高级算法复习

Advanced Algorithm, 20 Fall

2021 年 1 月 4 日

1 Cut

1.1 Min-Cut

Algorithm 1 Karger's min cut algorithm

Input: multigraph G(V,E)
Output: a cut

```
while |V| > 2 do
    choose a uniform e ∈ E
    contract(e)
end while
return remaining edges
```

该算法以不低于 $\frac{2}{n(n-1)}$ 的概率返回最小割，重复该算法 $\frac{n(n-1)}{2}$ 次，得不到最小割的概率满足：

$$\Pr[\text{fail}] \leq (1 - \frac{2}{n(n-1)})^{\frac{n(n-1)}{2}} < \frac{1}{e}$$

算法分析见课件或者讲义.

* **推论**: n 节点的图中独立最小割的个数至多是 $\frac{n(n-1)}{2}$

Algorithm 2 Fast Min Cut

Input: G

```
if |V| ≤ 6 then
    return a min-cut by brute force
else
    set t = ⌈1 + n/√2⌉
    G1 = Contract(G, t)
    G2 = Contract(G, t)
    return min(FastCut(G1), FastCut(G2))
end if
```

* **定理**: FastCut runs in time $O(n^2 logn)$ and returns a min-cut with probability $\Omega(\frac{1}{logn})$ 分析见课件

1.2 Max-Cut

Algorithm 3 Greedy

S=T=$\emptyset$
for i=1,2...n **do**
  v_i joins one of S,T to maximize current E(S,T)
end for

* **结论**: 贪心算法的近似率为 $\frac{1}{2}$

Algorithm 4 Random

S=T=$\emptyset$
for i=1,2...n **do**
  v_i joins one of S,T
end for

* **结论**: 随机算法的近似率为 $\frac{1}{2}$
* **重点**: 随机算法可以 Derandomization 成贪心算法. 见作业

2 Chernoff Bound

2.1 Balls into Bins

m 个小球随机独立的投入到 n 个桶中:

* 每个球所在的桶只有它自己一个球的概率 (生日问题)
  对于 $m = \sqrt{2nln\frac{1}{\epsilon}}$, Pr[no bin with more than two balls] = ϵ

* 没有空桶的概率 (收集问题)
  定理 * 设 X 为投入 n 个桶直到没有空桶所需要的球的个数，$E[X] = nH(n)$. (n 为调和级数)

* 桶中球的最大个数 (负载均衡问题)

  - $m = \theta(n)$,the max load is $O(\frac{\log n}{\log\log n})$ whp.
  - $m = n$, the max load is $O(\log\log n)$ whp.

2.2 Tail inequality

* Markov 不等式: $\Pr[X \geq t] \leq \frac{E[X]}{t}$

* Markov 一般形式: $\Pr[f(X) \geq t] \leq \frac{E[f(X)]}{t}$

* Chebyshev 不等式: $\Pr[|X - E[X]| \geq t] \leq \frac{Var[X]}{t^2}$

* Chernoff Bound 一般形式:

对于一系列独立随机变量 $X_1, X_2 \cdots X_n \in \{0,1\}$，令 $X = \sum_{i=1}^{n} X_i$，$\mu = E[X]$，那么有:

$$\forall \delta > 0, \Pr[X \geq (1+\delta)\mu] \leq (\frac{e^{\delta}}{(1+\delta)^{(1+\delta)}})^{\mu}$$

$$\forall 0 < \delta < 1, \Pr[X \leq (1-\delta)\mu] \leq (\frac{e^{-\delta}}{(1-\delta)^{(1-\delta)}})^{\mu}$$

* Chernoff Bound 实用形式:

对于一系列独立随机变量 $X_1, X_2 \cdots X_n \in \{0,1\}$，令 $X = \sum_{i=1}^{n} X_i$，$\mu = E[X]$，那么有:

$$\begin{aligned}
&\forall 0 < \delta < 1 \\
&\Pr[X \geq (1+\delta)\mu] \leq exp(-\frac{\mu\delta^2}{3}) \\
&\Pr[X \leq (1-\delta)\mu] \leq exp(-\frac{\mu\delta^2}{2}) \\
&\text{for } t \geq 2e\mu \\
&\Pr[X \geq t] \leq 2^{-t}
\end{aligned}$$

* Hoffeding 不等式

对于一系列独立随机变量 $X_1, X_2 \cdots X_n, X_i \in [a_i, b_i]$，令 $X = \sum_{i=1}^{n} X_i$，那么有：

$$\begin{aligned}
&\forall t > 0 \\
&\Pr[X \geq E[X] + t] \leq exp(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}) \\
&\Pr[X \leq E[X] - t] \leq exp(-\frac{2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2})
\end{aligned}$$

* Hoffeding 实用形式：

对于一系列独立随机变量 $X_1, X_2 \cdots X_n, X_i \in \{0,1\}$，令 $X = \sum_{i=1}^{n} X_i$，那么有：

$$\begin{aligned}
&\forall t > 0 \\
&\Pr[X \geq E[X] + t] \leq exp(-\frac{2t^2}{n}) \\
&\Pr[X \leq E[X] - t] \leq exp(-\frac{2t^2}{n})
\end{aligned}$$

* Hoffeding lemma：

对于任意随机变量 $X \in [a,b]$，若 $E[X] = 0$, 有:

$$E[e^{\lambda X}] \leq exp(\frac{\lambda^2(b-a)^2}{8})$$

3 Martingale

3.1 条件概率和条件期望

* $\Pr[A|B] = \frac{\Pr[A \wedge B]}{\Pr[B]}$

* $E[X|A] = \sum_x x \cdot \Pr[X = x|A]$
* $E[Y] = E[E[Y|X]]$
* $E[Y|Z] = E[E[Y|X,Z]|Z]$
* $E[E[f(X)g(X,Y)|X]] = E[f(X)E[g(X,Y)|X]]$

3.2 Martingales 的定义

一系列随机变量 $X_0, X_1 \cdots$ 是一个 Martingale，如果满足:

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

3.3 Azuma 不等式

设 $X_0, X_1 \cdots$ 是一个 martingale，且满足对于任意 $k \geq 1$，有:

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

则:

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

3.4 广义的 Martingale

一系列随机变量 $Y_0, Y_1 \cdots$ 是关于序列 $X_0, X_1 \cdots$ 的 martingale，如果满足:

$\forall i \geq 0$

* $Y_i$is a function of $X_0, X_1 \cdots X_i$
* $E[Y_{i+1}|X_0, X_1 \cdots X_i] = Y_i$

3.5 广义的 Azuma 不等式

设 $Y_0, Y_1 \cdots$ 是关于 $X_0, X_1 \cdots$ 的 martingale，且满足对于任意的 $k \geq 1$:

$$|Y_k - Y_{k-1} \leq c_k|$$

则有:

$$P(|Y_n - Y_0| \geq t) \leq 2exp(-\frac{t^2}{2\sum_{k=1}^{n}} c_k^2)$$

3.6 Doob Sequence

一个和序列 $X_1 \cdots X_n$ 有关的函数 f 的 Doob Sequence 是:

$$Y_i = E[f(X_1 \cdots X_n)|X_1, \cdots X_i]$$

特别的，$Y_0 = E[f(X_1, \cdots, X_n)], Y_n = f(X_1, \cdots, X_n)$

3.7 Doob Martingale

一个函数 f 的 Doob Sequence 是一个 martingale，就是说:

$$E[Y_i|X_1,\cdots X_{i-1}]=Y_{i-1}$$

4 Fingerprinting

4.1 有限域理论

Let S be a set, **closed** under binary operations $+$ (addition) and $\cdot$ (multiplication). It gives us the following algebraic structures if the corresponding set of axioms are satisfied.

Structures							Axioms	Operations
field	*commutative ring*	*ring*	*abelian group*	*group*	*monoid*	*semigroup*	1. **Addition** is **associative**: $\forall x,y,z\in S,(x+y)+z=x+(y+z)$.	$+$
							2. Existence of **additive identity** 0: $\forall x\in S, x+0=0+x=x$.	
							3. Everyone has an **additive inverse**: $\forall x\in S,\exists -x\in S$, s.t. $x+(-x)=(-x)+x=0$.	
							4. **Addition** is **commutative**: $\forall x,y\in S, x+y=y+x$.	
							5. Multiplication **distributes** over addition: $\forall x,y,z\in S, x\cdot(y+z)=x\cdot y+x\cdot z$ and $(y+z)\cdot x=y\cdot x+z\cdot x$.	$+,\cdot$
							6. **Multiplication** is **associative**: $\forall x,y,z\in S,(x\cdot y)\cdot z=x\cdot(y\cdot z)$.	$\cdot$
							7. Existence of **multiplicative identity** 1: $\forall x\in S, x\cdot 1=1\cdot x=x$.	
							8. **Multiplication** is **commutative**: $\forall x,y\in S, x\cdot y=y\cdot x$.	
							9. Every non-zero element has a **multiplicative inverse**: $\forall x\in S\setminus\{0\},\exists x^{-1}\in S$, s.t. $x\cdot x^{-1}=x^{-1}\cdot x=1$.	

图 1: 有限域理论

4.2 典型的域

- 无限域:$\mathbb{Q},\mathbb{R},\mathbb{C}$
- 有限域:
  * Prime Field $\mathbb{Z}_p$: 对于所有大于 1 的整数 n，$\mathbb{Z}_n=\{0,1,\cdots,n-1\}$ 在模 p 的意义下构成了 commutative ring，若 p 是素数，则 $\mathbb{Z}_p$ 是一个域。
  * Boolean arithmetics $GF(2)$：$\{0,1\}$ 在“异或”作为加法，“与”作为乘法时，构成域

4.3 多项式

设 $\mathbb{F}[x_1,x_2\cdots,x_n]$ 是域 $\mathbb{F}$ 上的 n 变量的多项式构成的环。那么对于 $f\in\mathbb{F}$:

$$f(x_1,x_2,\cdots,x_n)=\sum_{i1,i2\cdots,i_n\geq 0}a_{i_1,i_2\cdots,i_n}x_1^{i_1}x_2^{i_2}\cdots x_n^{i_n}$$

多项式 f 的度 (degree) 为: 满足 $a_{i_1,i_2\cdots,i_n}\neq 0$ 的最大的 $i_1+i_2+\cdots+i_n$

4.4 Polynomial Identity Testing(PIT)

问题，能否判定一个多项式是否恒为 0?

- 输入：$f\in\mathbb{F}[x_1,x_2\cdots,x_n]$ of degree d.
- 输出：$f\equiv 0$?

4.5 Schwartz-Zippel 定理

针对 PIT 问题，有 Schwartz-Zippel 定理：

$$f \not\equiv 0 \Longrightarrow \forall S \subset \mathbb{F},\ \Pr[f(r_1, r_2, \cdots, r_n) = 0] \leq \frac{\text{degree}(f)}{|S|}$$

4.6 Fingerprinting 问题定义

我们希望有这样的函数 $\textsc{Fing}$，它满足:

- 若 $X = Y$，则 $\textsc{Fing}(X) = \textsc{Fing}(Y)$
- 若 $X \neq Y$，我们希望 $\Pr[\textsc{Fing}(X) = \textsc{Fing}]$ 尽量小

那么可以将问题转化为 PIT 问题，并使用 Schwartz-Zippel 定理。

5 Hashing

5.1 估计集合大小

- 输入: 一个序列 $x_1, x_2 \cdots x_n \in \Omega$（有重复元素）
- 输出: $z = |\{x_1, x_2 \cdots x_n\}|$ 的估计值

假设有 k 个独立哈希函数 $h_1, h_2 \cdots h_k : \Omega \to [0, 1]$，设 $Y_j = \min_{1\leq i\leq n} h_j(x_i)$。设 $\bar{Y} = \frac{1}{k}\sum_{j=1}^{k} Y_j$。使用 $\widehat{Z} = \frac{1}{\bar{Y}} - 1$ 作为估计值。
可以证明（见课件）:

$$\Pr[\widehat{Z} > (1+\epsilon)z \text{ or } \widehat{Z} < (1-\epsilon)z] \leq \frac{4}{\epsilon^2 k} \leq \delta, \quad (\text{choose } k = \frac{4}{\epsilon^2 \delta})$$

5.2 其他应用

详情见课件

- Frequency Estimation
- Bloom Filters
- Heavy Hitters
- Count-Min Sketch

6 Nearest Neighbor Search(NNS)

6.1 Johnson-Lindenstrauss Theorem (JLT)

JLT 的原始描述: 对于任意的 ϵ，对于任意的 d 维空间的大小为 n 的点集 S, $S \in \mathbb{R}^d$, $|S| = n$，都有一个降维函数 $\Phi : \mathbb{R}^d \to \mathbb{R}^k$, $k = O(\epsilon^{-2} \log n)$，满足:

$$\forall x, y \in S, (1-\epsilon)\|x-y\|_2^2 \leq \|\Phi(x) - \Phi(y)\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2$$

JLT 的另一种描述方式: 对于任意的 ϵ，对于任意的 d 维空间的大小为 n 的点集 S, $S \in \mathbb{R}^d$，$|S| = n$，都存在一个矩阵 $A \in \mathbb{R}^{k\times d}$，$k = O(\epsilon^{-2}\log n)$，满足:

$$\forall x, y \in S, (1-\epsilon)\|x-y\|_2^2 \leq \|Ax - Ay\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2$$

JLT 的 Probabilistic method 描述: 对于一个随机的 $A \in \mathbb{R}^{k\times d}$

$$\Pr[\forall x, y \in S, (1-\epsilon)\|x-y\|_2^2 \leq \|Ax - Ay\|_2^2 \leq (1+\epsilon)\|x-y\|_2^2] \geq 1 - O(\frac{1}{n}) \quad (w.h.p)$$

JLT 的两种证明方法:

- 根据高斯分布来采样出 A，证明 $\forall$unit vector $u \in \mathbb{R}^d$，$\Pr[\|Au\|_2^2 - 1 \geq \epsilon] < \frac{1}{n^3}$
- Projection

6.2 NNS 的问题定义

发现最近邻在高维空间很难，考虑两种方法:

- 寻找近似近邻
- 使用 Local Sensitive Hash 进行降维

7 Greedy Algorithm

几个例子

- Set Cover
- Vertex Cover
- Scheduling
- Longest Processing Time(LPT)
- Online Scheduling

8 Dynamic Programming(DP)

3 个例子:

- 背包问题
- Bin-Packing
- Scheduling

近似求解的思想:

- Scaling and Rounding: 对原始数据进行粗化，得到新数据，新数据一般规模较小，可以求出新数据的最优解来近似原始数据的最优解.
- Grouping and Rounding: 和 Scaling and Rounding 类似.

9 Linear Programming(LP)

想法, 整数线性规划是 NP-Hard, 实数上的线性规划可以高效求解。所以可以进行 Relaxation and Rounding。

10 Prime Dual

理解强对偶性：对偶问题的上界是原问题的下界。例子: Metric Facility Location

11 Semidefinite Programming(SDP) & Sum of Square(SOS)

非线性约束的问题可以将变量从标量放到矢量，再 Rouding 到原始的空间中。

12 Lovasz Local Lemma(LLL)

12.1 LLL

LLL 的直观理解：定义一组坏事件，如果这些坏事件的独立性满足一些条件，那么如果每个坏事件单独发生的概率不高于一个量，这些坏事件以非 0 的概率都不发生。

LLL 版本 1:

m bad events $A_1, A_2 \cdots A_m$, every A_i is independent of all but $\leq d$other bad events

$$\forall i: \ \Pr[A_i] \leq \frac{1}{4d} \Longrightarrow \Pr[\bigwedge_{i=1}^{m} \bar{A}_i] > 0$$

LLL 版本 2:

m bad events $A_1, A_2 \cdots A_m$, every A_i is independent of all but $\leq d$other bad events

$$\forall i: \ \Pr[A_i] \leq \frac{1}{e(d+1)} \Longrightarrow \Pr[\bigwedge_{i=1}^{m} \bar{A}_i] > 0$$

注意: 这里的独立应该是如果每个坏事件都由一些变量来定义，那么独立就意味着不共享任何的变量。

12.2 Moser-Tardos Algorithm Framework

这是一种随机算法设计框架:

Algorithm 5 Moser-Tardos Algorithm

```
sample all X_1, ..., X_n
while ∃ an occurring bad event A_i:  do
    resample all X_j ∈ vbl(A_i)
end while
```

结论 *:

- 如果 $\forall i: \Pr[A_i] \leq \frac{1}{4d}$，则期望意义上经过 $m/(2d-1)$ 次重新采样后，会得到一个满足的解。
- 如果 $\forall i: \Pr[A_i] \leq \frac{1}{e(d+1)}$，则期望意义上经过 m/d 次重新采样后，会得到一个满足的解。