CMPT 210 2014 Practice Problems for Lectures 21–24

Birthdays, Tail Inequalities (Markov/Chebyshev/Chernoff), Sampling, Load Balancing

Unless otherwise stated, all random variables are discrete. Use exact expressions unless asked for a numerical approximation.

Problem 1: Matching Birthdays (L21)

In a class of n students, assume: (i) birthdays are uniform over $d = 365$ days (no leap years), (ii) students' birthdays are independent.

(a) Compute Pr(no two students share a birthday) as a product, and hence write Pr(at least one shared birthday)

(b) Let M be the number of *pairs* of students with matching birthdays. For $1 \le i < j \le n$, define the indicator r.v.

$$X_{i,j} = \begin{cases} 1, & \text{if students } i \text{ and } j \text{ share a birthday,} \\ 0, & \text{otherwise.} \end{cases}$$

Show that $M = \sum_{1 \le i < j \le n} X_{i,j}$, and compute $E[M]$.

(c) Show that the family $\{X_{i,j}\}$ is *not* mutually independent by giving a concrete conditional probability that changes after conditioning on two other indicators.

(d) Argue (carefully) that $X_{i,j}$ and $X_{i',j'}$ are pairwise independent when $\{i, j\} \cap \{i', j'\} = \varnothing$.

(e) Using that each $X_{i,j}$ is Bernoulli and (assuming the relevant pairwise independence), compute $\mathrm{Var}(M)$ in closed form.

(f) Plug in $n = 100$ to get formulas for $E[M]$ and $\sqrt{\mathrm{Var}(M)}$ (standard deviation).

Problem 2: Markov's Inequality + Shifting Trick (L21)

(a) (Proof) Let X be a non-negative r.v. Prove Markov's Inequality:

$$\Pr[X \geq x] \leq \frac{E[X]}{x} \qquad \text{for all } x > 0,$$

by introducing the indicator r.v. $I_{\{X \geq x\}}$ and comparing $xI_{\{X \geq x\}}$ to X.

(b) Let X be non-negative with $E[X] = 99.99$. Use Markov's Inequality to show

$$\Pr[X \geq 300] \leq \frac{1}{3}.$$

(c) (Shift to tighten) Suppose X is a r.v. such that $X \geq 100$ always and $E[X] = 150$. Define a shifted variable $Y := X - 100$ and use Markov's Inequality on Y to bound $\Pr[X \geq 200]$. Compare this bound to what you would get by applying Markov directly to X.

Problem 3: Chebyshev's Inequality (L21–L22)

(a) (Derivation from Markov) Starting from Markov's inequality, derive Chebyshev's inequality:

$$\Pr\left(|X - E[X]| \geq y\right) \leq \frac{\mathrm{Var}(X)}{y^2} \qquad \text{for all } y > 0,$$

by applying Markov to the non-negative r.v. $|X - E[X]|^2$.

(b) Let X be a r.v. with $E[X] = 100$ and standard deviation $\sigma_X = 15$. Use Chebyshev's inequality to bound $\Pr[X \geq 300]$.

(c) Let $X \sim \mathrm{Bin}(20, 0.75)$.

(i) Compute $E[X]$ and $\mathrm{Var}(X)$ using binomial formulas.

(ii) Use Chebyshev to give a lower bound on $\Pr(10 < X < 20)$ by rewriting it as $1 -$ Pr(bad event) for a suitable deviation from the mean.

Problem 4: Chernoff Bound for Coin Tossing (L22–L23)

Let T be the number of heads in $n = 1000$ independent tosses of a fair coin. Let T_i be the indicator of "toss i is heads", so $T = \sum_{i=1}^{1000} T_i$.

Chernoff bound (as used in lecture): for mutually independent $T_i \in [0, 1]$,

$$\Pr[T \geq cE[T]] \leq \exp\left(-\beta(c)\, E[T]\right), \qquad \beta(c) := c\ln(c) - c + 1, \quad c \geq 1.$$

(a) Compute $E[T]$.

(b) We want to upper-bound $\Pr[T \geq 1.2\, E[T]]$. Identify the value of c and write the Chernoff bound in the form $\exp(-\beta(c)E[T])$.

(c) Compute $\beta(1.2)$ (you may leave it as $1.2\ln(1.2) - 0.2$ if you want), and write a numerical approximation of the final bound.

(d) Compare with Chebyshev: compute $\mathrm{Var}(T)$ and use Chebyshev to upper-bound $\Pr[T \geq 1.2\, E[T]]$. Which bound is tighter here, and why (in one or two sentences)?

Problem 5: Chernoff Bound for Rare Events (Lottery-Style) (L23)

A game is played by $n = 10{,}000{,}000$ players. Each player independently wins with probability $p = 1/10000$. Let T_i be the indicator that player i wins, and let $T = \sum_{i=1}^{n} T_i$ be the total number of winners.

(a) Compute $E[T]$.

(b) Use the Chernoff bound to upper-bound $\Pr[T \geq 2000]$. (Hint: write 2000 as $cE[T]$ and identify c.)

(c) Compute $\beta(2) = 2\ln(2) - 1$ and write the resulting bound in the form $\exp(-\text{something})$. You may also convert it to a rough power of 10 if you wish.

(d) Give a short interpretation: what does a bound of the form $\exp(-\Theta(E[T]))$ say about the likelihood of being *twice* the mean when $E[T]$ is large?

Problem 6: Voter Poll Sample Size via Chebyshev (L23)

We wish to estimate an unknown fraction $p \in (0,1)$. We sample (with replacement) n people uniformly at random and define X_i to be the indicator that person i favors candidate A. Thus $\Pr(X_i = 1) = p$ and the X_i are mutually independent. Let $S_n = \sum_{i=1}^{n} X_i$ and $\hat{p} := S_n/n$.

(a) Identify the distribution of S_n and compute $E[\hat{p}]$.

(b) Compute $\mathrm{Var}(\hat{p})$ in terms of p and n.

(c) Using Chebyshev's inequality, show that

$$\Pr\left(|\hat{p} - p| \geq \varepsilon\right) \leq \frac{p(1-p)}{n\varepsilon^2}.$$

(d) We want $\Pr(|\hat{p} - p| < \varepsilon) \geq 1 - \delta$. Derive a sufficient condition on n in terms of ε, δ, and explain why

$$n \geq \frac{1}{4\varepsilon^2\delta}$$

is sufficient *without knowing* p.

(e) Plug in $\varepsilon = 0.02$ and $\delta = 0.01$ and compute the resulting n.

Problem 7: Pairwise Independent Sampling & WLLN (L23)

Let $G_1, \ldots, G_n$ be pairwise independent r.v.'s with common mean μ and variance σ^2. Define

$$S_n := \sum_{i=1}^{n} G_i, \qquad X_n := \frac{S_n}{n}.$$

(a) Compute $E[X_n]$.

(b) Using pairwise independence, compute $\mathrm{Var}(S_n)$ and hence $\mathrm{Var}(X_n)$.

(c) Apply Chebyshev to show

$$\Pr\left(|X_n - \mu| \geq \varepsilon\right) \leq \frac{\sigma^2}{n\varepsilon^2}.$$

(d) Use your bound to conclude (informally, one or two sentences) why $\lim_{n\to\infty} \Pr(|X_n - \mu| \leq \varepsilon) = 1$ for every fixed $\varepsilon > 0$ (a weak law of large numbers statement).

Problem 8: Comparing Tail Inequalities + Load Balancing (L24)

Part A: Markov vs. Chebyshev vs. Chernoff

Let $T_1, \ldots, T_n$ be r.v.'s with $T_i \in \{0, 1\}$ and $\Pr[T_i = 1] = p_i$. Let $T = \sum_{i=1}^{n} T_i$, so $E[T] = \sum_{i=1}^{n} p_i$.

(a) (No independence needed) Use Markov to show that for $c \geq 1$,

$$\Pr[T \geq cE[T]] \leq \frac{1}{c}.$$

(b) (Pairwise independence) Assume the T_i are pairwise independent. Show $\mathrm{Var}(T) = \sum_{i=1}^{n} p_i(1 - p_i)$ and derive the Chebyshev-style bound

$$\Pr[T \geq cE[T]] \leq \frac{\sum_{i=1}^{n} p_i(1 - p_i)}{(c-1)^2 (E[T])^2}.$$

(c) (Mutual independence) Assume the T_i are mutually independent and $T_i \in [0, 1]$. State the Chernoff bound in terms of $\beta(c) = c \ln(c) - c + 1$:

$$\Pr[T \geq cE[T]] \leq \exp\left(-\beta(c) E[T]\right).$$

(d) Specialize to $p_i = \frac{1}{2}$ for all i. Compare the three bounds for

$$\Pr\left(T \geq 0.6n\right).$$

(Hint: here $E[T] = 0.5n$ so $c = 1.2$. Write each bound as a function of n.)

Part B: Randomized Load Balancing (Fussbook)

A system receives $n = 24000$ tasks in each 10-minute interval. Each task is assigned independently and uniformly to one of m servers. Each server processes its assigned tasks sequentially.

Each task takes at most 1 second, and takes $1/4$ second on average.

A server is considered *overloaded* if it is assigned more than 600 seconds of work in the interval.

Let T be the total processing time assigned to *server 1* in the interval. For each task i, define

$$T_i := \begin{cases} (\text{processing time of task } i), & \text{if task } i \text{ is assigned to server 1,} \\ 0, & \text{otherwise.} \end{cases}$$

Thus $T = \sum_{i=1}^{n} T_i$ and $0 \leq T_i \leq 1$.

(a) Compute $E[T_i]$ and then $E[T]$ as a function of m.

(b) Show that the overload event $T > 600$ can be written as $T \geq cE[T]$ for a suitable c, and find c in terms of m.

(c) Using the Chernoff bound, write an upper bound on $\Pr[T \geq 600]$ in the form

$$\Pr[T \geq 600] \leq \exp\left(-\beta(c)\, E[T]\right).$$

(d) Use a union bound to show

$$\Pr(\text{some server is overloaded}) \leq m \Pr(T \geq 600).$$

(e) (Numerical check) Evaluate your bound for $m = 12$ and $m = 15$ (you may approximate $\beta(\cdot)$). Which m gives a meaningfully smaller upper bound?