高级算法设计与分析 Lecture 2

授课时间: 2020 年 2 月 24 日　授课教师: 孙晓明　记录人: 朱钦霖

1 概率基础

1.1 Markov Inequality

设 $r.v.\ X \geq 0$，则对 $\forall\ c > 0$，有

$$\Pr(X \geq c) \leq \frac{\mathbb{E}(X)}{c}$$

证明　假设 X 是连续型随机变量（离散型更易证明）

$$\begin{aligned}\mathbb{E}(X) &= \int_0^{+\infty} xf(x)dx \\ &\geq \int_c^{+\infty} xf(x)dx \\ &\geq \int_c^{+\infty} cf(x)dx \\ &= c\int_c^{+\infty} f(x)dx = c\Pr(X \geq c)\end{aligned}$$

□

例：设随机变量序列 $X_1, X_2, \ldots$ 互相独立，与 X 同分布，$\Pr(X=1)=p, \Pr(X=0)=1-p$。定义 T 为第一次出现 $X=1$ 时的实验次数，即 $T=\min\{t|X_t=1\}$.

T 的分布如式 $\Pr(T=k)=(1-p)^{k-1}p$，故

$$\begin{aligned}\mathbb{E}(T) &= \sum_{k=1}^{+\infty} k(1-p)^{k-1}p \\ &= p\sum_{k=1}^{+\infty}(x^k)'|_{x=1-p} \\ &= p(\sum_{k=1}^{+\infty} x^k)' \\ &= p(\frac{1}{1-x})' \\ &= p\frac{1}{(1-x)^2} = p\frac{1}{p^2} = \frac{1}{p}\end{aligned}$$

式中 $x=1-p$，求导运算也是对 x 而言。在 Markov Inequality 中设 $c=\frac{10}{p}$，可得 $\Pr(T \geq \frac{10}{p}) \leq \frac{1}{10}$。

1.2 Chebyshev Inequality

设 $r.v.\ X$，则对 $\forall\ c > 0$，有

$$\Pr(|X-\mathbb{E}(X)| \geq c) \leq \frac{Var(X)}{c^2}$$

证明

$$\begin{aligned}\Pr(|X-\mathbb{E}(X)|\geq c) &= \Pr((X-\mathbb{E}(X))^2\geq c^2)\\ &\leq \frac{\mathbb{E}((X-\mathbb{E}(X))^2)}{c^2} \quad \text{(根据 Markov Inequality)}\\ &= \frac{Var(X)}{c^2}\end{aligned}$$

□

例：设随机变量序列 $X_1, X_2, \ldots X_n$ 相互独立，与 X 同分布，$\Pr(X=0)=\Pr(X=1)=1/2$。定义 $S=X_1+X_2+\ldots+X_n$，有 $\mathbb{E}(S)=n/2$，$Var(S)=n/4$。由 Chebyshev Inequality 有 $\Pr(|S-\frac{n}{2}|\geq c)\leq\frac{n}{4c^2}$。若取 $c=5\sqrt{n}$，则有 $\Pr(|S-\frac{n}{2}|\geq 5\sqrt{n})\leq 1\%$。

2 随机采样

在离散集合 U 中有某个子集 T，欲估计 T 的大小或者 $p=|T|/|U|$。在 U 中独立均匀随机采样，如果样本属于 T，则定义随机变量 $X_i=1$，否则 $X_i=0$，于是得到独立同分布的 $X_1, X_2, \ldots, X_n$，$\Pr(X_i=1)=p, \Pr(X_i=0)=1-p$。定义 $\hat{p}=\frac{1}{n}\sum_{i=1}^{n}X_i$，易知 $\mathbb{E}(\hat{p})=p, Var(\hat{p})=\frac{nVar(X_1)}{n^2}=\frac{p(1-p)}{n}$。由 Chebyshev Inequality 有

$$\Pr(|\hat{p}-p|\geq\delta)\leq\frac{p(1-p)}{\delta^2 n}\leq\frac{1}{4\delta^2 n}$$

该式描述了 $\hat{p}$ 偏离 p 达到 δ 以上的概率，为保证此概率小于 ϵ，便有 $\frac{1}{4\delta^2 n}<\epsilon$ 即 $n>\frac{1}{4\delta^2\epsilon}$，有趣的是为达到绝对误差 $|\hat{p}-p|$ 足够小的要求，所需采样的数量 n 与总体 $|U|$ 和 $|T|$ 无关。如果对相对误差 $|1-\hat{p}/p|$ 有要求，后续课程将会给出进一步分析。

3 快速排序与 Las Vegas 算法

快速排序算法的运行时间是随机的，最差为 $O(n^2)$，平均时间复杂度为 $O(n\log n)$。该算法保证运行结果正确，运行时间随机。称运行时间随机（可能无穷）但运行结果保证正确的算法为 **Las Vegas** 算法。

4 矩阵乘积验证与 Monte Carlo 算法

矩阵乘积问题的算法输入为两个 $n\times n$ 矩阵 A, B，该算法复杂度针对行（列）数 n 定义，将两个元素相乘和相加视为一个原子操作。关于其算法复杂度的发展，见 Lecture 1 的 ppt。

矩阵乘积验证问题（Matrix Multiplication Verification, MMV）输入为 3 个 $n\times n$ 的矩阵 A, B, C，矩阵元素和其间的运算在素数 p 的同余类 $\mathbb{Z}_p$ 上，问题要求判定是否 $AB=C$。

MMV 的一个随机算法均匀随机地取向量 $\vec{x}\in\mathbb{Z}_p^n$，计算 $AB\vec{x}$ 和 $C\vec{x}$（耗时 $O(n^2)$）再比较两者是否相同（耗时 $O(n)$），如果相同则输出 1，否则输出 0。算法时间复杂度为 $O(n^2)$。当 $AB=C$ 时，对 $\forall x$ 算法总是得到 $AB\vec{x}=C\vec{x}$，故输出 1；当 $AB\neq C$ 时，$\exists\vec{x}$ 满足 $AB\vec{x}=C\vec{x}$，算法有概率随机选择这样的 $\vec{x}$ 最终错误地输出 0，下面分析出现该错误的概率：

设 $D = AB - C$，由于 $D \neq 0$，故 D 中至少有一个元素不为零，不妨设其为 d_{11}，有

$$
\begin{aligned}
&\Pr(AB\vec{x} = C\vec{x}) \\
=&\Pr(D\vec{x} = 0) \\
=&\Pr\big((d_{11}x_1 + d_{12}x_2 + \cdots + d_{1n}x_n = 0) \wedge \ldots \wedge (d_{n1}x_1 + \cdots + d_{nn}x_n = 0)\big) \\
\leq&\Pr(d_{11}x_1 + d_{12}x_2 + \cdots + d_{1n}x_n = 0) \\
=&\Pr(x_1 = -d_{11}^{-1}(d_{12}x_2 + \cdots + d_{1n}x_n)) \\
=&1/p
\end{aligned}
$$

综上 $\Pr(output = 1|AB \neq C) \leq 1/p, \Pr(output = 0|AB = C) = 0$，该算法是单边错误算法。

称算法结果有概率出错（单边错或双边错）但运行时间固定的算法为 **Monte Carlo** 算法。

5 复杂性类

对计算问题，通常用字符串将问题的输入编码。把字符串组成的集合称为一个语言，语言便可以用来表示一个有意义的计算问题，如（以下集合中的元素均为对要表示的对象进行编码所得的字符串）：

1. 连通图对应的语言：$CONN = \{G|G\text{是连通图}\}$
   非连通图对应的语言：$\overline{CONN} = \{G|G\text{不是连通图}\}$

2. 有完美匹配的图对应的语言：$PerfectMatching = \{G|\text{图 } G \text{ 中有完美匹配}\}$
   无完美匹配的图对应的语言：$\overline{PerfectMatching} = \{G|\text{图 } G \text{ 中没有完美匹配}\}$

3. 可 3 染色图对应的语言：$G3C = \{G|\text{图 G 能够被 3 染色}\}$
   不可 3 染色图对应的语言：$\overline{G3C} = \{G|\text{图 G 不能够被 3 染色}\}$

4. 有 Hamilton 回路的图对应的语言：$HC = \{G|\text{图 } G \text{ 中有 Hamilton 回路}\}$
   无 Hamilton 回路的图对应的语言：$\overline{HC} = \{G|\text{图 } G \text{ 中没有 Hamilton 回路}\}$

5. $3SAT = \{\phi|\phi\text{是合取范式，每个子句中只有三个文字，且}\phi\text{可满足}\}$；
   $\overline{3SAT} = \{\phi|\phi\text{是合取范式，每个子句中只有三个文字，且}\phi\text{不可满足}\}$

5.1 复杂性类

1. P(Polynomial-time)：图灵机可在多项式时间内解决的语言组成的语言类。

2. NP(Non-deterministic Polynomial-time)：称语言 L 属于语言类 NP，当且仅当对 L 存在多项式时间算法 A，L 的一个实例 x 是否属于 L 可以借助辅助证据 w 由 A 验证，即 $x \in L \iff \exists w, A(x, w) = 1$。

   例如 $\forall\ G \in G3C$，存在 G 的一个 3 染色方案，将其一同输入给算法 A 可以用多项式时间验证该染色方案确实对 G 合法，而对 $\forall G \notin G3C$，则任何 3 染色方案都不合法；再如 $\forall \phi \in 3SAT$，存在 ϕ 的一个成真赋值，将其一同输入给算法 A 可以用多项式时间验证该赋值确实使 ϕ 为真，而对 $\forall \phi \notin 3SAT$，任何赋值都不可使 ϕ 为真。所以 $G3C$ 和 $3SAT$ 都属于 NP。

3. co-NP：语言 L 的补语言 $\overline{L} \in \mathrm{NP}$，则 $L \in \text{co-NP}$。

   $G3C, 3SAT \in \mathrm{NP}$，故 $\overline{G3C}, \overline{3SAT} \in \text{co-NP}$。

5.2 随机复杂性类

1. RP(Radomized Polynomial-time)：语言 $L \in \mathrm{RP}$ 当且仅当存在随机多项式时间算法 A，对 L 的符合实例 $x \in L$, 随机数 r, $\Pr(A(x,r)=1) \geq 1/2$; 对实例 $x \notin L$, 随机数 r, $\Pr(A(x,r)=1)=0$。

2. co-RP：$L \in \text{co-RP}$ 当且仅当存在随机多项式时间算法 A，对 L 的符合实例 $x \in L$，随机数 r，$\Pr(A(x,r)=0)=0$；对实例 $x \notin L$，随机数 r，$\Pr(A(x,r)=0) \geq 1/2$。

按照定义，$MMV \in \text{co-RP}$，而 $\overline{MMV} \in \mathrm{RP}$。