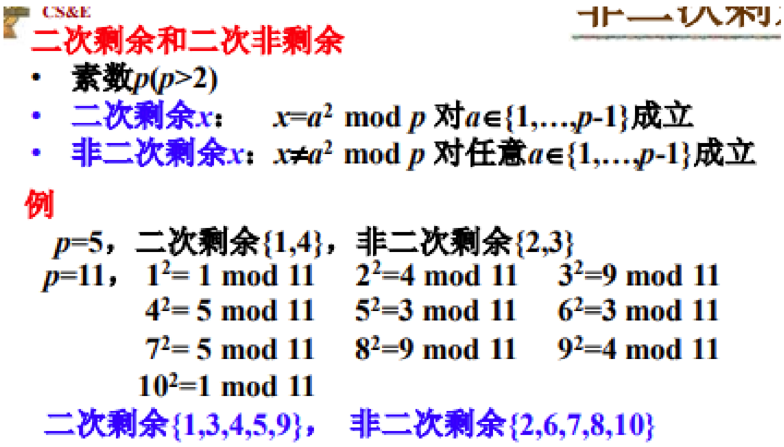
则有 $H(u,v)+H(v,u) = 2|E|R(u,v)$ (*)

由于 $H(u,v) = H(v,u)$

故 $H(u,v) = |E|R(u,v)$

基于随机抽样的算法

非二次剩余：



费马小定理：**费尔马小定理：** 若p是素数，则$a^{p-1}=1 \bmod p$对任意自然数a成立

二次剩余的判定方法：若x是二次剩余，则$x^{(p-1)/2} = 1 \bmod p$

非二次剩余的判定方法：若x是非二次剩余，则$x^{(p-1)/2} = -1 \bmod p$

水库抽样算法：

蒙特卡罗方法

定义

- 通过反复抽样完成计算的一大类算法
- 又称随机抽样方法或统计实验方法
- 用计算机实现的快速抽样和统计

缺点

- 计算结果存在统计误差
- 方法各要素需要仔细设计才能平衡统计误差和系统误差

步骤

- 构造或描述概率过程
  - 概率过程的数字特征与问题的解相关
  - 问题本身具有随机性，关键在于描述的准确性
  - 问题本身没有随机性，需要人为构造概率过程
- 实现从已知概率分布抽样

    - 随机数产生算法
    - 抽样质量决定方法是否有效
- 建立统计量作为问题的近似解
    - 无偏估计
    - 对实验结果进行考察、登记，得出问题的解

定理：如果$X_1,X_2,\dots,X_n$是独立同分布的示性变量，$E[X_i]=\mu$，

则 $n\ge\frac{3\ln(2/\delta)}{\varepsilon^2\mu}$ 时有

$$\Pr[\,|\frac{1}{n}\sum_{i=1}^{n}X_i-\mu|\le\varepsilon\mu\,]\ge 1-\delta$$

如上例，利用定理 容易建立样本数和近似程度之间的关系

如果随机算法的输出值X与问题的解V满足

$$\Pr[\,|X-V|\le\varepsilon V]\ge 1-\delta$$

则称该随机算法是一个(ε,δ)-近似

DNF满足性赋值计数问题

DNF满足性赋值计数问题

输入：文字$x_1,x_2,\dots,x_n$上DNF公式 $F=C_1\vee\dots\vee C_m$

输出：能使$C_1,\dots,C_m$之一被满足的$x_1,\dots,x_n$赋值的个数$c(F)$

1. $X=0$
2. For $k=1$ To N Do
3. 从$x_1,\dots,x_n$的2^n种可能赋值中均匀随机地抽取一个赋值
4. IF 所取赋值满足$C_1,\dots,C_m$中某个子句 Then $X=X+1$
5. 返回 $Y=(X/N)2^n$

算法实质：(1)用蒙特卡罗方法得到近似概率$X/N\approx c(F)/2^n$

(2)用近似概率乘以样本空间大小得到近似计数

$$X_i=\begin{cases}1 & \text{第}i\text{次抽取的随机样本满足}C_1,\dots,C_m\text{之一}\\ 0 & \text{否则}\end{cases}$$

$X_1,\dots,X_N$是独立同分布的两点分布 $\qquad X=\sum_{i=1}^{N}X_i$

$\Pr[X_i=1]=c(F)/2^n \quad E[X_i]=c(F)/2^n \quad E[X/N]=c(F)/2^n$

由Chernoff界可知

$\Pr[\,|X/N-c(F)/2^n|>\varepsilon\, c(F)/2^n\,]\le\delta \qquad N\ge 3\cdot 2^n\ln(2/\delta)/\varepsilon^2 c(F)$

$|X/N-c(F)/2^n|>\varepsilon\, c(F)/2^n \iff |Y-c(F)|>\varepsilon\, c(F)$

$\Pr[\,|Y-c(F)|>\varepsilon\, c(F)]\le\delta \qquad N\ge 3\cdot 2^n\ln(2/\delta)/\varepsilon^2 c(F)$

问题：抽样次数可能很大，尤其当$c(F)<<2^n$ 或 $c(F)=O(n^k)$

改造样本空间的必要性：

- 目标样本在样本空间内非常稀疏
- 需要很多次的抽样才能找到一个目标样本
- 在得到(ϵ,δ)近似需要海量的抽样次数

改造样本空间的方法：

- 找到样本空间的一个子空间，其大小易于计算
- 目标样本在子空间内稠密

- 实现子空间内的均匀抽样或根据已知分布抽样
- 建立(ϵ, δ)近似

第7章 概率方法与去随机化

概率论证法

引理: 设S是一个概率空间，X是S上的一个随机变量。
如果$E[X]=\mu$, 则$\Pr[X\geq\mu]>0$ 且 $\Pr[X\leq\mu]>0$.

两阶段概率论证

第一阶段

从概率空间抽样(样本不一定具有要求的性质)

第二阶段

修改样本使其具有要求的性质

在两阶段中结合期望论证得出结论

最大割问题

HIT CS&E

最大割的期望论证

概率空间

- 创建标记A,B
- $\forall v\in V$, 将v均匀随机地标记为A或B
- $S=\{v\in V \mid v$的标记为$A\}$ $V\text{-}S=\{v\in V \mid v$的标记为$B\}$

期望论证

- $\forall e\in E$, 端点标记相同的概率为1/2，不同的概率为1/2

$$X_e=\begin{cases}1 & e\text{的端点标记不同}\\ 0 & e\text{的端点标记相同}\end{cases}\qquad \Pr[X_e]=1/2 \quad E[X_e]=1/2$$

- $c(S)=\sum_{e\in E}X_e$
- $E[c(S)]=E[\sum_{e\in E}X_e]=\sum_{e\in E}E[X_e]=m/2$
- 存在大小至少为 $m/2$的割

最大割问题的Las Vegas算法

输入：连通图 $G=(V,E)$, 记$|E|=m$

输出：V的划分$S,V\text{-}S$使得介于S和$V\text{-}S$之间的边数$c(S)$最大

```
1. c←0, S←∅
2. For i=1 To m Do
3.     ∀v∈V, 以1/2的概率将v放入S_i
4.     c_i ←介于S_i和V-S_i之间的边条数
5.     If c_i>c Then c←c_i, S←S_i
6. 输出S,c
```

- 每一遍执行For循环，$c_i>m/2$的概率至少为$p\geq 1/(m/2+1)$
- $c>m/2$, 执行For循环的期望遍数为$1/p\leq m/2+1$
- 由Markov不等式可知
  For循环执行$m/2$遍， $c<m/2$的概率至多为1/2

作业

本章结束后，将该算法改造成确定型算法并进行分析

独立集算法

第一步：对顶点抽样

$\forall v \in V$，独立地以$1-1/d$的概率删除v及其邻边

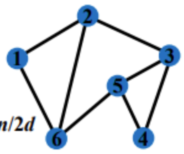
- 用X表示留下来的顶点个数
  X是随机变量

  $E[X] = n/d$　每个顶点以$1/d$概率留下

- 用Y表示留下来的边的数量
  Y也是随机变量

  边e留下 ⇔ e的端点均留下

  $E[Y] = m\cdot(1/d)^2 = (nd/2)\cdot(1/d)^2 = n/2d$

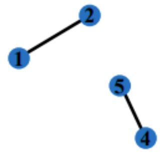


第二步：修改样本

对剩下的每条边，删除边及它的一个邻接顶点

- 最终剩下的顶点组成一个独立集
  相互之间没有边相连
- 最终剩下的顶点有X-Y个

  $E[X\text{-}Y] = E[X]\text{-}E[Y]$
  $= n/d - n/2d$
  $= n/2d$
  $= n^2/m$



二阶矩方法

定理：如果X是非负随机变量，则 $\Pr[X=0] \le \dfrac{E[X^2]}{(E[X])^2}$

证明：$\Pr[X=0] \le \Pr[\,|X-E[X]| \ge E[X]] \le \dfrac{E[X^2]}{(E[X])^2}$

定理：如果$X_i\,(i\ge 1)$ 是0-1随机变量且$X = \sum_{i=1}^{n} X_i$，则

$$\Pr[X>0] \ge \sum_{i=1}^{n} \frac{\Pr[X_i=1]}{E[X\,|X_i=1]}$$

证明思路：令
$$Y = \begin{cases} 1/X & X>0 \\ 0 & X=0 \end{cases}$$
则 $\Pr[X>0] = E[XY]$

然后根据$E[XY]$的定义式即可得出定理

lovasz局部引理

Lovasz Local Lemma

设$A_1,A_2,\ldots,A_n$是任意概率空间中的n个事件，这些事件的依赖图的度$\le d$，且$\Pr[A_i] \le p<1$对$i=1,2,\ldots,n$均成立。如果下列条件之一成立，则 $\Pr[\bigcap_{i=1}^{n} \overline{A_i}] > 0$

引理1. (Lovasz and Erdos 1973, 正式发表于1975)

$4pd<1$

引理2. (Lovasz 1977)

$ep(d+1)<1$

引理3. (Shearer 1985)

$p=1/2$　　$d=1$

或

$p < \dfrac{(d-1)^{d-1}}{d^d}$　　$d>1$

Lovasz Local Lemma

设$A_1,A_2,\ldots,A_n$是任意概率空间中的n个事件，这些事件的依赖图是有向图$G=(\{1,2,\ldots,n\},E)$。如果存在实数$x_1,x_2,\ldots x_n \in [0,1)$使得下式对$i=1,2,\ldots,n$均成立

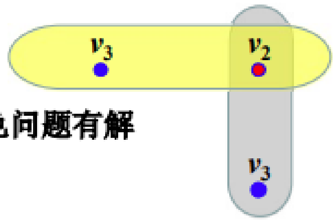
$$\Pr\left[\bigwedge_{i=1}^{n} \bar{A}_i\right] \le x_i \prod_{(i,j)\in E} (1-x_j)$$

则

$$\Pr\left[\bigwedge_{i=1}^{n} \bar{A}_i\right] \ge \prod_{i=1}^{n} (1-x_i)$$

推广定理

k-一致超图：每条边均包含k个顶点



k-一致超图的2-着色

边数少于2^{k-1}的k-一致超图的2-着色问题有解

推广(用Lovasz局部引理)

定理：设超图H的每条超边至少有k个顶点且每条超边至多与d条超边相交。如果$e(d+1) < 2^{k-1}$, 则H的2-着色问题有解。

算法式Lovasz引理

输入：一组随机变量$X_1,\ldots,X_n$

要避开的"坏"事件$A_1,\ldots,A_m$

输出：避开所有坏事件的随机变量取值

1. $x_1,x_2,\ldots,x_n \leftarrow X_1,\ldots,X_n$的一组随机赋值
2. while ($\exists i$: A_i发生) Do
3. 对A_i相关的随机变量重新随机赋值
4. 返回最终赋值

构造可满足的k-Sat的解

构造可满足性k-SAT的解

输入：变量$x_1,\ldots,x_n$上的CNF公式$F=C_1 \wedge C_2 \wedge \ldots \wedge C_m$

每个C_j是至多k个(否定)变量的或,即$|C_j| \le k$,

x_i,$\neg x_i$未同时出现在同一C_j中

每个布尔变量至多出现在$d \le 2^{k-2}$个子句中

输出：满足F的一组布尔变量赋值

1. $x_1,x_2,\ldots,x_n \leftarrow X_1,\ldots,X_n$的一组均匀随机赋值
2. while ($\exists j$: C_j未被满足) Do
3. 对C_j中的随机变量均匀随机地重新赋值
4. 返回最终赋值

去随机化

MAX-SAT随机算法去随机化

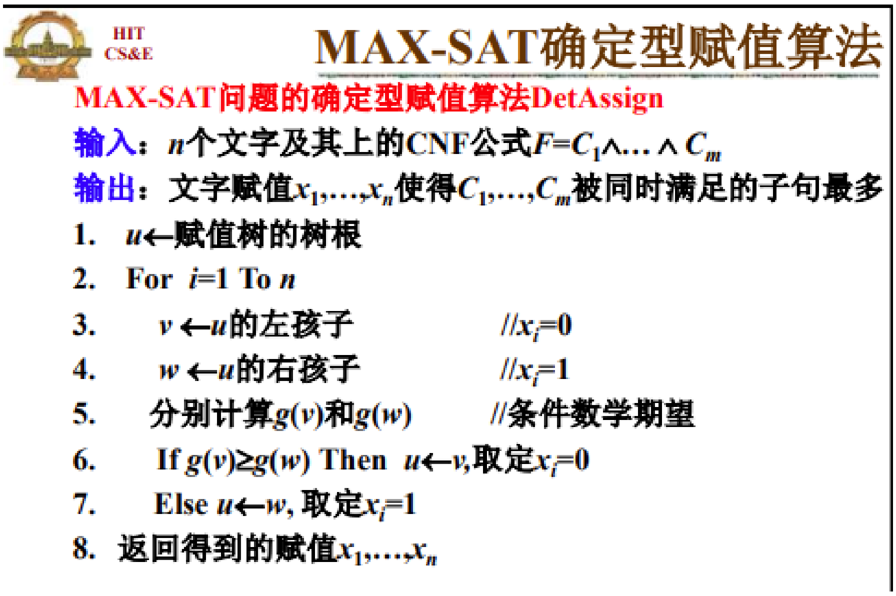
MAX-SAT问题的随机抽样算法RandSample

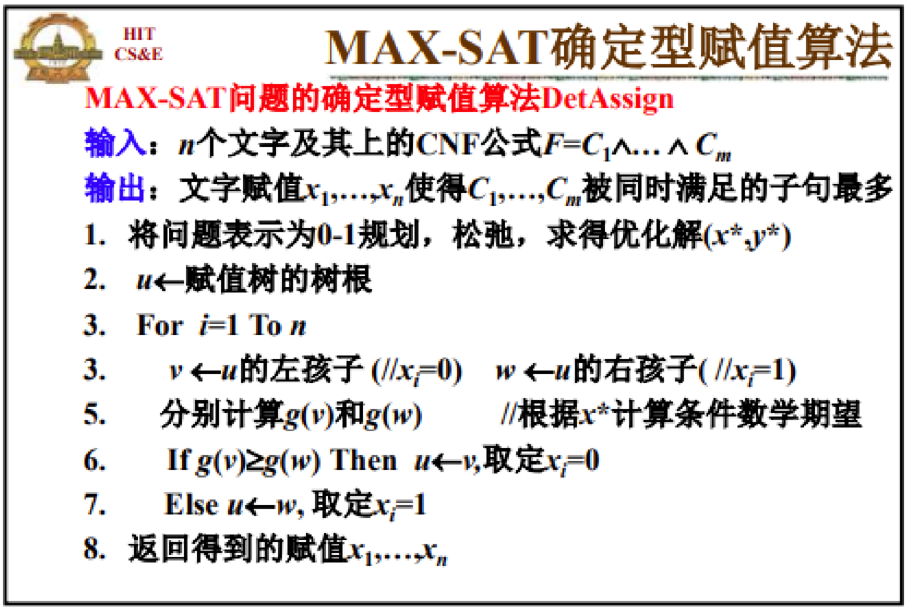
输入：n个文字及其上的CNF公式$F=C_1 \wedge \ldots \wedge C_m$

输出：文字赋值$x_1,\ldots,x_n$,使得$C_1,\ldots,C_m$被同时满足的子句最多

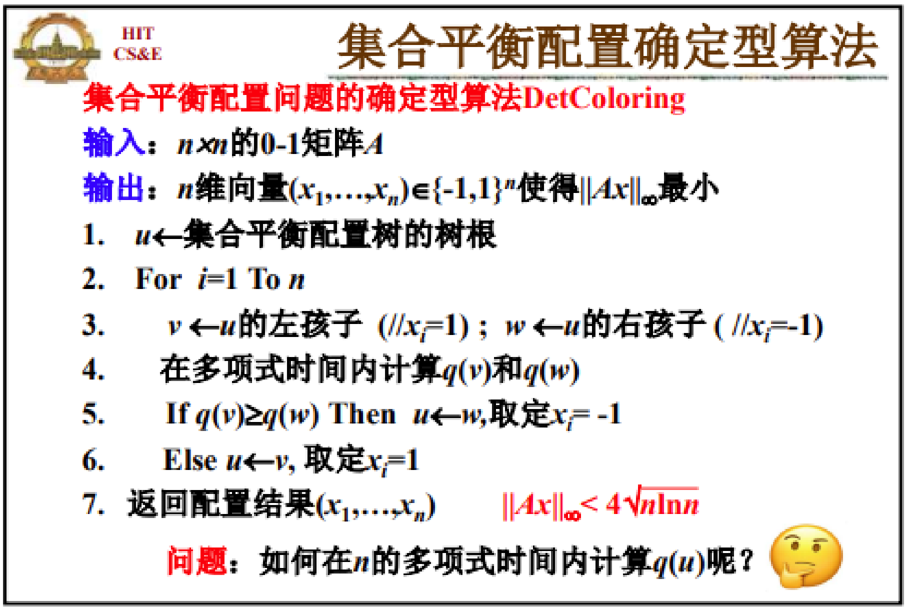
1. For i=1 To n Do
2. 第i个文字以概率1/2取真,以概率1/2取假
3. 返回1-2步得到的随机赋值

性能：$O(n)$时间E[2]-近似随机算法



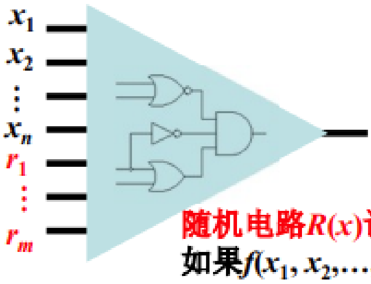


集合平衡配置随机算法去随机化



随机电路去随机化

随机电路R



$R(x_1,x_2,\ldots,x_n,r_1,r_2,\ldots,r_m)$

随机电路$R(x)$计算布尔函数f指的是

如果$f(x_1, x_2,\ldots, x_n)=1$，则

$$R(x_1, x_2,\ldots, x_n, r_1, r_2,\ldots, r_m) = 1$$

至少对$r=(r_1,r_2,\ldots,r_n)$的2^{m-1}个不同取值上成立

也就是说

$\Pr[R(x,r)=0] = 1$ 若$f(x)=0$

$\Pr[R(x,r)=1] \geq 1/2$ 若$f(x)=1$

*注：1/2可以是[1/2,1)中的任意数

去随机化过程

第一步：穷举所有x,r,列出$R(x,r)$的值

每列用$r=(r_1,\ldots,r_m)$的一组取值标定

2^m列

每行用$x=(x_1,\ldots,x_n)$的一组取值标定

2^n行

0	1	1	1	0	1	1	…	0
1	1	0	1	1	0	1	…	1
0	0	0	0	0	0	0	…	0
1	1	1	0	1	0	1	…	1
0	0	0	0	0	0	0	…	0
0	1	0	0	0	1	1	…	1
…	…	…	…	…	…	…	…	…
1	1	1	1	1	1	0	0	1

行x：恒等于0 $\Rightarrow f(x)=0$

行x：不恒等于0 $\Rightarrow f(x)=1$

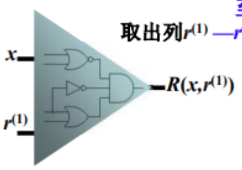
$\Rightarrow \Pr[D(x,r)=1] \geq 1/2$

$\Rightarrow$该行至少一半以上的元素等于1

第二步：删除全0行，剩下的矩阵一半以上元素为1

至少有一列$r^{(1)}$一半以上元素为1

取出列$r^{(1)}$ —$r^{(1)}$对应$r=(r_1,\ldots,r_m)$的一组具体取值



$D_1(x)=R(x,r^{(1)})$

2^m列

2^n行

0	1	1	1	$\boxed{0}$	1	1	…	0
1	1	0	1	$\boxed{1}$	0	1	…	1
~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~…~~	~~0~~
1	1	1	0	$\boxed{1}$	0	1	…	1
~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~…~~	~~0~~
0	1	0	0	$\boxed{0}$	1	1	…	1
…	…	…	…	…	…	…	…	…
1	1	1	1	$\boxed{1}$	1	0	0	1

一半以上的行不用再考虑

性质

在列$r^{(1)}$取1的所有行x上

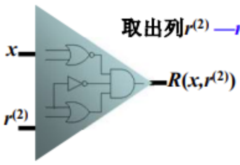
$D_1(x)$均能正确计算$f(x)$

HIT
CS&E

第三步：删除$r^{(1)}$中元素等于1的行，至多剩下一半的行

至少有一列$r^{(2)}$一半以上元素为1

取出列$r^{(2)}$ —$r^{(2)}$对应$r=(r_1,\ldots,r_m)$的一组具体取值



$D_2(x)=R(x,r^{(2)})$

2^m列

2^n行

0	1	1	1	0	1	1	…	0
~~1~~	~~1~~	~~0~~	~~1~~	~~1~~	~~0~~	~~1~~	~~…~~	~~1~~
~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~…~~	~~0~~
~~1~~	~~1~~	~~1~~	~~0~~	~~1~~	~~0~~	~~1~~	~~…~~	~~1~~
~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~0~~	~~…~~	~~0~~
0	1	0	0	0	1	1	…	1
…	…	…	…	…	…	…	…	…
~~1~~	~~1~~	~~1~~	~~1~~	~~1~~	~~1~~	~~0~~	~~0~~	~~1~~

又有一半以上的行不用再考虑

性质

在列$r^{(2)}$取1的所有行x上

$D_2(x)$均能正确计算$f(x)$

HIT
CS&E

重复第三步

- 每次至少删掉一半以上的行
- 经过k($k\le n$, $k\le m$)轮，所有行都将被删掉
- 得到$D_1(x),D_2(x),\ldots,D_k(x)$

重组得$D(x)$

- 将$D_1(x),D_2(x),\ldots,D_k(x)$用或门并联