大数据算法-2025 春

Lecture 1: 基础数学和统计工具

2025.2.25

Lecturer: 丁虎 *Scribe:* 黄震

本章介绍算法分析中常用的渐进性符号和一些概率不等式。关于更深刻的数学主题，参见 [1, 2].

1 渐进性分析与复杂度

假设 $f(n)$ 和 $g(n)$ 是两个关于问题规模 n 的函数，它们之间的渐进性关系定义如下：

- $f(n) = \Theta(g(n))$: $\exists n_0, c_1, c_2 > 0$, s.t., 当 $n > n_0$ 时，$c_1 g(n) \le f(n) \le c_2 g(n)$
- $f(n) = O(g(n))$: $\exists n_0, c_2 > 0$, s.t., 当 $n > n_0$ 时，$f(n) \le c_2 g(n)$
- $f(n) = o(g(n))$: $\forall c > 0, \exists n_0$, s.t., 当 $n > n_0$ 时，$f(n) < cg(n)$
- $f(n) = \Omega(g(n))$: $\exists n_0, c_1 > 0$, s.t., 当 $n > n_0$ 时，$f(n) \ge c_1 g(n)$
- $f(n) = \omega(g(n))$: $\forall c > 0, \exists n_0$, s.t., 当 $n > n_0$ 时，$f(n) > cg(n)$

渐进关系反应的是当 n 非常大时函数的量级关系，也可以理解为函数增长速率的大小关系。Θ 表示当 n 非常大时，两个函数的以相同的速度增长，属于同一量级，类似于 $=$ 关系。O 和 Ω 则类似于 $\le$ 和 $\ge$，o 和 ω 则类似于 $<$ 和 $>$

例如 $f(n) = 2n^2 + n + 4$，我们可以说 $f(n) = \Theta(n^2)$；也可以说 $f(n) = O(n^2)$ 或 $f(n) = O(n^3)$，可以是 $f(n) = o(n^2 \log n)$；还可以说 $f(n) = \Omega(n^2)$ 或 $f(n) = \Omega(n \log n)$ 或 $f(n) = \omega(n)$

Example 1.1. $T(n) = aT(n/b) + f(n)$，求 $T(n)$

Solution.

$$
\begin{aligned}
T(n) &= a(aT(n/b^2) + f(n/b)) + f(n) \\
&= a^2T(n/b^2) + af(n/b) + f(n) \\
&= a^{\log_b n} \cdot T(1) + \sum_{i=0}^{\log_b n-1} a^i \cdot f(\frac{n}{b^i}) \\
&= \Theta(n^{\log_b a}) + \sum_{i=0}^{\log_b n-1} a^i \cdot f(\frac{n}{b^i})
\end{aligned}
$$

(1) 当 $f(n) = \Theta(n^{log_b a})$ 时,

$$
\sum_{i=0}^{\log_b n-1} a^i \cdot f(\frac{n}{b^i}) = \sum_{i=0}^{\log_b n-1} a^i \cdot \Theta((\frac{n}{b^i})^{log_b a}) = \sum_{i=0}^{\log_b n-1} \Theta(n^{\log_b a}) = \Theta(n^{\log_b a} \log n)
$$

因此，$T(n) = \Theta(n^{\log_b a} \log n)$

(2) 当 $f(n) = O(n^{\log_b a - \epsilon})$ 时,

$$
\begin{aligned}
\sum_{i=0}^{\log_b n-1} a^i \cdot f(\frac{n}{b^i}) &= \sum_{i=0}^{\log_b n-1} a^i O((\frac{n}{b^i})^{\log_b a - \epsilon}) \\
&= \sum_{i=0}^{\log_b n-1} O(n^{\log_b a - \epsilon} \cdot (b^\epsilon)^i) \\
&= O(n^{\log_b a - \epsilon}) \cdot O(\frac{n^\epsilon - 1}{b^\epsilon - 1}) \\
&= O(n^{\log_b a - \epsilon}) \cdot O(n^\epsilon) \\
&= O(n^{log_b a})
\end{aligned}
$$

因此，$T(n) = O(n^{\log_b a})$

(3) 当 $f(n) = \Omega(n^{\log_b a + \epsilon})$, 且存在常数 $c < 1$ 使得当 n 足够大时有 $af(n/b) \leq cf(n)$ 时,

$$
\begin{aligned}
\sum_{i=0}^{\log_b n-1} a^i \cdot f(\frac{n}{b^i}) &\leq \sum_{i=0}^{\log_b n-1} c^i f(n) + O(1) \qquad (1) \\
&\leq f(n)\frac{1}{1-c} + O(1) \\
&= O(f(n))
\end{aligned}
$$

式 (1) 中的 O(1) 用于覆盖那些 n 不够大使得条件中不等式不成立的项。因此，$T(n) = O(f(n))$。结合 $T(n)$ 的定义可知 $T(n) = \Omega(f(n))$，综合来看，$T(n) = \Theta(f(n))$。 □

2 概率不等式

2.1 Markov's Inequality

Theorem 2.1. 给定随机变量 $X > 0$，$\forall k > 0$，有

$$\Pr[X \geq k \cdot \mathrm{E}[X]] \leq \frac{1}{k}$$

Proof.

$$\begin{aligned}
\mathrm{E}[X] &= \int_0^{\infty} x f_X(x) \mathrm{d}x \\
&\geq \int_{k \cdot \mathrm{E}[X]}^{\infty} x f_X(x) \mathrm{d}x \\
&\geq k \cdot \mathrm{E}[X] \int_{k \cdot \mathrm{E}[X]}^{\infty} f_X(x) \mathrm{d}x \\
&= k \cdot \mathrm{E}[X] \cdot \Pr[X \geq k \cdot \mathrm{E}[X]]
\end{aligned}$$

□

通过简单代换，Markov's Inequality 也可以表示为

$$\Pr[X \geq a] \leq \frac{\mathrm{E}[X]}{a}$$

2.2 Chebyshev's Inequality

Theorem 2.2. 给定随机变量 X，有

$$\Pr[|X - \mathrm{E}[X]|^2 \geq k \cdot \mathrm{Var}[X]] \leq \frac{1}{k}$$

或者表示为

$$\Pr[|X - \mathrm{E}[X]| \geq k] \leq \frac{\mathrm{Var}[X]}{k^2}$$

Proof. 将 $|X - \mathrm{E}[X]|^2$ 视为随机变量，直接利用 Markov's Inequality 即可证明。 □

2.3 Chernoff Bound

Theorem 2.3. 给定 n 个独立的随机变量 $X_1, \ldots, X_n$，满足 $0 \leq X_i \leq 1$。令 $S = \sum_{i=1}^n X_i$, $\mu = \mathrm{E}[S]$, $\delta \in (0,1)$, 则

$$\Pr[|S-\mu| \leq \delta\mu] \geq 1 - e^{-O(\frac{\delta^2\mu^2}{n})} \tag{2}$$

Proof. 我们在此处证明 $\Pr[S > (1+\delta)\mu]$ 部分，并假定考虑 X_i 为 Bernoulli 变量时（即 $\Pr[X_i = 1] = p_i, \Pr[X_i = 0] = 1 - p_i$）

$$\begin{aligned}
\Pr[S > (1+\delta)\mu] &= \Pr[e^{\lambda S} > e^{\lambda(1+\delta)\mu}] \\
&\leq \frac{1}{e^{\lambda(1+\delta)\mu}} \mathrm{E}[e^{\lambda S}] && (3) \\
&= \frac{1}{e^{\lambda(1+\delta)\mu}} \prod_{i=1}^n \mathrm{E}[e^{\lambda X_i}] && (4) \\
&= \frac{1}{e^{\lambda(1+\delta)\mu}} \prod_{i=1}^n (p_i e^\lambda + (1-p_i)) \\
&\leq \frac{1}{e^{\lambda(1+\delta)\mu}} \prod_{i=1}^n e^{p_i(e^\lambda - 1)} && (5) \\
&= \frac{e^{\mu(e^\lambda-1)}}{e^{\lambda(1+\delta)\mu}} \\
&= (\frac{e^\delta}{(1+\delta)^{1+\delta}})^\mu && (6)
\end{aligned}$$

上式中 (3) 可直接由 Markov's Inequality 得到。(4) 由 X_i 独立可得。(5) 利用不等式 $1 + x \leq e^x$ 进行放缩获得。令 $e^\lambda = 1 + \delta$ 代入得到 (6)。对上式最后的结果取对数分析：

$$\begin{aligned}
\ln(\frac{e^\delta}{(1+\delta)^{1+\delta}})^\mu &= \mu[\delta - (1+\delta)\ln(1+\delta)] \\
&\leq \mu[\delta - (1+\delta)\frac{\delta}{1+\delta/2}] && (7) \\
&= -\frac{\delta^2}{2+\delta}\mu
\end{aligned}$$

式 (7) 利用不等式 $\ln(1+x) \geq \frac{x}{1+x/2}$ 放缩可得。将上述结果结合，可得

$$\Pr[S > (1+\delta)\mu] \leq e^{-\frac{\delta^2}{2+\delta}\mu} \leq e^{-\frac{\delta^2}{3}\mu}$$

$\square$

关于 Chernoff Bound 的证明还可以使用其他的放缩方式，需要使用下述引理：

Lemma 2.4 (Hoeffding lemma)**.** 假设随机变量 $a \le X \le b$ 且 $\mathrm{E}[X] = 0$，那么对于 $\forall \lambda$，有

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

关于 Chernoff Bound 的另一种证明方式如下：

Proof. 同前述步骤，我们可以得到

$$\Pr[S > (1+\delta)\mu] \le \frac{1}{e^{\lambda(1+\delta)\mu}} \mathrm{E}[e^{\lambda S}]$$

使用 Hoeffding lemma 对 $\mathrm{E}[e^{\lambda S}]$ 进行放缩：

$$\begin{aligned}
\mathrm{E}[e^{\lambda S}] &= \prod_{i=1}^{n} \mathrm{E}[e^{\lambda X_i}] \\
&= \prod_{i=1}^{n} e^{\lambda p_i} \cdot \mathrm{E}[e^{\lambda (X_i - p_i)}] \\
&\le \prod_{i=1}^{n} e^{\lambda p_i} \cdot e^{\frac{\lambda^2}{8}} \qquad (8) \\
&= e^{\frac{\lambda^2}{8} n + \lambda \mu}
\end{aligned}$$

式 (8) 即为使用 Hoeffding 引理所得。将上述结果与之前分析相结合：

$$\Pr[S > (1+\delta)\mu] \le \frac{e^{\frac{\lambda^2}{8} n + \lambda\mu}}{e^{\lambda(1+\delta)\mu}} = e^{\frac{\lambda^2}{8} n - \lambda\delta\mu} = e^{-\frac{2\delta^2\mu^2}{n}}$$

上式最后一步中，令 $\lambda = \frac{4\delta\mu}{n}$ 即可得到对应结果。 $\square$

关于 $\Pr[S < (1-\delta)\mu]$ 的证明留作思考，后续考虑是否在此文档中补充。可以看到，两种证明方式最终获得结果在形式上有细微差异，但都可以写为式 (2) 的渐进形式。Chernoff Bound 表明 S 落在该区间之外的概率是和区间大小呈负指数关系。如果我们考虑的是 $\{X_i\}$ 的均值而不是它们之和时，使用 Chernoff Bound 会在 e 的负指数上引入随机变量个数 n，这意味着我们估计的精度是和采样个数呈负指数关系，这个结论远强于 Markov's Inequality 和 Chebyshev's Inequality （可以试着用这两个不等式分析 S 或均值并进行比对）

References

[1] R. Van Handel. Probability in high dimension. *Lecture Notes (Princeton University)*, 2(3):2–3, 2014.

[2] R. Vershynin. *Concentration of Sums of Independent Random Variables*, page 11–37. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2018.