高级算法设计

中国科学院计算技术研究所 孙晓明

2013-2014学年春季学期

目录

1 生日悖论(Birthday Paradox) 3

2 调查问卷的有效性 3

2.1 两个重要不等式 3

2.2 调查问卷的有效性 4

2.3 调查问卷的有效性(Cont'd) 4

3 随机算法验证矩阵乘法 5

4 随机算法的复杂度 6

5 Complexity Class 6

6 博弈论：Game Tree/AND OR Tree 9

7 Balls & Bins 10

7.1 基本知识 10

7.2 $m \sim \sqrt{n}$ 15

7.3 $m = n$ 16

7.4 总结 17

8 概率方法(Probability Method) 18

8.1 素数 18

8.2 Ramsey数 18

8.3 MAX-CUT 20

8.3.1 概率方法 20

8.3.2 2近似算法 21

8.4 独立集(Independent Set, IS) 22

8.5 AND OR Tree 22

8.6 MAX-SAT 23

8.7 25

9 代数化方法(Algebraic Method) 26

9.1 两数相等的判定 26

9.2 两个多项式相等的判定 27

9.3 Perfect Matching 27

9.4 交互式证明系统(Interactive Proof System) 28

10 随机游走(Random Walk) **29**
10.1 一般情形 29
10.2 将随机游走拓展到连通图上 30

11 电路问题 **31**

1 旅行商问题(TSP) **34**
1.1 W满足三角不等式TSP存在2近似 34
1.2 W满足三角不等式TSP存在$\frac{3}{2}$近似 35
1.3 一般情形下不存在常数近似 37

2 MAX-SAT **38**

3 Vertex Cover **40**
3.1 贪心算法 40
3.2 ILP & LP 41
3.3 极大perfect matching 41

第I部分：随机算法

1 生日悖论(Birthday Paradox)

问题描述：n个人，求解存在两个人同一天生日的概率。

求解之前先给出一个不等式：

$$e^x = 1 + x + \frac{x^2}{2!} + \cdots$$
$$e^{-x} = 1 - x + \frac{x^2}{2!} + \cdots$$

当$x \in [0,1]$时，$e^{-x} \geq 1 - x$.

$$\begin{aligned}
\Pr(E) &= 1 - \Pr\left(\overline{E}\right) \\
&= 1 - \prod_{i=1}^{n-1}\left(1 - \frac{i}{365}\right) \\
&\geq 1 - e^{\sum_{i=1}^{n-1}\left(-\frac{i}{365}\right)} \\
&= 1 - e^{-\frac{n(n-1)}{365\times 2}}.
\end{aligned}$$

当$n = 23$时存在两个人同一天生日的概率$\geq \frac{1}{2}$.

2 调查问卷的有效性

2.1 两个重要不等式

首先给出两个重要的不等式：马尔可夫不等式(Markov Inequality)和切比雪夫不等式(Chebyshev Inequality).

Markov Inequality: 随机变量$x \geq 0$, 常数$c > 0$, 则$Pr(x \geq c) \leq \frac{E(x)}{c}$.

证明：

$$\begin{aligned}
E(x) &= \sum aPr(x = a) \\
&= \sum_{a \geq c} aPr(x = a) + \sum_{a < c} aPr(x = a) \\
&\geq \sum_{a \geq c} aPr(x = a) \\
&\geq \sum_{a \geq c} cPr(x = a) \\
&= c\sum_{a \geq c} Pr(x = a) \\
&= cPr(x \geq c).
\end{aligned}$$

可得$Pr(x \geq c) \leq \frac{E(x)}{c}$.

Chebyshev Inequality: $Pr(|x - E(x)| \geq c) \leq \frac{Var(x)}{c^2}$.

证明：

$$\begin{aligned} Pr(|x - E(x)| \geq c) &= Pr((x - E(x))^2 \geq c^2) \\ &\leq \frac{E((x - E(x))^2)}{c^2} \\ &= \frac{Var(x)}{c^2}. \end{aligned}$$

2.2 调查问卷的有效性

通过发放问卷调查民众支持率，调查n个人得到支持率p, 为了保证调查问卷的有效性(p与其期望值$E(p)$之差的绝对值不超过5%这一事件的概率$\geq$ 95%)，n应满足什么条件。

$$\begin{aligned} \Pr\left(|p - E(p)| \leq 0.05\right) &= \Pr\left(\left|\frac{\sum x_i}{n} - E\left(\frac{\sum x_i}{n}\right)\right| \leq 0.05\right) \\ &= 1 - \Pr\left(\left|\frac{\sum x_i}{n} - E\left(\frac{\sum x_i}{n}\right)\right| > 0.05\right) \\ &\geq 1 - \frac{Var\left(\frac{\sum x_i}{n}\right)}{0.05^2} \\ &= 1 - \frac{Var\left(x_i\right)}{n0.05^2} \\ &= 1 - \frac{p(1-p)}{n0.05^2} \\ &\geq 1 - \frac{\frac{1}{4}}{n0.05^2} \\ &= 1 - \frac{100}{n} = 0.95. \end{aligned}$$

解得$n = 2000$, 注意到n的取值与总体人数无关。

2.3 调查问卷的有效性(Cont'd)

2.2节中的差值上限5%是绝对误差，当p的值很小时容易被5%(绝对误差)淹没。引入如下的度量方式：

$$Pr\left(\widehat{p} \in [0.99p, 1.01p]\right) \geq 1 - \varepsilon$$

其中$\widehat{p} = \frac{x_1 + x_2 + \cdots + x_n}{n}$.

$$\begin{aligned} Pr\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - p\right| \leq 0.01p\right) &= 1 - Pr\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - p\right| > 0.01p\right) \\ &= 1 - Pr\left(|x_1 + x_2 + \cdots + x_n - np| > 0.01np\right). \end{aligned} \tag{2.1}$$

令$\delta = 0.01$, 利用Chernoff's Bound可得

$$\begin{aligned}(2.1)\text{式} &\geq 1 - 2e^{-\frac{\delta^2}{3}np} \\ &\geq 1 - \varepsilon.\end{aligned}$$

即

$$2e^{-\frac{\delta^2}{3}np} \leq \varepsilon.$$

解得

$$n \geq \left\lceil \frac{3\ln\frac{2}{\varepsilon}}{\delta^2 p} \right\rceil.$$

3 随机算法验证矩阵乘法

关于算法复杂度的几个符号：

- $O(n^2) \leq cn^2$
- $\Omega(n) \geq cn$
- $\Theta(n^2) = O(n^2)\ \&\ \Omega(n^2)$

问题提出：n阶矩阵A,B,C(均定义在有限域F_2上), 判断AB是否等于C.

如果直接求解AB, 复杂度为$O(n^3)$. 通过矩阵分块可以降到$O(n^{2.73})$, 近年来最优的复杂度为$O(n^{2.373})$.

下面给出随机算法：

(1) 随机取$x \in \{0,1\}^n$;

(2) 验证$A(Bx)$是否等于Cx. 若成立则说$AB = C$, 否则不等。

算法复杂度为$O(n^2)$, 出错的概率

$$\begin{aligned}Pr(error) &= Pr_x(AB \neq C, ABx = Cx) \\ &= Pr_x(AB \neq C, (AB - C)x = 0) \\ &= Pr_x(D \neq 0, Dx = 0).\end{aligned} \tag{3.1}$$

由$D \neq 0$, 通过变换使得$D_{11} = 1$. 将(3.1)按照全概率公式展开得

$$(3.1)\text{式} = \sum_{b_2,\cdots,b_n \in \{0,1\}^n} Pr(Dx = 0 | x_2 = b_2, \cdots, x_n = b_n) Pr(x_2 = b_2, \cdots, x_n = b_n). \tag{3.2}$$

其中

$$
\begin{aligned}
Pr(Dx=0|x_2=b_2,\cdots,x_n=b_n) &= \prod_{i=1}^{n} Pr\left(\sum_{j=1}^{n} D_{ij}x_j=0|x_2=b_2,\cdots,x_n=b_n\right) \\
&\le Pr(D_{11}x_1+\cdots+D_{1n}x_n=0|x_2=b_2,\cdots,x_n=b_n) \\
&= Pr(x_1+c=0) \\
&= \frac{1}{2}.
\end{aligned} \tag{3.3}
$$

代入(3.2)中并结合(3.1)可得$Pr_x(D\neq 0, Dx=0)\le\frac{1}{2}$, 即出错的概率$\le\frac{1}{2}$.

显然这个算法属于**单边错误**，即如果$A(Bx)=Cx$, 不一定有$AB=C$, 但是如果$A(Bx)\neq Cx$, 那么一定有$AB\neq C$. 为了提高算法的正确性，可以多次随机取$x\in\{0,1\}^n$. 比如重复10次，若有一次$A(Bx)\neq Cx$, 则可以直接断定$AB\neq C$; 否则就可以以很高的概率说$AB=C$. 在这种情况下，若$AB\neq C$, 随机选择10个x都满足$A(Bx)=Cx$的概率$\le(\frac{1}{2})^{10}<\frac{1}{1000}$, 即出错的概率$<\frac{1}{1000}$.

4 随机算法的复杂度

随机算法就是在确定性算法上进行一个distribution.

$$
\begin{aligned}
&Time(A,I) \\
&\max_{I\in\widetilde{I}}\ E(Time(A,I)) \\
&\min_{\overrightarrow{u}}\left(\max_{I\in\widetilde{I}} E\left(Time(A,I)\right)\right)=\min_{\overrightarrow{u}}\max_{I\in\widetilde{I}}\left(\overrightarrow{u}\cdot Time(A,I)\right)
\end{aligned}
$$

5 Complexity Class

- **P**: Polynomial, 多项式时间可计算
- **NP**: Non-deterministic Polynomial, 多项式时间可判定
- **BPP**: Bounded error Probabilistic Polynomial
- **RP**: Randomize Polynomial
- **co-RP**: complementary-RP
- **ZPP**: Zero-error Probabilistic Polynomial

(1)

$$
\begin{aligned}
&L\in NP,\ \exists A,\ O(n^c) \\
&\forall x\in L,\ \exists y,\ A(x,y)=1 \\
&\forall x\notin L,\ \forall y,\ A(x,y)=0
\end{aligned}
$$

(2)

$$L \in BPP,\ \exists A,\ O(n^c)$$
$$\forall x \in L,\ Pr(A(x) = 1) \geq p$$
$$\forall x \notin L,\ Pr(A(x) = 0) \geq p$$
$$\left(p \in (\frac{1}{2}, 1]\right)$$

(3)

$$L \in RP,\ \exists A,\ O(n^c)$$
$$\forall x \in L,\ Pr(A(x) = 1) \geq p$$
$$\forall x \notin L,\ Pr(A(x) = 0) = 1$$
$$\left(p \in \left[\frac{1}{2}, 1\right]\right)$$

(4)

$$L \in co - RP,\ \exists A,\ O(n^c)$$
$$\forall x \in L,\ Pr(A(x) = 1) = 1$$
$$\forall x \notin L,\ Pr(A(x) = 0) \geq p$$
$$\left(p \in \left[\frac{1}{2}, 1\right]\right)$$

(5)

$$L \in ZPP,\ \exists A,\ O(n^c)$$
$$\forall x \in L,\ Pr(A(x) = 1) = 1$$
$$\forall x \notin L,\ Pr(A(x) = 0) = 1$$

定理1： $BPP_{2/3} = BPP_{0.99}$.

证明：根据集合论知识，需要证明$BPP_{2/3} \supseteq BPP_{0.99}$和$BPP_{2/3} \subseteq BPP_{0.99}$成立。显然$BPP_{2/3} \supseteq BPP_{0.99}$, 只需证明$BPP_{2/3} \subseteq BPP_{0.99}$.

设$L \in BPP_{2/3}$, 则∃算法A满足

$$x \in L \Rightarrow Pr(A(x) = 1) \geq \frac{2}{3}$$
$$x \notin L \Rightarrow PR(A(x) = 0) \geq \frac{2}{3}$$

现在需要证明$L \in BPP_{0.99}$, 即等价于∃算法A'满足

$$x \in L \Rightarrow Pr(A'(x) = 1) \geq 0.99$$
$$x \notin L \Rightarrow Pr(A'(x) = 0) \geq 0.99$$

关键是通过算法A构造A', 这里我们只证明$x \in L$的情况。

对于任意的$x \in L$, 执行n次$A(x)$, 结果分别定义为$A_1(x), \cdots, A_n(x)$, 显然$A_i(x)$是相互独立的随机变量且有

$$Pr(A_i(x) = 1) = p \geq \frac{2}{3}.$$

定义变量

$$X = \sum_{i=1}^{n} A_i(x)$$

则$E(X) = np$. 如果$X > \frac{n}{2}$, 则$A'(x) = 1$; 否则$A'(x) = 0$.

$$\begin{aligned} Pr(A'(x) = 1) &= Pr\left(X > \frac{n}{2}\right) \\ &= 1 - Pr\left(X \leq \frac{n}{2}\right). \end{aligned} \tag{5.1}$$

根据Chernoff's Bound, $\forall 0 < \delta < 1$,

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

令$(1-\delta)E(X) = \frac{n}{2}$, 可得$\delta = 1 - \frac{1}{2p}$, 代入上式得

$$\begin{aligned} Pr\left(X \leq \frac{n}{2}\right) &\leq e^{-\frac{\left(1-\frac{1}{2p}\right)^2}{2}E(X)} \\ &= e^{-\frac{n}{2}\left(p+\frac{1}{4p}-1\right)} \\ &\leq e^{-\frac{n}{48}}. \end{aligned} \tag{5.2}$$

注意到在(5.2)中，由$p \in \left[\frac{2}{3}, 1\right)$不难得到$p + \frac{1}{4p} - 1 \geq \frac{1}{24}$.

根据(5.1)(5.2)可得

$$Pr(A'(x) = 1) \geq 1 - e^{-\frac{n}{48}} \geq 0.99.$$

解得$n \geq 96ln10$.

因此我们可以构造算法A', 当$x \in L$时确保$Pr(A'(x) = 1) \geq 0.99$成立。不失一般性，当$x \notin L$时，$Pr(A'(x) = 0) \geq 0.99$成立。

更一般地，对于$a, b \in \left(\frac{1}{2}, 1\right)$, 有$BPP_a = BPP_b$成立。不再赘述。

定理2： $ZPP = RP \bigcap co - RP$.

证明：即证$ZPP \subseteq RP \bigcap co - RP$和$ZPP \supseteq RP \bigcap co - RP$成立。

1) $\forall L \in ZPP$, $\exists A$使得

$$x \in L, Pr(A(x) = 1) = 1$$

$$x \notin L, Pr(A(x) = 0) = 1$$

显然$L \in RP$且$L \in co - RP$, 即$L \in RP \bigcap co - RP$, 从而有$ZPP \subseteq RP \bigcap co - RP$.

2) $\forall L \in RP \bigcap co-RP$,

$\exists A_1$使得

$$x \in L, Pr(A_1(x) = 1) \geq \frac{1}{2}$$
$$x \notin L, Pr(A_1(x) = 0) = 1$$

$\exists A_2$使得

$$x \in L, Pr(A_2(x) = 1) = 1$$
$$x \notin L, Pr(A_2(x) = 0) \geq \frac{1}{2}$$

构造算法$\widetilde{A}$: 同时运行$A_1(x)$和$A_2(x)$, 若$A_1(x) = 1$则必有$x \in L$, 输出1; 若$A_2(x) = 0$则必有$x \notin L$, 输出0; 若$A_1(x) = 0$且$A_2(x) = 1$则不能确定。

设$Pr(x \in L) = p$, 则$Pr(x \notin L) = 1 - p$,

$$Pr(A_1(x) = 0, A_2(x) = 1) < \frac{1}{2}p + \frac{1}{2}(1 - p) = \frac{1}{2}.$$

若不能确定，可以重新选取随机数继续运行A_1和A_2. 从上述概率可知运行n次之后不能确定结果的概率$< \left(\frac{1}{2}\right)^n$. 因此

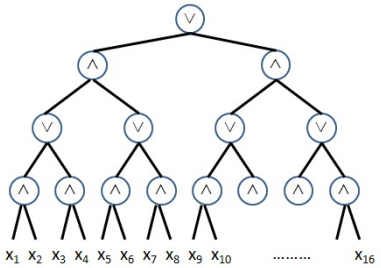
$$x \in L, Pr(\widetilde{A}(x) = 1) = 1$$
$$x \notin L, Pr(\widetilde{A}(x) = 0) = 1$$

即$L \in ZPP$, 从而有$ZPP \supseteq RP \bigcap co-RP$成立。

证毕。

6 博弈论：Game Tree/AND OR Tree

建立一个深度为$2n$的AND OR Tree, 图中所示为$n = 2$. 利用确定性算法计算$f(x_1, x_2, \cdots, x_{4^n})$的复杂度为$4^n$.



$$
\begin{aligned}
&T(n) = max\{T_0(n), T_1(n)\}.\\
&T_1(n) \le \frac{1}{2} * 2 * T_1(n-1) + \frac{1}{2}\left[2T_1(n-1) + \frac{1}{2}T_0(n-1) + \frac{1}{2}\left(T_1(n-1) + T_0(n-1)\right)\right]\\
&\qquad = \frac{9}{4}T_1(n-1) + \frac{1}{2}T_0(n-1).\\
&T_0(n) \le 2\left[\frac{1}{2}T_0(n-1) + \frac{1}{2}\left(T_1(n-1) + T_0(n-1)\right)\right]\\
&\qquad = T_1(n-1) + 2T_0(n-1).
\end{aligned}
$$

$$
\begin{pmatrix} T_1(n) \\ T_0(n) \end{pmatrix} \le \begin{pmatrix} \frac{9}{4} & \frac{1}{2} \\ 1 & 2 \end{pmatrix} \begin{pmatrix} T_1(n-1) \\ T_0(n-1) \end{pmatrix} \le \begin{pmatrix} \frac{9}{4} & \frac{1}{2} \\ 1 & 2 \end{pmatrix}^{n-1} \begin{pmatrix} T_1(1) \\ T_0(1) \end{pmatrix} = U \begin{pmatrix} \lambda_1^{n-1} & \\ & \lambda_2^{n-1} \end{pmatrix} U^{-1} \begin{pmatrix} T_1(1) \\ T_0(1) \end{pmatrix}.
$$

解得特征值$\lambda_{1,2} = \frac{17 \pm \sqrt{33}}{8}$. 令$4^n = N$, $O\left(\left(\frac{17+\sqrt{33}}{8}\right)^n\right) = O\left(N^{\log_4 \frac{17+\sqrt{33}}{8}}\right)$.

7 Balls & Bins

问题描述：将m个球随机放入到n个盒子中，最后每个盒子中的球数分别为$x_1, \cdots, x_n$. 定义$K = \max\limits_{1 \le i \le n} x_i$. 该模型可以用于建模哈希(Hash)函数，生日悖论(Birthday Paradox)以及负载平衡(Workload Balance)问题。

7.1 基本知识

Union Bound:

$$
Pr(\exists x_i) = Pr(x_1 \cup x_2 \cup \cdots \cup x_n) \le Pr(x_1) + Pr(x_2) + \cdots + Pr(x_n).
$$

Chernoff's Bound:

$x_1, x_2, \cdots, x_n$是独立(不必同分布)的0-1随机变量，$Pr(x_i = 1) = p_i$, 定义随机变量$X = x_1 + x_2 + \cdots + x_n$, $E(X) = \sum p_i = \mu$.

(1) $\forall \delta > 0$, $Pr(X \ge (1+\delta)\mu) \le \left[\frac{e^\delta}{(1+\delta)^{1+\delta}}\right]^\mu$;

(2) $\forall 0 < \delta < 1$, $Pr(X \le (1-\delta)\mu) \le \left[\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right]^\mu$.

证明：

(1)

$$
\begin{aligned}
Pr(X \ge (1+\delta)\mu) &= Pr\left(e^X \ge e^{(1+\delta)\mu}\right)\\
&= Pr\left(e^{\lambda X} \ge e^{\lambda(1+\delta)\mu}\right) \qquad (\lambda > 0)\\
&\le \frac{E\left(e^{\lambda X}\right)}{e^{\lambda(1+\delta)\mu}}. \qquad (\text{根据Markov Inequatity})
\end{aligned} \tag{7.1}
$$

$$
\begin{aligned}
E\left(e^{\lambda X}\right) &= E\left(\prod_{i=1}^{n} e^{\lambda x_i}\right) \\
&= \prod_{i=1}^{n} E\left(e^{\lambda x_i}\right) \\
&= \prod_{i=1}^{n} \left((1-p_i) + p_i e^{\lambda}\right) \\
&= \prod_{i=1}^{n} \left(1 + p_i\left(e^{\lambda} - 1\right)\right).
\end{aligned} \tag{7.2}
$$

由$e^x \geq 1 + x,\ x \geq 0$得

$$
\begin{aligned}
E\left(e^{\lambda X}\right) &\leq \prod_{i=1}^{n} e^{p_i\left(e^{\lambda}-1\right)} \\
&= e^{\sum_{i=1}^{n} p_i\left(e^{\lambda}-1\right)} \\
&= e^{\left(e^{\lambda}-1\right)\sum_{i=1}^{n} p_i} \\
&= e^{\left(e^{\lambda}-1\right)\mu}.
\end{aligned} \tag{7.3}
$$

结合(7.1)(7.3)可得，

$$
\begin{aligned}
Pr(X \geq (1+\delta)\mu) &\leq \frac{e^{\left(e^{\lambda}-1\right)\mu}}{e^{\lambda(1+\delta)\mu}} \\
&= \left[e^{e^{\lambda}-1-\lambda(1+\delta)}\right]^{\mu}.
\end{aligned} \tag{7.4}
$$

令$g(\lambda) = e^{\lambda} - 1 - \lambda(1+\delta)$, 则$g'(\lambda) = e^{\lambda} - (1+\delta)$. 由$g'(\lambda) = 0$解得

$$
\lambda_0 = ln(1+\delta).
$$

将λ_0的值代入(7.4)中得

$$
\begin{aligned}
\left[e^{e^{\lambda}-1-\lambda(1+\delta)}\right]^{\mu} &= \left[e^{\delta-(1+\delta)ln(1+\delta)}\right]^{\mu} \\
&= \left[\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right]^{\mu}.
\end{aligned} \tag{7.5}
$$

由(7.4)(7.5)得

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

证毕。

(2) 同理可证明(2)成立。

Chernoff's Bound (Cont'd):

$\forall 0 < \delta < 1$,

(1) $\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \leq e^{-\frac{\delta^2}{3}}$;

(2) $\frac{e^{-\delta}}{(1-\delta)^{1-\delta}} \leq e^{-\frac{\delta^2}{2}}$;

(3) $Pr\left(|X-\mu| \geq \delta\mu\right) \leq 2e^{-\frac{\delta^2}{3}\mu}$.

证明：

(1)

$$ln\frac{e^{\delta}}{(1+\delta)^{1+\delta}} = \delta - (1+\delta)ln(1+\delta) \tag{7.6}$$

将$ln(1+x)$泰勒展开得

$$ln(1+x) = x - \frac{x^2}{2} + \frac{x^3}{3} - \frac{x^4}{4} + \cdots$$

则

$$\begin{aligned} \delta - (1+\delta)ln(1+\delta) &= \delta - (1+\delta)\left(\delta - \frac{\delta^2}{2} + \frac{\delta^3}{3} - \frac{\delta^4}{4} + \cdots\right) \\ &= -\frac{\delta^2}{2} + \frac{\delta^3}{6} - \frac{\delta^4}{12} + \frac{\delta^5}{20} - \cdots \\ &\leq -\frac{\delta^2}{2} + \frac{\delta^3}{6} \\ &\leq -\frac{\delta^2}{3}. \end{aligned} \tag{7.7}$$

由(7.6)(7.7)得

$$\frac{e^{\delta}}{(1+\delta)^{1+\delta}} \leq e^{-\frac{\delta^2}{3}}.$$

证毕。

(2)

$$\begin{aligned} ln\frac{e^{-\delta}}{(1-\delta)^{1-\delta}} &= -\delta - (1-\delta)ln(1-\delta) \\ &= -\delta - (1-\delta)\left(-\delta - \frac{\delta^2}{2} - \frac{\delta^3}{3} - \cdots\right) \\ &= -\frac{\delta^2}{2} - \frac{\delta^3}{6} - \frac{\delta^4}{12} - \cdots \\ &\leq -\frac{\delta^2}{2}. \end{aligned} \tag{7.8}$$

即

$$\frac{e^{-\delta}}{(1-\delta)^{1-\delta}} \leq e^{-\frac{\delta^2}{2}}.$$

证毕。

(3) 根据Chernoff's Bound, $\forall 0 < \delta < 1$,

$$Pr(X \geq (1+\delta)\mu) \leq e^{-\frac{\delta^2}{3}\mu}$$
$$Pr(X \leq (1-\delta)\mu) \leq e^{-\frac{\delta^2}{2}\mu}$$

$$\begin{aligned} Pr\left(|X-\mu| \geq \delta\mu\right) &= Pr(X \geq (1+\delta)\mu \cup X \leq (1-\delta)\mu) \\ &\leq Pr(X \geq (1+\delta)\mu) + Pr(X \leq (1-\delta)\mu) \\ &\leq e^{-\frac{\delta^2}{3}\mu} + e^{-\frac{\delta^2}{2}\mu} \\ &\leq 2e^{-\frac{\delta^2}{3}\mu}. \end{aligned}$$

证毕。

Stirling's Approximation:

$$\sqrt{2\pi n}\left(\frac{n}{e}\right)^n \leq n! \leq e^{\frac{1}{12n}}\sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$
$$n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n.$$

定理1：对于任意的$0 \leq m \leq n$, $\left(\frac{n}{m}\right)^m \leq \binom{n}{m} \leq \left(\frac{en}{m}\right)^m$.

证明：首先给出一个引理：如果$0 < m \leq n$并且$0 < k < m$, 则$\frac{n-k}{m-k} \geq \frac{n}{m}$.
先证明这个引理：

$$\begin{aligned} \frac{n-k}{m-k} - \frac{n}{m} &= \frac{m(n-k) - n(m-k)}{(m-k)m} \\ &= \frac{(n-m)k}{(m-k)m} \geq 0. \end{aligned}$$

证毕。

$$\begin{aligned} \binom{n}{m} &= \frac{n(n-1)\cdots(n-m+1)}{m!} \\ &= \frac{n(n-1)\cdots(n-m+1)}{m(m-1)\cdots(m-m+1)} \\ &= \frac{n}{m}\frac{n-1}{m-1}\cdots\frac{n-(m-1)}{m-(m-1)}. \end{aligned}$$

根据引理可得

$$\frac{n}{m}\frac{n-1}{m-1}\cdots\frac{n-(m-1)}{m-(m-1)} \geq \left(\frac{n}{m}\right)^m.$$

即证明了

$$\binom{n}{m} \geq \left(\frac{n}{m}\right)^m.$$

下面用两种方法证明$\binom{n}{m} \leq \left(\frac{en}{m}\right)^m$成立。

(1)

$$\binom{n}{m} = \frac{n!}{m!(n-m)!} \leq \frac{n^m}{m!}. \tag{7.9}$$

根据Stirling's Approximation可得$\frac{1}{m!} \leq \frac{1}{\sqrt{2\pi m}\left(\frac{m}{e}\right)^m}$，代入上式得

$$\frac{n^m}{m!} \leq \frac{1}{\sqrt{2\pi m}}\left(\frac{en}{m}\right)^m \leq \left(\frac{en}{m}\right)^m.$$

(2) 根据e^x的泰勒展开式可得

$$e^m = \sum_{i=0}^{\infty} \frac{m^i}{i!} \geq \frac{m^m}{m!}.$$

变形得$\frac{1}{m!} \leq \left(\frac{e}{m}\right)^m$, 结合(7.9)式即可得$\binom{n}{m} \leq \left(\frac{en}{m}\right)^m$.

综上即证明了定理成立。

定理2：在Bins & Balls问题中，令$m = n$, 定义随机变量

$$y_i = \begin{cases} 1, & x_i \geq \frac{\ln n}{10 \ln \ln n} \\ 0, & otherwise \end{cases}$$

则随机变量y_i, y_j负相关，即对于任意的$1 \leq i < j \leq n$, $Cov(y_i, y_j) \leq 0$.

证明：不失一般性，令$k = \frac{lnn}{10lnlnn}$.

$$\begin{aligned} Cov(y_i, y_j) &= E(y_i y_j) - E(y_i)E(y_j) \\ &= p(y_i y_j = 1) - p(y_i = 1)p(y_j = 1) \\ &= p(x_i \geq k, x_j \geq k) - p(x_i \geq k)p(x_j \geq k) \\ &= p(x_j \geq k)[p(x_i \geq k | x_j \geq k) - p(x_i \geq k)]. \end{aligned} \tag{7.10}$$

利用全概率公式，将$p(x_i \geq k)$展开得

$$p(x_i \geq k) = p(x_i \geq k | x_j \geq k)p(x_j \geq k) + p(x_i \geq k | x_j < k)p(x_j < k). \tag{7.11}$$

将(7.11)代入到(7.10)中得

$$\begin{aligned} Cov(y_i, y_j) &= p(x_j \geq k)[p(x_i \geq k | x_j \geq k)(1 - p(x_j \geq k)) - p(x_i \geq k | x_j < k)p(x_j < k)] \\ &= p(x_j \geq k)[p(x_i \geq k | x_j \geq k)p(x_j < k) - p(x_i \geq k | x_j < k)p(x_j < k)] \\ &= p(x_j \geq k)p(x_j < k)[p(x_i \geq k | x_j \geq k) - p(x_i \geq k | x_j < k)]. \end{aligned}$$

因此只需证明

$$p(x_i \geq k | x_j \geq k) \leq p(x_i \geq k | x_j < k). \tag{7.12}$$

定义函数$f(l)=p(x_i\geq k|x_j=l)$, 显然$f(l)$是关于l的减函数。

$$\begin{aligned}p(x_i\geq k|x_j\geq k)&=\sum_{l\geq k}\frac{p(x_j=l)}{p(x_j\geq k)}p(x_i\geq k|x_j=l)\\p(x_i\geq k|x_j<k)&=\sum_{l<k}\frac{p(x_j=l)}{p(x_j<k)}p(x_i\geq k|x_j=l)\end{aligned}\tag{7.13}$$

注意到当$l_1\geq k, l_2<k$时

$$p(x_i\geq k|x_j=l_1)<p(x_i\geq k|x_j=l_2).$$

另

$$\begin{aligned}\sum_{l\geq k}\frac{p(x_j=l)}{p(x_j\geq k)}&=1\\\sum_{l<k}\frac{p(x_j=l)}{p(x_j<k)}&=1\end{aligned}$$

所以我们可以得到

$$\begin{aligned}\sum_{l\geq k}\frac{p(x_j=l)}{p(x_j\geq k)}p(x_i\geq k|x_j=l)&\leq p(x_i\geq k|x_j=k)\\\sum_{l<k}\frac{p(x_j=l)}{p(x_j<k)}p(x_i\geq k|x_j=l)&> p(x_i\geq k|x_j=k)\end{aligned}\tag{7.14}$$

这样就证明了(7.12)成立，即证明了结论。

7.2 $m\sim\sqrt{n}$

当$m\sim\sqrt{n}$, 比如$m=\frac{\sqrt{n}}{10}$时

$$\begin{aligned}Pr(K=1)&=1-Pr(K\geq 2)\\&=1-\sum_{j=2}^{m}Pr(K=j).\end{aligned}\tag{7.15}$$

$$\begin{aligned}Pr(K=j)&\leq \Pr\left(\exists_i x_i=j\right)\\&=Pr(x_1=j\cup x_2=j\cup\cdots\cup x_n=j)\\&\leq nPr(x_1=j)\\&=n\binom{m}{j}\left(\frac{1}{n}\right)^j\left(1-\frac{1}{n}\right)^{m-j}\\&\leq n\binom{m}{j}\left(\frac{1}{n}\right)^j\\&\leq n\left(\frac{m}{n}\right)^j\\&=n\left(\frac{1}{10\sqrt{n}}\right)^j.\end{aligned}\tag{7.16}$$

代入(7.15)中可得

$$\begin{aligned}\sum_{j=2}^{m} Pr(K=j) &\leq \sum_{j=2}^{m} n\left(\frac{1}{10\sqrt{n}}\right)^{j} \\ &\leq n\frac{\frac{1}{100n}}{1-\frac{1}{10\sqrt{n}}} \\ &\leq \frac{1}{90}.\end{aligned}$$

即$Pr(K=1) \geq 1-\frac{1}{90}$.

可以看出当$m \sim \sqrt{n}$时，K将以很高的概率等于1.

7.3 $m = n$

$$\begin{aligned}Pr(K < 10lnn) &= 1 - Pr(K \geq 10lnn) \\ &= 1 - Pr(\exists x_i \geq 10lnn).\end{aligned}$$

$$
\begin{aligned}
Pr(\exists x_i \geq 10lnn) &= Pr(x_1 \geq 10lnn \cup x_2 \geq 10lnn \cup \cdots \cup x_n \geq 10lnn) \\
&\leq \sum_{i=1}^{n} Pr(x_i \geq 10lnn) \\
&= nPr(x_i \geq 10lnn) \\
&= n \sum_{j=10lnn}^{n} Pr(x_i = j) \\
&= n \sum_{j=10lnn}^{n} \binom{n}{j} \left(\frac{1}{n}\right)^j \left(1 - \frac{1}{n}\right)^{n-j} \\
&\leq n \sum_{j=10lnn}^{n} \binom{n}{j} \left(\frac{1}{n}\right)^j \\
&\leq n \sum_{j=10lnn}^{n} \left(\frac{en}{j}\right)^j \frac{1}{n^j} \\
&= n \sum \left(\frac{e}{j}\right)^j \\
&\leq n \sum \left(\frac{1}{3lnn}\right)^j \\
&\leq \frac{n\left(\frac{1}{3lnn}\right)^{10lnn}}{1 - \frac{1}{3lnn}} \\
&\leq 2n \left(\frac{1}{3lnn}\right)^{10lnn} \\
&= 2n \frac{1}{3^{10lnn} n^{10lnlnn}} \\
&= o(1).
\end{aligned}
$$

由此可得

$$Pr(K < 1olnn) \geq 1 - o(1).$$

即当$m = n$时，K以很高的概率$< 10lnn$.

7.4 总结

当m取不同的量级时，对应的K值会相应地产生变化。对应表格如下：

m	$\sqrt{n}$	n	$nlnn$
K	$K > 1$	$\Theta\left(\frac{lnn}{lnlnn}\right)$	$\Theta\left(\frac{m}{n}\right)$

8 概率方法(Probability Method)

8.1 素数

(a) 素数定理：$\pi(n) \sim \Theta\left(\frac{n}{lnn}\right)$, 其中$\pi(n)$表示$< n$ 的素数个数。

(b) 素数的个数是无穷的。

证明：用两种方法来证明。

(1) 最经典的证明是欧几里得方法。反证法，假设素数一共有n个, 分别为$p_1, p_2, \cdots, p_n$. 定义数

$$p = p_1 p_2 \cdots p_n + 1.$$

易证p不能被任一p_i整除，即p为素数。又$p \neq p_i$, 这与假设矛盾，即证明了素数的无穷性。

(2) 引入费马数$F_n = 2^{2^n}$, 则

$$F_{n+1} = 2^{2^{n+1}} = (F_n - 1)^2 + 1.$$

$gcd\,(F_n, F_{n+1}) = 1$, 可得F_n与F_{n+1}没有相同的素因子，由此证明了素数是无穷的。

8.2 Ramsey数

经典问题是说6个人中至少有3个人互相认识或者互相不认识。这个问题可以转换成图论来求解，6个人对应6个结点，任意两个人之间有一条边相连，红边或蓝边，红边表示认识，蓝边表示不认识。最后形成了一个由6个结点组成的完全图，只需证明其中含有红色或蓝色三角形。

定义Ramsey数：对一个图的边进行**二染色**(红色或蓝色)，$R(m,n)$表示出现红色K_m或蓝色K_n所需的最少结点数，其中K_i表示i个结点中任意两个结点之间的边同色。

据此有以下Ramsey数：

$$R(3,3) = 6$$
$$R(2,n) = n$$
$$R(3,4) = 9$$

定理1：$R(n,n) \leq 2^{2n}$.

证明：首先给出一个引理。

引理：$R(s,t) \leq R(s-1,t) + R(s,t-1)$.

先证明这个引理成立，分两种情况：

(1) 存在结点V至少和$R(s-1,t)$个结点连接红边，这$R(s-1,t)$个结点形成红色K_{s-1}或者蓝色K_t. 红色的K_{s-1}和V形成了红色的K_s.

(2) 存在结点V'至少和$R(s,t-1)$个结点连接蓝边，这$R(s,t-1)$个结点形成红色K_s或者蓝色K_{t-1}. 蓝色的K_{t-1}和V形成了蓝色的K_t.

至此就证明了引理，下面通过归纳法证明：

$$R(s,t) \leq \binom{s+t-2}{s-1}.$$

$r=s=2$时，显然有$R(2,2) \leq \binom{2+2-2}{2-1} = 2.$

假设$R(s-1,t)$和$R(s,t-1)$满足不等式，则

$$\begin{aligned} R(s,t) &\leq R(s-1,t) + R(s,t-1) \\ &\leq \binom{s+t-3}{s-2} + \binom{s+t-3}{s-1} \\ &= \binom{s+t-2}{s-1}. \end{aligned}$$

证毕。

当$s=t=n$时，$R(n,n) \leq \binom{2n-2}{n-1}$. 根据String's Approximation可得，

$$\begin{aligned} \binom{2n-2}{n-1} &= \frac{(2n-2)!}{(n-1)!(n-1)!} \\ &\leq \sqrt{2\pi(2n-2)} \left(\frac{2n-2}{e}\right)^{2n-2} \frac{1}{2\pi(n-1)} \left(\frac{e}{n-1}\right)^{2n-2} \\ &= \frac{1}{\sqrt{\pi(n-1)}} 2^{2n-2} \\ &\leq 2^{2n}. \end{aligned}$$

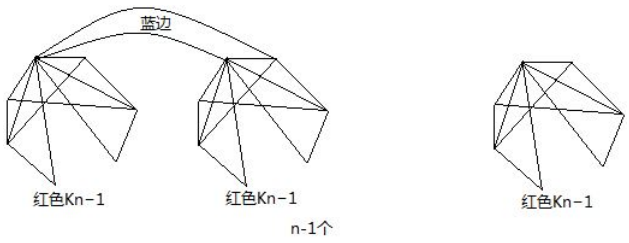
证毕。

定理2： $R(n,n) > (n-1)^2$.

证明：构造图如下

如图所示有$n-1$个红色K_{n-1}, 不同K_{n-1}之间的结点连接蓝边。显然这个由$(n-1)^2$个结点组成的完全图不包含红色和蓝色的K_n. 即证明了$R(n,n) > (n-1)^2$.

定理3： $R(n,n) > 2^{\frac{n}{2}}$.



证明：取N个点，则一共有$\binom{N}{2}$条边。对每条边随机染色，即以$\frac{1}{2}$的概率染成红色，以$\frac{1}{2}$的概率染成蓝色。记生成的图为G.

$$Pr(G\text{中不出现红色}K_n\text{且不出现蓝色}K_n) = 1 - Pr(\text{出现红色}K_n \cup \text{出现蓝色}K_n). \tag{8.1}$$

$$\begin{aligned} Pr(\text{出现红色}K_n \cup \text{出现蓝色}K_n) &\leq 2Pr(\text{出现红色}K_n) \\ &\leq 2Pr(\text{某个}K_n)\binom{N}{n} \\ &= 2\frac{1}{2^{C_n^2}}C_N^n \\ &\leq 2\left(\frac{eN}{n}\right)^n \frac{1}{2^{\frac{n^2-n}{2}}}. \end{aligned} \tag{8.2}$$

对(8.2)式右端以2为底取对数，得

$$\log_2 2\left(\frac{eN}{n}\right)^n \frac{1}{2^{\frac{n^2-n}{2}}} = 1 + n(log_2 e + log_2 N - log_2 n) - \frac{n^2}{2} + \frac{n}{2} \tag{8.3}$$

代入$N = 2^{\frac{n}{2}}$, 得

$$(8.3)\text{式} = 1 + n\left(log_2 e + \frac{1}{2}\right) - nlog_2 n < 0. \tag{8.4}$$

可知取对数之前(8.2)式右端<1, 事实上是<<1, 即

$$Pr(\text{出现红色}K_n \cup \text{出现蓝色}K_n) << 1.$$

综上，当$N = 2^{\frac{n}{2}}$时，图G中将以概率$1 - o(1)$不出现红色K_n且不出现蓝色K_n, 即以概率方法证明了$R(n,n) > 2^{\frac{n}{2}}$.

8.3 MAX-CUT

8.3.1 概率方法

最大割是NP完全问题，这里用概率方法阐述一下。

将图G的结点分成两个集合L和R. 边uv的两个结点u, v有四种分布情况，分别是：$u \in L, v \in L$; $u \in L, v \in R$; $u \in R, v \in L$; $u \in R, v \in R$. 因此uv被cut的概率=$\frac{1}{2}$.

定义割边数为$C(L,R)$, 则有

$$E\left(C(L,R)\right)=\sum_{e}\frac{1}{2}=\frac{m}{2}.$$

割边的期望值为$\frac{m}{2}$, 则一定存在某种割使得割边数目$\geq\frac{m}{2}$.

8.3.2 2近似算法

(a) 算法描述

> (1) 任意初始化割$S\bigcup\overline{S}$.
>
> (2) 对于任一结点v如果有少于$\frac{1}{2}$的边通过割，就将其移到割的另一个集合中。
>
> (3) 如果没有这样的结点存在，则终止算法，否则回到(2).

(b) 复杂度分析

显然，通过每一步迭代，通过割的边数至少增加1. 由于最大割最多是图的边数，因此算法经过线性次数的迭代一定会终止。

(c) 正确性证明

算法终止后，我们得到集合S和$\overline{S}$, 分别写作$S=\{x_1,\cdots,x_s\}$和$\overline{S}=\{y_1,\cdots,y_t\}$. $deg(x)$表示顶点x的度数。

根据**握手定理**(Handshaking lemma)可得

$$\sum_{i=1}^{s}deg(x_i)+\sum_{j=1}^{t}deg(y_j)=2m.$$

其中m表示边数。通过上述算法可知

$$2|C|\geq\frac{1}{2}\sum_{i=1}^{s}deg(x_i)+\frac{1}{2}\sum_{j=1}^{t}deg(y_j)=m.$$

即

$$|C|\geq\frac{m}{2}\geq\frac{1}{2}opt.$$

证毕。

8.4 独立集(Independent Set, IS)

给定图$G=(V,E)$, 独立集$V'\subseteq V$并且V'中任两个结点之间没有边相连。求解最大独立集(MIS)是NP完全问题。

定理1：图G满足$|V|=n$, 点的平均度数为d, 则$\#MIS\geq\frac{n}{2d}$.

证明：利用概率方法。将任一结点v以概率$\frac{1}{d}$放入集合S中，这样形成的S其中可能有边存在，令$m(S)$表示S中形成的边数。

$$E(|S|)=\frac{n}{d}$$
$$E(m(S))=\sum_{e}\frac{1}{d^2}=\frac{1}{d^2}\frac{nd}{2}=\frac{n}{2d}$$
$$E(|S|-m(S))=\frac{n}{d}-\frac{n}{2d}=\frac{n}{2d}$$

所以$\exists S_0$满足

$$|S_0|-m(S_0)\geq\frac{n}{2d}.$$

显然有$\#MIS\geq\frac{n}{2d}$. 证毕。

8.5 AND OR Tree

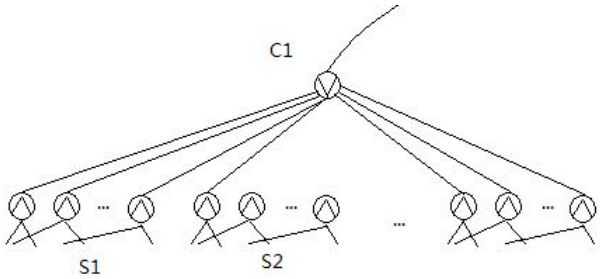
定义

$$maj(x_1\sim x_n)=\begin{cases}1, & \sum x_i>\frac{n}{2}\\ 0, & otherwise\end{cases}$$

$$Th_k(x_1\sim x_n)=\begin{cases}1, & \sum x_i\geq k\\ 0, & otherwise\end{cases}$$

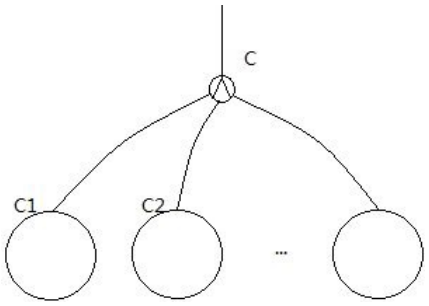
定理1：$\exists$深度为3, 多项式时间的电路C实现$Th_k(k=log_2n+1)$.

证明：将$\{x_1,x_2,\cdots,x_n\}$随机分到log_2n个集合中，分别记为$S_1,S_2,\cdots,S_{log_2n}$. 将每个集合中任意两个元素AND($\wedge$)一次，之后将$log_2n$个集合OR($\vee$)一次，得到$C_1$. 如下图所示



执行这样的操作n^5次，对应为$C_1\sim C_{n^5}$. 将C_1到C_{n^5}进行AND($\wedge$)操作，得到最终的C.

$$Pr(C=Th_k)=1-Pr(C\neq Th_k).$$



其中

$$
\begin{aligned}
Pr(C \neq Th_k) &= Pr(C = 1, Th_k = 0) \\
&= Pr(C(x_1, \cdots, x_n) = 1, x_1 + \cdots + x_n \leq log_2 n) \\
&\leq \sum_{x_1+\cdots+x_n \leq log_2 n} Pr(C(x) = 1) \\
&= \left[\binom{n}{0} + \binom{n}{1} + \cdots + \binom{n}{\log_2 n} \right] Pr(C_i(x_1, \cdots, x_n) = 1)^{n^5} \\
&\leq 2^n Pr(C_i(x_1, \cdots, x_n) = 1)^{n^5} \\
&= 2^n \left[1 - \left(1 - \frac{1}{\log_2 n}\right)\left(1 - \frac{2}{\log_2 n}\right) \cdots \left(1 - \frac{\log_2 n - 1}{\log_2 n}\right) \right]^{n^5} \\
&= 2^n \left[1 - \frac{(log_2 n)!}{(\log_2 n)^{\log_2 n}} \right]^{n^5} \\
&\leq 2^n \left[1 - \frac{\sqrt{2\pi \log_2 n}\left(\frac{\log_2 n}{e}\right)^{\log_2 n}}{(\log_2 n)^{\log_2 n}} \right]^{n^5} \qquad \text{(根据Stirling's Approximation)} \\
&= 2^n \left[1 - \frac{\sqrt{2\pi \log_2 n}}{e^{\log_2 n}} \right]^{n^5} \\
&\leq 2^n e^{-\frac{\sqrt{2\pi \log_2 n}}{e^{\log_2 n}} n^5} \qquad \left(\text{根据} \lim_{x\to\infty} \left(1 - \frac{1}{x}\right)^x = \frac{1}{e} \right) \\
&= \frac{2^n}{e^{\sqrt{2\pi \log_2 n} n^{5 - \frac{\log_2 n}{\ln n}}}} \qquad \left(\frac{\log_2 n}{\ln n} = \frac{1/\log_n 2}{1/\log_n e} = \frac{\log_n e}{\log_n 2} = \log_2 e < 2 \right) \\
&\leq \frac{2^n}{e^{n^3}} \\
&= o(1).
\end{aligned}
$$

即$Pr(C = Th_k) = 1 - o(1)$. 证毕。

8.6 MAX-SAT

问题描述：给出一个SAT实例(比如3SAT)，求能最多满足的子句数。

使用概率方法，令$Pr(x_i = 1) = Pr(x_i = 0) = \frac{1}{2}$. 随机变量$z_i$对应第$i$个子句，为

真则$z_i = 1$, 反之$z_i = 0$. $Z = \sum_{i=1}^{m} z_i$.

$$\begin{aligned} E(Z) &= \sum_{i=1}^{m} E(z_i) \\ &= \sum_{i=1}^{m} \left(1 - \frac{1}{2^{k_i}}\right) \\ &\geq \frac{m}{2}. \end{aligned}$$

定理1： $E(X) = \underset{Y}{E}\left(\underset{X}{E}(X|Y)\right)$.

证明：

$$\begin{aligned} \text{右端} &= \sum_{y} Pr(Y = y)E(X|Y = y) \\ &= \sum_{y} Pr(Y = y)\left[\sum_{x} xPr(X = x|Y = y)\right] \\ &= \sum_{y}\sum_{x} Pr(Y = y)xPr(X = x|Y = y) \\ &= \sum_{y}\sum_{x} xPr(X = x, Y = y) \\ &= \sum_{x} x\sum_{y} Pr(X = x, Y = y) \\ &= \sum_{x} xPr(X = x) \\ &= E(X). \end{aligned}$$

利用条件概率实现确定性算法。

$$\begin{aligned} E(Z) &= \underset{x_1}{E}\left(\underset{Z}{E}(Z|x_1)\right) \\ &= \Pr(x_1 = 1)E(Z|x_1 = 1) + \Pr(x_1 = 0)E(Z|x_1 = 0) \\ &= \frac{1}{2}\left[E(Z|x_1 = 1) + E(Z|x_1 = 0)\right] \\ &\geq \frac{m}{2}. \end{aligned} \tag{8.5}$$

由(8.5)可得

$$max\{E(Z|x_1 = 1), E(Z|x_1 = 0)\} \geq \frac{m}{2}.$$

分别计算$E(Z|x_1 = 1)$和$E(Z|x_1 = 0)$的值，通过比较可固定x_1的值，依次下去可分别确定$x_2, \cdots, x_n$的值。

8.7

$$L \in RP, \exists A$$
$$x \in L, Pr(A(x) = 1) \geq \frac{1}{2}$$
$$x \notin L, Pr(A(x) = 0) = 1$$

N为素数，从集合$\{1, \cdots, N\}$中随机选择两个数a, b. 利用算法A构造$\widetilde{A}$:

$$A(x, (a+b) mod\ N)$$
$$A(x, (2a+b) mod\ N)$$
$$\cdots$$
$$A(x, (ta+b) mod\ N)$$

若$\exists A_i(x_i, (ia+b) mod\ N) = 1$则输出1, 否则输出0.

首先证明$ia+b, ja+b$两两独立。

$$Pr(ia+b=s, ja+b=t) = \frac{\#\{(a,b)|ia+b=s, ja+b=t\}}{N^2}.$$

当$i \neq j$时，方程组$\begin{cases} ia+b=s \\ ja+b=t \end{cases}$有唯一解。由此可得

$$Pr(ia+b=s, ja+b=t) = \frac{1}{N^2}.$$

又

$$Pr(ia+b=s) = \frac{1}{N}$$
$$Pr(ja+b=t) = \frac{1}{N}$$

得

$$Pr(ia+b=s, ja+b=t) = Pr(ia+b=s)Pr(ja+b=t).$$

即证明了$ia+b, ja+b$两两独立。

算法$\widetilde{A}$出错的概率，当$x \in L$

$$Pr(\widetilde{A}=0) = Pr(A_1(x)=0, \cdots, A_t(x)=0). \tag{8.6}$$

令随机变量$Z_1, \cdots, Z_t$分别对应$A_1(x), \cdots, A_t(x)$, $Z = Z_1 + \cdots + Z_t$.

$$\begin{aligned} \text{式(8.6)} &= Pr(Z_1 + \cdots + Z_t = 0) \\ &= Pr(Z=0) \\ &= Pr(Z - E(Z) = -E(Z)) \\ &= Pr(|Z - E(Z)| = E(Z)) \\ &\leq \frac{Var(Z)}{E^2(Z)}. \end{aligned} \tag{8.7}$$

$$Var(Z) = \sum_{i=1}^{t} Var(Z_i) + 2\sum_{i,j} Cov(Z_i, Z_j) = tVar(Z_i). \tag{8.8}$$

$$\begin{aligned} \frac{Var(Z)}{E^2(Z)} &= \frac{tVar(Z_i)}{t^2E^2(Z_i)} \\ &= \frac{p(1-p)}{tp^2} \\ &\leq \frac{\frac{1}{4}}{t\left(\frac{1}{2}\right)^2} \\ &= \frac{1}{t}. \end{aligned} \tag{8.9}$$

由(8.6)(8.7)(8.9)可得，当$x \in L$时

$$Pr(\widetilde{A} = 0) \leq \frac{1}{t}.$$

假设运行一次算法A需要时间T, 则运行t次A需要时间tT, 正确率$\geq 1 - \frac{1}{t}$.

9 代数化方法(Algebraic Method)

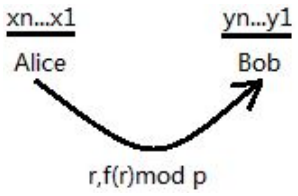
9.1 两数相等的判定

问题描述：两个数x和y比特位均为n, 分别表示为$x_n \cdots x_1$和$y_n \cdots y_1$, 判断其是否相等。

如果直接判断，复杂度为$O(n)$. 采用代数化方法，构造多项式

$$f(z) = x_n z^{n-1} + x_{n-1} z^{n-2} + \cdots + x_2 z^1 + x_1.$$
$$g(z) = y_n z^{n-1} + y_{n-1} z^{n-2} + \cdots + y_2 z^1 + y_1.$$

在$4n - 8n$中取素数p, 从$\{1, 2, \cdots, p-1\}$中随机取一个数r(比特数量级为$O(logn)$), 判断$f(r)$是否等于$g(r)$. 用下列情形描述：



注意到通信的比特量级为$O(logn)$.

下面证明该算法的高效性。出错的情况，当$x \neq y$时

$$\begin{aligned} Pr(f(r) = g(r)) &= Pr(f(r) - g(r) = 0) \\ &= Pr(h(r) = 0). \qquad (h(z) \triangleq f(z) - g(z)) \end{aligned}$$

$h(r)$是$n-1$次的多项式，根据代数基本定理可得$h(r) = 0$最多有$n-1$个实数根。因此

$$Pr(h(r) = 0) \leq \frac{n-1}{p-1} < \frac{1}{4}.$$

9.2 两个多项式相等的判定

Schwartz-Zippel lemma: 多项式$Q(x_1, \cdots, x_n)$不恒等于零，次数$deg(Q) = d$. $r_1, \cdots, r_n \in S$, 则

$$Pr(Q(r_1, \cdots, r_n) = 0) \leq \frac{d}{|S|}.$$

证明：用数学归纳法。显然$n = 1$时成立。假设$n - 1$时成立，则为n时

$$Q(x_1, \cdots, x_n) = P_{d_1}(x_2, \cdots, x_n){x_1}^{d_1} + P_{d_1-1}(x_2, \cdots, x_n){x_1}^{d_1-1} + \cdots + P_0(x_2, \cdots, x_n){x_1}^0.$$

$$\begin{aligned} Pr(Q = 0) = &Pr\left(P_{d_1}(x_2, \cdots, x_n) = 0\right) Pr\left(Q = 0 | P_{d_1}(x_2, \cdots, x_n) = 0\right) \\ &+ Pr\left(P_{d_1}(x_2, \cdots, x_n) \neq 0\right) Pr\left(Q = 0 | P_{d_1}(x_2, \cdots, x_n) \neq 0\right). \end{aligned} \tag{9.1}$$

注意到

$$\begin{aligned} &Pr\left(P_{d_1}(x_2, \cdots, x_n) = 0\right) \leq \frac{d - d_1}{|S|} \\ &Pr\left(Q = 0 | P_{d_1}(x_2, \cdots, x_n) = 0\right) \leq 1 \\ &Pr\left(P_{d_1}(x_2, \cdots, x_n) \neq 0\right) \leq 1 \\ &Pr\left(Q = 0 | P_{d_1}(x_2, \cdots, x_n) \neq 0\right) \leq \frac{d_1}{|S|} \end{aligned} \tag{9.2}$$

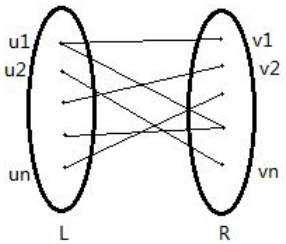
代入到(9.1)可得

$$Pr(Q = 0) \leq \frac{d - d_1}{|S|} + \frac{d_1}{|S|} = \frac{d}{|S|}.$$

证毕。

9.3 Perfect Matching

如下图所示的二分图，存在Perfect Matching的充分必要条件是$\Gamma(S) \geq |S|$.



下面利用代数化方法。首先写出上图对应的邻接矩阵A, 然后将其中值为1的元素改成变量x_{ij}, 得到矩阵A', 则图中存在Perfect Matching的充分必要条件是$det(A') \neq 0$.

证明：若图中存在Perfect Matching, 则对应矩阵A'中一定存在n个不同行且不同列的非零元素，导致$det(A') \neq 0$. 反之易证。

9.4 交互式证明系统(Interactive Proof System)

问题描述：给定3SAT问题φ, 则φ为真的赋值数为$\#\{x|\varphi(x)=T\}$. 现在给定一个具体的

$$\varphi(x_1,\cdots,x_n)=(x_1\vee\overline{x_2}\vee x_3)\wedge(\overline{x_1}\vee x_3\vee\overline{x_5})\wedge\cdots$$

Alice要向Bob(BPP的图灵机)证明φ为真的取值数为M.

采用交互式证明。将φ写成多项式形式

$$f(x_1,\cdots,x_n)=(1-(1-x_1)x_2(1-x_3))(1-x_1(1-x_3)x_5)\cdots$$

Alice即向Bob证明

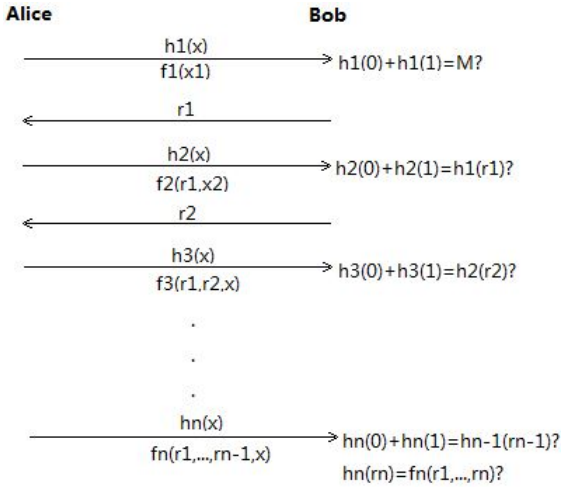
$$M=\sum_{x_1,\cdots,x_n:0}^{1}f(x_1,\cdots,x_n).$$

如下图所示，随机取值来自集合$S=\{1,2,\cdots,N\}$并且

$$f_1(x_1)=\sum_{x_2,\cdots,x_n=0}^{1}f(x_1,\cdots,x_n)$$

$$f_2(x_1,x_2)=\sum_{x_3,\cdots,x_n=0}^{1}f(x_1,\cdots,x_n)$$

$$\cdots$$



Bob只需对结果进行验证并随机地从S中选择数。如果所有的验证都能通过，则我们可以以很高的概率证明$\#3SAT=M$.

证明：

$$Pr(\text{所有验证均通过，}M\neq\sum f)$$
$$=Pr(\text{所有验证均通过，}(h_n\neq f_n)\vee(h_n=f_n,h_{n-1}\neq f_{n-1})\vee\cdots\vee(h_n,h_{n-1},\cdots,h_1\neq f_1)).\tag{9.3}$$

根据Union Bound,

$$\begin{aligned}\text{式}(9.3) &\leq \frac{deg(x_n)}{|S|} + \frac{deg(x_{n-1})}{|S|} + \cdots + \frac{deg(x_1)}{|S|} \\ &\leq \frac{3mn}{|S|} \\ &= \frac{O(mn)}{|S|}.\end{aligned}$$

通过给定一个较大的$|S|$, 可以确保

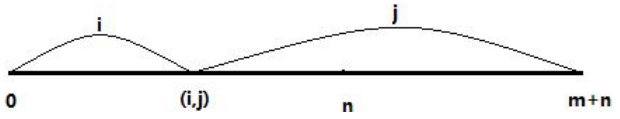
$$Pr(\text{所有验证均通过，} M \neq \sum f) = o(1).$$

证毕。

10 随机游走(Random Walk)

10.1 一般情形

问题描述：一个赌徒初始时有本钱n, 他每赌一次钱将以$\frac{1}{2}$的概率增加1以$\frac{1}{2}$的概率减少1. 他不停地赌下去，直到钱输光或者赚了m就停止(这里0和$m+n$可看做数轴上的吸收壁)。示意图如下



定义$p(i,j)$表示初始位置为i并最终在$m+n$处停止(被吸收)的概率，则

$$\begin{aligned}&p(0, m+n) = 0 \\ &p(m+n, 0) = 1 \\ &p(i,j) = \frac{1}{2}p(i+1, j-1) + \frac{1}{2}p(i-1, j+1)\end{aligned}$$

解得

$$p(i,j) = \frac{i}{m+n}.$$

定义$T(i,j)$表示初始位置为i并到达吸收壁需要的时间。令$m=n$, 则

$$\begin{aligned}&T(0, 2n) = 0 \\ &T(2n, 0) = 0 \\ &T(i,j) = 1 + \frac{1}{2}T(i-1, j+1) + \frac{1}{2}T(i+1, j-1)\end{aligned}$$

具体解法：

$$
\begin{aligned}
&T(2n-2) = 2T(2n-1) - 2\\
&T(2n-3) = 3T(2n-1) - 6\\
&T(2n-4) = 4T(2n-1) - 12\\
&\cdots\\
&T(2n-(2n-2)) = (2n-2)T(2n-1) - (2n-3)(2n-2)\\
&T(2n-(2n-1)) = (2n-1)T(2n-1) - (2n-2)(2n-1)
\end{aligned}
$$

即

$$
\begin{aligned}
T(2) &= (2n-2)T(2n-1) - (2n-3)(2n-2)\\
T(1) &= (2n-1)T(2n-1) - (2n-2)(2n-1)
\end{aligned}
\tag{10.1}
$$

又

$$
2T(1) = T(2) + 2 \tag{10.2}
$$

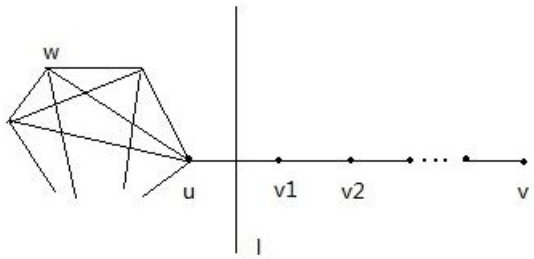
由(10.1)(10.2)解得$T(2n-1) = 2n-1$, 则

$$
\begin{aligned}
T(n) &= T(2n-n)\\
&= nT(2n-1) - (n-1)n\\
&= n^2.
\end{aligned}
$$

即$T(n,n) = n^2$.

10.2 将随机游走拓展到连通图上

给定一个无向连通图G, 结点u为出发点，v为吸收壁。定义d_u表示u的度数，u以概率$\frac{1}{d_u}$向邻接的结点随机移动，一旦到达v就停止。设$h_{u,v}$表示u到v的期望时间，自然地我们会考虑$h_{u,v}$是否等于$h_{v,u}$(此时v为出发点，u为吸收壁)。直观上感觉是相等的，但实际上是不等的。下面通过一个例子给出证明。



如上图所示，直线l左边是n个结点组成的完全图，右边同样有n个结点。根据已知有$h_{v,u}=n^2$, 对$h_{u,v}$有

$$h_{u,v}=\frac{1}{n}h_{v_1,v}+\frac{n-1}{n}h_{w,v}+1 \tag{10.3}$$

$$h_{w,v}=\frac{1}{n-1}h_{u,v}+\frac{n-2}{n-1}h_{w,v}+1 \tag{10.4}$$

$$h_{v_1,v}=\frac{1}{2}h_{u,v}+\frac{1}{2}h_{v_2,v}+1 \tag{10.5}$$

由(10.4)可得

$$h_{w,v}=h_{u,v}+(n-1) \tag{10.6}$$

由(10.5)可得

$$h_{v_1,v}-h_{v_2,v}=h_{u,v}-h_{v_1,v}+2 \tag{10.7}$$

由(10.3)得

$$nh_{u,v}=h_{v_1,v}+(n-1)h_{w,v}+n$$

化简可得

$$h_{u,v}-h_{v_1,v}=(n-1)(h_{w,v}-h_{u,v})+n$$

根据(10.6)得

$$h_{u,v}-h_{v_1,v}=(n-1)^2+n=n^2-n+1 \tag{10.8}$$

由(10.7)和(10.8)可得$h_{u,v}=\Theta\left(n^3\right)$.

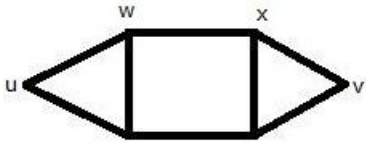
显然$h_{u,v}$和$h_{v,u}$具有不同的量级，即$h_{u,v}\neq h_{v,u}$.

11 电路问题

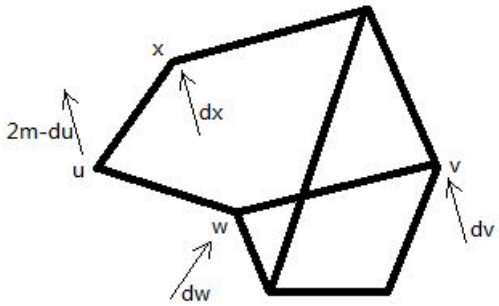
如下图所示，每个边表示阻值为1的电阻。利用电路的有关知识进行分析，可知$R_{u,v}=\frac{3}{2}$. 根据上一部分的相关内容，我们可以得出下式：

$$\begin{cases} h_{u,v}=1+h_{w,v} \\ h_{w,v}=1+\frac{1}{3}h_{u,v}+\frac{1}{3}h_{w,v}+\frac{1}{3}h_{x,v} \\ h_{x,v}=1+\frac{1}{3}h_{w,v}+\frac{1}{3}h_{x,v} \end{cases}$$

解得$h_{u,v}=12$.



$\frac{h_{u,v}}{R_{u,v}}=8$, 这个值刚好等于边数$m$, 猜想是否有$h_{u,v}=mR_{u,v}$成立。答案是否，但是有$h_{u,v}+h_{v,u}=2mR_{u,v}$.



证明：构造如图所示的电路(称为电路I)，满足u处的电势为0, 即$U_u = 0$. u以外的所有其他结点均流入自身度数单位的电流，由电路的平衡可知u 流出电流$2m - d_u$.

根据结点处的电流平衡可得

$$\begin{cases} \sum\limits_{x\in\Gamma(w)} (U_w - U_x) = d_w \ (\forall w \neq u) \\ U_u = 0 \end{cases} \tag{11.1}$$

再根据首达时间可得

$$\begin{cases} h_{w,u} = 1 + \frac{1}{d_w} \sum\limits_{x\in\Gamma(w)} h_{x,u} \ (\forall w \neq u) \\ h_{u,u} = 0 \end{cases}$$

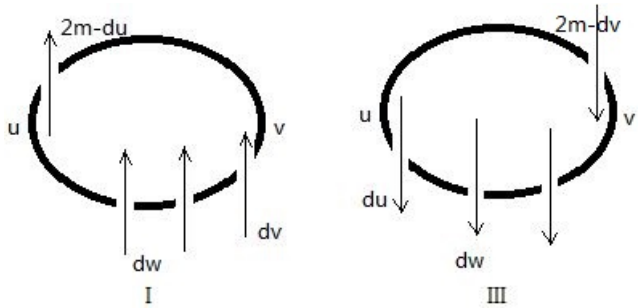
整理得

$$\begin{cases} d_w = \sum\limits_{x\in\Gamma(w)} (h_{w,u} - h_{x,u}) \ (\forall w \neq u) \\ h_{u,u} = 0 \end{cases} \tag{11.2}$$

将(11.1)(11.2)对应起来并根据解的唯一性可得$U_v = h_{v,u}$.

同理再构造另外一个电路(称为电路II)，满足$U_v = 0$. v以外的所有其他结点均流入自身度数单位的电流，u流出电流$2m - d_v$. 类似地可得$\forall w \neq v, U_u = h_{u,v}$. 将该电路的电流方向取反得到新的电路(称为电路III)，则有$U_u = -h_{u,v}$.

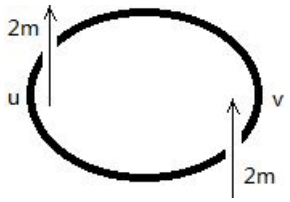
电路I和III的示意图如下：



对应相加起来可得

再根据欧姆定律可得

$$2mR_{u,v} = h_{v,u} - 0 + 0 - (-h_{u,v}).$$



即

$$h_{u,v} + h_{v,u} = 2mR_{u,v}.$$

证毕。

第II部分：近似算法

近似算法常用于求解NP问题。定义近似比为

$$\frac{A(I)}{opt(I)}$$

其中I为问题实例，A为近似算法，opt为最优算法。

1 旅行商问题(TSP)

首先阐述一些基本概念。

哈密尔顿路径：在一个给定的图中，一条路径经过所有顶点有且仅有一次。

哈密尔顿圈：上述路径构成一个圈，即起点和终点重合，其余所有结点都只被访问一次。**判断一个给定的图$G=(V,E)$中是否存在哈密尔顿圈是NP完全问题。**

旅行商问题(TSP)：给定一个**完全图**$G=(V,E)$, 求解图中边权之和最小的**哈密尔顿圈。TSP是NP完全问题。**

设给定的图为G, W为对应的边权，则W满足三种情形：

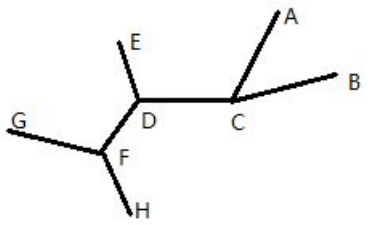
(1) 欧几里得空间。

(2) 三角不等式，这是最常讨论的情况。存在2近似和$\frac{3}{2}$近似。

(3) 一般情形，不存在常数近似。

1.1 W满足三角不等式TSP存在2近似

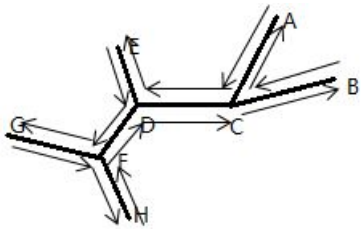
我们在这一小节阐述2近似算法A. 先在G上取得一个最小生成树(Kruskal算法或Prim算法)T, 如下图所示：



则T的边权之和为$w(T)$. 从A出发按照下图所示的路线构成了一个回路，显然该回路的长度为$2w(T)$.

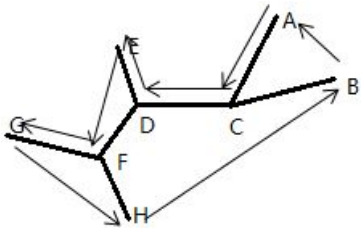
为了得到哈密尔顿圈(cycle), 我们从起点出发走到某个邻接结点，一直下去直到某个结点u的所有邻接结点都已走过，则u选择一个还未走过的结点v, 从v继续下去直到回到起点。对应示意图如下，这样我们就得到了TSP的一个解，即$A(I)$. 根据三角不等式可得

$$A(I) \leq 2w(T). \tag{1.1}$$



令$opt(I)$表示最优解，显然$opt(I)$是一个圈，去掉其中的任意一条边就得到了一个树T', 显然$w(T') \geq w(T)$. 又$opt(I) \geq w(T')$, 可得

$$opt(I) \geq w(T). \tag{1.2}$$

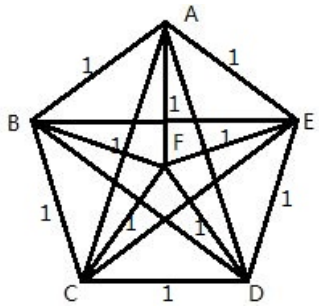


结合(1.1)(1.2)可得

$$A(I) \leq 2opt(I).$$

即证明了算法A是2近似的。

下面通过一个实例说明这个算法是紧的(tight). 如图所示为n(这里以$n = 6$为例)个结点组成的完全图，满足AB=BC=CD=DE=EA=AF=BF=CF=DF=EF=1, 其余边为2.

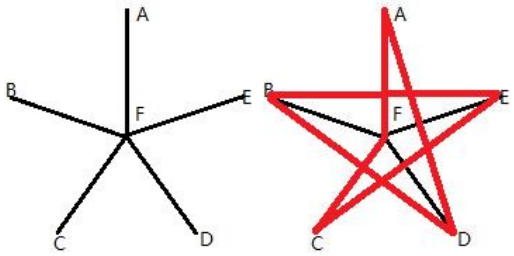


不难得到$opt(I) = n$, 对应图中为ABCDEFA. 根据算法A取最小生成树(左图)并按右图所示执行A, 即为AFCEBDA. 不难得到$A(I) = 2 * (n-2) + 2 = 2n - 2$.

近似比$\frac{A(I)}{opt(I)} = 2 - \frac{2}{n}$, 任给一个$2 - \varepsilon$, 总可以找到一个$n$满足，即证明了该2近似算法是紧的。

1.2 W满足三角不等式TSP存在$\frac{3}{2}$近似

同样先找出最小生成树T, 标出T中度数为奇数的结点，设这些结点构成集合O. 根据握手定理可得$|O|$为偶数，因此O中结点形成的完全图一定存在perfect matching,



我们可以以多项式时间找出一个最小权值的perfect matching, 设为M. 定义N为O上的TSP最优解，N_1是从N中取出的任一perfect matching, N_2是N_1在N上的补集，显然N_2也是O上的一个perfect matching, 则有

$$w(M) \leq min\{w(N_1), w(N_2)\} \tag{1.3}$$

又

$$w(N_1) + w(N_2) = w(N) \tag{1.4}$$

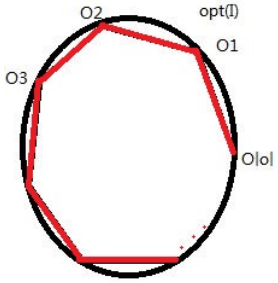
可得

$$min\{w(N_1), w(N_2)\} \leq \frac{1}{2}w(N) \tag{1.5}$$

由(1.3)(1.5)得

$$w(M) \leq \frac{1}{2}w(N). \tag{1.6}$$

设$opt(I)$为最优解，我们在$opt(I)$中只考虑集合O中的结点。如下图所示，这样我们就得到了一个O上的哈密尔顿圈M'.



根据三角不等式可得

$$w(M') \leq opt(I) \tag{1.7}$$

又由于N是O上的TSP最优解，所以

$$w(N) \leq w(M'). \tag{1.8}$$

由(1.6)(1.7)(1.8)可得

$$w(M) \leq \frac{1}{2}opt(I) \tag{1.9}$$

又

$$w(T) \leq opt(I) \tag{1.10}$$

得

$$w(T) + w(M) \leq \frac{3}{2} opt(I). \tag{1.11}$$

在T上添加M之后所得的图每个结点的度数均为偶数，因此存在欧拉回路E. 在E上执行1.1节中的做法，即从E的起点出发走到某个邻接结点，一直下去直到某个结点u的所有邻接结点都已走过，则u选择一个还未走过的结点v, 从v继续下去直到回到起点。经过这样的处理得到最终的$A(I)$, 根据三角不等式可得

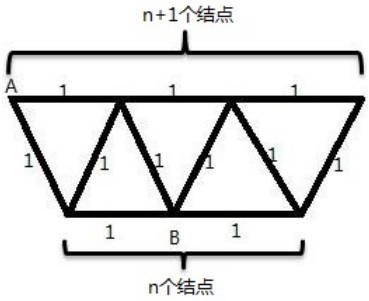
$$A(I) \leq w(T) + w(M). \tag{1.12}$$

由(1.11)(1.12)得

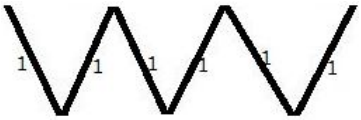
$$A(I) \leq \frac{3}{2} opt(I). \tag{1.13}$$

证毕。

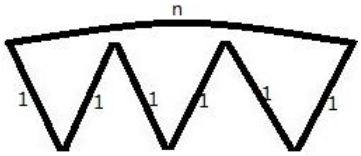
下面证明该$\frac{3}{2}$近似算法是紧的。给出一个具体的例子如下。



定义没有边直接相连的两个结点(设为A,B)之间的边权$w(A, B)$ =AB之间的最短路径(显然满足三角不等式)。显然$opt(I) = 2n + 1$. 对应最小生成树如下：



根据上述算法，最终得到的哈密尔顿圈为：



即$A(I) = n + 2n = 3n$, 近似比$= \frac{3n}{2n+1}$. 证毕。

1.3 一般情形下不存在常数近似

给定图$G = (V, E)$, 假设存在多项式时间算法A为100近似，即有$A(G) \leq 100 opt(G)$. 构造$\widetilde{G} = (V, E')$, 满足对任一$e \in E'$, 若$e \in E$则$w(e) = 1$, 否则$w(e) = 200n$, 即$\widetilde{G}$**是一个完全图**。

若G中存在哈密尔顿圈，则$opt(\widetilde{G}) = n$, $A(\widetilde{G}) \leq 100n$;

若G中不存在哈密尔顿圈，则$A(\widetilde{G}) > 200n$.

可以看到通过在$\widetilde{G}$上执行算法A, 可以判定图G中是否存在哈密尔顿圈。即多项式时间算法A解决了NP完全问题，即证明了P=NP, 产生矛盾。证毕。

2 MAX-SAT

问题描述：给出$\varphi = c_1 \wedge c_2 \wedge \cdots \wedge c_m$, 找出可满足的最多的$c_i$数目。

Alg 1.考虑用随机算法，对变量x_i以$\frac{1}{2}$的概率取1以$\frac{1}{2}$的概率取0. 子句

$$c_i = \overline{x_{i_1}} \vee x_{i_2} \vee \cdots \vee x_{i_k}$$

为真的概率等于$1 - \frac{1}{2^k}$. 定义随机变量Y_i表示第i个子句是否为真，X表示m个子句中为真的数目，则

$$E(X) = E(Y_1) + E(Y_2) + \cdots + E(Y_m) = (1 - \frac{1}{2^k})m.$$

显然$opt(I) \leq m$, 即该近似算法的期望近似比$\geq 1 - \frac{1}{2^k}$.

Alg 2.将该问题化为整数规划(ILP):

对于某个$c_i = \overline{x_1} \vee x_2 \vee \overline{x_3} \vee x_4$, 对应的$y_i = (1 - x_1) + x_2 + (1 - x_3) + x_4$.

$$\begin{aligned}
&\max\ y_1 + \cdots + y_m \\
&s.t. \\
&\begin{cases}
\cdots \\
(1 - x_1) + x_2 + (1 - x_3) + x_4 \geq y_i (\forall i) \\
\cdots \\
x_1, \cdots, x_n \in \{0, 1\} \\
y_1, \cdots, y_m \in \{0, 1\}
\end{cases}
\end{aligned}$$

化成线性规划(LP)形式：

$$\begin{aligned}
&\max\ y_1 + \cdots + y_m \\
&s.t. \\
&\begin{cases}
\cdots \\
(1 - x_1) + x_2 + (1 - x_3) + x_4 \geq y_i (\forall i) \\
\cdots \\
x_1, \cdots, x_n \in [0, 1] \\
y_1, \cdots, y_m \in [0, 1]
\end{cases}
\end{aligned}$$

可在多项式时间内解得opt_{LP}, 则$opt_{ILP} \leq opt_{LP}$. 经过LP求解得$x_1^*, x_2^*, \cdots, x_n^*$及$y_1^*, y_2^*, \cdots, y_m^*$, 其中$x_i^*, y_j^* \in [0, 1]$. 以$x_i^*$的概率设置$x_i$为1, 以$y_j^*$的概率设置$y_j$为1.

对于$c_i = \overline{x_1} \vee x_2 \vee \overline{x_3} \vee x_4$有，

$$y_i^* \leq (1 - x_1^*) + x_2^* + (1 - x_3^*) + x_4^*.$$

左端为1的概率即为y_i^*, 右端为1的概率

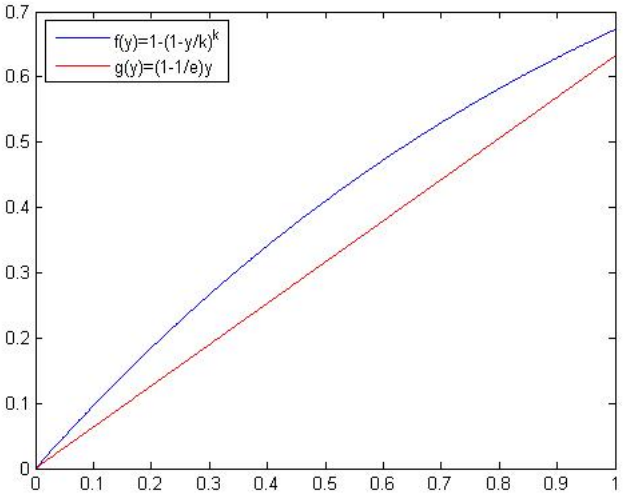
$$\begin{aligned} p &= 1 - x_1^*(1 - x_2^*)x_3^*(1 - x_4^*) \\ &\geq 1 - \left[\frac{x_1^* + (1 - x_2^*) + x_3^* + (1 - x_4^*)}{4}\right]^4 \\ &\geq 1 - \left(\frac{4 - y_i^*}{4}\right)^4 \\ &= 1 - \left(1 - \frac{y_i^*}{4}\right)^4. \end{aligned}$$

令$f(y) = 1 - \left(1 - \frac{y}{k}\right)^k$, 则

$$\begin{aligned} f'(y) &= \left(1 - \frac{y}{k}\right)^k \geq 0 \\ f''(y) &= (k-1)\left(1 - \frac{y}{k}\right)^{k-2}\left(-\frac{1}{k}\right) < 0 \end{aligned}$$

且

$$f(1) = 1 - \left(1 - \frac{1}{k}\right)^k \geq 1 - \frac{1}{e}.$$



因此

$$1 - \left(1 - \frac{y}{k}\right)^k \geq \left(1 - \frac{1}{e}\right) y.$$

$$\Pr(c_i=1) \geq 1-\left(1-\frac{y_i^*}{k}\right)^k \geq \begin{cases} y_i^* & ,k=1 \\ y_i^* - \frac{{y_i^*}^2}{4} \geq \frac{3y_i^*}{4} & ,k=2 \\ y_i^* - \frac{{y_i^*}^2}{3} + \frac{{y_i^*}^3}{27} \geq \frac{2y_i^*}{3} & ,k=3 \\ \left(1-\frac{1}{e}\right) y_i^* & ,k \geq 4 \end{cases}$$

可满足的期望值

$$\sum_{i=1}^{m} \Pr(c_i=1) \geq \sum \left(1-\frac{1}{e}\right) y_i^* = \left(1-\frac{1}{e}\right) \sum y_i^* = \left(1-\frac{1}{e}\right) opt_{LP} \geq \left(1-\frac{1}{e}\right) opt_{ILP}.$$

近似比$\geq 1-\frac{1}{e}$.

对比Alg 1和Alg 2:

k	1	2	3	$\cdots$	
Alg 1的近似比	$\frac{1}{2}$	$\frac{3}{4}$	$\frac{7}{8}$	$\cdots$	1
Alg 2的近似比	1	$\frac{3}{4}$	$\frac{2}{3}$	$\cdots$	$1-\frac{1}{e}$

考虑同时运行两个算法，取最优解，则至少取得$\frac{3}{4}$的近似比。

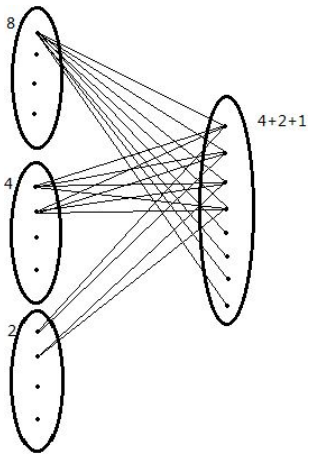
3 Vertex Cover

问题形式化：给定图$G=(V,E)$,

$$\begin{aligned} &\min\ S \subseteq V \\ &s.t.\ (u,v) \in E,\ u \in S\ or\ v \in S \end{aligned}$$

3.1 贪心算法

找出度数最大的结点v_1, 在G中去除v_1及其连接的所有边得到图G', 在G'中继续此操作直到图中没有边存在。该算法有时会很差，近似比量级为$\Theta(log|V|)$. 下面给出一个具体例子。



如图所示，右边有$N = 2^n$(图中以$n = 3$为例)个结点，左边有n组(每组有2^{n-1}个结点)结点。每组结点内部没有边相连，图中左边与右边结点连接规则为：

左边第一组每个结点度数为2^n(使右边每个结点度数增加2^{n-1}), 左边第二组每个结点度数为2^{n-1}(使右边每个结点度数增加2^{n-2}), $\cdots$, 直到第n组每个结点的度数为2^1(使右边每个结点度数增加2^0).

可知$opt(I) = 2^n$, $A(I) = n2^{n-1}$, 近似比$= \frac{n}{2} = \frac{1}{2}logN \sim \Theta(log|V|)$.

3.2 ILP & LP

写成ILP形式：

$$\min \sum x_i$$

$$s.t.$$

$$\begin{cases} \forall (u,v) \in E,\ x_u + x_v \geq 1 \\ x_1, \cdots, x_n \in \{0,1\} \end{cases}$$

对应LP形式则将约束条件改为$x_i \in [0,1]$, 得到LP的最优解为

$$x_1^*, \cdots, x_n^*$$

直接四舍五入得

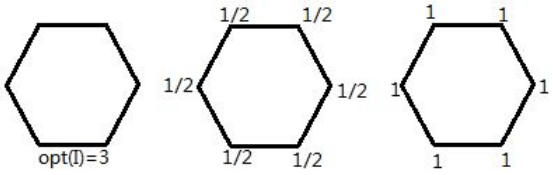
$$y_1, \cdots, y_n.$$

由$x_u^* + x_v^* \geq 1$可得x_u^*与x_v^*至少有一个≥ 0.5, 也即y_u与y_v中至少有一个等于1, 这样就满足了约束条件(和≥ 1).

$$\sum y_i \leq \sum 2x_i^* = 2opt_{LP} \leq 2opt_{ILP}.$$

即得到了2近似算法。

下面通过一个例子说明该2近似是紧的。

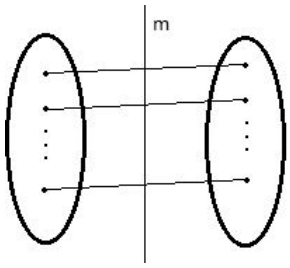


如图所示，最优解为3, 根据IP计算出来每个结点的值为$\frac{1}{2}$, 经过四舍五入每个结点值为1, 即$A(I) = 6$, 近似比=2.

3.3 极大perfect matching

如下图所示，在图G上找到一个极大的perfect matching, 割边的数目为m, 取这m条边对应的$2m$个结点为$A(I)$. 显然$opt(I) \geq m$,

$$A(I) = 2m \leq 2opt(I).$$



即该算法是2近似的。