تمرین پنجم

امیرحسین مهدی‌نژاد
شماره دانشجویی ۸۱۰۸۰۰۰۵۸
mahdinejad@ut.ac.ir

۱

اثبات می‌کنیم Y و Z متغیرهای پواسون مستقل با میانگین μp و $\mu(1-p)$ به ترتیب باشند. می‌دانیم X متغیر پواسون با میانگین μ تعداد ارورهای گرامری در یک صفحه که با احتمال p و ارورهای املایی با احتمال $1-p$ را نشان می‌دهد.
همچنین Y و Z متغیرهای تصادفی باشند که به ترتیب ارورهای گرامری و املایی را در یک صفحه از کتاب نمایندگی می‌کنند.
اگر تعداد ارورهای یک صفحه n باشد، تعداد ارورهای گرامری از توزیع $Y \sim Bin(n,\ p)$ است. احتمال کل برای داشتن k ارور گرامری برابر است با:

$$P(Y=k) = \sum_{n=k}^{\infty} P(Y=k|X=n)P(X=n) = \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \frac{e^{-\mu}\mu^n}{n!}$$

$$= \frac{e^{-\mu}(\mu p)^k}{k!} \sum_{n=k}^{\infty} \frac{\mu^{n-k}}{(n-k)!} (1-p)^{n-k}$$

از فرمول $\sum_{j=0}^{\infty} \frac{x^j}{j!} = e^x$ نتیجه می‌شود:

$$P(Y=k) = \frac{e^{-\mu}(\mu p)^k}{k!} \sum_{j=0}^{\infty} \frac{(\mu - p\mu)^j}{j!} = \frac{e^{-\mu}(\mu p)^k}{k!} e^{\mu - p\mu} = \frac{e^{-p\mu}(\mu p)^k}{k!}$$

به همین ترتیب داریم $Z \sim Poisson(\mu(1-p))$.
برای مستقل بودن Y و Z داریم:

$$P(Y=y,\ Z=z) = P(Y=y,\ X=y+z)$$

$$= P(Y=y\,|\,X=y+z)P(X=y+z) = \frac{(y+z)!}{y!z!} p^y (1-p)^z \frac{e^{-\mu}\mu^{y+z}}{(y+z)!}$$

$$= \frac{e^{-p\mu}(p\mu)^{y+z}}{y!} \cdot \frac{e^{-(1-p)\mu}((1-p)\mu)^{y+z}}{z!}$$

نتیجه می‌شود $P(Y=y,\ Z=z) = P(Y=y)P(Z=z)$ که Y و Z متغیرهای پواسون مستقل با میانگین‌های به ترتیب μp و $\mu(1-p)$ هستند.

---

۳ الف:

فرض کنیم X متغیر تصادفی تعداد $k-gap$ها در n سطل باشد.

اگر یک $k - gap$ از سطل i شروع می‌شد $X_i = 1$ و در غیر این‌صورت $X_i = 0$ قرار می‌دهیم. حال می‌توان گفت $X = \sum_{i=0}^{n-k} X_i$ و امید ریاضی را به صورت زیر نوشت:

$$E[X] = E[\sum_{i=0}^{n-k} X_i] = \sum_{i=0}^{n-k} E[X_i]$$

از طرفی

$$E[X_i] = P[X_i = 1]$$

که همان احتمال نیوفتادن توپ b به ازای تمام bهای 1 تا m در سطل iام تا $i + k - 1$ام یا به بیانی دیگر ضرب احتمالات افتادن توپ b در یکی از $n - k$ سطل دیگر است که با $(\frac{n-k}{n})^m$ برابر می‌شود. نتیجه می‌گیریم:

$$E[X] = \sum_{i=0}^{n-k} (\frac{n-k}{n})^m = (n - k + 1)(\frac{n-k}{n})^m$$

ب:

به ازای $0 \le i \le n - k$، متغیر تصادفی $X = \sum_{i=0}^{n-k} X_i$ همان تعداد $k - gap$ها در n سطل پس از پرتاب m توپ است. اگر x_i تعداد توپ‌ها در سطل i بوده و متغیر تصادفی پواسون مستقل متناظر آن y_i باشد، به همین شکل متغیر تصادفی پواسون مستقل متناظر با X_i را Y_i در نظر می‌گیریم. از طرفی می‌دانیم:

$$X_i = f(x_i, x_{i+1}, \cdots, x_{i+k-1}) = 1\, if\, (\sum_{j=i}^{i+k-1} x_j = 0)\, otherwise\, 0$$

به همین ترتیب Y_i به صورت تابع $f(y_i, y_{i+1}, \cdots, y_{i+k-1})$ قابل تعریف است.

در نظر بگیریم $= \sum_{j=0}^{n/k} X_{i+jk}$ $Z_i = X_i + X_{i+2k} + \cdots$ و همین‌طور $Z_i^P = Y_i + Y_{i+k} + Y_{i+2k} + \cdots$ $= \sum_{j=0}^{n/k} Y_{i+jk}$ باشد. بدین شکل می‌توان تعداد کل $k - gap$ها را به صورت $X = \sum_{i=0}^{k-1} Z_i$ بیان کرد.

از آنجا که Z_i تابعی از X_i که خود آن تابعی از x_i است و از طرفی $E[Z_i] = \frac{E[X]}{k}$ می‌توان تابعی مثل g را به این شکل نوشت:

$$g(x_0, \cdots, x_{n-1}) = 1\, if\, (Z_0 \ge (1 + \delta)\frac{E[X]}{k})\, otherwise\, 0$$

$$g(y_0, \cdots, y_{n-1}) = 1\, if\, (Z_0^P \ge (1 + \delta)\frac{E[X]}{k})\, otherwise\, 0$$

فلذا امید ریاضی را به شکل زیر بدست می‌آوریم:

$$E[g(x_0, \cdots, x_{n-1})] = P[Z_0 \geq (1+\delta)\frac{E[X]}{k}]$$
$$E[g(y_0, \cdots, y_{n-1})] = P[{Z_0}^p \geq (1+\delta)\frac{E[X]}{k}]$$
$$\rightarrow E[g(x_0, \cdots, x_{n-1})] \leq e\sqrt{m}E[g(y_0, \cdots, y_{n-1})]$$

در نتیجه $P[Z_0 \geq (1+\delta)\frac{E[X]}{k}] \leq e\sqrt{m}P[{Z_0}^P \geq (1+\delta)\frac{E[X]}{k}]$ است که با استفاده از کران چرنوف روی جمع آزمایشات مستقل پواسون می‌توان نوشت:

$$P[{Z_0}^P \geq (1+\delta)\frac{E[X]}{k}] \leq e^{-E[X]\delta^2/3k}$$

در نهایت اگر $X \geq (1+\delta)E[X]$ باشد، حتما یک i وجود دارد که $Z_i \geq (1+\delta)\frac{E[X]}{k}$ و داریم:

$$P[X \geq (1+\delta)E[X]] \leq P[\bigcup_{j=0}^{k-1} Z_i \geq (1+\delta)\frac{E[X]}{k}]$$
$$\leq \sum_{j=0}^{k-1} P[Z_i \geq (1+\delta)\frac{E[X]}{k}]$$
$$\leq ke\sqrt{m}P[{Z_i}^P \geq (1+\delta)\frac{E[X]}{k}]$$
$$= ke^{1-E[X]\delta^2/3k}\sqrt{m}$$

که همان خواسته‌ی سوال است.

۵

یک $Bloom\ filter$ شامل آرایه‌ای nبیتی از $A[0]$ تا $A[n-1]$ است که اگر آیتمی به آن افزوده شود شمارنده‌ی متناظرش زیاد شده و در صورت حذف، کم می‌شود.

قصد داریم برای احتمال رخداد خطا در مدت زمان t بار عملیات $insertion$ یا $deletion$ یک کران تعیین کنیم. در هر لحظه، تعداد عناصر داخل مجموعه را حداکثر برابر با n، شمارنده‌ها را m و توابع $hash$ را k در نظر می‌گیریم.

اگر $c(i)$ را عدد متناظر با iمین شمارنده در نظر بگیریم، احتمال اینکه iمین شمارنده j بار افزایش یابد برابر است با:

$$P(C(i) = j) = \binom{nk}{j}(\frac{1}{m})^j(1-\frac{1}{m})^{nk-j}$$

لذا احتمال اینکه مقدار شمارنده حداقل j باشد با $mP(c(i) \geq j)$ و کران آن به این صورت خواهد بود:

$$P(c(i) \geq j) \leq \binom{nk}{j}(\frac{1}{m^j}) \leq (\frac{enk}{jm})^j$$

با دانستن اینکه نرخ خطای $false - positive$ برابر $k = \frac{(\ln 2)m}{n}$ است می‌توان نوشت:

$$P(\max_i c(i) \geq j) \leq m(\frac{e\ln 2}{j})^j$$

وقتی شمارنده ۴ بیتی است به ازای مقادیر بیشتر از ۱۶ سرریز خواهد کرد لذا:

$$P(\max_i\ c(i)\ \geq 16) \leq 1.37\ \times 10^{-15} \times m$$

لذا به ازای t این مقدار حداکثر $1.37\ \times 10^{-15}\ \times mt$ خواهد بود.

۷

در نظر بگیریم $p\ =\ f(n)\ =\ o(n^{-2/4})$ باشد. برای هر مجموعه از k راس C_i که $1\ \leq\ i\ \leq \binom{n}{k}$ متغیر X_i را نشان‌دهنده‌ی اینکه C_i یک کلیک ۵تایی باشد در نظر می‌گیریم. با دانستن $X\ =\ \sum_{i=1}^{\binom{n}{5}} X_i$ داریم:

$$E[X]\ =\binom{n}{5}p^{10}$$

از طرفی $E[X]\ =\ o(n^5).o(n^{-5})\ =\ o(1)$ و برای nهای بزرگ $E[x]\ <\ \varepsilon$ است. احتمال اینکه زیرگرافی از $G_{n,p}$ یک کلیک ۵تایی داشته باشد با دانستن $P[X\ \geq\ 1]\ \leq\ E[X]\ <\ \varepsilon$ کمتر از ε خواهد بود.

می‌دانیم $E[X|X_j\ =\ 1]\ =\ \sum_{i=1}^{\binom{n}{5}} E[X_i|X_j\ =\ 1]$ است. حالتی که C_i و C_j دو راس مشترک داشته باشند را در نظر می‌گیریم؛ تعداد $\binom{5}{2}$ راه وجود دارد که این رئوس و $\binom{n-5}{3}$ راه که رئوس باقی‌مانده را انتخاب کنیم. حال ۳ راس غیرمشترک را اضافه می‌کنیم و باید ۹ یال جدید بیاوریم تا $X_i\ =\ 1$ و $X_j\ =\ 1$ شود. لذا عبارت قبلی را در p^9 ضرب می‌کنیم.

این کار را برای ۱، ۲، ۳ و ۵ راس مشترک انجام می‌دهیم تا به عبارت زیر می‌رسیم:

$$P[X\ <\ 0] \geq \frac{\binom{n}{5}p^{10}}{\binom{n-5}{0}+\binom{n-5}{5}p^{10}+5\binom{n-5}{4}p^{10}+10\binom{n-5}{3}p^{9}+10\binom{n-5}{2}p^{7}+5\binom{n-5}{1}p^{4}}$$

با در نظر گرفتن $p\ =\ f(n)\ =\ \omega(n^{-1/2})$ وقتی $n\ \rightarrow\ \infty$ این مقدار از پایین به ۱ نزدیک می‌شود.

۸

الف:

گراف تصادفی را به صورت $G_{n,N}$ به طوری که به ازای یک ثابت $c\ >\ 0$ مقدار $N\ =\ cn$ باشد در نظر می‌گیریم. تعداد رئوس مجزا را با X نمایش می‌دهیم. از نامساوی مارکوف داریم:

$$P(X\ \geq\ 1)\ \leq\ E[X]$$

امید ریاضی گراف تصادفی به ازای $p\ =\ cn$ بدین شکل محاسبه می‌شود:

$$E[x]\ \ =\ \ n(1\ -\ p)^{n-1} = n(1\ -\ cn)^{n-1} = n(1\ -\frac{1}{n})^{n}$$

ب:

به ازای هر $t \geq 0$ و برای یک $\lambda > 0$ و با فرض $|X_k - X_{k-1}| \leq c_k$ برای $X_0, \ \cdots, \ X_n$ داریم:

$$P(|X_t - X_0| \geq \lambda) \leq 2e^{-\frac{\lambda^2}{2\sum_{k=1}^{t} c_k^2}}$$

به همین ترتیب داریم $P(|X_t - X_0| \geq \lambda c\sqrt{t}) \leq 2e^{-\frac{\lambda^2}{2}}$ که از این طریق اثبات می‌کنیم:

$$\begin{aligned} P(|X - \frac{cn}{e}| \geq \lambda\sqrt{cn} + 1) &\leq P(|X - \frac{cn}{e}| \geq 2\lambda\sqrt{cn}) \\ &\leq P(|X - E[X]| \geq 2\lambda\sqrt{cn}) \\ &\leq 2\frac{[e^{\alpha(X_t - X_0)}]}{e^4} \\ &\leq 2e^{-\frac{2\lambda^2}{4}} \end{aligned}$$

در نهایت به $P(|X - E[X]| \geq 2\lambda\sqrt{cn}) \leq 2e^{-\frac{\lambda^2}{2}}$ می‌رسیم که همان خواسته‌ی سوال است.